1 MÁTRIXOK ÉS MÁTRIXMŰVELETEK

Definíció. *$n, m > 0$ egész számok. Egy A $m \times n$ típusú (valós) mátrixon valós a_{ij} számok alábbi táblázatát értjük:*

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1j} & \dots & a_{1n} \\ \vdots & \vdots & & \vdots & & \vdots \\ a_{i1} & a_{i2} & \dots & a_{ij} & \dots & a_{in} \\ \vdots & \vdots & & \vdots & & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mj} & \dots & a_{mn} \end{bmatrix}.$$

a_{ij} az A mátrix i-edik sorában és j-edik oszlopában álló mátrixelem.
Mátrixok szokásos jelölései még:

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1j} & \dots & a_{1n} \\ \vdots & \vdots & & \vdots & & \vdots \\ a_{i1} & a_{i2} & \dots & a_{ij} & \dots & a_{in} \\ \vdots & \vdots & & \vdots & & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mj} & \dots & a_{mn} \end{pmatrix}$$

$$A = [a_{ij}]_{i,j=1}^{m,n}, \qquad A = (a_{ij})_{i,j=1}^{m,n}.$$

Az $m \times n$ típusú valós mátrixok halmaza: $\mathbb{R}^{m\times n}$
$A \in \mathbb{R}^{m\times n}$ *négyzetes* (kvadratikus), ha $m = n$.
Ekkor a tömör megadási módok:

$$A = [a_{ij}]_{i,j=1}^{n}, \qquad A = (a_{ij})_{i,j=1}^{n}.$$

A mátrixok közti fontosabb műveletek:

1. **Összeadás:** $A, B \in \mathbb{R}^{m\times n}$,

$$C = A + B \in \mathbb{R}^{m\times n} \Leftrightarrow c_{ij} = a_{ij} + b_{ij} \quad (i = 1, \dots, m,\ j = 1, \dots, n).$$

Az összeadásra fennáll, hogy

$$A + B = B + A, \quad (A + B) + C = A + (B + C).$$

2. **Számmal való szorzás:** $A \in \mathbb{R}^{m\times n}$, λ valós szám,

$$C = \lambda A \in \mathbb{R}^{m\times n} \Leftrightarrow c_{ij} = \lambda a_{ij} \quad (i = 1, \dots, m,\ j = 1, \dots, n).$$

A számmal való szorzásra fennáll, hogy

$$\lambda(\mu A) = (\lambda\mu)A, \quad (\lambda + \mu)A = \lambda A + \mu A.$$

Megállapodás szerint $\lambda A = A\lambda$.

3. Transzponálás (tükrözés): $A \in \mathbb{R}^{m\times n}$,

$$C = A^T \in \mathbb{R}^{n\times m} \Leftrightarrow c_{ij} = a_{ji} \quad (i = 1,\dots,n,\ j = 1,\dots,m).$$

A transzponálásra fennáll, hogy

$$(A^T)^T = A, \quad (A+B)^T = A^T + B^T.$$

Az A mátrix *szimmetrikus*, ha $A^T = A$.

4. Szorzás: $A \in \mathbb{R}^{m\times k}$, $B \in \mathbb{R}^{k\times n}$,

$$C = AB \in \mathbb{R}^{m\times n} \Leftrightarrow c_{ij} = \sum_{t=1}^{k} a_{it}b_{tj} \quad (i = 1,\dots,m,\ j = 1,\dots,n).$$

A szorzatmátrix (i,j) indexű elemét úgy kapjuk, hogy az i-edik sort szorozzuk a j-edik oszloppal, azaz

$$c_{ij} = [a_{i1},\dots,a_{ik}] \begin{bmatrix} b_{1j} \\ \vdots \\ b_{kj} \end{bmatrix}.$$

A mátrixszorzás fontos tulajdonságai:

$$\begin{aligned} (AB)C &= A(BC), \\ A(B+C) &= AB + AC, \\ (A+B)C &= AC + BC, \\ (AB)^T &= B^T A^T. \end{aligned}$$

A szorzás nem kommutatív, tehát általában

$$AB \neq BA. \tag{1}$$

Konvenció: A mátrix és mátrix-vektor műveletek felírásánál feltesszük, hogy az ott szereplő mátrixok, ill. vektorok méretei olyanok, amelyek lehetővé teszik az adott műveletet.

Definíció. *Az egyetlen sorból, vagy egyetlen oszlopból álló mátrixot vektornak nevezzük.*

Sorvektor: $x = [x_1,\dots,x_n]$.

Oszlopvektor:

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^n,$$

$\mathbb{R}^n$ az n komponensű oszlopvektorok halmaza (tulajdonképpen $\mathbb{R}^n \equiv \mathbb{R}^{n\times 1}$).

Más lehetséges alakok:

oszlopvektor: $x = [x_1,\dots,x_n]^T$,

sorvektor

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}^T \in \mathbb{R}^n.$$

i-edik egységvektor: i-edik komponense 1, a többi pedig 0, azaz

$$e_i = [0, \ldots, 0, 1, 0, \ldots, 0]^T \in \mathbb{R}^n.$$

Definíció. *$x, y \in \mathbb{R}^n$ skaláris szorzata*

$$x^T y = \sum_{i=1}^{n} x_i y_i.$$

1.1 Mátrixok részekre bontása (particionálása)

Az A $m \times k$ típusú mátrixot sorok szerint particionáljuk, ha

$$A = \begin{bmatrix} a_1^T \\ \vdots \\ a_i^T \\ \vdots \\ a_m^T \end{bmatrix},$$

ahol $a_i^T = [a_{i1}, \ldots, a_{ik}]$ az i-edik (k-dimenziós) sorvektort jelöli.

A B $k \times n$ típusú mátrixot oszlopok szerint particionáljuk, ha

$$B = [b_1, \ldots, b_n],$$

ahol

$$b_i = \begin{bmatrix} b_{1i} \\ \vdots \\ b_{ki} \end{bmatrix}$$

az i-edik (k-dimenziós) oszlopvektort jelöli.

A fenti particionálások felhasználásával

$$AB = \left[a_i^T b_j\right]_{i,j=1}^{m,n}.$$

Tehát AB a sorok és oszlopok skalárszorzataiból álló mátrix.

Az AB mátrixszorzatot felírhatjuk még a következő alakokban is

$$AB = [Ab_1, \ldots, Ab_n], \quad AB = \begin{bmatrix} a_1^T B \\ \vdots \\ a_m^T B \end{bmatrix}.$$

Általános partícionálás az $A \in \mathbb{R}^{m\times n}$ mátrix esetén

$$A = \begin{bmatrix} A_{11} & \dots & A_{1j} & \dots & A_{1r} \\ \vdots & & \vdots & & \vdots \\ A_{i1} & \dots & A_{ij} & \dots & A_{ir} \\ \vdots & & \vdots & & \vdots \\ A_{p1} & \dots & A_{pj} & \dots & A_{pr} \end{bmatrix},$$

ahol $A_{ij} \in \mathbb{R}^{m_i\times n_j}$ $(i = 1, \dots, p,\ j = 1, \dots, r)$ és

$$\sum_{i=1}^{p} m_i = m, \quad \sum_{j=1}^{r} n_j = n.$$

Az azonos sorban álló blokkok sorainak száma azonos.

Az azonos oszlopban álló blokkok oszlopainak száma azonos.

Azonosan particionált mátrixok összegzését és skalárral való szorzását blokkonként végezhetjük, úgy mintha a blokkok számok lennének.

Particionált mátrixok blokkonkénti szorzásánál az első mátrix oszlopok szerinti particionálása meg kell, hogy egyezzen a második tényező sorok szerinti particionálásával.

Például az

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \in \mathbb{R}^{m\times n}, \quad B = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} \in \mathbb{R}^{n\times p}$$

particionált mátrixok szorzata akkor lehetséges, ha $A_{11} \in \mathbb{R}^{r\times s}$ és $B_{11} \in \mathbb{R}^{s\times\sigma}$. Ekkor

$$C = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} = \begin{bmatrix} A_{11}B_{11} + A_{12}B_{21} & A_{11}B_{12} + A_{12}B_{22} \\ A_{21}B_{11} + A_{22}B_{21} & A_{21}B_{12} + A_{22}B_{22} \end{bmatrix}.$$

1.2 Speciális mátrixok

Definíció. *Az $I \in \mathbb{R}^{n\times n}$ mátrix egységmátrix, ha*

$$I = \begin{bmatrix} 1 & 0 & \dots & \dots & 0 \\ 0 & 1 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & 1 & 0 \\ 0 & \dots & \dots & 0 & 1 \end{bmatrix}.$$

Az egységmátrixra fennáll, hogy minden $A \in \mathbb{R}^{n\times n}$ esetén

$$AI = IA = A.$$

Definíció. *A $D \in \mathbb{R}^{n\times n}$ diagonálmátrix, ha*

$$D = \begin{bmatrix} d_1 & 0 & \dots & \dots & 0 \\ 0 & d_2 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & d_{n-1} & 0 \\ 0 & \dots & \dots & 0 & d_n \end{bmatrix}.$$

A diagonálmátrixra fennáll, hogy minden $A \in \mathbb{R}^{n\times m}$ és $B \in \mathbb{R}^{m\times n}$ esetén

$$DA = D \begin{bmatrix} a_1^T \\ \vdots \\ a_n^T \end{bmatrix} = \begin{bmatrix} d_1 a_1^T \\ \vdots \\ d_n a_n^T \end{bmatrix}, \quad BD = [b_1, \dots, b_n]\, D = [d_1 b_1, \dots, d_n b_n].$$

A D diagonálmátrixot $\operatorname{diag}(d_1, \dots, d_n)$, vagy $\operatorname{diag}(d_i)$ $(i = 1, \dots n)$ is jelölheti.
Definíció. *Az $0 \in \mathbb{R}^{m\times n}$ mátrix zérusmátrix, ha minden eleme 0, azaz*

$$0 = \begin{bmatrix} 0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 0 \end{bmatrix}.$$

A zérusmátrixra fennáll, hogy minden A mátrix esetén

$$A + 0 = A, \quad A0 = 0.$$

1.3 Mátrixok inverze és determinánsa

Definíció. *Az $X \in \mathbb{R}^{n\times n}$ mátrixot az $A \in \mathbb{R}^{n\times n}$ mátrix inverzének nevezzük, ha $AX = XA = I$.*

Ha az X inverz mátrix létezik, akkor egyértelmű. Jelölése $A^{-1} = X$.
Az inverz mátrix tulajdonságai:

$$\left(A^{-1}\right)^{-1} = A, \quad (AB)^{-1} = B^{-1}A^{-1}, \quad \left(A^T\right)^{-1} = \left(A^{-1}\right)^T := A^{-T}.$$

Jelölje $A(i)$ azt az $(n-1)\times(n-1)$-es részmátrixot, amelyet az

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ \vdots & \vdots & & \vdots \\ a_{i1} & a_{i2} & \dots & a_{in} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix}$$

mátrixból az első oszlop és az i-edik sor elhagyásával kapunk.
Definíció. *Az $A \in \mathbb{R}^{n\times n}$ $(n \geq 2)$ négyzetes mátrix determinánsát a*

$$\det(A) = a_{11}a_{22} - a_{12}a_{21}, \quad n = 2$$

$$\det(A) = \sum_{i=1}^{n} (-1)^{i+1} a_{i1} \det(A(i)), \quad n \geq 3.$$

előírások definiálják.

Megjegyezzük, hogy az egy elemű $[a_{11}]$ mátrix determinánsán az a_{11} értéket értjük.

Tétel. *Az $A \in \mathbb{R}^{n\times n}$ mátrixnak akkor és csak akkor van inverze, ha* $\det(A) \neq 0$.

1.4 Vektorok és mátrixok normája

Definíció. *Az $f : \mathbb{R}^n \to \mathbb{R}$ függvényt vektornormának nevezzük, ha*

$$\begin{aligned} f(x) &\geq 0 \quad (\forall x \in \mathbb{R}^n), \quad f(x) = 0 \Leftrightarrow x = 0, && (2)\\ f(\lambda x) &= |\lambda| f(x) \quad (\forall x \in \mathbb{R}^n,\ \forall \lambda \in \mathbb{R}), && (3)\\ f(x+y) &\leq f(x) + f(y) \quad (\forall x, y \in \mathbb{R}^n). && (4) \end{aligned}$$

A vektornorma szokásos jelölése: $\|x\|$.

A fontosabb vektornormák a következők:

$$\begin{aligned} \|x\|_1 &= \sum_{i=1}^{n} |x_i|, && (5)\\ \|x\|_2 &= \left(\sum_{i=1}^{n} x_i^2\right)^{\frac{1}{2}} \quad \text{(euklideszi norma)}, && (6)\\ \|x\|_\infty &= \max_{1\leq i\leq n} |x_i| \quad \text{(maximum norma)}. && (7) \end{aligned}$$

Definíció. *Az $x, y \in \mathbb{R}^n$ $(x, y \neq 0)$ vektorok szöge θ, amelynek koszinuszát a*

$$\cos(\theta) = \frac{x^T y}{\|x\|_2 \|y\|_2}$$

összefüggés definiálja.

Definíció. *Az $f : \mathbb{R}^{n\times n} \to \mathbb{R}$ függvényt mátrixnormának nevezzük, ha*

$$\begin{aligned} f(A) &\geq 0 \quad \left(\forall A \in \mathbb{R}^{n\times n}\right), \quad f(A) = 0 \Leftrightarrow A = 0, && (8)\\ f(\lambda A) &= |\lambda| f(A) \quad \left(\forall A \in \mathbb{R}^{n\times n},\ \forall \lambda \in \mathbb{R}\right), && (9)\\ f(A+B) &\leq f(A) + f(B) \quad \left(\forall A, B \in \mathbb{R}^{n\times n}\right). && (10) \end{aligned}$$

A mátrixnorma szokásos jelölése: $\|A\|$.

A leggyakrabban használt mátrixnormák:

$$\|A\|_1 = \max_{1\leq j\leq n} \sum_{i=1}^{n} |a_{ij}| \quad \text{(oszlopösszeg norma)}, \qquad (11)$$

$$\|A\|_2 = \{A^T A \text{ legnagyobb sajátértéke}\}^{\frac{1}{2}} \quad \text{(spektrálnorma)}, \qquad (12)$$

$$\|A\|_\infty = \max_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}| \quad \text{(sorösszeg norma)}\,, \tag{13}$$

$$\|A\|_F = \left(\sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij}^2 \right)^{\frac{1}{2}} \quad \text{(Frobenius norma)}\,. \tag{14}$$

Definíció. *A* $\|.\|_M : \mathbb{R}^{n\times n} \to \mathbb{R}$ *mátrixnormát a* $\|.\|_V : \mathbb{R}^n \to \mathbb{R}$ *vektornorma által indukált mátrixnormának nevezzük, ha*

$$\|A\|_M = \max\left\{ \|Ax\|_V : \|x\|_V = 1 \right\}. \tag{15}$$

Tétel. *Indukált mátrixnormában* $\|AB\| \le \|A\|\,\|B\|$ *(*$\forall A, B \in \mathbb{R}^{n\times n}$*).*

Bizonyítás. Először igazoljuk, hogy indukált mátrixnormában

$$\|Ax\| \le \|A\|\,\|x\| \quad (x \in \mathbb{R}^n)\,.$$

Ha $x \neq 0$, az indukált mátrixnorma definíciója alapján

$$\|Ax\| = \left\| A\,\|x\|\,\frac{x}{\|x\|} \right\| = \|x\| \left\| A\frac{x}{\|x\|} \right\| \le \|x\|\,\|A\|\,,$$

ahonnan

$$\|ABx\| \le \|A\|\,\|Bx\| \le \|A\|\,\|B\|\,\|x\|$$

és a tétel állítása következik. □

Megjegyezzük, hogy az állítás nem minden mátrixnormára igaz

Az indukált normákhoz hasonló (de velük nem azonos) fogalom a következő.

Definíció. *A* $\|.\|_M : \mathbb{R}^{n\times n} \to \mathbb{R}$ *mátrixnorma kompatibilis a* $\|.\|_V : \mathbb{R}^n \to \mathbb{R}$ *vektornormával, ha* $\|Ax\|_V \le \|A\|_M\,\|x\|_V$.

Az $\|A\|_F$ Frobenius norma kompatibilis a $\|x\|_2$ vektornormával.

Az indukált mátrixnormák az őket indukáló vektornormákkal kompatibilisek.

2 LINEÁRIS EGYENLERENDSZEREK MEGOLDÁSA

A lineáris egyenletrendszerek általános alakja m egyenlet és n ismeretlen esetén:

$$\begin{array}{rcl} a_{11}x_1 + \ldots + a_{1j}x_j + \ldots + a_{1n}x_n &=& b_1 \\ & \vdots & \\ a_{i1}x_1 + \ldots + a_{ij}x_j + \ldots + a_{in}x_n &=& b_i \\ & \vdots & \\ a_{m1}x_1 + \ldots + a_{mj}x_j + \ldots + a_{mn}x_n &=& b_m \end{array} \tag{16}$$

Tömörebb alakban:

$$Ax = b, \tag{17}$$

ahol

$$A = [a_{ij}]_{i,j=1}^{m,n} \in \mathbb{R}^{m\times n},\ x \in \mathbb{R}^n,\ b \in \mathbb{R}^m.$$

Ha $m < n$, akkor az egyenletrendszer *alulhatározott.*

Ha $m > n$, akkor az egyenletrendszer *túlhatározott.*

Ha $m = n$, akkor az egyenletrendszer *négyzetes.*

Az egyenletrendszerek geometriai tartalma:

Definíció. *Az $\mathbb{R}^n$ euklideszi tér d ($d \in \mathbb{R}^n$) normálvektorú és $x_0 \in \mathbb{R}^n$ ponton átmenő hipersíkját az*

$$(x - x_0)^T d = 0 \tag{18}$$

egyenletet kielégítő $x \in \mathbb{R}^n$ pontok határozzák meg.

A hipersík egyenlete más formában:

$$x^T d = x_0^T d. \tag{19}$$

Felhasználva az

$$A = \begin{bmatrix} a_1^T \\ \vdots \\ a_i^T \\ \vdots \\ a_m^T \end{bmatrix} \quad \left(a_i^T = [a_{i1}, \ldots, a_{in}]\right)$$

felbontást, az $Ax = b$ egyenletrendszer ekvivalens alakja:

$$\begin{aligned} a_1^T x &= b_1 \\ &\vdots \\ a_m^T x &= b_m \end{aligned} \tag{20}$$

Tehát a lineáris egyenletrendszer megoldása m hipersík közös része.

Három eset lehetséges:

(i) az egyenletrendszernek nincs megoldása,

(ii) az egyenletrendszernek pontosan egy megoldása van,

(iii) az egyenletrendszernek végtelen sok megoldása van.

Definíció. *Ha az $Ax = b$ lineáris egyenletrendszernek legalább egy megoldása van, akkor az egyenletrendszert konzisztensnek nevezzük. Ha az egyenletrendszernek nincs megoldása, akkor az egyenletrendszert inkonzisztensnek nevezzük.*

Például az $x + 2y = 1$, $x + 2y = 4$ egyenletrendszer inkonzisztens.

Az $Ax = b$ egyenletrendszert felírható az ekvivalens

$$\sum_{i=1}^{n} x_i a_i = x_1 a_1 + \ldots + x_n a_n = b$$

alakban is, ahol a_i az A mátrix i-edik oszlopa.

Definíció. *A $\sum_{i=1}^n x_i a_i$ összeg az $\{a_i\}_{i=1}^n$ vektorok lineáris kombinációja.*

Tétel. *Az egyenletrendszer akkor és csak akkor oldható meg, ha b kifejezhető az A oszlopvektorainak lineáris kombinációjaként.*

Definíció. *Az $\{a_i\}_{i=1}^k \subseteq \mathbb{R}^m$ vektorok lineárisan összefüggők, ha létezik $x \in \mathbb{R}^k$ $(x \neq 0)$, hogy*

$$\sum_{i=1}^{k} x_i a_i = 0. \tag{21}$$

Ha nincs ilyen $x \neq 0$ vektor, akkor az $\{a_i\}_{i=1}^k$ vektorok lineárisan függetlenek.

A megoldhatóság egy "másik jellemzését" adhatjuk a *rang* fogalmával:

$$\operatorname{rank}(A) = \begin{array}{c}\text{lineárisan független oszlop- vagy sorvektorok} \\ \text{maximális száma}\end{array} \tag{22}$$

1. Az $Ax = b$ egyenletrendszer akkor és csak akkor megoldható, ha $\operatorname{rank}(A) = \operatorname{rank}([A, b])$.
2. Ha $\operatorname{rank}(A) = \operatorname{rank}([A, b]) = n$, akkor az $Ax = b$ egyenletrendszernek pontosan egy megoldása van.

A továbbiakban csak négyzetes egyenletrendszerekkel foglalkozunk. Feltesszük tehát, hogy $m = n$. Ismert a következő

Tétel. *Az $Ax = b$ $(A \in \mathbb{R}^{n\times n}, b \in \mathbb{R}^n)$ egyenletrendszernek akkor és csak akkor van pontosan egy megoldása, ha létezik A^{-1}. Ekkor a megoldás $x = A^{-1}b$.*

Tétel. *Az $Ax = 0$ $(A \in \mathbb{R}^{n\times n})$ homogén lineáris egyenletrendszernek akkor és csak akkor van $x \neq 0$ nemtriviális megoldása, ha* $\det(A) = 0$.

2.1 Háromszögmátrixú egyenletrendszerek

Definíció. *Az $A = [a_{ij}]_{i,j=1}^n$ mátrix alsó háromszög alakú, ha minden $i < j$ esetén $a_{ij} = 0$.*

Sematikusan

$$\begin{bmatrix} * & 0 & \dots & \dots & 0 \\ * & * & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \vdots \\ \vdots & & & * & 0 \\ * & \dots & \dots & * & * \end{bmatrix}.$$

Definíció. *Az $A = [a_{ij}]_{i,j=1}^n$ mátrix felső háromszög alakú, ha minden $i > j$ esetén $a_{ij} = 0$.*

A felső háromszögmátrixok alakja:

$$\begin{bmatrix} * & * & \dots & \dots & * \\ 0 & * & & & \vdots \\ \vdots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & * & * \\ 0 & \dots & \dots & 0 & * \end{bmatrix}.$$

Megjegyzés. *A diagonálmátrixok (köztük a zérusmátrix) egyidejűleg alsó és felső háromszögmátrixok is.*

Alsó vagy felső háromszögmátrixok esetén $\det(A) = a_{11}a_{22}\dots a_{nn}$.

2.2 A háromszögmátrixú egyenletrendszerek megoldása

1. Tekintsük az

$$\begin{array}{ccccccccc} a_{11}x_1 & & & & & & & = & b_1 \\ \vdots & \ddots & & & & & & & \vdots \\ a_{i1}x_1+ & \dots & +a_{ii}x_i & & & & & = & b_i \\ \vdots & & \vdots & \ddots & & & & & \vdots \\ a_{n1}x_1+ & \dots & +a_{ni}x_i & \dots & +a_{nn}x_n & & & = & b_n \end{array}$$

alsó háromszögmátrixú egyenletrendszert!

Az egyenletrendszer akkor és csak akkor oldható meg egyértelműen, ha $a_{11} \neq 0,\dots, a_{nn} \neq 0$.

Az alsó háromszögmátrixú egyenletrendszer megoldását adja a következő algoritmus:

$$\begin{array}{l} x_1 = b_1/a_{11} \\ \textbf{for } i = 2 : n \\ \quad x_i = (b_i - \sum_{j=1}^{i-1} a_{ij}x_j)/a_{ii} \\ \textbf{end} \end{array} \tag{23}$$

2. Tekintsük most az

$$\begin{array}{ccccccccc} a_{11}x_1+ & \dots & +a_{1i}x_i+ & \dots & +a_{1n}x_n & = & b_1 \\ \ddots & \vdots & & & \vdots & & \vdots \\ & a_{ii}x_i+ & \dots & +a_{in}x_n & = & b_i \\ & & \ddots & \vdots & & \vdots \\ & & & a_{nn}x_n & = & b_n \end{array}$$

felső háromszögmátrixú egyenletrendszert! Az egyenletrendszer akkor és csak akkor oldható meg egyértelműen, ha $a_{11} \neq 0,\dots, a_{nn} \neq 0$. A felső háromszögmátrixú egyenletrendszer megoldását a következő, ún. *visszahelyettesítő* algoritmus adja:

$$\begin{array}{l} x_n = b_n/a_{nn} \\ \textbf{for } i = n-1 : -1 : 1 \\ \quad x_i = (b_i - \sum_{j=i+1}^{n} a_{ij}x_j)/a_{ii} \\ \textbf{end} \end{array} \tag{24}$$

2.3 A Gauss-módszer

A Gauss-féle eliminációs módszer két fázisból áll:

I. Azonos átalakításokkal az $Ax = b$ egyenletrendszert felső háromszög alakúra hozzuk:

II. A kapott felső háromszögmátrixú egyenletrendszert a (24) Algoritmussal megoldjuk.

A felső háromszög alakra hozás:

$$\begin{array}{rcl} a_{11}x_1 + \ldots + a_{1j}x_j + \ldots + a_{1n}x_n & = & b_1 \\ & \vdots & \\ a_{i1}x_1 + \ldots + a_{ij}x_j + \ldots + a_{in}x_n & = & b_i \\ & \vdots & \\ a_{n1}x_1 + \ldots + a_{nj}x_j + \ldots + a_{nn}x_n & = & b_n \end{array}$$

Ha $a_{11} \neq 0$, akkor az a_{11} alatti x_1 együtthatókat nullává tesszük (kinullázzuk) úgy, hogy az i-edik sorból kivonjuk $(i = 2, \ldots, n)$ az első sor γ-szorosát:

$$(a_{i1} - \gamma a_{11})x_1 + (a_{i2} - \gamma a_{12})x_2 + \ldots + (a_{in} - \gamma a_{1n})x_n = b_i - \gamma b_1. \qquad (25)$$

Az $a_{i1} - \gamma a_{11} = 0$ feltételből kapjuk, hogy $\gamma = \frac{a_{i1}}{a_{11}}$.

Így az első oszlop a_{11} alatti kinullázását a

$$\left.\begin{array}{l} \gamma = a_{i1}/a_{11} \\ a_{ij} = a_{ij} - \gamma a_{1j} \quad (j = 2, \ldots, n) \\ b_i = b_i - \gamma b_1 \end{array}\right\} \quad (i = 2, \ldots, n) \qquad (26)$$

algoritmussal végezzük.

Vegyük észre, hogy az algoritmus felülírja az A mátrix $2 \leq i, j \leq n$ indexű és a b vektor $2 \leq i \leq n$ indexű elemeit (a 0-kat viszont feleslegesen nem írja be az alsó háromszög részbe).

A felülírt elemeknél is megtartva az eredeti jelölést, a kapott ekvivalens egyenletrendszer alakja:

$$\begin{array}{ccccccccc} a_{11}x_1 & + & a_{12}x_2 & + & \ldots & + & a_{1n}x_n & = & b_1 \\ & & a_{22}x_2 & + & \ldots & + & a_{2n}x_n & = & b_2 \\ & & \vdots & & & & \vdots & & \vdots \\ & & a_{n2}x_2 & + & \ldots & + & a_{nn}x_n & = & b_n \end{array} . \qquad (27)$$

Ezt szétbonthatjuk az n ismeretlent tartalmazó első egyenletre és az $n-1$ ismeretlent tartalmazó kisebb $(n-1)\times(n-1)$-es egyenletrendszerre. Ha $a_{22} \neq 0$, akkor a kisebb egyenletrendszeren megismételjük az előző lépést és így tovább.

Tegyük fel, hogy a $(k-1)$-edik oszlopban a kinullázást már elvégeztük és az

$$\begin{array}{ccccccccccc}
a_{11}x_1+ & \dots & \dots & +a_{1k}x_k & + & \dots & + & a_{1n}x_n & = & b_1 \\
 & \ddots & & \vdots & & & & \vdots & & \vdots \\
 & & \ddots & \vdots & & & & \vdots & & \vdots \\
 & & & a_{kk}x_k & + & \dots & + & a_{kn}x_n & = & b_k \\
 & & & \vdots & & & & \vdots & & \vdots \\
 & & & a_{ik}x_k & + & \dots & + & a_{in}x_n & = & b_i \\
 & & & \vdots & & & & \vdots & & \vdots \\
 & & & a_{nk}x_k & + & \dots & + & a_{nn}x_n & = & b_n
\end{array}$$

egyenletrendszert kaptuk. Ha $a_{kk} \neq 0$, akkor kinullázzuk az a_{kk} alatti x_k együtthatókat. Az i-edik sorból a k-adik sort γ-szorosát kivonva az

$$(a_{ik} - \gamma a_{kk})x_k + (a_{i,k+1} - \gamma a_{k,k+1})x_{k+1} + \ldots + (a_{in} - \gamma a_{kn})x_n = b_i - \gamma b_k \quad (28)$$

egyenlet adódik. Az $a_{ik} - \gamma a_{kk} = 0$ feltételből kapjuk, hogy $\gamma = \frac{a_{ik}}{a_{kk}}$.

A k-adik oszlop a_{kk} alatti kinullázását tehát a

$$\left.\begin{array}{l}
\gamma = a_{ik}/a_{kk} \\
a_{ij} = a_{ij} - \gamma a_{kj} \quad (j = k+1, \ldots, n) \\
b_i = b_i - \gamma b_k
\end{array}\right\} \quad (i = k+1, \ldots, n)$$

algoritmussal végezhetjük el.

A kinullázást mindaddig folytathatjuk, amíg az $a_{kk} \neq 0$ és $k \leq n-1$ feltételek fennállnak. Ha sikerül az A mátrixot felső háromszög alakra hozni, akkor a (24) Algoritmust alkalmazzuk (visszahelyettesítünk). A következőkben $A\,(i,j)$ az A mátrix a_{ij} elemét jelöli.

A GAUSS-MÓDSZER:

I. (eliminációs) fázis:

for $k = 1 : n-1$

for $i = k+1 : n$

$\gamma = A(i,k)/A(k,k)$

$A(i, k+1:n) = A(i, k+1:n) - \gamma * A(k, k+1:n)$

$b_i = b_i - \gamma b_k$

end

end

II. (visszahelyettesítő) fázis:

$x_n = b_n/a_{nn}$

for $i = n-1 : -1 : 1$

$x_i = (b_i - \sum_{j=i+1}^{n} a_{ij}x_j)/a_{ii}$

end

2.4 A Gauss-módszer műveletigénye

A szükséges aritmetikai műveletszám (műveletigény) az egyenletrendszer megoldó eljárások fontos minőségi jellemzője, mert az ilyen algoritmusok számítógépideje nagyjából arányos az aritmetikai műveletigénnyel.

A szorzás és az osztás nagyjából azonos időigényűek. Ezért ezeket összevonva, azonos multiplikatív műveletként (jele M) számoljuk.

A kivonás és összeadás műveleteknél hasonló a helyzet. Itt a közös alapegység az additív művelet (jele A).

Az I. fázis k-adik lépésében a műveletek és műveletszámok a következők:

for $i = k+1 : n$

$\gamma = A(i,k)/A(k,k)$	$\Rightarrow$	M
$A(i, k+1:n) = A(i, k+1:n) - \gamma * A(k, k+1:n)$	$\Rightarrow$	$(n-k)(M+A)$
$b_i = b_i - \gamma b_k$	$\Rightarrow$	$M+A$

end

A ciklus műveletigénye összesen

$$(n-k)\left(M + (n-k)(M+A) + M + A\right),$$

azaz

$$\Big((n-k)^2 + 2(n-k)\Big)M + \Big((n-k)^2 + (n-k)\Big)A.$$

Ezt összegezve a $k = 1, \ldots, n-1$ lépésekre kapjuk, hogy az I. fázis műveletigénye

$$\sum_{i=1}^{n-1}(i^2+2i)M + \sum_{i=1}^{n-1}(i^2+i)A.$$

Felhasználva, hogy

$$\sum_{i=1}^{k} i^2 = \frac{k(k+1)(2k+1)}{6}$$

kapjuk, hogy

$$\left(\frac{(n-1)n(2n-1)}{6}+(n-1)n\right)M+\left(\frac{(n-1)n(2n-1)}{6}+\frac{(n-1)n}{2}\right)A,$$

illetve

$$\left(\frac{n^3}{3}+\frac{n^2}{2}-\frac{5}{6}n\right)M+\left(\frac{n^3}{3}-\frac{n}{3}\right)A.$$

A II. fázis műveletigénye:

$$\begin{array}{lll} x_n = b_n/a_{nn} & \Rightarrow & M \\ \textbf{for } i = n-1:-1:1 & & \\ \quad x_i = (b_i - \sum_{j=i+1}^{n} a_{ij}x_j)/a_{ii} & \Rightarrow & (n-i+1)M+(n-i)A \\ \textbf{end} & & \end{array}$$

Összegezve az $i = n-1:-1:1$ lépések műveletigényét kapjuk, hogy a II. fázis összköltsége

$$M+\sum_{j=2}^{n} jM+\sum_{j=1}^{n-1} jA=\left(\frac{n^2}{2}+\frac{n}{2}\right)M+(\frac{n^2}{2}-\frac{n}{2})A.$$

Az I. és II. fázis költségét összeadva kapjuk a Gauss-módszer számítási összköltségét:

$$\left(\frac{n^3}{3}+n^2-\frac{n}{3}\right)M+\left(\frac{n^3}{3}+\frac{n^2}{2}-\frac{5}{6}n\right)A.$$

Nagy n értékekre az $\frac{n^3}{3}$ együttható válik dominánssá mindkét zárójeles kifejezésben.

Definíció. *Az $f(n)$ mennyiség $O(n^\gamma)$ nagyságrendű, ha van olyan $K>0$ szám, hogy $\|f(n)\| \le Kn^\gamma$ $(\forall n \in N)$.*

Tétel. *A Gauss-módszer műveletigénye $\frac{n^3}{3}+O(n^2)$ multiplikatív és ugyanennyi $\frac{n^3}{3}+O(n^2)$ additív művelet.*

Definíció (C.B. Moler). *1 (régi) flop az a számítási munka, amely az $s = s + x * y$ művelet (1 összeadás + 1 szorzás) elvégzéséhez kell.*

Például az $x^T y$ skalárszorzat $(x, y \in \mathbb{R}^n)$ kiszámítása az

$$s = 0, \quad s = s + x_i y_i \quad (i = 1, \ldots, n)$$

algoritmussal n flop műveletigényű.

Definíció. *1 (új) flop az a számítási munka, amely egy $+, -, *, /$ aritmetikai művelet elvégzéséhez kell.*

Tétel. *A Gauss-módszer műveletigénye $\frac{n^3}{3}+O(n^2)$ régi flop.*

Az $Ax = b$ alakú $n \times n$-es egyenletrendszerek megoldásához szükséges műveletigény gyors mátrixinvertáló eljárásokkal $O(n^{2.808})$ régi flop.

2.5 A főelemkiválasztásos Gauss-módszer

A Gauss-módszer I. fázisában előfordulhat, mondjuk a k-adik lépésben, hogy

$$a_{kk} = 0.$$

Például a

$$\begin{array}{rcrcrcr} & & 4x_2 & + & x_3 & = & 9 \\ x_1 & + & x_2 & + & 3x_3 & = & 6 \\ 2x_1 & - & 2x_2 & + & x_3 & = & -1 \end{array}$$

rendszernél $a_{11} = 0$.

Ilyen esetekben a sorok, vagy az oszlopok felcserélésével megkísérelhetjük elérni, hogy az a_{kk} helyére zérustól különböző elem kerüljön.

A fenti esetben például az első és harmadik sor felcserélésével kapjuk, hogy

$$\begin{array}{rcrcrcr} 2x_1 & - & 2x_2 & + & x_3 & = & -1 \\ x_1 & + & x_2 & + & 3x_3 & = & 6 \\ & & 4x_2 & + & x_3 & = & 9 \end{array}$$

Az első és második oszlop oszlop cseréjével pedig azt kapjuk, hogy

$$\begin{array}{rcrcrcr} 4x_2 & & & + & x_3 & = & 9 \\ x_2 & + & x_1 & + & 3x_3 & = & 6 \\ 2x_2 & - & 2x_1 & + & x_3 & = & -1 \end{array}$$

A sorok cseréjénél az egyenletek (és b megfelelő komponenseinek) sorrendje, az oszlopok cseréjénél pedig a változók sorrendje változik meg.

Általában, így az előző példában is, több választási lehetőségünk is van sor-, vagy oszlopcserére.

Ha azonban az $a_{kk} = 0$ elem alatt minden együttható zérus, akkor az $[a_{ij}]_{i,j=1}^{n,k}$ részmátrix oszlopai lineárisan összefüggők, A szinguláris és az eliminációs eljárás sorcserével sem folytatható. Hasonló a helyzet, ha a_{kk} sorában, tőle jobbra, minden együttható zérus, mert ekkor A ismét szinguláris.

Az a_{kk} elemet k-adik *pivot, vagy főelemnek* nevezzük. A sorok felcserélését úgy, hogy az új pivot elem zérustól különböző legyen, *pivotálási*, vagy *főelemkiválasztási eljárásnak* nevezzük.

A pivot elem megválasztása nagymértékben befolyásolja az eredmények megbízhatóságát.

Példa:

$$\begin{array}{rcrcr} 10^{-17}x & + & y & = & 1 \\ x & + & y & = & 2 \end{array}$$

Ha ezt a pivotálás nélküli Gauss-módszerrel számítógépen megoldjuk, akkor (15 tizedesjegy pontosságú MATLAB számítások esetén) az $x = 0$, $y = 1$ közelítő eredményt kapjuk. A helyes eredmény: $x = \frac{1}{1-10^{-17}}$, $y = 1 - \frac{10^{-17}}{1-10^{-17}}$. Az első és a második egyenlet felcserélésével kapott

$$\begin{array}{rcrcr} x & + & y & = & 2 \\ 10^{-17}x & + & y & = & 1 \end{array}$$

egyenletrendszeren ugyanaz a módszer az $x = 1$, $y = 1$ közelítő megoldást adja. Ez utóbbi közel van a pontos megoldáshoz, míg az első eredmény katasztrófálisan eltér.

Általában is igaz, hogy a közelítő megoldás pontosságát nagymértékben javítja a helyesen megválasztott pivotálás. Pivot elemnek nagy abszolút értékű elemet kell választani.

Két alapvető pivotálási stratégiát használunk.

Részleges főelemkiválasztás: A k-adik lépésben a k-adik oszlop a_{jk} ($k \leq j \leq n$) elemei közül kiválasztjuk a maximális abszolút értékűt. Ha ennek indexe i, akkor a k-adik és az i-edik sort felcseréljük. A pivotálás után teljesül, hogy

$$|a_{kk}| = \max_{k \leq i \leq n} |a_{ik}| .$$

Teljes főelemkiválasztás: A k-adik lépésben az a_{ij} ($k \leq i, j \leq n$) mátrixelemek közül kiválasztjuk a maximális abszolút értékűt. Ha ennek indexe (i, j), akkor a k-adik és az i-edik sort, valamint a k-adik oszlopot és j-edik oszlopot felcseréljük. A pivotálás után teljesül, hogy

$$|a_{kk}| = \max_{k \leq i,j \leq n} |a_{ij}| .$$

Megjegyezzük, hogy oszlopcsere esetén változócsere is történik.

A főelemkiválasztásos Gauss-módszer esetén az I. fázis minden lépésében pivotálást hajtunk végre. A teljes főelemkiválasztást biztonságos stratégiának tekintjük.

A részleges főelemkiválasztás egyéb technikákkal kiegészítve ugyancsak biztonságos és kevesebb extra aritmetikai műveletet igényel. Ezért a gyakorlatban inkább ezt használjuk. Ekkor az I. fázis alakja algoritmikus formában

for $k = 1 : n - 1$
  Határozzuk meg a t indexet, hogy $|A(t,k)| = \max_{k \leq i \leq n} |A(i,k)|$.
  if $k \neq t$
    Cseréljük fel a k-adik és t-edik sort!
  end
  for $i = k + 1 : n$
    $\gamma = A(i,k) / A(k,k)$
    $A(i, k+1 : n) = A(i, k+1 : n) - \gamma * A(k, k+1 : n)$
    $b_i = b_i - \gamma b_k$
  end
end

3 A KLASSZIKUS HIBASZÁMÍTÁS ELEMEI

A klasszikus hibaszámítás alapmodellje a következő:

A pontos értékeket nem ismerjük, csak adott hibakorlátú közelítéseiket. A közelítő értékekkel pontosan végzett műveletek eredményét az ismeretlen

elméleti eredmény közelítésének tekintjük és azt vizsgáljuk, hogy mekkora a közelítés hibája.

Például a $\sqrt{2} \approx 1.41$ közelítés hibája legfeljebb 0.01.

Jelölések és elnevezések:

x pontos érték,
a az x közelítése ($a \approx x$),
$\Delta a = x - a$ a közelítés hibája,
δa az a közelítő érték abszolút hibakorlátja, ha fennáll

$$|x - a| = |\Delta a| \leq \delta a.$$

3.1 Az aritmetikai műveletek abszolút hibái

Legyen x és y két pontos érték, a az x, b pedig az y közelítése. Tegyük fel, hogy az a és b közelítések abszolút hibakorlátai δa, ill. δb, azaz

$$|x - a| = |\Delta a| \leq \delta a, \quad |y - b| = |\Delta b| \leq \delta b.$$

Jelölje $\Diamond$ a $+, -, *, /$ műveletek bármelyikét. Az $a \Diamond b$ művelet eredményét az $x \Diamond y$ elméleti eredmény közelítésének tekintjük és a

$$|\Delta\,(a \Diamond b)| = |(x \Diamond y) - (a \Diamond b)| \leq \delta\,(a \Diamond b)$$

mennyiségre keresünk becsléseket, ahol $\Delta\,(a \Diamond b)$ a művelet hibáját, $\delta(a \Diamond b)$ pedig abszolút hibakorlátját jelöli.

Tétel. *Igazak a következő becslések:*

$$\delta\,(a + b) \leq \delta a + \delta b, \tag{29}$$
$$\delta\,(a - b) \leq \delta a + \delta b. \tag{30}$$

Bizonyítás. Az összeg esetében fennáll, hogy

$$\begin{aligned} |(x + y) - (a + b)| &= |(x - a) + (y - b)| \\ &\leq |x - a| + |y - b| = |\Delta a| + |\Delta b| \leq \delta a + \delta b, \end{aligned}$$

amiből a fenti állítás következik. A különbség esetében hasonlóan kapjuk, hogy

$$\begin{aligned} |(x - y) - (a - b)| &= |(x - a) - (y - b)| \\ &\leq |x - a| + |y - b| = |\Delta a| + |\Delta b| \leq \delta a + \delta b, \end{aligned}$$

ami bizonyítandó volt. □

A szorzat abszolút hibakorlátjára kapjuk, hogy

$$\begin{aligned} |xy - ab| &= |(a + \Delta a)(b + \Delta b) - ab| \\ &= |a\Delta b + b\Delta a + \Delta a \Delta b| \leq |a|\,\delta b + |b|\,\delta a + \delta a \delta b. \end{aligned}$$

Ha $|a| \gg \delta a$ és $|b| \gg \delta b$, akkor a $\delta a \delta b$ másodrendű hibatagot elhanyagolhatjuk és a

$$\delta(ab) \approx |a|\,\delta b + |b|\,\delta a \tag{31}$$

közelítő becslést kapjuk.

Az osztás esetén azt kapjuk, hogy

$$\begin{aligned}\left|\frac{x}{y}-\frac{a}{b}\right| &= \left|\frac{a+\Delta a}{b+\Delta b}-\frac{a}{b}\right| \\ &= \left|\frac{-a\Delta b+b\Delta a}{b(b+\Delta b)}\right| \le \frac{|a|\,|\Delta b|+|b|\,|\Delta a|}{|b|^2\,(1-\frac{|\Delta b|}{|b|})} \le \frac{|a|\,\delta b+|b|\,\delta a}{|b|^2\,(1-\frac{\delta b}{|b|})}.\end{aligned}$$

Ha $|b| \gg \delta b$, akkor a nevezőben lévő $\frac{\delta b}{|b|}$ tagot elhanyagolhatjuk és a

$$\delta(a/b) \approx \frac{|a|\,\delta b+|b|\,\delta a}{|b|^2} \tag{32}$$

közelítő becslést kapjuk.

Tétel. *Fennállnak az alábbi közelítő egyenlőségek:*

$$\delta(ab) \approx |a|\,\delta b+|b|\,\delta a \quad (|a| \gg \delta a,\ |b| \gg \delta b)\,, \tag{33}$$

$$\delta(a/b) \approx \frac{|a|\,\delta b+|b|\,\delta a}{|b|^2} \quad (b \neq 0,\ |b| \gg \delta b)\,. \tag{34}$$

Figyeljük meg, hogy osztás abszolút hibakorlátja 0-hoz közeli b esetén rendkívül nagy lehet!

3.2 Függvényértékek hibája

1. Legyen $f : \mathbb{R} \to \mathbb{R}$ legalább kétszer folyonosan differenciálható függvény, $x \approx a$. Az $f(x)$ helyett $f(a)$-t számoljuk. Az

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(\xi)}{2}(x-a)^2 \quad (\xi \in (a-\delta a, a+\delta a))$$

másodrendű Taylor-formulából kapjuk, hogy

$$|f(x)-f(a)| = \left|f'(a)(x-a)+\frac{f''(\xi)}{2}(x-a)^2\right| \le |f'(a)|\,\delta a + M(\delta a)^2,$$

ahol $M \ge \frac{1}{2}\,|f''(x)|$ $(x \in [a-\delta a, a+\delta a])$. A másodrendű $M(\delta a)^2$ tagot elhanyagolva kapjuk, hogy a függvénybehelyettesítés abszolút hibája

$$\delta\left(f\left(a\right)\right) \approx |f'(a)|\,\delta a. \tag{35}$$

2. Legyen $f : \mathbb{R}^n \to \mathbb{R}$ legalább kétszer folytonosan differenciálható függvény, $x, a \in \mathbb{R}^n$ és $x \approx a$. Legyen $\Delta a = x - a$, $|x_i - a_i| = |\Delta a_i| \le \delta a_i$ $(i = 1, \ldots, n)$ és $\delta a = [\delta a_1, \ldots, \delta a_n]^T$. Az $f(x)$ helyett az $f(a)$ függvényértéket számoljuk. A többváltozós

$$f(x) = f(a) + \nabla f(a)^T(x-a) + \frac{1}{2}(x-a)^T H(a+\xi(x-a))(x-a) \quad (0 < \xi < 1)\,,$$

Taylor-formulát használjuk, ahol $\nabla f(x) = \left[\frac{\partial f(x)}{\partial x_1}, \ldots, \frac{\partial f(x)}{\partial x_n}\right]^T$ és

$$H(x) = \left[\frac{\partial^2 f(x)}{\partial x_i \partial x_j}\right]_{i,j=1}^{n}$$

az ún. Hesse-mátrix. Tegyük fel, hogy $\|H(a+\xi(x-a))\| \leq M$. Ekkor az $\left|x^T y\right| \leq \|x\|_2 \|y\|_2$ és $\|Ax\|_2 \leq \|A\|_F \|x\|_2$ egyenlőtlenségek alapján

$$\begin{aligned}\left|(x-a)^T H(a+\xi(x-a))(x-a)\right| &\leq \|x-a\| \|H(a+\xi(x-a))(x-a)\| \\ &\leq \|H(a+\xi(x-a))\| \|x-a\|^2.\end{aligned}$$

Ezt felhasználva kapjuk, hogy

$$\begin{aligned}|f(x)-f(a)| &\leq \left|\nabla f(a)^T(x-a)\right| + \frac{1}{2}M\|x-a\|^2 \\ &\leq \sum_{i=1}^{n}\left|\frac{\partial f(a)}{\partial x_i}\right|\delta a_i + \frac{1}{2}M\|x-a\|^2\end{aligned}$$

ahonnan a másodrendű $M\|x-a\|^2 = M\|\Delta a\|^2 \leq M\sum_{i=1}^{n}(\delta a_i)^2$ tagot elhanyagolva kapjuk a

$$\delta(f(a)) \approx \sum_{i=1}^{n}\left|\frac{\partial f(a)}{\partial x_i}\right|\delta a_i \tag{36}$$

becslést.

3.3 Az aritmetikai műveletek relatív hibái

Az abszolút hiba sok esetben semmitmondó. Például egy 0.001 nagyságrendű elméleti mennyiség 0.05 abszolút hibakorlátú közelítése nem sokat ér. A $\pi \approx 22/7$ közelítés sok esetben jó lehet, de például a csillagászatban már bizonyosan nem.

Definíció. *Az x szám valamely a közelítő értékének relatív hibája a $\frac{\delta a}{|x|}$ menynyiség.*

Az x pontos érték általában nem ismeretes, ezért a $\frac{\delta a}{|x|}$ helyett a $\frac{\delta a}{|a|}$ közelítést használjuk. Ennek hibájára fennáll, hogy

$$\left|\frac{\delta a}{|x|} - \frac{\delta a}{|a|}\right| = \delta a\frac{||a|-|x||}{|a|\,|x|} \leq \delta a\frac{|a-x|}{|a|\,|x|} \leq \frac{(\delta a)^2}{|a|\,|x|}. \tag{37}$$

A hiba elhanyagolható, ha $|x|$ és $|a|$ lényegesen nagyobb a másodrendű $(\delta a)^2$ menynyiségnél.

Tétel.

$$\frac{\delta(a+b)}{|a+b|} = \max\left\{\frac{\delta a}{|a|}, \frac{\delta b}{|b|}\right\} \quad (ab>0), \tag{38}$$

$$\frac{\delta(a-b)}{|a-b|} = \frac{\delta a+\delta b}{|a-b|} \quad (ab>0), \tag{39}$$

$$\frac{\delta(ab)}{|ab|} \approx \frac{\delta a}{|a|} + \frac{\delta b}{|b|}, \tag{40}$$

$$\frac{\delta(\frac{a}{b})}{\left|\frac{a}{b}\right|} \approx \frac{\delta a}{|a|} + \frac{\delta b}{|b|}. \tag{41}$$

Egymáshoz közeli a és b esetén a kivonás relatív hibája rendkívül nagy lehet!

Példa. *Számítsuk ki a* $\sqrt{1996} - \sqrt{1995}$ *mennyiséget, ha ismertek a* $\sqrt{1996} \approx 44.67$ *és* $\sqrt{1995} \approx 44.66$ *közelítő értékek, amelyek közös abszolút hibakorlátja* 0.01*, a közös relatív hibakorlát pedig* 0.022%*.*

A kivonás elvégzésével kapjuk, hogy $\sqrt{1996}-\sqrt{1995} \approx 0.01$, amelynek relatív hibakorlátja az általános formulából

$$\frac{0.01 + 0.01}{0.01} = 2,$$

azaz 200%. Most lehetőségünk van az elméleti relatív hiba kiszámolására is, ami "csak" 10.66%. Ez a valóságos hiba is jelentős mértékű, a kiinduló adatok hibájához képest kb. 5×10^2-szoros.

A különbség képzését elkerülhetjük a

$$\sqrt{1996} - \sqrt{1995} = \frac{1996 - 1995}{\sqrt{1996} + \sqrt{1995}} = \frac{1}{\sqrt{1996} + \sqrt{1995}} \approx \frac{1}{89.33} \approx 0.01119$$

átalakítással. A számláló pontos érték. A nevező abszolút hibája 0.02, a hányados relatív hibája pedig $0.02/89.33 \approx 0.00022 = 0.022\%$. Ez összhangban van a kiinduló adatok relatív hibáival és lényegesen kisebb, mint amit a közvetlen kivonásnál kaptunk.

Hasonló fogásokat lehet alkalmazni más esetekben is.

3.4 Függvényértékek relatív hibája és a kondíciószám

Legyen ismét $f : \mathbb{R} \to \mathbb{R}$ kétszer folytonosan differenciálható és $x \approx a$. Az $f(x)$ pontos érték helyett számított $f(a)$ érték relatív hibája

$$\frac{\delta(f(a))}{|f(a)|} \approx \frac{|f'(a)|\,\delta a}{|f(a)|}. \tag{42}$$

Érdekesebb ez a mennyiség az a közelítő érték relatív hibájával való összehasonlításban. A

$$\frac{|f(a + \Delta a) - f(a)|}{|f(a)|} : \frac{|\Delta a|}{|a|}$$

mennyiség, amely a relatív hibák hányadosa, azt méri, hogy az a adatban fellépő bizonytalanságot az $f(a)$ függvénybe való behelyettesítés mennyire "nagyítja" meg. Egyszerű átalakításokkal adódik, hogy

$$\frac{|f(a + \Delta a) - f(a)|}{|f(a)|} : \frac{|\Delta a|}{|a|} \approx \frac{|f'(a)|\,|\Delta a|}{|f(a)|} \cdot \frac{|a|}{|\Delta a|} = \frac{|f'(a)|\,|a|}{|f(a)|}.$$

Definíció. *A*

$$c(f,a) = \frac{|f'(a)|\,|a|}{|f(a)|} \tag{43}$$

mennyiséget az $f : \mathbb{R} \to \mathbb{R}$ függvény a pontbeli kondíciószámának nevezzük.

Egy függvényt numerikusan instabilnak, vagy rosszul kondicionáltnak nevezünk, ha nagy a kondíciószáma. A függvény stabil, vagy jól kondicionált, ha a kondíciószám kicsi. Természetesen a kicsi és nagy jelző relatív. Mint később látni fogjuk, ezek a relatív jelzők adott feladat esetén a rendelkezésre álló számítógép aritmetikájától és a közelítés megkövetelt pontosságától függenek.

Példa. Az $f(x) = 1 + \sqrt{x-1}$ és $x > 1$. Ekkor

$$c(f,x) = \frac{|x|}{2\left(\sqrt{x-1} + (x-1)\right)},$$

ami tetszőlegesen nagy lehet, ha x elég közel van 1-hez. Ezért a példa függvénye numerikusan instabil. Ha bevezetjük az új $x = 1 + t$ változót, akkor kapjuk, hogy $g(t) = f(1+t) = 1 + \sqrt{t}$. Ennek a függvénynek a $t > 0$ helyen vett kondíciószáma

$$c(g,t) = \frac{\sqrt{t}}{2 + 2\sqrt{t}}.$$

Ha $t \approx 0$, azaz $x \approx 1$, akkor a kondíciószám kicsi marad. Tehát stabilizáltuk a számítást egy egyszerű átalakítással.

Definíció. *Egy $f(x)$ mennyiség $O\left(g(x)^{\gamma}\right)$ nagyságrendű, ha alkalmas $K > 0$ számmal fennáll $|f(x)| \leq K\,|g(x)|^{\gamma}$ minden szóbajövő x értékre. Egy $h(x)$ mennyiség $o(g(x))$ nagyságrendű, ha $\frac{\|h(x)\|}{g(x)} \to 0$ $(x \to x_0)$.*

Tekintsük az $F = [f_1, \ldots f_n]^T : \mathbb{R}^m \to \mathbb{R}^n$ többváltozós függvényt. Az F'-t a következőképpen értelmezzük:

$$F'(x) = \left[\frac{\partial f_i}{\partial x_j}\right]_{i,j=1}^{n,m}$$

Legyen most F olyan, amelyre fennáll, hogy

$$F(x) = F(a) + F'(a)(x-a) + o(\|x-a\|), \quad (x \to a).$$

Az $F(x)$ helyett $F(a)$-t számoljuk. Az $F(a)$ relatív hibája az a vektor relatív hibájához viszonyítva a következő

$$\frac{\|F(x) - F(a)\|}{\|F(a)\|} : \frac{\|x-a\|}{\|a\|}.$$

A "nagyítási szám" korlátja az a pont egy $\varepsilon > 0$ sugarú nyílt környezetében a

$$c(F,a,\varepsilon) = \sup\left\{ \frac{\|F(x) - F(a)\|}{\|F(a)\|} : \frac{\|x-a\|}{\|a\|} \mid \|x-a\| < \varepsilon,\ x \neq a \right\}$$

mennyiség. Az

$$\frac{\|F(x) - F(a)\|}{\|F(a)\|} : \frac{\|x-a\|}{\|a\|} = \frac{(\|F'(a)(x-a)\| + o(\|x-a\|))}{\|x-a\|} \cdot \frac{\|a\|}{\|F(a)\|}.$$

és

$$\frac{\|F'(a)(x-a)\|}{\|x-a\|} \leq \|F'(a)\|$$

összefüggések miatt

$$\lim_{\varepsilon \to 0} c(F,a,\varepsilon) = \frac{\|a\| \, \|F'(a)\|}{\|F(a)\|}.$$

Definíció. *Az $F : \mathbb{R}^m \to \mathbb{R}^n$ függvény valamely a pontbeli kondíciószáma a*

$$c(F,a) = \frac{\|a\| \, \|F'(a)\|}{\|F(a)\|} \tag{44}$$

mennyiség.

Definiáljuk az $F : \mathbb{R}^n \to \mathbb{R}^n$ leképezést az $Ay = x$ egyenletrendszer megoldásával, azaz legyen $F(x) = A^{-1}x$ ($A \in \mathbb{R}^{n\times n}$, $\det(A) \neq 0$). Ekkor $F' \equiv A^{-1}$ és

$$c(A^{-1},a) = \frac{\|a\| \left\|A^{-1}\right\|}{\|A^{-1}a\|} = \frac{\|Ay\| \left\|A^{-1}\right\|}{\|y\|} \leq \|A\| \left\|A^{-1}\right\| \quad (Ay = a).$$

A jobboldali felső korlátot az A mátrix kondíciószámának hívjuk. Ez a korlát pontos, mert létezik olyan $a \in \mathbb{R}^n$, hogy $c(A^{-1},a) = \|A\| \left\|A^{-1}\right\|$.

3.5 Direkt és inverz hibák

Vizsgáljuk egy $f(x)$ függvényérték kiszámítását.

Definíció. Ha az $\hat{y}$ közelítést számoljuk a pontos $y = f(x)$ érték helyett, akkor a *direkt (forward) hiba* $\Delta y = \hat{y} - y$.

Definíció. Ha egy $x + \Delta x$ értékre fennáll, hogy $\hat{y} = f(x + \Delta x)$, azaz $\hat{y}$ a perturbált (megváltoztatott) $x + \Delta x$ értékhez tartozó pontos függvényérték, akkor a Δx értéket *inverz (backward) hibának* nevezzük.

A kétfajta hibát mutatja a következő ábra:

Az inverz hiba elemzését és becslését *inverz hibaanalízisnek* nevezzük. Ha több inverz hiba is létezik, akkor a legkisebb inverz hiba meghatározása az érdekes.

Definíció. *Az $y = f(x)$ értéket számító algoritmust inverz stabilnak nevezzük, ha bármely x értékre olyan $\hat{y}$ számított értéket ad, amelyre a Δx inverz hiba kicsi.*

A "kicsi" jelző környezetfüggő.

Vizsgáljuk most a direkt és az inverz hiba kapcsolatát.

Tegyük fel, hogy $\hat{y} = f(x + \Delta x)$ és f kétszer folytonosan differenciálható. Ekkor

$$\hat{y} - y = f(x + \Delta x) - f(x) = f'(x)\,\Delta x + \frac{f''(x + \vartheta \Delta x)}{2!}(\Delta x)^2 \quad (\vartheta \in (0,1))$$

és a számított megoldás relatív hibája

$$\frac{\hat{y} - y}{y} = \left(\frac{x f'(x)}{f(x)}\right)\frac{\Delta x}{x} + O\left((\Delta x)^2\right).$$

Innen kapjuk az alábbi, hibaszámítási ökölszabálynak is nevezett

$$\frac{\delta(\hat{y})}{|y|} \leq c(f,x)\,\frac{\delta(x)}{|x|} \tag{45}$$

közelítő egyenlőtlenséget, amely szóban kifejezve a következő:

$$\text{relatív direkt hiba} \leq \text{kondíciószám} \times \text{relatív inverz hiba}. \tag{46}$$

Az egyenlőtlenség azt mutatja, hogy egy rosszul kondicionált probléma számított megoldásának nagy lehet a (relatív) direkt hibája. Egy algoritmust *direkt stabilnak* nevezünk, ha a direkt hiba kicsi. Egy direkt stabil módszer nem feltétlenül inverz stabil. Ha az inverz hiba és a kondíciószám kicsi, akkor az algoritmus direkt stabil.

Példa. Vizsgáljuk az $f(x) = \log x$ függvényt! Ennek kondíciószáma $c(f,x) = c(x) = 1/|\log x|$, amely $x \approx 1$ esetén nagy. Tehát az $x \approx 1$ értékekre a relatív direkt hiba nagy lesz.

4 A LEBEGŐPONTOS HIBAANALÍZIS

A digitális számítógépek egy F véges számhalmazt ábrázolnak és az aritmetikai műveleteket ezekkel a számokkal végzik. Ha a művelet eredménye az F halmazbeli szám, akkor a művelet eredményét pontosan kapjuk meg. Egyébként pedig három eset léphet fel:

- kerekítés ábrázolható (nemzérus) számhoz,
- alulcsordulás (kerekítés 0-hoz),
- túlcsordulás.

Definíció. *A lebegőpontos számok halmaza*

$$F(\beta,t,L,U) = \\ = \left\{\pm m \times \beta^e \mid \tfrac{1}{\beta} \leq m < 1,\ m = 0.d_1d_2\ldots d_t,\ L \leq e \leq U\right\} \cup \{0\}, \tag{47}$$

ahol

- *β a számrendszer alapja,*
- *m a lebegőpontos szám mantisszája a β alapú számrendszerben,*
- *e az ábrázolt szám kitevője (karakterisztikája, exponense),*
- *t a mantissza hossza (az aritmetika pontossága),*
- *L a legkisebb kitevő (alulcsordulási határ kitevője),*
- *U a legnagyobb kitevő (túlcsordulási határ kitevője).*

A három leggyakrabban használt számrendszer a következő:

elnevezés	β	felhasználás
bináris	2	legtöbb számítógép
decimális	10	legtöbb számológép
hexadecimális	16	IBM és hasonló nagyszámítógépek

A mantisszát felírhatjuk az

$$m = 0.d_1 d_2 \dots d_t = \frac{d_1}{\beta} + \frac{d_2}{\beta^2} + \dots + \frac{d_t}{\beta^t} \tag{48}$$

alakban is.

Innen látható, hogy az $\frac{1}{\beta} \le m < 1$ feltétel miatt az első jegyre teljesülnie kell az $1 \le d_1 \le \beta - 1$ egyenlőtlenségnek. A többi számjegyre fennáll, hogy $0 \le d_i \le \beta - 1$ $(i = 2, \dots, t)$. Az ilyen számrendszereket *normalizáltaknak* nevezzük.

A 0 jegyet és a tizedespontot értelemszerűen nem szokás ábrázolni. Ha $\beta = 2$, akkor az első jegy csak az 1 lehet, amelyet szintén nem ábrázolnak.

A (48) felírást használva a $F = F(\beta, t, L, U)$ halmazt megadhatjuk az

$$F = \left\{ \pm \left(\frac{d_1}{\beta} + \frac{d_2}{\beta^2} + \dots + \frac{d_t}{\beta^t} \right) \beta^e \mid \ L \le e \le U \right\} \cup \{0\} \tag{49}$$

alakban is, ahol $0 \le d_i \le \beta - 1$ $(i = 1, \dots, t)$ és $1 \le d_1$.

Példa. Határozzuk meg az $F(\beta, t, L, U)$ halmaz elemeinek számát! A mantissza t számjegye közül az első $\beta - 1$ féle lehet (0 nem!), a többi viszont a $0, 1, \dots, \beta - 1$ bármelyike. Előjeltől eltekintve tehát $(\beta - 1)\beta^{t-1}$ különböző mantissza állítható össze. Az $L \le e \le U$ miatt $U - L + 1$ különböző kitevőnk van. Könnyű belátni, hogy ha két szám akár a mantisszában, akár a kitevőben (vagy mindkettőben) különbözik, akkor nem egyenlőek. Mostmár az előjelet, valamint a zérust is beszámítva, kapjuk az elemek számát: $2(\beta - 1)\beta^{t-1}(U - L + 1) + 1$.

Az F halmaz elemei nem egyenletesen helyezkednek el a számegyenesen! Például $\beta = 2$, $t = 3$, $L = -1$ és $U = 2$ esetén a 33 elemű F halmaz pozitív része

$$\left\{ \frac{1}{4}, \frac{5}{16}, \frac{6}{16}, \frac{7}{16}, \frac{1}{2}, \frac{5}{8}, \frac{6}{8}, \frac{7}{8}, 1, \frac{10}{8}, \frac{12}{8}, \frac{14}{8}, 2, \frac{20}{8}, 3, \frac{28}{8} \right\}.$$

Általában, a β alapú számrendszerben az m mantisszára fennáll, hogy

$$\frac{1}{\beta} \le m = \frac{d_1}{\beta} + \frac{d_2}{\beta^2} + \dots + \frac{d_t}{\beta^t} \le \frac{\beta - 1}{\beta} + \frac{\beta - 1}{\beta^2} + \dots + \frac{\beta - 1}{\beta^t} = 1 - \frac{1}{\beta^t}.$$

Az $\left[\frac{1}{\beta}, 1\right]$ intervallumba eső F-beli szomszédos számok távolsága β^{-t}. Minthogy az F halmaz elemeit a $\pm m \times \beta^e$ számok alkotják, a szomszédos F-beli számok távolsága az exponens értékének megfelelően változik. A szomszédos elemek legnagyobb távolsága β^{U-t}, a legkisebb pedig β^{L-t}.

Az ábrázolható számok nagyságrendjét adja meg a következő

Tétel. *Ha $a \in F$, $a \neq 0$, akkor $M_L \le |a| \le M_U$, ahol*

$$M_L = \beta^{L-1}, \quad M_U = \beta^U (1 - \beta^{-t}).$$

Bizonyítás. Tetszőleges $a \in F$, $a \neq 0$ számra fennáll, hogy

$$|a| = m\beta^e \quad \left(m \in \left[\frac{1}{\beta}, 1 - \frac{1}{\beta^t} \right] \right),$$

ahonnan $L \le e \le U$ miatt

$$\beta^{L-1} \le \frac{1}{\beta}\beta^e \le m\beta^e \le \beta^e\left(1-\beta^{-t}\right) \le \beta^U\left(1-\beta^{-t}\right).$$

Ezzel az állitást igazoltuk. □

Legyen $a, b \in F$ és jelölje $\Diamond$ a négy aritmetikai művelet $(+, -, *, /)$ bármelyikét. A következő esetek lehetségesek:

(1) $a\Diamond b \in F$ (pontos eredmény),

(2) $|a\Diamond b| > M_U$ (aritmetikai túlcsordulás),

(3) $0 < |a\Diamond b| < M_L$ (aritmetikai alulcsordulás),

(4) $a\Diamond b \notin F$, $M_L < |a\Diamond b| < M_U$ (nem ábrázolható eredmény).

Az utolsó két esetben a lebegőpontos aritmetika az $a\Diamond b$ eredményhez hozzárendeli a legközelebbi F-beli számot. Ha két szomszédos F-beli szám az $a\Diamond b$ eredménytől egyformán távol van, akkor általában a nagyobbik számhoz kerekítünk.

Például ötjegyű decimális aritmetika esetén a 2.6457513 számot a 2.6458 számhoz kerekítjük.

Legyen $G = [-M_U, M_U]$. Világos, hogy $F \subset G$. Legyen $x \in G$. A kerekítéssel x-hez rendelt F-beli számot jelölje $fl(x)$. Az $x \to fl(x)$ leképezést *kerekítésnek* nevezzük. Legyen $u = \frac{1}{2}\beta^{1-t}$ az *egységnyi kerekítés mértéke*. Igaz a

Tétel. *Ha $x \in G$, akkor*

$$fl(x) = x(1+\varepsilon), \quad |\varepsilon| \le u.$$

Bizonyítás. Az általánosság megszorítása nélkül feltehetjük, hogy $x > 0$. Tegyük fel, hogy az x számot közrefogó szomszédos F-beli számok $m_1\beta^e$ és $m_2\beta^e$. Ekkor fennáll, hogy

$$m_1\beta^e \le x \le m_2\beta^e,$$

ahol $\frac{1}{\beta} \le m_1 < m_2 \le 1-\beta^{-t}$ és $m_2 - m_1 = \beta^{-t}$. Mármost $fl(x) = m_1\beta^e$, vagy $fl(x) = m_2\beta^e$. Bármelyik választás esetében igaz, hogy

$$|fl(x) - x| \le \frac{|m_2 - m_1|}{2}\beta^e = \frac{\beta^{e-t}}{2}.$$

Ezért a kerekítés relatív hibájára fennáll, hogy

$$\frac{|fl(x)-x|}{|x|} \le \frac{|fl(x)-x|}{m_1\beta^e} \le \frac{\beta^{e-t}}{2m_1\beta^e} = \frac{\beta^{-t}}{2m_1} \le \frac{1}{2}\beta^{1-t} = u.$$

Fennáll tehát, hogy $fl(x) - x = \lambda x u$, ahol $|\lambda| \le 1$. Ezt átrendezve kapjuk, hogy

$$fl(x) = x(1+\lambda u),$$

ahol az $\varepsilon = \lambda u$ számra teljesül, hogy $|\varepsilon| = |\lambda u| \le u$. Ha $(1-\beta^{-t})\beta^e \le x \le \beta^e$, akkor a bizonyítás az $m_2 = 1$ választással érvényben marad. Ezzel a tételt maradéktalanul igazoltuk. □

A tétel tulajdonképpen azt mondja ki, hogy a lebegőpontos aritmetikában a kerekítés relatív hibája korlátos és ez a korlát u, az egységnyi kerekítés mértéke.

A $\epsilon_M = 2u = \beta^{1-t}$ értéket szokás *gépi epszilonnak* is nevezni. Az ϵ_M az 1 és a hozzá legközelebbi 1-nél nagyobb szám távolsága. Bináris alap esetén az

$x = 1$
while $1 + x > 1$
  $x = x/2$
end

algoritmussal határozhatjuk meg $\epsilon_M/2$ értékét. A MATLAB rendszerben $\epsilon_M \approx 2.2204 \times 10^{-16}$.

A lebegőpontos aritmetikai műveletek eredményére vonatkozóan a következő feltevéssel élünk (szabvány modell):

$$fl(a \Diamond b) = (a \Diamond b)\,(1 + \varepsilon)\,, \quad |\varepsilon| \le u \quad (a, b \in F)\,. \tag{50}$$

Az IEEE aritmetikai szabvány, amelyet később ismertetünk, kielégíti ezt a feltevést. A feltevés fontos következménye, hogy $a \Diamond b \neq 0$ esetén a műveletek relatív hibájára ugyancsak teljesül, hogy

$$\frac{|fl(a \Diamond b) - (a \Diamond b)|}{|a \Diamond b|} \le u.$$

Tehát az aritmetikai műveletek relatív hibája kicsi.

Vannak bizonyos lebegőpontos aritmetikák, amelyek nem elégítik ki a (50) feltevést. Ennek az az oka, hogy a kivonásnál nincs egy ún. ellenőrző jegyük.

Az egyszerűség kedvéért vizsgáljuk az $1 - 0.111$ különbséget háromjegyű bináris aritmetikában. Az első lépésben a kitevőket azonos értékre hozzuk

$$\begin{array}{rccccccc} & 2 & \times & 0 & . & 1 & 0 & 0 \\ - & 2 & \times & 0 & . & 0 & 1 & 1 \quad 1 \end{array}$$

Ha a számítást négy értékes jegyre végezzük, akkor az eredmény

$$\begin{array}{rccccccc} & 2^1 & \times & 0 & . & 1 & 0 & 0 \\ - & 2^1 & \times & 0 & . & 0 & 1 & 1 \quad 1 \\ \hline & 2^1 & \times & 0 & . & 0 & 0 & 0 \quad 1 \end{array}$$

amelyből a normalizált eredmény $2^{-2} \times 0.100$. Vegyük észre, hogy a kivonásra került szám nem normalizált, mert első jegye 0. A felhasznált ideiglenes negyedik mantisszajegyet, ellenőrző jegynek nevezzük. Ha nincs ilyen ellenőrző jegy, akkor a megfelelő számítások

$$\begin{array}{rccccccc} & 2^1 & \times & 0 & . & 1 & 0 & 0 \\ - & 2^1 & \times & 0 & . & 0 & 1 & 1 \\ \hline & 2^1 & \times & 0 & . & 0 & 0 & 1 \end{array}$$

amelyből a normalizált eredmény $2^{-1} \times 0.100$. Ennek relatív hibája 100%. Nincs ellenőrző jegye a CRAY szuperszámítógépeknek, valamint egy sor zsebkalkulátornak.

Ha nincs ellenőrző jegy, akkor a műveletek eredményeire az

$$fl(x \pm y) = x(1+\alpha) \pm y(1+\beta), \quad |\alpha|, |\beta| \le u, \tag{51}$$

$$fl(x \diamondsuit y) = (x \diamondsuit y)(1+\delta), \quad |\delta| \le u, \quad \diamondsuit = *, /. \tag{52}$$

összefüggések teljesülnek.

Tegyük fel a továbbiakban, hogy van ellenőrző jegy a kivonásnál és teljesül a (50) feltevés. Vezessük be a következő jelöléseket:

$$|z| = [|z_1|, \ldots, |z_n|]^T \quad (z \in \mathbb{R}^n), \tag{53}$$

$$|A| = [|a_{ij}|]_{i,j=1}^{m,n} \quad \left(A \in \mathbb{R}^{m \times n}\right), \tag{54}$$

$$A \le B \Leftrightarrow a_{ij} \le b_{ij} \quad \left(A, B \in \mathbb{R}^{m \times n}\right). \tag{55}$$

Igazolhatók az alábbi eredmények, ahol E az aktuális művelet hibáját (hibamátrixát) jelöli:

$$\left|fl\left(x^T y\right) - x^T y\right| \le 1.01nu\,|x|^T |y| \quad (nu \le 0.01), \tag{56}$$

$$fl(\alpha A) = \alpha A + E \quad (|E| \le u\,|\alpha A|), \tag{57}$$

$$fl(A+B) = (A+B) + E \quad (|E| \le u\,|A+B|), \tag{58}$$

$$fl(AB) = AB + E \quad \left(|E| \le nu\,|A|\,|B| + O\left(u^2\right)\right). \tag{59}$$

A szabvány modellnek eleget tevő lebegőpontos aritmetikáknak számos sajátos tulajdonsága van. Fontos tulajdonságuk, hogy az összeadás a kerekítés miatt nem asszociatív.

Ezt mutatják a következő MATLAB példák, amelyeket a *format long e* utasítás kiadása után ellenőrízhetünk.

Példa. MATLAB rendszerben

$$fl\left(fl\left(10^{-16}+1\right)-1\right) = 0, \quad fl\left(10^{-16} + fl\left(1-1\right)\right) = 10^{-16}.$$

Példa. Ha $a = 1$, $b = c = 3 \times 10^{-16}$, akkor a MATLAB 6.1 rendszerben Pentium 4 processzorú számítógépen

$$(a+b)+c \neq a+(b+c),$$

amelyet az $((a+b)+c)-(a+(b+c))$ utasítás segítségével ellenőrizhetünk.

Nagyszámú adat összegzésénél a kommutativitással (tulajdonképpen asszociativitással) is probléma lehet. Vizsgáljuk most a $\sum_{i=1}^{n} x_i$ összeg kiszámítását! A természetes algoritmus az ún. rekurzív összegzés:

$s = 0$
for $i = 1 : n$
  $s = s + x_i$
end

Példa. Számítsuk ki az

$$s_n = 1 + \sum_{i=1}^{n} \frac{1}{i^2 + i}$$

összeget $n = 4999$ esetén. A rekurzív összegzéssel kapott MATLAB eredmény

$$1.999800000000002e + 000.$$

Ha az összegzést fordított (azaz nagyság szerint növekedő) sorrendben végezzük, akkor az eredmény

$$1.999800000000000e + 000.$$

Ha a kétféle képpen kapott értékeket összevetjük az elméleti $s_n = 2 - \frac{1}{n+1}$ összeggel, akkor láthatjuk, hogy a második összegzés adott pontos eredményt. Ennek magyarázata az, hogy amikor a kisebb tagokkal kezdjük, akkor ezek összegei értékes jegyeket érnek a végső eredményben.

Nagy mennyiségű, előjelben és nagyságrendben eltérő szám nagy pontosságú összeadása nem egyszerű feladat. A következő algoritmus, amely az egyik legérdekesebb ilyen célra kifejlesztett eljárás, W. Kahantól származik.

A KOMPENZÁLT ÖSSZEGZÉS ALGORITMUSA:

$s = 0$
$e = 0$
for $j = 1 : n$
  $temp = s$
  $y = x(j) + e$
  $s = temp + y$
  $e = (temp - s) + y$
end

Példa. Kahan algoritmusát az

$$s_{4999} = \sum_{j=1}^{4999} \frac{1}{j^2 + j}$$

összegre az $x(j) = 1/\left(j^2 + j\right)$ szereposztással alkalmazva a MATLAB 6.1 a pontos 9.998000000000000e-001 értéket adja.

4.1 A lebegőpontos aritmetikai szabvány

Az ANSI/IEEE Std 754-1985 bináris ($\beta = 2$) lebegőpontos aritmetikai szabványt 1985-ben hozták nyilvánosságra.

A szabvány specifikálja

- az alapvető lebegőpontos műveleteket,
- összehasonlításokat,
- kerekítési módokat,
- az aritmetikai kivételeket és kezelésüket,
- a különböző aritmetikai formák közti konverziót.

A négyzetgyökvonás az alapvető műveletek közé tartozik. A szabvány nem mond semmit az exponenciális és transzcendens függvényekről.

A szabvány két fő lebegőpontos formátumot ismer: az egyszeres és a dupla pontosságút.

típus	méret	mantissza	e	u	$[M_L, M_U] \approx$
egyszeres	32 bit	23+1 bit	8 bit	$2^{-24} \approx 5.96 \times 10^{-8}$	$10^{\pm 38}$
dupla	64 bit	52+1 bit	11 bit	$2^{-53} \approx 1.11 \times 10^{-16}$	$10^{\pm 308}$

Mindkét formátumban egy bitet az előjelnek tartanak fenn. Minthogy a lebegőpontos számok normalizálva vannak és az első jegy mindig 1, ez a jegy nincs tárolva. A mantisszában szereplő +1 ezt a rejtett bitet jelzi.

Az aritmetikai kivételek kezelése a következő:

0

y=$\pm m \times \beta^{L-t}$, $0 < m < \beta^{t-1}$ alakú számok.

Az IEEE aritmetika zárt rendszer. Minden aritmetikai műveletnek van matematikailag értelmes vagy értelmetlen eredménye. A kivételes műveletek esetén jelzést ad ki, amely után a számításokat előírásszerűen folytatja. Az IEEE aritmetikai szabvány kielégíti a (50) modellt.

Az IEEE szabvány korai hardver megvalósításai közül ki kell emelni az Intel 80x87 matematikai koprocesszorokat, a DEC Alpha, a HP (Precision Architecture), az IBM RS/6000, az INMOS T800 és T900 processzorokat, a Motorola (680x0) és Sun (SPARCstation) processzorokat, valamint a HP tudományos kalkulátorait.

Végül megjegyezzük a következőket. Egyszeres pontosság esetén a mantissza hossza kb. 7 értékes jegyet enged meg a tizes számrendszerbe átszámolva. Ugyanez dupla pontosság esetén kb. 16 értékes jegyet jelent. Létezik még egy 80 biten ábrázolt, ún. kiterjesztett pontosság is, ahol $t = 63$, a kitevő pedig 15 bites.

5 LINEÁRIS EGYENLETRENDSZEREK HIBAANALÍZISE

Az $Ax = b$ egyenletrendszer elméleti megoldását x, a közelítő megoldásokat pedig $\hat{x}$ jelöli.

A közelítő megoldás direkt hibája $\Delta x = \hat{x} - x$.

Az $r = r(y) = Ay - b$ mennyiséget *reziduális hibának* nevezzük.

Az elméleti megoldás esetén $r(x) = 0$, a közelítő megoldás esetén pedig

$$r(\hat{x}) = A\hat{x} - b = A(\hat{x} - x) = A\Delta x.$$

Az inverz hiba meghatározásához különféle modelleket használunk. A legáltalánosabb esetben feltesszük, hogy az $\hat{x}$ számított megoldás kielégíti az $\hat{A}\hat{x} = \hat{b}$

egyenletrendszert, ahol $\hat{A} = A + \Delta A$ és $\hat{b} = b + \Delta b$. A ΔA és Δb mennyiségeket inverz hibáknak nevezzük.

Meg kell különböztetnünk a probléma érzékenységét és a megoldó algoritmusok stabilitását.

Adott probléma érzékenységén a megoldás változásának mértékét értjük a probléma (input) paramétereinek függvényében.

Adott algoritmus érzékenységén, vagy stabilitásán a számítási hibák végeredményre gyakorolt hatásának mértékét értjük.

Egy problémát, vagy algoritmust annál stabilabbnak tekintünk mennél kisebb az input paraméterek, ill. számítási hibák megoldásra (számított megoldásra) gyakorolt hatása. Az érzékenység, ill. stabilitás fogalmának egyik jellemzési formája a korábban látott kondíciószám, amely az eltérések relatív hibáit hasonlítja össze.

Algoritmusok felhasználásának a következő általános elveit lehet megfogalmazni:

- A gyakorlatban csak stabil (jól kondicionált) algoritmusokat használunk.
- Instabil (inkorrekt kitűzésű), vagy rosszul kondicionált feladatot általános célú algoritmusokkal általában nem tudunk megoldani.

5.1 Érzékenységvizsgálat

Tegyük fel, hogy az $Ax = b$ egyenlet helyett a perturbált

$$A\hat{x} = b + \Delta b \tag{60}$$

egyenletrendszert oldjuk meg. Legyen $\widehat{x} = x + \Delta x$ és vizsgáljuk a két megoldás eltérését!

Tétel. *Ha A nemszinguláris és $b \neq 0$, akkor*

$$\frac{\|\Delta x\|}{\|x\|} \leq \operatorname{cond}(A) \frac{\|\Delta b\|}{\|b\|}, \tag{61}$$

ahol $\operatorname{cond}(A) = \|A\| \left\|A^{-1}\right\|$ *az A mátrix ún. kondíciószáma.*

Bizonyítás. Az $A\hat{x} = A(x + \Delta x) = b + \Delta b$ egyenlőségből $\Delta x = A^{-1}\Delta b$, ahonnan $\|\Delta x\| \leq \left\|A^{-1}\right\| \|\Delta b\|$ következik. Másrészt $\|b\| \leq \|A\| \|x\|$, ahonnan $\frac{1}{\|x\|} \leq \frac{\|A\|}{\|b\|}$. A két egyenlőtlenséget összeszorozva kapjuk, hogy

$$\frac{\|\Delta x\|}{\|x\|} \leq \|A\| \left\|A^{-1}\right\| \frac{\|\Delta b\|}{\|b\|},$$

ami éppen a bizonyítandó állítás. □

Az A mátrix kondíciószáma erősen befolyásolhatja az $\widehat{x}$ perturbált megoldás relatív hibáját. Egy rendszert jól kondicionáltnak nevezünk, ha $\operatorname{cond}(A)$ kicsi és rosszul kondicionáltnak nevezünk, ha $\operatorname{cond}(A)$ nagy.

A nagy és kicsi jelzők relatívak és környezetfüggők. A kondíciószám függ a normától. Ha a normától való függés lényeges, akkor ezt külön jelöljük. Ennek megfelelően például $\operatorname{cond}_\infty(A) = \|A\|_\infty \left\|A^{-1}\right\|_\infty$.

A kondicionáltság egy lehetséges geometriai jellemzését adja a következő példa.

Példa. A

$$\begin{array}{rcrcl} 1000x_1 & + & 999x_2 & = & b_1 \\ 999x_1 & + & 998x_2 & = & b_2 \end{array}$$

egyenletrendszer rosszul kondicionált ($\mathrm{cond}_\infty(A) = 3.99 \times 10^6$). A két egyenes majdnem párhuzamos. Ezért, ha perturbáljuk a jobboldalt, az új metszéspont messze lesz az előzőtől.

A most vizsgált modellben az inverz hiba Δb, a Tétel pedig a relatív direkt hibára ad becslést. Ez teljes összhangban van a hibaszámítási ökölszabállyal. A tétel állítása az $r(\hat{x}) = A\hat{x} - b = \hat{b} - b = \Delta b$ összefüggés miatt átírható az ekvivalens

$$\frac{\|\hat{x} - x\|}{\|x\|} \leq \mathrm{cond}(A) \frac{\|r(\hat{x})\|}{\|b\|} \tag{62}$$

alakba. Az egyenlőtlenség jelentése az, hogy az $\hat{x}$ perturbált megoldás relatív hibája kicsi, ha A kondíciószáma kicsi és az $\|r(\hat{x})\| / \|b\|$ relatív reziduális hiba kicsi. Ha azonban a rendszer rosszul kondicionált, akkor ez nem szükségképpen igaz.

Példa. Vizsgáljuk az $Ax = b$ egyenletrendszert, ahol

$$A = \begin{bmatrix} 1+\epsilon & 1 \\ 1 & 1 \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad x = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

Legyen $\hat{x} = \begin{bmatrix} 2 \\ -1 \end{bmatrix}$. Ekkor $r = \begin{bmatrix} 2\epsilon \\ 0 \end{bmatrix}$ és $\|r\|_\infty / \|b\|_\infty = 2\epsilon$, de $\|\hat{x} - x\|_\infty / \|x\|_\infty = 2$.

Tegyük fel, hogy az $Ax = b$ egyenlet helyett a perturbált

$$(A + \Delta A)\hat{x} = b + \Delta b \tag{63}$$

egyenletrendszert oldjuk meg. Legyen $\widehat{x} = x + \Delta x$. Igaz a következő

Tétel. *Ha A nemszinguláris,* $\mathrm{cond}(A) \frac{\|\Delta A\|}{\|A\|} < 1$ *és* $b \neq 0$*, akkor*

$$\frac{\|\Delta x\|}{\|x\|} \leq \frac{\mathrm{cond}(A)\left(\frac{\|\Delta A\|}{\|A\|} + \frac{\|\Delta b\|}{\|b\|}\right)}{1 - \mathrm{cond}(A)\frac{\|\Delta A\|}{\|A\|}}. \tag{64}$$

A tételt nem igazoljuk. A tételből következik a következő un. "ökölszabály".

Ökölszabály. *Tegyük fel, hogy $Ax = b$. Ha A és b elemeit s decimális jegyre pontosak és* $\mathrm{cond}(A) \sim 10^t$*, ahol $t < s$, akkor a számított megoldás kb. $s - t$ jegyre lesz pontos.*

Az ökölszabály következő heurisztikus levezetését adjuk. A feltevések miatt

$$\frac{\|\Delta A\|}{\|A\|} \approx 10^{-s}, \quad \frac{\|\Delta b\|}{\|b\|} \approx 10^{-s}$$

és

$$\mathrm{cond}(A)\frac{\|\Delta A\|}{\|A\|} \approx 10^{t-s} \ll 1.$$

Ezért feltehetjük, hogy $1-\mathrm{cond}(A)\frac{\|\Delta A\|}{\|A\|} \approx 1$. A fenti tétel alapján

$$\frac{\|\Delta x\|}{\|x\|} \approx \mathrm{cond}(A)\left(\frac{\|\Delta A\|}{\|A\|}+\frac{\|\Delta b\|}{\|b\|}\right) \approx 10^{t-s},$$

amiből az ökölszabály következik.

A Tétel

$$cond\,(A)\,\frac{\|\Delta A\|}{\|A\|} < 1$$

feltételének jelentése szemléletes: azt biztosítja, hogy az $A+\Delta A$ mátrix ne legyen szinguláris.

A $\mathrm{cond}(A)\frac{\|\Delta A\|}{\|A\|} < 1$ egyenlőtlenség ugyanis ekvivalens a $\|\Delta A\| < \frac{1}{\|A^{-1}\|}$ feltétellel és az A nemszinguláris mátrix legközelebbi szinguláris mátrixtól való távolsága éppen $\frac{1}{\|A^{-1}\|}$. Igaz a következő

Tétel *(Eckart-Young-Gastinel). Legyen $A \in \mathbb{R}^{n\times n}$ nemszinguláris, $P \in \mathbb{R}^{n\times n}$ pedig tetszőleges szinguláris mátrix. Akkor fennáll, hogy*

$$\|A-P\| \geq \frac{1}{\|A^{-1}\|}. \tag{65}$$

Létezik továbbá olyan P szinguláris mátrix, amelyre egyenlőség áll fenn.

5.2 Wilkinson tétele

Wilkinson igazolta, hogy *az $Ax=b$ egyenletrendszer Gauss-módszerrel lebegőpontos aritmetikában kapott $\hat{x}$ közelítő megoldása kielégíti az*

$$(A+\Delta A)\,\widehat{x} = b \tag{66}$$

egyenletrendszert, ahol

$$\|\Delta A\|_\infty \leq 8n^3\rho_n\,\|A\|_\infty\,u + O(u^2). \tag{67}$$

A ρ_n a pivot elemek *növekedési tényezője*. Minthogy a gyakorlatban ρ_n kicsi, a

$$\frac{\|\Delta A\|_\infty}{\|A\|_\infty} \leq 8n^3\rho_n u + O(u^2)$$

relatív inverz hiba is az.

A Gauss-elimináció "gyengén stabil" mind a teljes, mind pedig a parciális főelemválasztás esetén.

A Wilkinson tételből kapjuk, hogy

$$\mathrm{cond}_\infty(A)\frac{\|\Delta A\|_\infty}{\|A\|_\infty} \leq 8n^3\rho_n \mathrm{cond}_\infty\,(A)\,u + O\left(u^2\right).$$

Kis kondíciószám esetén feltehetjük, hogy $1-\text{cond}_\infty(A)\frac{\|\Delta A\|_\infty}{\|A\|_\infty} \approx 1$. Az előző szakaszbeli Tétel felhasználásával ($\Delta b = 0$ eset) a direkt hibára az alábbi közelítő becslést kapjuk:

$$\frac{\|\Delta x\|_\infty}{\|x\|_\infty} \leq 8n^3 \rho_n \text{cond}_\infty(A)\, u. \tag{68}$$

Ez az ökölszabály helyességét támasztja alá a Gauss-módszer esetén.

Tekintsük a következő példát, amelynek együtthatóit pontosan tudjuk ábrázolni:

$$888445x_1 + 887112x_2 = 1,$$
$$887112x_1 + 885781x_2 = 0.$$

Itt $\text{cond}(A)_\infty$ ugyan nagy, de $\text{cond}_\infty(A)\frac{\|\Delta A\|_\infty}{\|A\|_\infty}$ elhanyagolható az 1 mellett. A feladat pontos megoldása $x_1 = 885781$, $x_2 = -887112$. A MATLAB által adott közelítő megoldás $\hat{x}_1 = 885827.23$, $\hat{x}_2 = -887158.30$, amelynek relatív hibája

$$\frac{\|x - \hat{x}\|_\infty}{\|x\|_\infty} = 5.22 \times 10^{-5}.$$

Minthogy $s \approx 16$ és $\text{cond}(A)_\infty \approx 3.15 \times 10^{12}$ az eredmény lényegében megfelel a Wilkinson tételnek, ill. az ökölszabálynak. A Wilkinson tétel az inverz hiba mértékére a

$$\|\Delta A\|_\infty \leq 1.26 \times 10^{-8}$$

becslést adja.

5.3 Utólagos hibabecslések

A valamilyen módszerrel kapott közelítő megoldás hibájának utólagos becslésére azért van szükség, hogy valamilyen adatunk legyen az eredmény megbizhatóságáról.

Tétel *(Auchmuty). Jelölje $\widehat{x}$ az $Ax = b$ egyenletrendszer valamilyen módon kiszámított közelítő megoldását. Ekkor igaz, hogy*

$$\|x - \hat{x}\|_2 = \frac{c\,\|r(\hat{x})\|_2^2}{\|A^T r(\hat{x})\|_2},$$

ahol $c \geq 1$ konstans, amely A-tól függ.

Megemlítjük, hogy a c hibakonstans értékét nem befolyásolja az $\hat{x} - x$ hibavektor nagysága, csak az iránya. Igaz továbbá, hogy

$$C_2(A) = \frac{1}{2}\left(\text{cond}_2(A) + \frac{1}{\text{cond}_2(A)}\right) \leq \text{cond}_2(A).$$

A számítógépes tapasztalatok azt mutatják, hogy c nem túl nagy szám, gyakran $c \leq 100$. Az összefüggés alapján a reziduális és az $\|r(\hat{x})\|_2^2 / \left\|A^T r(\hat{x})\right\|_2$ hányados meghatározásával nagyságrendileg helyesen becsülhetjük a közelítő megoldás abszolút hibáját.

5.4 A közelítő megoldás iteratív javítása

Az eljárás első ismert alkalmazása Fox, Goodwin, Turing és Wilkinson nevéhez fűződik (1946). Jelölje $\hat{x}$ az $Ax = b$ egyenletrendszer közelítő megoldását, ϕ pedig a közelítő módszert.. Legyen $r(y) = Ay - b$ az y pontbeli reziduális hiba. Az $\hat{x}$ közelítő megoldás pontosságát a következő iteratív eljárással lehet javítani.

AZ ITERATÍV JAVÍTÁS ALGORITMUSA:

$i = 1$, $x_1 = \hat{x}$

for $i = 1, \ldots$

$r = Ax_i - b$

Számítsuk ki az $Ad = r$ egyenletrendszer $\hat{d}$ közelítő megoldását a ϕ-módszer segítségével!

$x_{i+1} = x_i - \hat{d}$

Ha $\left\|\hat{d}\right\| / \|x_i\| < tol$, akkor vége.

end

Az eljárás különféle változatai ismertek. Általában a Gauss módszer LU-felbontáson alapuló változatát szokták használni.

Jankowski és Wozniakowski az iteratív javitást tetszőleges olyan ϕ módszerre vizsgálták, amely az $Ax = b$ egyenletrendszer 1-nél kisebb relatív hibájú $\hat{x}$ közelítését állítja elő, azaz amelyre $\|\hat{x} - x\| \leq q \|x\|$ $(q < 1)$.

Igazolták, hogy az iteratív javítás még egyszeres pontosságú lebegőpontos aritmetikában is javítja a közelítő megoldás pontosságát és a ϕ módszert gyengén stabillá teszi.

6 INTERPOLÁCIÓ

Az interpoláció feladata a következő:

Adottak/ismertek/mértek az $y = f(x)$ $(f : \mathbb{R} \to \mathbb{R})$ függvény

$$a \leq x_1 < x_2 < \ldots < x_n \leq b \tag{69}$$

pontokban felvett értékei, az

$$y_i = f(x_i) \quad (i = 1, \ldots, n) \tag{70}$$

függvényértékek.

Az $f(x)$ függvényt, amely lehet ismert, vagy akár ismeretlen is, egy olyan, általában könnyen számítható $h(x)$ függvénnyel közelítjük (vagy helyettesítjük), amelyre fennáll, hogy

$$y_i = h(x_i) \quad (i = 1, \ldots, n). \tag{71}$$

Az $\{x_i\}_{i=1}^n$ pontokat *interpolációs alappontoknak*, az (71) feltételt *interpolációs feltételnek* nevezzük.

Az interpolációs feltétel teljesülése esetén azt reméljük, hogy a

$$h(x) = h\left(x; \{x_i\}_{i=1}^n, \{y_i\}_{i=1}^n\right)$$

interpoláló függvény az (x_i, x_{i+1}) intervallumokban jól közelíti az $f(x)$ függvényt.

A $h(x)$ függvény megválasztásától függően beszélünk különböző típusú interpolációkról.

Ha a $h(x)$ függvénnyel $f(x)$-et az (x_1, x_n) intervallumon kívül közelítjük, akkor *extrapolációról* beszélünk.

6.1 A lineáris interpoláció

A lineáris interpoláció esetén a $h(x)$ függvény alakja

$$h(x) = a_1\phi_1(x) + a_2\phi_2(x) + \ldots + a_n\phi_n(x) = \sum_{i=1}^{n} a_i\phi_i(x), \tag{72}$$

ahol a $\phi_i : [a, b] \to \mathbb{R}$ $(i = 1, \ldots, n)$ bázisfüggvények adottak. Az ismeretlen $a_1, \ldots, a_n$ együtthatókat az interpolációs feltételből határozhatjuk meg. Ekkor teljesülnie kell az alábbi n feltételnek

$$\begin{array}{rcl} a_1\phi_1(x_1) + a_2\phi_2(x_1) + \ldots + a_n\phi_n(x_1) & = & f(x_1), \\ \vdots & \vdots & \vdots \\ a_1\phi_1(x_n) + a_2\phi_2(x_n) + \ldots + a_n\phi_n(x_n) & = & f(x_n), \end{array} \tag{73}$$

amely lineáris egyenletrendszer az ismeretlen $a_1, \ldots, a_n$ együtthatókra nézve. Legyen

$$B = [\phi_j(x_i)]_{i,j=1}^{n} \tag{74}$$

és

$$a = [a_1, \ldots, a_n]^T, \quad c = [f(x_1), \ldots, f(x_n)]^T. \tag{75}$$

A fenti feltétel tömör alakban

$$Ba = c. \tag{76}$$

Ha $\det(B) \neq 0$, akkor az egyenletrendszernek pontosan egy megoldása van: $a = B^{-1}c$.

A gyakorlatban sokféle $\{\phi_i(x)\}_{i=1}^{n}$ bázisfüggvényt alkalmaznak. Az egyik legfontosabb a

$$\phi_1(x) = 1,\ \phi_2(x) = x,\ \ldots\ ,\ \phi_n(x) = x^{n-1} \tag{77}$$

függvényrendszer, amely a *Lagrange-féle interpolációs feladatot* definiálja. Ekkor az interpolációs feladat mátrixa

$$B = \begin{bmatrix} 1 & x_1 & \ldots & x_1^{n-1} \\ \vdots & \vdots & & \vdots \\ 1 & x_n & \ldots & x_n^{n-1} \end{bmatrix} \tag{78}$$

az ún. Vandermonde-féle mátrix, amely a $\det(B) = \prod_{1 \leq i < j \leq n}(x_j - x_i)$ összefüggés miatt nemszinguláris. Tehát a Lagrange-féle interpolációs feladatnak egyértelmű megoldása van.

További fontos esetek a következők. A trigonometrikus interpolációt a

$$\phi_1(x) = 1,\ \phi_{2k}(x) = \sin(kx),\ \phi_{2k+1}(x) = \cos(kx) \quad \left(k = 1, \ldots, \frac{n-1}{2}\right) \tag{79}$$

függvényrendszer ($n = 2k+1$, $[a,b] = [-\pi, \pi]$), az exponenciális interpolációt pedig a

$$\phi_i(x) = e^{\lambda_i x} \quad (i = 1, \ldots, n,\ \lambda_1 < \lambda_2 < \ldots < \lambda_n) \tag{80}$$

függvényrendszer definiálja. Racionális törtfüggvényeket használ a

$$\phi_i(x) = \frac{1}{a_i + x} \quad (i = 1, \ldots, n,\ 0 < a_1 < \ldots < a_n) \tag{81}$$

függvényrendszer. Itt fel kell tennünk, hogy $x + a_1 > 0$. Ez könnyen teljesül, ha $x \in [a,b]$ és $a + a_1 > 0$.

Nem minden $\{\phi_i(x)\}_{i=1}^n$ függvényrendszer és $x_1 < x_2 < \ldots < x_n$ alappontrendszer esetén van megoldása a lineáris interpolációs feladatnak.

Példa. Legyen $\phi_1(x) = 1$, $\phi_2(x) = x^2$, $x_1 = -1$, $x_2 = 1$. Ekkor

$$B = \begin{bmatrix} 1 & (-1)^2 \\ 1 & 1 \end{bmatrix}, \quad \det(B) = 0.$$

A lineáris interpolációs feladat mátrixa sok esetben rosszul kondicionált. Ilyenkor speciális technikákat, vagy más típusú interpolációt kell használni.

6.2 A Lagrange-féle interpolációs feladat

A feladat szokásos megfogalmazása a következő.

Adottak az $x_1 < x_2 < \ldots < x_n$ alappontok és az $y_i = f(x_i)$ $(i = 1, \ldots, n)$ függvényértékek. Határozzuk meg azt a legfeljebb $(n-1)$-edfokú

$$p(x) = a_0 + a_1 x + \ldots + a_{n-1} x^{n-1} \tag{82}$$

polinomot, amelyre teljesül a

$$y_i = p(x_i) \quad (i = 1, \ldots, n) \tag{83}$$

interpolációs feltétel.

A *Lagrange-féle* interpolációs polinom létezését és egyértelműségét már beláttuk. A polinom többféle ekvivalens alakban is felírható. Különösen fontos azonban a Lagrange-féle előállítás. Legyen

$$l_i(x) = \prod_{k=1, k\neq i}^{n} \frac{x - x_k}{x_i - x_k} \quad (i = 1, \ldots, n) \tag{84}$$

az i-edik *Lagrange-féle alappolinom*. Ekkor az interpolációs polinom előáll

$$p(x) = \sum_{i=1}^{n} y_i l_i(x) \tag{85}$$

alakban. Ennek igazolására vegyük észre, hogy

$$l_i(x_j) = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases} \tag{86}$$

és

$$p(x_j) = \sum_{i=1}^{n} y_i l_i(x_j) = y_j l_j(x_j) = y_j \quad (j = 1, \ldots, n). \tag{87}$$

A Lagrange-féle interpolációs polinom hibájára vonatkozik a következő

Tétel *(Cauchy). Ha* $f \in C^n[a,b]$, $[x_1, x_n] \subset [a,b]$ *és* $x \in [a,b]$, *akkor*

$$f(x) - p(x) = \frac{f^{(n)}(\xi_x)}{n!}(x - x_1)(x - x_2)\ldots(x - x_n), \tag{88}$$

ahol $\xi_x = \xi(x)$ *az* x *és az* x_1, x_n *pontok által kifeszített intervallumban van.*

Bizonyítás. Ha van i, hogy $x = x_i$, akkor állításunk triviális. Egyébként legyen $\omega(x) = (x - x_1)(x - x_2)\ldots(x - x_n)$ és tekintsük a következő segédfüggvényt:

$$W(t) = f(t) - p(t) - [f(x) - p(x)]\frac{\omega(t)}{\omega(x)}. \tag{89}$$

A $W(t) \in C^n[a,b]$ függvénynek van $n+1$ gyökhelye: $x, x_1, \ldots, x_n$. A Rolle–tétel miatt $W(t)$ bármely két gyökhelye között a $W'(t)$ deriváltfüggvénynek zérushelye van. Ezért $W'(t)$-nek legalább n zérushelye van. Hasonlóképpen okoskodva belátható, hogy $W''(t)$-nek legalább $n-1$, $W^{(3)}(t)$-nek legalább $n-2$ zérushelye van, és így tovább. Végül $W^{(n)}(t)$-nek is van legalább egy zérushelye, amit jelöljön ξ_x. Minthogy $p^{(n)}(t) \equiv 0$ és $\omega^{(n)}(t) \equiv n!$, azért

$$W^{(n)}(\xi_x) = f^{(n)}(\xi_x) - [f(x) - p(x)]\frac{n!}{\omega(x)} = 0, \tag{90}$$

ahonnan átrendezéssel kapjuk a tétel állítását. □

Következmény. *Ha* $\left|f^{(n)}(x)\right| \leq M_n$ *(*$x \in [a,b]$*), akkor*

$$|f(x) - p(x)| \leq \frac{M_n}{n!}(b - a)^n. \tag{91}$$

Konkrét n esetén szélsőértékszámítással élesebb becslés is levezethető.

Példa. Hány ekvidisztáns alappontban kell megadnunk a $\sin x$ függvény táblázatát a $\left[0, \frac{\pi}{2}\right]$ intervallumon ahhoz, hogy a közbülső pontokban lineáris Lagrange-interpolációt használva az elkövetett hiba legfeljebb $\varepsilon = 10^{-4}$ legyen? Vezessük be a $h = x_{i+1} - x_i$ jelölést. A Cauchy-tétel következménye alapján olyan h-t keresünk, melyre $M_2 h^2/2 \leq 10^{-4}$. Mivel $(\sin x)'' = -\sin x$, választhatjuk az $M_2 = 1$ értéket. Ezzel $h \leq \sqrt{2}/100$, $n \geq \frac{\pi}{2h}$ miatt $n \geq 112$ adódik. Ha viszont a hibakorlátot az

$$|f(x) - p(x)| \leq \frac{M_2}{2} \max |(x - x_i)(x - x_{i+1})|$$

becslésből közvetlenül vezetjük le szélsőértékszámítással, akkor az élesebb,

$$|f(x) - p(x)| \leq \frac{M_2 h^2}{8}$$

eredményt kapjuk. Ez alapján kiderül, hogy $n = 28$ pont is elég.

Az interpolációs eljárásoktól elvárjuk, hogy a pontok számának növelése esetén a közelítés hibája csökken. Ez azonban nem minden esetben van így, amint azt Runge példája is mutatja.

Példa (Runge). *Ábrázoljuk az* $f(x) = \frac{1}{1+x^2}$ *függvényt a* $[-5, 5]$ *intervallumon és* n *különböző értékeire* ($n = 11, 17, \ldots$) *az* $f(x)$ *függvényhez és az* $x_i = -5 + 10\frac{i-1}{n-1}$ ($i = 1, \ldots, n$) *alappontokhoz tartozó Lagrange-féle interpolációs polinomot! Mit tapasztal, ha* n *értéke nő? Tapasztalja-e ugyanezt a jelenséget, ha a számításokat a* $[-3, 3]$ *intervallumon végzi?*

Nagy n-ek esetén numerikus instabilitás is felléphet. Ennek illusztrálására tegyük fel, hogy az $y_i = f(x_i)$ függvényértékeket ε_i hibával ismerjük ($i = 1, \ldots, n$). Ekkor az elméleti

$$p(x) = \sum_{i=1}^{n} f(x_i) l_i(x)$$

Lagrange-interpolációs polinom helyett a perturbált

$$\tilde{p}(x) = \sum_{i=1}^{n} (f(x_i) + \varepsilon_i) l_i(x)$$

polinommal számolunk. A kettő eltérésére teljesül, hogy

$$\delta\left(p\left(x\right)\right) = |\tilde{p}(x) - p(x)| = \left|\sum_{i=1}^{n} \varepsilon_i l_i(x)\right| \leq \sum_{i=1}^{n} |\varepsilon_i| \, |l_i(x)| \leq \left(\max_{1 \leq i \leq n} |\varepsilon_i|\right) \sum_{i=1}^{n} |l_i(x)| \, .$$

Ez a becslés pontos. Igazolható, hogy

$$\sum_{i=1}^{n} |l_i(x)| > \frac{2}{\pi} \log n + c, \tag{92}$$

ahol c konstans. Ha n elég nagy, akkor a $\delta\left(p\left(x\right)\right)$ perturbációs hiba is nagy lesz.

A divergencia és numerikus instabilitás miatt sok esetben más típusú interpolációs technikákat használunk.

Példa. Közelítsük másodfokú függvénnyel az $f(x) = \cos(\frac{\pi}{2}x)$ függvényt a$[-1, 1]$-ben az $x_1 = -1, x_2 = 0, x_3 = 1$ pontokra támaszkodva! $f(x) \approx p(x) = A_1 + A_2 x + A_3 x^2$. Az együtthatókra felírható az

$$\begin{aligned} A_1 - A_2 + A_3 &= 0 \\ A_1 &= 1 \\ A_1 + A_2 + A_3 &= 0 \end{aligned}$$

egyenletrendszer. Innen $p(x) = 1 - x^2$. Természetesen ugyanezt kapjuk az $l_i(x)$ Lagrange-függvényekkel is. $f(x_1) = f(x_3) = 0$ miatt elég az $l_2(x)$-t meghatározni, ez $1 - x^2$, ami jelen esetben a $p(x)$ polinommal megegyezik. A közelítés hibáját

$$h \leq \frac{M_3}{3!} \max_{-1 \leq x \leq 1} |(x+1)x(x-1)|$$

becsli, ahol M_3 az $|f'''(x)|$ maximuma, jelen esetben $\pi^3/8$. Szélsőértékszámítással adódik, hogy

$$\max_{-1 \leq x \leq 1} |(x+1)x(x-1)| = \frac{8}{27},$$

azaz $h \leq \pi^3/216 \simeq 0.15$.

6.3 Harmadfokú szplájn interpoláció

A szplájn interpoláció is a lineáris interpolációk közé tartozik alkalmasan megválasztott $\{\phi_i\}$ bázisfüggvény-rendszerrel.

A szplájn interpoláció esetén a $h(x)$ interpoláló függvényt szakaszonként adjuk meg speciális csatlakozási feltételekkel.

Az $a = x_1 < x_2 < \ldots < x_n = b$ alappontokhoz, illetve az $y_i = f(x_i)$ $(i = 1, \ldots, n)$ függvényértékekhez olyan $S(x)$ függvényt keresünk, amely kielégíti a következő feltételeket:

(i) $S(x) = S_i(x)$ $(x \in [x_i, x_{i+1}])$, ahol $S_i(x)$ legfeljebb harmadfokú polinom $(i = 1, \ldots, n-1)$,

(ii) $S(x_i) = y_i$ $(i = 1, \ldots, n)$,

(iii) $S_i(x_{i+1}) = S_{i+1}(x_{i+1})$ $(i = 1, \ldots, n-2)$,

(iv) $S_i'(x_{i+1}) = S_{i+1}'(x_{i+1})$ $(i = 1, \ldots, n-2)$,

(v) $S_i''(x_{i+1}) = S_{i+1}''(x_{i+1})$ $(i = 1, \ldots, n-2)$,

(vi) $S''(x_1) = A_n$, $S''(x_n) = B_n$.

Az (ii)-(iii) feltételek együttesen azt jelentik, hogy az $S(x)$ függvény folytonos az $[a, b]$ intervallumon. Az (iv)-(v) feltételek azt mondják ki, hogy $S'(x)$ és $S''(x)$ folytonos. Ha $A_n = B_n = 0$, akkor az $S(x)$ függvényt *természetes szplájn*nak nevezzük.

Legyen $h_i = x_{i+1} - x_i$ az i-edik részintervallum hossza $(i = 1, \ldots, n-1)$. A $S(x)$ szplájnt az $[x_i, x_{i+1}]$ intervallumon a

$$S(x) = S_i(x) = a_i + b_i(x - x_i) + c_i(x - x_i)^2 + d_i(x - x_i)^3 \tag{93}$$

alakban keressük $(i = 1, \ldots, n-1)$. Az (ii)-(vi) feltételek felhasználásával az ismeretlen a_i, b_i, c_i és d_i együtthatókat a következőképpen határozhatjuk meg.

Az (ii), azaz az $S(x_i) = S_i(x_i) = y_i$ interpolációs feltétel miatt $a_i = y_i$ $(i = 1, \ldots, n-1)$. Az (iii) csatlakozási feltétel alakja

$$S_i(x_{i+1}) = y_i + b_i h_i + c_i h_i^2 + d_i h_i^3 = y_{i+1} \quad (i = 1, \ldots, n-1), \tag{94}$$

Az (iv) csatlakozási feltétel alakja

$$S_i'(x_{i+1}) = b_i + 2c_i h_i + 3d_i h_i^2 = b_{i+1} = S_{i+1}'(x_{i+1}) \quad (i = 1, \ldots, n-2). \tag{95}$$

Hasonlóképpen kapjuk, hogy a (v) feltétel alakja

$$S_i''(x_{i+1}) = 2c_i + 6d_i h_i = 2c_{i+1} = S_{i+1}''(x_{i+1}) \quad (i = 1, \ldots, n-2).$$

Ebből az egyenlőség-láncból a

$$c_{i+1} = c_i + 3d_i h_i \quad (i = 1, \ldots, n-2)$$

összefüggéseket kapjuk. A (vi) végpont-feltétel alakja

$$S''(x_1) = 2c_1 = A_n, \quad S''(x_n) = 2c_{n-1} + 6d_{n-1}h_{n-1} = B_n. \tag{96}$$

Így kapjuk, hogy

$$d_i = \frac{c_{i+1} - c_i}{3h_i} \quad (i = 1, \ldots, n-2), \quad d_{n-1} = \frac{B_n - 2c_{n-1}}{6h_{n-1}}. \tag{97}$$

Mindent összevetve a $3(n-1)$ ismeretlenre az (94)-(97) összefüggések összesen $3(n-1)$ egyenletet adnak. Az (94) egyenletből b_i-t kifejezhetjük a következőképpen:

$$b_i = \frac{y_{i+1} - y_i}{h_i} - c_i h_i - d_i h_i^2 \quad (i = 1, \ldots, n-1). \tag{98}$$

A (96)-(98) összefüggéseket felhasználva a szplájn előállítható a $c_1, c_2, \ldots, c_{n-1}$ együtthatók ismeretében. A (97) összefüggést a (98)–ba beírva kapjuk, hogy $i = 1, \ldots, n-2$ esetén

$$b_i = \frac{y_{i+1} - y_i}{h_i} - c_i h_i - \frac{c_{i+1} - c_i}{3h_i} h_i^2 = \frac{y_{i+1} - y_i}{h_i} - \frac{2c_i + c_{i+1}}{3} h_i. \tag{99}$$

Hasonlóképpen kapjuk, hogy

$$\begin{aligned} b_{n-1} &= \frac{y_n - y_{n-1}}{h_{n-1}} - c_{n-1}h_{n-1} - \frac{B_n - 2c_{n-1}}{6h_{n-1}} h_{n-1}^2 \\ &= \frac{y_n - y_{n-1}}{h_{n-1}} - \frac{4c_{n-1} - B_n}{6} h_{n-1}. \end{aligned} \tag{100}$$

Legyen $\Delta_i = (y_{i+1} - y_i)/h_i$ és helyettesítsük a b_i és d_i együtthatókra vonatkozó összefüggéseket a (95) egyenlőségbe:

$$\begin{aligned} \Delta_{i+1} - \frac{2c_{i+1} + c_{i+2}}{3} h_{i+1} &= \Delta_i - \frac{2c_i + c_{i+1}}{3} h_i + 2c_i h_i + 3\frac{c_{i+1} - c_i}{3h_i} h_i^2 \\ (i &= 1, \ldots, n-3), \end{aligned}$$

$$\begin{aligned} \Delta_{n-1} - \frac{4c_{n-1} - B_n}{6} h_{n-1} &= \Delta_{n-2} - \frac{2c_{n-2} + c_{n-1}}{3} h_{n-2} + \\ &\quad + 2c_{n-2}h_{n-2} + 3\frac{c_{n-1} - c_{n-2}}{3h_{n-2}} h_{n-2}^2. \end{aligned}$$

Az összefüggések átrendezésével kapjuk, hogy

$$h_i c_i + 2(h_i + h_{i+1})c_{i+1} + h_{i+1}c_{i+2} = 3\left(\Delta_{i+1} - \Delta_i\right), \tag{101}$$

$(i = 1, \ldots, n-3)$ és

$$h_{n-2}c_{n-2} + 2(h_{n-2} + h_{n-1})c_{n-1} = 3\left(\Delta_{n-1} - \Delta_{n-2} + \frac{h_{n-1}}{6}B_n\right). \tag{102}$$

Figyelembevéve, hogy $c_1 = A_n/2$, az első egyenlet $(i = 1)$ átmegy a

$$2\left(h_1 + h_2\right)c_2 + h_2 c_3 = 3\left(\Delta_2 - \Delta_1 - \frac{h_1}{2}A_n\right) \tag{103}$$

alakba. Az i-edik egyenletet $(h_i + h_{i+1})$-el osztva kapjuk, hogy

$$2c_2 + \lambda_1 c_3 = \frac{3}{h_1 + h_2}\left(\Delta_2 - \Delta_1 - \frac{h_1}{2}A_n\right), \tag{104}$$

$$\mu_i c_i + 2c_{i+1} + \lambda_i c_{i+2} = \frac{3}{h_i + h_{i+1}}\left(\Delta_{i+1} - \Delta_i\right) \quad (i = 2, \ldots, n-3), \tag{105}$$

$$\mu_{n-2}c_{n-2} + 2c_{n-1} = \frac{3}{h_{n-2} + h_{n-1}}\left(\Delta_{n-1} - \Delta_{n-2} + \frac{h_{n-1}}{6}B_n\right), \tag{106}$$

ahol $\lambda_i = h_{i+1}/\left(h_i + h_{i+1}\right)$, $\mu_i = 1 - \lambda_i$ $(i = 1, \ldots, n-2)$. Ez egy $n-2$ ismeretlenes lineáris egyenletrendszer a $c_2, c_3, \ldots, c_{n-1}$ ismeretlenekre. Az egyenletrendszer mátrixa $n > 4$ esetén tehát

$$A = \begin{bmatrix} 2 & \lambda_1 & 0 & \ldots & 0 \\ \mu_2 & 2 & \lambda_2 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \mu_{n-3} & 2 & \lambda_{n-3} \\ 0 & \ldots & 0 & \mu_{n-2} & 2 \end{bmatrix}$$

egy három átlóból álló sávmátrix, amely $O(n)$ régi flop művelettel numerikusan stabilan megoldható a Gauss-módszer egy speciális változatával.

Tétel. *Az (i)-(vi) feltételekkel meghatározott $S(x)$ szplájn létezik és egyértelmű. Ha $f \in C^2[a,b]$, akkor létezik $K > 0$ konstans , hogy*

$$|f(x) - S(x)| \le K\left(\max_{1 \le i \le n-1} h_i\right)^2 \quad (x \in [a,b]). \tag{107}$$

Ha $f \in C^3[a,b]$, $A_n = f''(x_1)$ és $B_n = f''(x_n)$, akkor létezik $\widetilde{K} > 0$ konstans, hogy

$$|f(x) - S(x)| \le \widetilde{K}\left(\max_{1 \le i \le n-1} h_i\right)^3 \quad (x \in [a,b]). \tag{108}$$

Ha az $f''(x_1)$ és $f''(x_n)$ információk nem állnak rendelkezésre, akkor a természetes szplájnt definiáló $A_n = B_n = 0$ választással élünk. A közelítés hibájára ekkor a (107) becslés érvényes.

A természetes szplájn az

$$\int_a^b \left[s''(x)\right]^2 dx \to \min \quad (s(x_i) = y_i,\ i = 1, \ldots, n) \tag{109}$$

feltételes szélsőértékfeladat megoldása.

A természetes szplájn mechanikai tartalma: Legyen adott egy rugalmas rúd (szplájn), amely átmegy az (x_i, y_i) pontokon (tengelyeken). A legkisebb deformációs energia mechanikai elve miatt a szplájn azt az alakot veszi fel, amely a fenti (közelítő) kifejezést minimalizálja.

A (vi) végpont-feltétel helyett más kikötések is lehetségesek. Ilyenek például az

$$S'(x_1) = a_1, \quad S'(x_n) = b_1, \tag{110}$$

ill. az

$$S^{(i)}(x_1) = S^{(i)}(x_n) \quad (i = 1, 2) \tag{111}$$

végpont-feltételek. Az előbbi feltételek az ún. *teljes szplájn*t, míg az utóbbiak az ún. *periodikus szplájn*t definiálják.

A szplájn függvények előnyei: gyors és numerikusan stabil kiszámítás, nagyon jó közelítési tulajdonságok.

Hátrányuk: a bonyolult megadás, amely számítógépek használata esetén nem jelent komoly problémát.

6.4 A MATLAB interpolációs eljárásai

- interp1
  - spline
  - ppval

7 NUMERIKUS DERIVÁLÁS

A numerikus deriválás alapproblémája az $f : \mathbb{R} \to \mathbb{R}$ függvény deriváltjának kiszámítása egy vagy több adott pontban.

A probléma kézenfekvő megoldása: az $f(x)$ függvényt egy $h(x)$ függvénnyel (lineáris interpolációval, szplájn-interpolációval, stb.) közelítjük és az $f'(x)$ közelítésének a közelítő függvény $h'(x)$ deriváltját tekintjük.

Sematikusan: ha $f(x) \approx h(x)$, akkor $f'(x) \approx h'(x)$ és általában $f^{(j)}(x) \approx h^{(j)}(x)$.

7.1 A Lagrange-interpoláció esete

Adott $x_1 < x_2 < \ldots < x_n$ alappontok és $y_i = f(x_i)$ $(i = 1, \ldots, n)$ függvényértékek esetén az $f(x)$ függvény Lagrange-féle interpolációs polinomja

$$p(x) = \sum_{i=1}^{n} y_i l_i(x).$$

Az $f(x)$ függény x-pontbeli j-edik deriváltjának közelítését az

$$f^{(j)}(x) \approx p^{(j)}(x) = \sum_{i=1}^{n} y_i l_i^{(j)}(x) \tag{112}$$

összefüggés adja meg, amelynek hibájára fennáll az

$$\left| f^{(j)}(x) - p^{(j)}(x) \right| \leq \sum_{i=0}^{j} \frac{j!}{(j-i)!(n+i)!} \max_{x \in [a,b]} \left| f^{(n+i)}(x) \right| \left| \omega^{(j-i)}(x) \right| \tag{113}$$

egyenlőtlenség, ahol $x, x_1, x_n \in [a, b]$ és $\omega(x) = (x - x_1)(x - x_2) \ldots (x - x_n)$.

Legyen $n = 2k + 1$, az alappontok pedig legyenek

$$t - kh, \ldots, t - h, t, t + h, \ldots, t + kh. \tag{114}$$

Ha $p(x)$ az $f(x)$ függvény ezen pontokra támaszkodó Lagrange-féle interpolációs polinomja, akkor igaz, hogy

$$|f'(t) - p'(t)| < \frac{Kh^{2k}}{\binom{2k}{k}(2k+1)}, \tag{115}$$

ahol K az $\left| f^{(2k+1)}(x) \right|$ korlátja a $[t - kh, t + kh]$ intervallumon.

Az $n = 3$ esetben az

$$f'(t) \approx \frac{1}{2h} \left[f(t+h) - f(t-h) \right], \tag{116}$$

az $n = 5$ esetben pedig az

$$f'(t) \approx \frac{1}{12h} \left[f(t-2h) - 8f(t-h) + 8f(t+h) - f(t+2h) \right] \tag{117}$$

közelítő formulát kapjuk.

A közelítés hibája az első esetben $O\left(h^2\right)$, a második esetben pedig $O\left(h^4\right)$. Végül megjegyezzük, hogy nagy n értékekre Lagrange-interpoláción alapuló numerikus deriválást ritkán alkalmaznak a fellépő numerikus instabilitás miatt.

7.2 Közelítés differencia hányadosokkal

A differencia hányadosok alkalmazása a közelítő deriválás legelterjedtebb módszere. Legcélszerűbben a Taylor-sorfejtés felhasználásával vezethetünk le közelítő formulákat.

Tegyük fel, hogy $f \in C^2$ és írjuk fel $f(x)$ másodfokú Taylor-polinomját:

$$f(x+h) = f(x) + hf'(x) + \frac{h^2}{2}f''(\xi) \quad (x < \xi < x+h).$$

Egyszerű számolással adódik, hogy

$$\frac{f(x+h) - f(x)}{h} = f'(x) + \frac{h}{2}f''(\xi),$$

ahonnan a

$$f'(x) \approx \frac{f(x+h) - f(x)}{h} \tag{118}$$

közelítést kapjuk, amelynek hibája $O(h)$.

Ennél pontosabb közelítést kaphatunk, ha $f \in C^3$. Legyen

$$f(x+h) = f(x) + hf'(x) + \frac{h^2}{2}f''(x) + \frac{h^3}{6}f^{(3)}(\xi_1), \quad x < \xi_1 < x+h,$$

$$f(x-h) = f(x) - hf'(x) + \frac{h^2}{2}f''(x) - \frac{h^3}{6}f^{(3)}(\xi_2), \quad x-h < \xi_2 < x.$$

Kivonással és átrendezéssel kapjuk, hogy

$$\frac{f(x+h) - f(x-h)}{2h} = f'(x) + \frac{h^2}{12}[f^{(3)}(\xi_1) + f^{(3)}(\xi_2)] = f'(x) + \frac{h^2}{6}f^{(3)}(\xi_3),$$

ahol $x-h < \xi_3 < x+h$. Az ebből adódó

$$f'(x) \approx \frac{f(x+h) - f(x-h)}{2h} \tag{119}$$

közelítés hibája $O(h^2)$ nagyságrendű.

Magasabbrendű deriváltak

$$f^{(j)}(x) \approx \frac{1}{h^j} \sum_{i=-m}^{n} c_i f(x+ih) \tag{120}$$

közelítéseit hasonló módon lehet megkonstruálni. Általában (120) alakú közelítéseket keresünk, ahol $n+m \geq j$. Az ismeretlen c_i együtthatókat a pontossági követelményből vezethetjük le. Legyen

$$f(x+ih) = \sum_{l=0}^{j} f^{(l)}(x) \frac{i^l h^l}{l!} + O\left(h^{j+1}\right)$$

és

$$\sum_{i=-m}^{n} c_i f(x+ih) = \sum_{i=-m}^{n} c_i \left(\sum_{l=0}^{j} f^{(l)}(x) \frac{i^l h^l}{l!} \right) + O\left(h^{j+1}\right).$$

Átrendezéssel kapjuk, hogy

$$\sum_{i=-m}^{n} c_i f(x+ih) = \sum_{l=0}^{j} f^{(l)}(x) \frac{h^l}{l!} \left(\sum_{i=-m}^{n} c_i i^l \right) + O\left(h^{j+1}\right).$$

Ha most fennáll, hogy

$$\sum_{i=-m}^{n} c_i i^l = 0 \quad (0 \le l \le j-1), \quad \sum_{i=-m}^{n} c_i i^j = j!,$$

akkor

$$\sum_{i=-m}^{n} c_i f(x+ih) = f^{(j)}(x) h^j + O\left(h^{j+1}\right).$$

Innen az $O(h)$ pontosságú (120) formulát kapjuk.

A második derivált ismert közelítései az

$$f''(x) \approx \frac{f(x) - 2f(x+h) + f(x+2h)}{h^2} \tag{121}$$

és az

$$f''(x) \approx \frac{f(x+h) - 2f(x) + f(x-h)}{h^2} \tag{122}$$

képletek. Az utóbbi hibája $O\left(h^2\right)$. Általában is igaz, hogy az ún. *centrális differencia formulák* ($m = n$ eset) egy nagyságrenddel jobb közelítést adnak.

A deriválás instabil művelet.

Legyen $f(x)$ tetszőleges differenciálható függvény. Ha ezt a függvényt a differenciálható $\eta(x)$ függvénynyel megváltoztatjuk (perturbáljuk), akkor a deriváltak megváltozása

$$|f'(x) - (f'(x) + \eta'(x))| = |\eta'(x)|.$$

Megmutatjuk, hogy tetszőlegesen kis $\eta(x)$ esetén is lehet $\eta'(x)$ nagy. Legyen $\eta(x) = \epsilon \sin\left(\frac{x}{\epsilon^2}\right)$, amelyre fennáll, hogy $|\eta(x)| \le \epsilon$. Minthogy $\eta'(x) = \frac{1}{\epsilon} \cos\left(\frac{x}{\epsilon^2}\right)$, az $x = 0$ pontban a derivált megváltozása $|\eta'(0)| = \frac{1}{\epsilon}$. Ha $\epsilon \to 0$, akkor ez tetszőleges nagy lehet.

Vizsgáljuk most a perturbációk hatását a numerikus deriválás esetén. Legyen

$$D_h(f(x)) = (f(x+h) - f(x))/h.$$

Ennek hibája $|D_h(f(x)) - f'(x)| \le (K/2)h$, ahol $K = \max_{x \in [x, x+h]} |f''(x)|$. Tegyük fel, hogy $f(x)$ helyett az $\widetilde{f}(x)$ perturbált függvénnyel számolunk, amelyre teljesül, hogy $\left|\widetilde{f}(t) - f(t)\right| \le \epsilon/2$ ($t \in [x, x+h]$). Ekkor $f'(x)$ közelítésének hibája

$$\left| D_h\left(\widetilde{f}(x) - f(x)\right) \right| \le \epsilon/h$$

miatt

$$\left|D_h\left(\widetilde{f}\,(x)\right) - f'\,(x)\right| = \left|D_h\left(f\,(x)\right) - f'\,(x) + D_h\left(\widetilde{f}\,(x) - f\,(x)\right)\right| \leq \frac{K}{2}h + \frac{\epsilon}{h}.$$

Ha rögzített ϵ esetén $h \to 0$, akkor a hibakorlát végtelenhez tart. Tehát h és ϵ megválasztása célszerűen nem független egymástól. A most kapott hibakorlátot a $h = \sqrt{2\epsilon/K}$ választás minimalizálja. Ekkor a korlát értéke $\sqrt{2K\epsilon}$. Ha K értéke, vagy jó becslése nem ismert, akkor a $h = c_1\sqrt{\epsilon}$ választást használjuk egy c_1 tapasztalati konstanssal. Ez választás a becslés $O\left(\sqrt{\epsilon}\right)$ nagyságrendjét tekintve helyes eredményt szolgáltat. Dupla pontosságú lebegőpontos aritmetikában $\epsilon \geq \epsilon_M \approx 2.2204 \times 10^{-16}$. A $\epsilon = \epsilon_M$ esetben a hiba mértéke 10^{-8} nagyságrendű.

8 NUMERIKUS INTEGRÁLÁS

Numerikus integrálást (numerikus kvadraturát) általában akkor végzünk, ha

- a primitív függvény nem ismert,
- vagy nem állítható elő könnyen,
- ha az $f(x)$ függvénynek csak véges sok értéke ismert.

A numerikus eljárások alapötlete sematikusan a következő:

$$f(x) \approx h(x) \Rightarrow \int_a^b f(x)dx \approx \int_a^b h(x)dx. \tag{123}$$

8.1 Interpolációs eljárások

Legyen adott $a \leq x_1 < x_2 < \ldots < x_n \leq b$ és $y_i = f(x_i)$ $(i = 1, \ldots, n)$. Az $f\,(x)$ függvényt a Lagrange-féle interpolációs polinommal közelítve kapjuk, hogy

$$\int_a^b f(x)dx \approx \int_a^b p(x)dx = \int_a^b \left[\sum_{i=1}^n y_i l_i(x)\right] dx = \sum_{i=1}^n y_i \int_a^b l_i(x)dx. \tag{124}$$

Az ilyen közelítéseket *interpolációs típusú kvadratura (integráló) formuláknak* nevezzük. A közelítés hibájára $f \in C^n\,[a,b]$ esetén a Cauchy-tétel alapján fennáll, hogy

$$R_n\,(f) = \int_a^b f(x)dx - \int_a^b p(x)dx = \frac{1}{n!}\int_a^b f^{(n)}(\xi_x)(x-x_1)(x-x_2)\ldots(x-x_n)dx.$$

Ha $\left|f^{(n)}(x)\right| \leq M_n$ $(x \in [a,b])$, akkor

$$|R_n\,(f)| = \left|\int_a^b f(x)dx - \int_a^b p(x)dx\right| \leq \frac{M_n}{n!}(b-a)^{n+1}. \tag{125}$$

Nagy n értékekre ez a közelítés igen rosszul viselkedhet.

Helyette ún. *összetett kvadratura (integráló) formulákat* használunk.

Ezek lényege: az $[a,b]$ intervallumot felosztjuk részintervallumokra, az egyes részintervallumokra alkalmazunk egy előre rögzített kvadraturaformulát, és az így kapott részeredményeket összegezzük.

8.2 A trapézformula

Legyen $x_1 = a$ és $x_2 = b$. A két pontra támaszkodó elsőfokú Lagrange-féle interpolációs polinom

$$p(x) = f(x_1)\frac{x - x_2}{x_1 - x_2} + f(x_2)\frac{x - x_1}{x_2 - x_1}, \tag{126}$$

amelynek határozott integrálja

$$\begin{aligned}\int_{x_1}^{x_2} p(x)dx &= \left[f(x_1)\frac{(x-x_2)^2}{2(x_1-x_2)} + f(x_2)\frac{(x-x_1)^2}{2(x_2-x_1)}\right]_{x_1}^{x_2} \\ &= \frac{x_2 - x_1}{2}\left[f(x_1) + f(x_2)\right].\end{aligned}$$

Ha $f(x_1)$ és $f(x_2)$ előjele azonos, akkor ez az eredmény az $(x_1, 0)$, $(x_2, 0)$, $(x_2, f(x_2))$, $(x_1, f(x_1))$ pontok által határolt trapéz területe. A kapott

$$\int_a^b f(x)\,dx \approx \frac{b-a}{2}\left[f(a) + f(b)\right] \tag{127}$$

közelítés hibájára fennáll, hogy

$$\left|\int_a^b f(x)dx - \frac{b-a}{2}\left[f(a) + f(b)\right]\right| \leq \frac{M_2}{12}(b-a)^3. \tag{128}$$

Ha az $[a, b]$ intervallumot felbontjuk az

$$a = x_1 < x_2 < \ldots < x_{n+1} = b \tag{129}$$

pontokkal n részintervallumra, akkor az *összetett trapézformula* a következő:

$$\int_a^b f(x)dx \approx T_n(f) = \sum_{i=1}^{n} \frac{x_{i+1} - x_i}{2}\left[f(x_i) + f(x_{i+1})\right]. \tag{130}$$

A képlet hibájára $f \in C^2[a, b]$ esetén fennáll, hogy

$$\left|\int_a^b f(x)dx - \sum_{i=1}^{n} \frac{x_{i+1} - x_i}{2}\left[f(x_i) + f(x_{i+1})\right]\right| \leq \frac{M_2}{12}\sum_{i=1}^{n}(x_{i+1} - x_i)^3. \tag{131}$$

Ha az alappontok ekvidisztánsak, azaz $x_i = x_1 + (i-1)h$ ($h = \frac{b-a}{n}$, $i = 1, \ldots, n+1$), akkor a képlet alakja egyszerűsödik:

$$\int_a^b f(x)dx \approx T_n(f) = \frac{h}{2}\left[f(x_1) + 2\sum_{i=2}^{n} f(x_i) + f(x_{n+1})\right]. \tag{132}$$

A képlet hibájára pedig $nh = b - a$ miatt fennáll, hogy

$$\left|\int_a^b f(x)dx - T_n(f)\right| \leq \frac{M_2(b-a)h^2}{12} = \frac{M_2(b-a)^3}{12n^2}. \tag{133}$$

8.3 A Simpson formula

Legyen $x_1 = a$, $x_2 = \frac{a+b}{2}$ és $x_3 = b$. Tekintsük a három pontra támaszkodó másodfokú Lagrange-féle interpolációs polinomot:

$$\begin{aligned} p(x) &= f(x_1)\frac{(x-x_2)(x-x_3)}{(x_1-x_2)(x_2-x_3)} + f(x_2)\frac{(x-x_1)(x-x_3)}{(x_2-x_1)(x_2-x_3)} + \\ &\quad + f(x_3)\frac{(x-x_1)(x-x_2)}{(x_3-x_1)(x_3-x_2)}. \end{aligned}$$

Ennek az $[a,b]$ intervallumon vett integrálja adja a következő közelítő formulát

$$\int_a^b f(x)dx \approx \frac{b-a}{6}\left[f(a) + 4f(\frac{a+b}{2}) + f(b)\right], \tag{134}$$

amelynek hibájára $f \in C^4[a,b]$ esetén fennáll, hogy

$$\left|\int_a^b f(x)dx - \frac{b-a}{6}\left[f(a) + 4f(\frac{a+b}{2}) + f(b)\right]\right| \leq M_4\frac{(b-a)^5}{2880}. \tag{135}$$

Ha az $[a,b]$ intervallumot itt is felbontjuk az

$$a = x_1 < x_2 < \ldots < x_{n+1} = b$$

pontokkal n részintervallumra, akkor az *összetett Simpson formula* a következő:

$$\int_a^b f(x)dx \approx S_n(f) = \sum_{i=1}^{n} \frac{x_{i+1}-x_i}{6}\left[f(x_i) + 4f(\frac{x_i+x_{i+1}}{2}) + f(x_{i+1})\right].$$

Ennek hibájára fennáll, hogy

$$\left|\int_a^b f(x)dx - S_n(f)\right| \leq \frac{M_4}{2880}\sum_{i=1}^{n}(x_{i+1}-x_i)^5.$$

Ha az alappontok ekvidisztánsak, azaz $x_i = x_1+(i-1)h$ ($h = \frac{b-a}{n}$, $i = 1,\ldots,n+1$), akkor a képlet alakja

$$S_n(f) = \frac{h}{6}\left[f(x_1) + 2\sum_{i=2}^{n} f(x_i) + 4\sum_{i=1}^{n} f(x_i + \frac{h}{2}) + f(x_{n+1})\right], \tag{136}$$

amelynek képlethibájára fennáll, hogy

$$\left|\int_a^b f(x)dx - S_n(f)\right| \leq \frac{M_4(b-a)}{2880}h^4 = \frac{M_4(b-a)^5}{2880n^4}. \tag{137}$$

8.4 Kvadraturaformulák hibáinak utólagos becslése

Alapgondolata egyszerű, az extrapoláció (Runge-féle szabály) általános ötletét használja.

Elvégezzük a numerikus integrálást n és $2n$ részintervallum esetén. Ha fennáll, hogy

$$|T_n(f) - T_{2n}(f)| \leq \varepsilon, \tag{138}$$

akkor a $T_{2n}(f)$ közelítést ε pontosságúnak fogadjuk el.

Ugyanezt csináljuk a Simpson-formula esetén is. Igazolhatók a következő állítások.

Tétel *(Rownland-Miel). Ha $f''(x)$ előjele állandó, akkor az összetett trapézmódszer hibájára fennáll, hogy*

$$\left| \int_a^b f(x)dx - T_{2n}(f) \right| \leq |T_n(f) - T_{2n}(f)| . \tag{139}$$

Tétel *(Rownland-Miel). Ha $f^{(4)}(x)$ előjele állandó, akkor az összetett Simpson-formula hibájára fennáll, hogy*

$$\left| \int_a^b f(x)dx - S_{2n}(f) \right| \leq |S_n(f) - S_{2n}(f)| . \tag{140}$$

8.5 A MATLAB kvadratura eljárása

- quad (adaptív Simpson)

9 NEMLINEÁRIS EGYENLETEK

Az

$$f(x) = 0 \quad (\ f : \mathbb{R} \to \mathbb{R}) \tag{141}$$

alakú egyenletek közelítő megoldási módszereit vizsgáljuk. Az $x^* \in \mathbb{R}$ elemet az egyenlet megoldásának nevezzük, ha $f(x^*) = 0$. Minden esetben feltesszük, hogy $f(x)$ folytonos.

Az $f(x) = 0$ egyenlet javasolt megoldási módszerei egy, az x^* megoldáshoz konvergáló $\{x_i\}_{i=0}^{\infty}$ sorozatot képeznek. A konvergencia sebességét az alábbi módon jellemezhetjük.

Definíció. *Az $\{x_i\}_{i=0}^{\infty} \in \mathbb{R}$ $(n \geq 1)$ sorozat lineáris sebességgel konvergál egy $x^* \in \mathbb{R}$ határértékhez, ha létezik egy $0 \leq q < 1$ konstans úgy, hogy $|x_i - x^*| \leq q\,|x_{i-1} - x^*|$ $(i \geq 1)$.*

A lineáris konvergencia sebesség esetén fennáll, hogy

$$|x_i - x^*| \leq q\,|x_{i-1} - x^*| \leq q^2\,|x_{i-2} - x^*| \leq \ldots \leq q^i\,|x_0 - x^*| , \tag{142}$$

tehát az x_i közelítések hibáit egy nullához tartó mértani sorozat tagjaival tudjuk felülről becsülni.

Definíció. *Az* $\{x_i\}_{i=0}^{\infty} \in \mathbb{R}$ *(*$n \geq 1$*) sorozat p-ed rendű sebességgel (*$p > 1$*) konvergál egy* $x^* \in \mathbb{R}$ *határértékhez, ha létezik egy* $\gamma > 0$ *konstans úgy, hogy* $|x_i - x^*| \leq \gamma |x_{i-1} - x^*|^p$ *(*$i \geq 1$*).*

A p-edrendű konvergencia sebesség a lineárisnál lényegesen gyorsabb. Teljes indukcióval igazolhatjuk, hogy

$$|x_i - x^*| \leq \frac{1}{\sqrt[p-1]{\gamma}} \left(\sqrt[p-1]{\gamma} \, |x_0 - x^*| \right)^{p^i} \quad (i \geq 1) \,. \tag{143}$$

Ha $q = \sqrt[p-1]{\gamma} \, |x_0 - x^*| < 1$, akkor az x_i közelítések hibáit a $\left\{ cq^{p^i} \right\}$ nullához tartó sorozat becsli felülről ($c = 1/\sqrt[p-1]{\gamma}$). Ez nyilvánvalóan gyorsabban tart 0-hoz mint a cq^i mértani sorozat.

9.1 Az intervallumfelező eljárás

Tegyük fel, hogy $f : \mathbb{R} \to \mathbb{R}$ folytonos az $[a, b]$ intervallumon és fennáll, hogy

$$f(a) f(b) < 0. \tag{144}$$

Ekkor a Bolzano-tétel miatt az $f(x) = 0$ egyenletnek van legalább egy $x^* \in (a, b)$ gyöke. Ezt a Bolzano–tétel bizonyításából ismert eljárással kaphatjuk meg.

Legyen $c = (a + b)/2$ és vizsgáljuk az $f(c)$ értékét. Ha $f(a) f(c) < 0$, akkor az $[a, c]$ intervallumban van gyök. Egyébként a $[c, b]$ intervallum tartalmaz gyököt. Az új intervallumot újra megfelezzük és így tovább. Az egymásba skatulyázott zárt intervallumok ráhúzódnak az egyenlet egy gyökére.

Algoritmikus formában:

$[a_1, b_1] = [a, b]$,

$c_i = (a_i + b_i)/2$,

$$[a_{i+1}, b_{i+1}] = \begin{cases} [a_i, c_i] \,, & \text{ha } f(a_i) f(c_i) < 0 \\ [c_i, b_i] \,, & \text{egyébként} \end{cases} \,, \quad (i = 1, 2, \ldots) \,.$$

Az x^* gyököt az $[a_i, b_i]$ intervallum tetszőleges y pontjával közelíthetjük. Az y közelítés hibájára fennáll, hogy

$$|x^* - y| \leq \max\{y - a_i, b_i - y\}. \tag{145}$$

A $\max\{y - a_i, b_i - y\}$ korlát akkor a legkisebb, ha $y = \frac{a_i + b_i}{2}$. Ezért az x^* gyök i-edik közelítéseként általában az $x_i = (a_i + b_i)/2$ felezőpontot használjuk. Nyilván

$$|x^* - x_i| \leq \frac{b_i - a_i}{2} = \frac{b - a}{2^i} \quad (i = 1, 2, \ldots). \tag{146}$$

Az algoritmust akkor állítjuk le, ha a közelítés hibája kisebb, mint egy előre megadott $\varepsilon > 0$ hibakorlát.

AZ INTERVALLUMFELEZŐ ALGORITMUS:

input $[a, b] \,, \varepsilon > 0$.

while $b - a > 2\varepsilon$

$x = (a + b)/2$

if $f(a) f(x) < 0$

$b = x$
else
$a = x$
end
end
$x = (a+b)/2$

Megjegyezzük, hogy az $\{x_i\}$ sorozat csak folytonos $f(x)$ esetén konvergál biztosan az x^* gyökhöz. (Egyébként még a gyök létezése sem garantált.)

Példa. Legyen $f(x) = 4\left(1-x^2\right) - e^x = 0$ és határozzuk meg a gyököket. Az egyenletnek a $[-1,0]$, ill. $[0,1]$ intervallumokon vannak gyökei, ui.

$$f(-1)\,f(0) = -3e^{-1} < 0, \quad f(0)\,f(1) = -3e < 0.$$

A felező módszer tehát alkalmazható. Az $\varepsilon = 10^{-6}$ pontosságú közelítéshez szükséges lépések számát mindkét intervallum esetén a $|x_i - x^*| \le \frac{b-a}{2^i} = \frac{1}{2^i} \le \varepsilon$ egyenlőtlenség megoldása adja. Eszerint $i \ge -\log\varepsilon/\log 2 \approx 19.93$, azaz $i = 20$ lépés szükséges ($x_1 \approx -0.950455$, $x_2 = 0.703439$).

9.2 A fixpont iterációs módszer

A módszert az $f(x) = x - g(x) = 0$ alakú vagy ilyen alakra hozott egyenletek esetén alkalmazzuk. Az $f(x) = 0$ egyenlet ekvivalens az

$$x = g(x) \tag{147}$$

egyenlettel. Az x^* pontot a $g(x)$ leképezés fixpontjának nevezzük, ha $x^* = g(x^*)$. Leképezések fixpontjára vonatkozik a következő

Tétel. *Ha $g \in C[a,b]$ és $a \le g(x) \le b$ minden $x \in [a,b]$ esetén, akkor a $g(x)$ függvénynek az $[a,b]$ intervallumon van fixpontja.*

Bizonyítás. Feltehetjük, hogy $g(a) > a$ és $g(b) < b$. Legyen $h(x) = g(x) - x$. Ekkor $h(x)$ folytonos $[a,b]$-n és $h(a) > 0$ és $h(b) < 0$. Ezért a $h(x)$ függvénynek van $\xi \in (a,b)$ gyöke, azaz $h(\xi) = g(\xi) - \xi = 0$. Tehát ξ fixpont. □

dtbpF8.4899cm7.161cm0ptnumfix.wmf

Definíció. *A $g \in C[a,b]$ függvény kontrakció az $[a,b]$ intervallumon, ha létezik $0 \le q < 1$ úgy, hogy*

$$|g(x) - g(y)| \le q\,|x-y|, \quad x, y \in [a,b]. \tag{148}$$

Példa. A $g(x) = x^2$ függvény kontrakció a $\left[0, \frac{1}{4}\right]$ intervallumon, ui.

$$\left|x^2 - y^2\right| = \le 1/2\underbrace{|x+y|}\,|x-y| \le \frac{1}{2}\,|x-y|, \quad x, y \in \left[0, \frac{1}{4}\right].$$

Példa. A $g(x) = x^2$ függvény nem kontrakció a $[0,1]$ intervallumon, ui. $x, y \in \left[\frac{3}{4}, 1\right]$ esetén

$$\left|x^2 - y^2\right| = \ge 3/2\underbrace{|x+y|}\,|x-y| \ge \frac{3}{2}\,|x-y| > |x-y| \quad (x \ne y)$$

Tétel. *Ha* $g \in C[a,b]$, $a \leq g(x) \leq b$ *(*$x \in [a,b]$*) és* $g(x)$ *kontrakció* $[a,b]$*-n, akkor pontosan egy fixpont létezik* $[a,b]$*-ben.*

Bizonyítás. Az előző Tétel a fixpont létezését bizonyítja. Tegyük fel, hogy $x^*, y^* \in [a,b]$ fixpontok és $x^* \neq y^*$. Ekkor fennáll, hogy

$$|x^* - y^*| = |g(x^*) - g(y^*)| \leq q\,|x^* - y^*|,$$

ahonnan osztással az $1 \leq q < 1$ ellentmondást kapjuk. Tehát csak egy fixpont van. □

Tétel. *A* $g \in C^1[a,b]$ *függvény kontrakció az* $[a,b]$ *intervallumon, ha*

$$\max_{x \in [a,b]} |g'(x)| = q < 1. \tag{149}$$

Bizonyítás. A Lagrange-tétel alapján

$$|g(x) - g(y)| = |g'(\xi)(x-y)| = |g'(\xi)|\,|x-y| \leq q\,|x-y|. \ \square$$

A következő tétel megadja a fixpont iterációs módszert és a konvergenciájára vonatkozó feltételeket.

Tétel. *Legyen* $g \in C[a,b]$ *olyan, hogy* $a \leq g(x) \leq b$ *(*$x \in [a,b]$*) és tegyük fel, hogy* $g(x)$ *kontrakció* $[a,b]$*-n. Ekkor minden* $x_0 \in [a,b]$ *esetén az*

$$x_{i+1} = g(x_i) \quad (i = 0, 1, \ldots) \tag{150}$$

iteráció sorozat lineáris sebességgel konvergál az x^* *fixponthoz, azaz*

$$|x_i - x^*| \leq q^i\,|x_0 - x^*| \quad (i = 0, 1, \ldots). \tag{151}$$

Bizonyítás. Minthogy $g(x) \in [a,b]$ minden $x \in [a,b]$-re, azért $\{x_i\}_{i=0}^{\infty} \subset [a,b]$. Igaz a következő becslés:

$$|x_{n+1} - x_n| = |g(x_n) - g(x_{n-1})| \leq q\,|x_n - x_{n-1}| \leq \ldots \leq q^n\,|x_1 - x_0|.$$

Ennek segítségével belátjuk, hogy $\{x_i\}_{i=0}^{\infty}$ Cauchy-sorozat, azaz $|x_m - x_n| \to 0$, hacsak $m, n \to \infty$. Egyszerű becsléssel kapjuk, hogy $m > n$ esetén

$$\begin{gathered} |x_m - x_n| \leq |x_m - x_{m-1}| + \ldots + |x_{n+1} - x_n| \leq \\ \leq \left(q^{m-n-1} + \ldots + 1\right)|x_{n+1} - x_n| \leq \tfrac{1-q^{m-n}}{1-q}\,|x_{n+1} - x_n| \leq \\ \leq q^n\,\tfrac{1-q^{m-n}}{1-q}\,|x_1 - x_0| \leq \tfrac{q^n}{1-q}\,|x_1 - x_0|, \end{gathered}$$

ahonnan $m, n \to \infty$ esetén $|x_m - x_n| \to 0$ következik. Az $\{x_i\}_{i=0}^{\infty} \subset [a,b]$ Cauchy-sorozat egyúttal konvergens is, tehát létezik $x^* \in [a,b]$ határértéke. A $g(x)$ függvény folytonossága miatt fennáll a

$$\begin{array}{ccc} x_{i+1} & = & g(x_i) \\ \downarrow & & \downarrow \\ x^* & = & g(x^*) \end{array}$$

diagram helyessége. Tehát x^* fixpont. A fenti egyenlőtlenség-láncból az $x_m \to x^*$ határátmenettel kapjuk, hogy

$$|x_n - x^*| \leq \frac{q^n}{1-q} |x_1 - x_0|, \quad n \geq 0.$$

A lineáris konvergencia ebből már következik. A tételben szereplő becslést az

$$|x_n - x^*| = |g(x_{n-1}) - g(x^*)| \leq q |x_{n-1} - x^*| \leq \ldots \leq q^n |x_0 - x^*|$$

egyenlőtlenség láncból kapjuk. □

A FIXPONT ITERÁCIÓS ELJÁRÁS ALGORITMUSA:

Input x_0, $\varepsilon > 0$.
while kilépési feltétel=hamis
$x_{i+1} = g(x_i)$,
$i = i + 1$
end

Ha ismert a q érték, vagy egy jó becslése, akkor az $\varepsilon > 0$ pontosság eléréséhez szükséges iterációk számát a

$$\frac{q^n}{1-q} |x_1 - x_0| \leq \varepsilon \tag{152}$$

egyenlőtlenség megoldásával kaphatjuk meg. Alkalmazva az $||a| - |b|| \leq |a - b|$ és

$$|x^* - x_n| - |x_n - x_{n+1}| \leq |x^* - x_{n+1}| \leq q |x^* - x_n|$$

egyenlőtlenségeket kapjuk, hogy

$$|x^* - x_n| \leq \frac{1}{1-q} |x_{n+1} - x_n|. \tag{153}$$

Ha teljesül, hogy

$$|x_{n+1} - x_n| \leq (1-q)\varepsilon, \tag{154}$$

akkor az x_n közelítés abszolút hibája kisebb mint ε. Ekkor az iterációt leállíthatjuk, abból kiléphetünk.

Ha ismerjük a q értékét és fennáll a konvergencia, akkor a fenti egyenlőtlenségek alapján megállapíthatjuk a szükséges iterációk számát, vagy a megoldáshoz való közelséget.

Általában azonban nem ez a helyzet. Vagy nem tudjuk q pontos értékét, vagy azt nem tudjuk, hogy egy adott x_0 pont olyan pont-e, amelyből indulva a konvergencia garantálható.

Ennek ellenére általános az alábbi kilépési feltételek használata:

$$(B) \quad |x_{i+1} - x_i| \leq c_2\varepsilon; \qquad (C) \quad i = i_{\max}. \tag{155}$$

Minthogy a feltételek egyike sem garantálja az $|x_{i+1} - x^*| \leq \varepsilon$ feltétel teljesülését, célszerű az (B) és (C) feltételt együtt használni.

A konvergencia tétel feltételeit, a kontraktivitást, de különösen az $a \leq g(x) \leq b$ feltételt általában nem könnyű biztosítani.

Az $f(x) = 0$ alakban megadott egyenletek átírása az $x = g(x)$ formára igen könnyű, ui. $x = x - f(x)$ ilyen alak. A kontraktivitás biztosítása azonban korántsem egyszerű feladat. Sok esetben az ekvivalens

$$x = x - \alpha f(x) \tag{156}$$

fixpont feladatot vizsgáljuk, ahol az α konstanst vagy $\alpha(x)$ függvényt úgy választjuk meg, hogy a $g(x) = x - \alpha f(x)$ függvény kontrakció legyen. Ilyen tulajdonképpen a következő szakaszban ismertetésre kerülő Newton-módszer is.

Végül megjegyezzük, hogy lebegőpontos aritmetikában előfordulhat, hogy a sorozat nem konvergál a fixponthoz, hanem körülötte "beciklizál".

9.3 A Newton-módszer

Tegyük fel, hogy $f : \mathbb{R} \to \mathbb{R}$ folytonosan differenciálható. A módszer lényege, hogy az x_i pontban a függvényhez érintőt húzunk és ennek az érintőnek a zérushelye adja meg a keresett gyök $(i+1)$-edik közelítését, azaz x_{i+1}-et. Az érintő iránytangense $f'(x_i)$ és egyenlete

$$y - f(x_i) = f'(x_i)(x - x_i). \tag{157}$$

Az $y = 0$ egyenlet megoldása:

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)}, \tag{158}$$

feltéve, hogy $f'(x_i) \neq 0$. E képlethez eljuthatunk egy kissé más érveléssel is. Nevezetesen $f(x)$-et linearizáljuk az x_i pontban, azaz közelítjük az elsőfokú Taylor-polinomjával:

$$f(x) \approx f(x_i) + f'(x_i)(x - x_i). \tag{159}$$

Ezután az $f(x) = 0$ egyenlet helyettesítjük a $f(x_i) + f'(x_i)(x - x_i) = 0$ egyenlettel, amelynek gyöke közelíti az $f(x) = 0$ egyenlet gyökét.

A Newton-módszer tehát a következő. Adott egy $x_0 \in \mathbb{R}$ kezdeti közelítés és képezzük az

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)} \quad (i = 0, 1, \ldots) \tag{160}$$

sorozatot.

Vegyük észre: a Newton-módszer tulajdonképpen az $x = x - f(x)/f'(x)$ fixpont feladatra alkalmazott iterációs eljárás. A Newton-módszer konvergenciájára vonatkozik az alábbi

Tétel. *Legyen $f : (a, b) \to \mathbb{R}$ kétszer folytonosan differenciálható, $|f''(x)| \leq \gamma$ és $|f'(x)| \geq \rho > 0$ $(x \in (a, b))$. Ha az $f(x) = 0$ egyenletnek van egy x^* gyöke*

az (a,b) intervallumban, akkor van egy olyan $\eta > 0$ szám, hogy $|x_0 - x^*| < \eta$ esetén $x_i \to x^*$ $(i \to +\infty)$ és

$$|x_{i+1} - x^*| \leq \frac{\gamma}{2\rho} |x_i - x^*|^2 \quad (i = 0, 1, \ldots). \tag{161}$$

Bizonyítás. Legyen $\eta_1 = \min\{x^* - a, b - x^*\} > 0$. Tekintsük a

$$f(x^*) = f(x_i) + f'(x_i)(x^* - x_i) + (1/2) f''(\xi_i)(x^* - x_i)^2$$

Taylor-sort, ahonnan $f(x^*) = 0$ miatt

$$f(x_i) = -f'(x_i)(x^* - x_i) - (1/2) f''(\xi_i)(x^* - x_i)^2$$

következik. Behelyettesítéssel kapjuk, hogy

$$x_{i+1} = x_i + (x^* - x_i) + \frac{1}{2}\frac{f''(\xi_i)}{f'(x_i)}(x_i - x^*)^2,$$

azaz

$$x_{i+1} - x^* = \frac{1}{2}\frac{f''(\xi_i)}{f'(x_i)}(x_i - x^*)^2.$$

Innen a $|f''(x)| \leq \gamma$ és $|f'(x)| \geq \rho > 0$ $(x \in (a,b))$ feltevések miatt

$$|x_{i+1} - x^*| \leq \frac{\gamma}{2\rho} |x_i - x^*|^2$$

következik, feltéve, hogy $x_i \in (a,b)$. Ha $|x_0 - x^*| < \eta = \min\{\eta_1, 2\rho/\gamma\}$, akkor

$$|x_1 - x^*| \leq \left(\frac{\gamma}{2\rho}|x_0 - x^*|\right)|x_0 - x^*| \leq |x_0 - x^*| < \eta$$

miatt $x_1 \in (x^* - \eta, x^* + \eta) \subset (a,b)$. Hasonlóan folytatva könnyen igazolhatjuk, hogy $x_i \in (a,b)$ és

$$|x_{i+1} - x^*| \leq \frac{2\rho}{\gamma}\left(\frac{\gamma}{2\rho}|x_0 - x^*|\right)^{2^{i+1}} \qquad (i = 0, 1, \ldots).$$

Tehát $|x_0 - x^*| < \eta$ esetben az $\{x_i\}_{i=0}^{\infty}$ sorozat másodrendben konvergens. □

Azt mondjuk, hogy a Newton-módszer konvergenciája lokális, mert az x_1 kezdeti közelítésnek az x^* gyök "közelében" kell lennie. A Newton-módszer másodrendű konvergenciáját az alábbi megjegyzéssel jellemezhetjük. Tegyük fel, hogy $x^* \neq 0$. Ekkor fennáll, hogy

$$\frac{|x_{i+1} - x^*|}{|x^*|} \leq \frac{\gamma|x^*|}{2\rho}\left(\frac{|x_i - x^*|}{|x^*|}\right)^2 \leq \frac{\gamma \max\{|a|, |b|\}}{2\rho}\left(\frac{|x_i - x^*|}{|x^*|}\right)^2. \tag{162}$$

Ez azt jelenti, hogy az x_{i+1} közelítés relatív hibája az i-edik közelítés relatív hibájának négyzete. Ha tehát beállt a közelítés első két tizedesjegye, akkor dupla pontosságú aritmetikában 3-4 lépésben beáll az elérhető legnagyobb pontosság.

Kilépési feltételek. A Newton-módszert elvileg akkor kell megállítanunk, amikor elértünk egy adott $\varepsilon > 0$ pontosságú közelítést, azaz fennáll, hogy

$$|x_i - x^*| \leq \varepsilon. \tag{163}$$

A gyök ismerete nélkül ezt a hibát ténylegesen becsülni nem tudjuk. Ezért különböző heurisztikus kilépési feltételeket használunk. A leggyakoribbak:

$$(A) \quad |f(x_i)| \leq \varepsilon_1; \quad (B) \quad |x_{i+1} - x_i| \leq \varepsilon_2; \quad (C) \quad i = i_{\max}. \tag{164}$$

A feltételek egyikének teljesülése sem garantálja a (163) pontossági feltétel teljesülését. Ezért célszerűbb a három feltételt együtt használni.

A függvényközelítéseknél korábban látott

$$x_{k+1} = \frac{1}{2}\left(x_k + \frac{a}{x_k}\right) \quad (k = 0, 1, \ldots) \tag{165}$$

négyzetgyök algoritmus nem más, mint a Newton-módszer az $f(x) = x^2 - a = 0$ egyenletre alkalmazva. Felmerül a kérédés, hogy milyen x_0 értékekre lesz az eljárás konvergens? Igaz a következő, Fourier-től származó

Tétel. *Legyen $f \in C^2[a,b]$, $f'(x) \neq 0$ és $f''(x) \neq 0$, ha $x \in [a,b]$. Tegyük fel, hogy létezik $x^* \in (a,b)$ gyök. Ha az $x_0 \in [a,b]$ pont olyan, hogy $f(x_0) f''(x_0) > 0$, akkor a Newton-módszer monoton konvergál az x^* megoldáshoz.*

A tételben szereplő $[a,b]$ végtelen intervallum is lehet. Esetünkben $f(x) = x^2 - a$ és $f''(x) = 2$. Ezért $x_0 > \sqrt{a}$ esetén $f(x_0) f''(x_0) > 0$. Tehát ilyen x_0 értékekre a Newton-módszer biztosan konvergál.

10 DIFFERENCIÁLEGYENLETEK KÖZELÍTŐ MEGOLDÁSA

Az

$$y' = f(x,y), \quad y(x_0) = y_0 \quad \left(f : \mathbb{R} \times \mathbb{R}^\ell \to \mathbb{R}^\ell\right) \tag{166}$$

alakú kezdetiérték feladatokat vizsgáljuk, ahol $f(x,y) = [f_1(x,y), \ldots, f_\ell(x,y)]^T$ ($x \in \mathbb{R}$, $y \in \mathbb{R}^\ell$) folytonos a

$$D = \{(x,y) : \|x - x_0\|_\infty < \kappa_x, \ \|y - y_0\|_\infty < \kappa_y\} \subseteq \mathbb{R}^{\ell+1} \tag{167}$$

nyílt tartományon, κ_x és κ_y pozitív konstansok és létezik olyan $L > 0$ konstans, hogy

$$\|f(x,y) - f(x,z)\|_\infty \leq L \|y - z\|_\infty \quad ((x,y),(x,z) \in D). \tag{168}$$

Ekkor minden $(x_0, y_0) \in D$ esetén a kezdetiérték feladatnak létezik pontosan egy $y(x) = [y_1(x), \ldots, y_\ell(x)]^T$ megoldása valamely $[x_0, B]$ intervallumon, azaz

$$y'(x) = f(x, y(x)), \quad x \in [x_0, B]. \tag{169}$$

A feladat numerikus megoldásán a következőt értjük. A megoldást egy $[x_0, b]$ intervallum ($b \leq B$) diszkrét pontjaiban keressük. Ezek a pontok legyenek

$$x_0 = t_0 < t_1 < \ldots < t_j < \ldots < t_N = b. \tag{170}$$

A $\{t_i\}_{i=0}^N$ alappont halmazt az $[x_0, b]$ intervallum felosztásának nevezzük. A felosztás ekvidisztans (egyentávolságú), ha

$$t_i = t_0 + ih \quad (i = 0, 1, \ldots, N), \quad h = \frac{b - t_0}{N}. \tag{171}$$

Az $y(x)$ elméleti megoldás t_i pontbeli közelítését jelölje y_i. Értelemszerűen $y(t_0) = y_0$. A $h_i = t_{i+1} - t_i > 0$ mennyiséget i-edik lépéshossznak nevezzük.

10.1 Az explicit Euler-módszer

A Cauchy-feladat (egyik) jellegzetessége az, hogy ha egy x pontban ismert az $y(x)$ megoldás vektor, akkor ismert az $y'(x) = f(x, y(x))$ derivált vektor is. A vektor komponensekre fennálló $y_i(x + h) \approx y_i(x) + hy_i'(x) = y_i(x) + h_i f(x, y(x))$ $(i = 1, \ldots, \ell)$ elsőrendű közelítéseket (tkp. érintőket) vektor formában felírva kapjuk az

$$y(x+h) = \begin{bmatrix} y_1(x+h) \\ \vdots \\ y_t(x+h) \\ \vdots \\ y_\ell(x+h) \end{bmatrix} \approx y(x) + hy'(x) = y(x) + hf(x, y(x))$$

közelítést. Ha az x pontban az $y(x) \approx \hat{y}$ közelítő érték ismert, akkor a fenti képlet átmegy az

$$y(x+h) \approx \hat{y} + hf(x, \hat{y})$$

közelítésbe. Az Euler-módszer alapgondolata ezek után a következő: A $t_1 = t_0 + h_0$ pontban közelítsük az $y(t_1)$ elméleti megoldást a megoldásgörbe (x_0, y_0) pontbeli "érintőjével", azaz legyen

$$y(t_0 + h_0) \approx y_0 + h_0 y'(t_0) = y_1 = y_0 + h_0 f(t_0, y_0). \tag{172}$$

A t_1 pontbeli y_1 közelítést felhasználva kapjuk, hogy

$$y(t_2) \approx y(t_1) + h_1 f(t_1, y(t_1)) \approx y_2 = y_1 + h_1 f(t_1, y_1). \tag{173}$$

Az eljárást folytatva kapjuk hogy

$$y_{i+1} = y_i + h_i f(x_i, y_i) \quad (i = 0, 1, \ldots, N-1), \tag{174}$$

ahol $y_i \approx y(t_i)$. Ezt a képletet nevezzük explicit Euler-módszernek. Grafikusan ábrázolva az eljárást skalár differenciálegyenlet ($f : \mathbb{R}^2 \to \mathbb{R}$) esetén a következő

ábrát nyerhetjük:dtbpF10.3461cm7.798cm0pteuler1u.wmfA rombusszal jelölt és egyenes szakaszokkal összekötött pontok az y_i $(i = 0, 1, 2)$ közelítő megoldások, amelyeken átmennek az $y'(x) = f(x, y)$, $y(t_i) = y_i$ differenciálegyenletek megoldásai $(t_i = 1 + 0.5(i-1)$, $i = 0, 1, 2)$.

A következőkben az eljárás hibáját elemezzük.

Definíció. *Az* $T(y(x), h) = y(x+h) - (y(x) + hf(x, y(x)))$ *mennyiséget az* x *pontbeli lokális hibának nevezzük.*

Definíció. *Az* $e_i = y_i - y(t_i)$ *hibát az* i*-edik pontbeli globális hibának nevezzük* $(i = 0, 1, \ldots, N)$*.*

A definíciók és a (168) feltételek alapján fennáll, hogy

$$\begin{aligned} \|y_{i+1} - y(t_{i+1})\|_\infty &= \|y_i + h_i f(t_i, y_i) - [y(t_i) + h_i f(t_i, y(t_i)) + T(y(t_i), h_i)]\|_\infty \\ &\leq \|y_i - y(t_i) + h_i [f(t_i, y_i) - f(t_i, y(t_i))]\|_\infty + \|T(y(t_i), h_i)\|_\infty \\ &\leq (1 + h_i L) \|y_i - y(t_i)\|_\infty + \|T(y(t_i), h_i)\|_\infty , \end{aligned}$$

azaz

$$\|e_{i+1}\|_\infty \leq (1 + h_i L) \|e_i\|_\infty + \|T(y(t_i), h_i)\|_\infty \quad (i = 0, 1, \ldots, N-1). \tag{175}$$

A globális hiba vizsgálatához két további eredményre van szükségünk. Igaz a következő

Lemma. *A* $\delta_{i+1} \leq \alpha_i \delta_i + \gamma_i$, $(\alpha_i \geq 0$, $i = 0, 1, \ldots, n)$ *egyenlőtlenség megoldása zárt alakban*

$$\delta_{n+1} \leq \left(\prod_{j=0}^{n} \alpha_j \right) \delta_0 + \sum_{i=0}^{n} \left(\prod_{j=i+1}^{n} \alpha_j \right) \gamma_i, \quad n \geq 0. \tag{176}$$

Bizonyítás. Az $n = 0$ esetben a rekurzió alapján,

$$\delta_1 \leq \alpha_0 \delta_0 + \gamma_0,$$

a képlet alapján pedig ($\prod_{j=l}^{i} f(j) = 1$, ha $i < l$)

$$\delta_1 \leq \left(\prod_{j=0}^{0} \alpha_j \right) \delta_0 + \sum_{i=0}^{0} \left(\prod_{j=1}^{0} \alpha_j \right) \gamma_i = \alpha_0 \delta_0 + \gamma_0.$$

Feltéve, hogy az állítás valamely $n \geq 0$ értékre igaz, igazoljuk helyességét az $n+1$ értékre is. Ekkor kapjuk, hogy

$$\begin{gathered} \delta_{n+2} \leq \alpha_{n+1} \delta_{n+1} + \gamma_{n+1} \leq \\ \leq \alpha_{n+1} \left(\textstyle\prod_{j=0}^{n} \alpha_j \right) \delta_0 + \alpha_{n+1} \sum_{i=0}^{n} \left(\textstyle\prod_{j=i+1}^{n} \alpha_j \right) \gamma_i + \gamma_{n+1} \leq \\ \leq \left(\textstyle\prod_{j=0}^{n+1} \alpha_j \right) \delta_0 + \sum_{i=0}^{n+1} \left(\textstyle\prod_{j=i+1}^{n+1} \alpha_j \right) \gamma_i, \end{gathered}$$

ami bizonyítandó volt. □

Ha a lemmát az $\alpha_j = 1 + h_j L$, $\gamma_j = \|T(y(t_j), h_j)\|_\infty$ szereposztással alkalmazzuk, akkor az

$$\|e_{n+1}\|_\infty \leq \left[\prod_{j=0}^{n}(1+h_jL)\right]\|e_0\|_\infty + \sum_{i=0}^{n}\left[\prod_{j=i+1}^{n}(1+h_jL)\right]\|T(y(t_i),h_i)\|_\infty \tag{177}$$

egyenlőtlenséget kapjuk. Vizsgáljuk most meg az $T(y(x),h)$ lokális hiba nagyságát. Tegyük fel, hogy $y(x)\in C^2[x_0,b]$. Ekkor $y(x+h)=y(x)+hy'(x)+R_1(x,h)$, amelynek maradéktagjára teljesül, hogy

$$\|R_1(x,h)\|_\infty \leq h^2K_2 \quad \left(2K_2=\max_{1\leq i\leq \ell}\max_{x\in[x_0,b]}|y_i''(x)|\right). \tag{178}$$

A képlethiba ennek megfelelően

$$T(y(x),h)=(y(x)+hy'(x)+R_1(x,h))-(y(x)+h\overbrace{f(x,y(x))}^{y'(x)})=R_1(x,h),$$

ahonnan $\|T(y(x),h)\|_\infty \leq K_2h^2$ $(x,x+h\in[x_0,b])$ és

$$|T(y(t_i),h_i)|\leq K_2h_i^2 \quad (i=0,1,\ldots,N-1) \tag{179}$$

következik. Az $1+x\leq e^x$ $(x\geq 0)$ egyenlőtlenség figyelembevételével kapjuk, hogy $1+h_jL\leq e^{Lh_j}$ és

$$\prod_{j=i+1}^{n}(1+h_jL)\leq\prod_{j=i+1}^{n}e^{Lh_j}=e^{L\sum_{j=i+1}^{n}h_i}\leq e^{L(b-x_0)}. \tag{180}$$

Ha ezt a (177) képletbe behelyettesítjük, akkor kapjuk, hogy

$$\|e_{n+1}\|_\infty \leq e^{L(b-x_0)}\|e_0\|_\infty + \sum_{i=0}^{n}e^{L(b-x_0)}K_2h_i^2. \tag{181}$$

A nyilvánvaló

$$\sum_{i=0}^{n}h_i^2\leq\left(\max_{0\leq i\leq n}h_i\right)\sum_{i=0}^{n}h_i\leq(b-x_0)\max_{0\leq i\leq n}h_i,$$

egyenlőtlenség miatt

$$\|e_{n+1}\|_\infty \leq e^{L(b-x_0)}\left(\|e_0\|_\infty+(b-x_0)K_2\max_{0\leq i\leq n}h_i\right) \quad (n=0,1,\ldots,N-1). \tag{182}$$

Mivel $e_0=y_0-y(x_0)=0$, igaz a következő

Tétel. *Az Euler-módszer globális hibájára $y(x)\in C^2[x_0,b]$ esetén fennáll, hogy*

$$\max_{1\leq i\leq N}\|e_i\|_\infty\leq(b-x_0)K_2e^{L(b-x_0)}\max_{0\leq i\leq N-1}h_i. \tag{183}$$

Az Euler-módszer hibája a legnagyobb lépéshosszal arányos. Ha a $\{t_i\}_{i=0}^N$ felosztás minden határon túl finomodik, azaz $N \to \infty$ és $\max_{0\le i\le N} h_i \to 0$ egyidejűleg teljesül, akkor fennáll, hogy

$$\max_{1\le i\le N} \|e_i\|_\infty \to 0, \quad N \to \infty. \tag{184}$$

Tehát az Euler-módszer (elsőrendben) konvergens.

A konvergencia jellegét mutatja a következő ábra, amelyen az $y' = 2y/x + 2x^3$, $y(1) = 2$ Cauchy-probléma elméleti és Euler-megoldásai láthatók $h = 1, 1/2, 1/4, 1/8$ esetén.dtbpF10.3593cm7.8002cm0pteuler2u.wmf

10.2 Explicit egylépéses módszerek

Az Euler-módszernek számtalan hatékonyabb továbbfejlesztése ismeretes. Ezek közül az egyik legfontosabb az explicit egylépéses módszerek osztálya, amelynek alakja

$$y_{i+1} = y_i + h_i \phi(t_i, y_i, h_i) \quad (i = 0, 1, \ldots, N-1). \tag{185}$$

Az $f(x,y)$ függvénytől függő $\phi(x,y,h)$ ($\phi : \mathbb{R}\times\mathbb{R}^\ell\times\mathbb{R} \to \mathbb{R}^\ell$) növekményfüggvény minden változójában folytonos, az y változóban kielégíti a

$$\|\phi(x,y,h) – \phi(x,z,h)\|_\infty \le K \|y – z\|_\infty, \left(K \ge 0, (x,y), (x,z) \in D, |h| \le \widehat{h}\right) \tag{186}$$

feltételt és

$$\phi(x,y,0) = f(x,y). \tag{187}$$

Az Euler-módszer a $\phi(x,y,h) = f(x,y)$ speciális esetnek felel meg. Az egylépéses módszerek x pontbeli lokális hibáját az

$$T(y(x), h) = y(x+h) - (y(x) + h\phi(x, y(x), h)) \tag{188}$$

mennyiség definiálja. Az egylépéses módszer p-edrendű, ha létezik $K_p > 0$ konstans, hogy

$$\|T(y(x), h)\|_\infty \le K_p h^{p+1}, \quad x, x+h \in [x_0, b]. \tag{189}$$

Az Euler-módszerhez hasonlóan beláthatjuk, hogy az egylépéses módszerek globális hibájára is fennáll az alábbi egyenlőtlenség

$$\|e_{n+1}\|_\infty \le (1 + h_n K)\|e_n\|_\infty + \|T(y(t_n), h_n)\|_\infty \quad (n = 0, 1, \ldots, N-1). \tag{190}$$

Ha az egylépéses módszer p-edrendű, akkor az Euler-módszerhez hasonlóan igazolhatjuk, hogy a globális hibára fennáll a

$$\max_{1\le i\le N} \|e_i\|_\infty \le c \left(\max_{0\le i\le N-1} h_i\right)^p \tag{191}$$

egyenlőtlenség, ahol $c > 0$ konstans. Tehát a p-edrendű egylépéses módszer p-edrendben konvergál.

Fontos megjegyezni, hogy egy p-edrendben konvergáló módszer általában csak olyan kezdetiérték feladatok esetében konvergál p-edrendben, amelyek megoldása folytonosan differenciálható legalább $(p+1)$-szer. Ha a differenciálegyenlet elméleti megoldása ennél kevesebbszer differenciálható, akkor a konvergencia rendje is csökken.

A p-edrendű egylépéses módszer ui. az $y(x)$ megoldás függvényt $p+1$-ed rendű hibával közelíti. Ez azt jelenti, hogy az $y(x+h)$ és $y(x)+h\phi(x,y(x),h)$ függvényeket az x pont körül sorbafejtve, a két sorfejtés első $p+1$ tagja megegyezik. A legkézenfekvőbb ilyen tulajdonságú formula az ún. *Taylor-sor módszer*, ahol

$$\phi(x,y(x),h)=\sum_{i=1}^{p} y^{(i)}(x)\,\frac{h^{i-1}}{i!}. \tag{192}$$

Az explicit Runge-Kutta módszerek elkerülik a Taylor-sor együtthatóinak meghatározását és csak az y és $f(x,y)$ információkat használják. Szokásos alakjuk

$$\begin{gathered} y_{n+1}=y_n+h_n\textstyle\sum_{i=1}^{m} c_i k_i,\\ k_1=f(x_n,y_n),\\ k_i=f\left(x_n+a_i h_n,\, y_n+h_n\textstyle\sum_{j=1}^{i-1} b_{ij}k_j\right)\quad (i=2,\dots,m). \end{gathered} \tag{193}$$

Igazolható, hogy $\sum_{i=1}^{m} c_i = 1$ esetén a $\phi(x,y,0)=f(x,y)$ feltétel teljesül. Általában feltesszük még, hogy

$$a_i=\sum_{j=1}^{i-1} b_{ij}\quad (i=2,\dots,m). \tag{194}$$

Az explicit Runge-Kutta módszereket az alábbi mátrix sémában is meg lehet adni:

$$\begin{array}{c|ccccc} 0 & & & & & \\ a_2 & b_{21} & & & & \\ \vdots & \vdots & \ddots & & & \\ a_m & b_{m1} & \dots & b_{m,m-1} & & \\ \hline & c_1 & \dots & c_{m-1} & c_m \end{array} \tag{195}$$

Legnevezetesebb az alábbi negyedrendű Runge-Kutta módszer:

$$\begin{array}{c|cccc} 0 & & & & \\ 1/2 & 1/2 & & & \\ 1/2 & 0 & 1/2 & & \\ 1 & 0 & 0 & 1 & \\ \hline & 1/6 & 1/3 & 1/3 & 1/6 \end{array}$$

Az Euler-módszer a

$$\begin{array}{c|c} 0 & \\ \hline & 1 \end{array}$$

sémának felel meg.

A numerikus módszerekkel kapott közelítő megoldásokra általában a

$$\max_{1\leq i\leq N} \|e_i\|_\infty \leq \epsilon \tag{196}$$

globális hiba korlátot írjuk elő. Könnyen igazolható, hogy

$$\|T\left(y\left(t_n\right),h_n\right)\|_\infty \leq h_n\epsilon \quad (n=0,\ldots,N-1) \tag{197}$$

esetén alkalmas $\widehat{c}>0$ konstanssal fennáll a

$$\max_{1\leq n\leq N-1} \|e_i\|_\infty \leq \widehat{c}\epsilon \tag{198}$$

egyenlőtlenség. Ez a $\widehat{c}$ szám többnyire nem ismert. A (196) feltételt úgy próbáljuk meg teljesíteni, hogy a lokális hibákra egy

$$\|T\left(y\left(t_n\right),h_n\right)\|_\infty \leq h_n\epsilon/q \quad (i=0,\ldots,N-1) \tag{199}$$

alakú feltételt írunk elő, ahol $q>0$ egy tapasztalati konstans. Ha a (199) feltétel teljesül, akkor a számított közelítő megoldásokat $\epsilon>0$ hibájúnak fogadjuk el.

A lokális hibára vonatkozó (199) feltétel felveti a képlethiba becslésének kérdését. Számos módszer esetén analitikus és numerikus becslések is adhatók. Itt a két legjobban bevált, ill. leggyakrabban használt módszert ismertetjük.

A lépésfelezéses hibabecslés (Runge-féle szabály). A t_n pontból kiindulva végezzünk el két lépést a $h_n/2$ lépéshosszal is. Igy a t_n+h_n pontban kapunk egy második $\hat{z}_{n+1}$ közelítést is. Igazolható, hogy

$$T\left(y\left(t_n\right),h_n\right) \approx \frac{\hat{z}_{n+1}-y_{n+1}}{2^p-1}, \tag{200}$$

ahol p a módszer rendje.

Párosított Runge-Kutta formulák. Az ilyen módszereknél egy p-ed és egy $(p+1)$-ed rendű Runge-Kutta formulát úgy választanak meg, hogy az alacsonyabb rendű formulához tartozó k_i értékek egyúttal a magasabb rendű formulában is szerepelnek és a két formulával kapott közelítő megoldások különbsége becsli az alacsonyabbrendű módszer lokális hibáját. Sematikusan ábrázolva

$$\begin{array}{c|cccc} 0 & & & & \\ a_2 & b_{21} & & & \\ \vdots & \vdots & \ddots & & \\ a_m & b_{m1} & \ldots & b_{m,m-1} & \\ \hline & c_1 & \ldots & c_{m-1} & c_m \\ & d_1 & \ldots & d_{m-1} & d_m \end{array} \tag{201}$$

Az alacsonyabbrendű formula:

$$y_{n+1} = y_n + h_n\left(c_1k_1+\ldots+c_mk_m\right). \tag{202}$$

A magasabbrendű formula:

$$\hat{y}_{n+1} = y_n + h_n \left(d_1 k_1 + \ldots + d_m k_m \right). \tag{203}$$

A lokális hiba becslése:

$$T\left(y\left(t_n \right), h_n \right) \approx h_n \sum_{i=1}^{m} \left(d_i - c_i \right) k_i. \tag{204}$$

Az ilyen formulák előnye a lépésfelezéssel szemben az, hogy csak egy pontra támaszkodó információt használnak fel. Igazolható, hogy $m \leq 5$ esetén ezek a becslések nem lehetnek aszimptotikusan pontosak.

Az egyik legismertebb formula az ún. 2(3)-as Runge-Kutta-Fehlberg képlet (a MATLAB ode23.m eljárása):

0			
1	1		
1/2	1/4	1/4	
	1/2	1/2	0
	1/6	1/6	4/6

A gyakorlatban a magasabbrendű formula közelítésével folytatják az eljárást.

Aszimptotikusan pontos England alábbi 4(5)-ös formulája, ahol $m = 6$.

0						
1/2	1/2					
1/2	1/4	1/4				
1	0	−1	2			
2/3	7/27	10/27	0	1/27		
1/5	28/625	−125/625	546/625	54/625	−378/625	
	1/6	0	4/6	1/6	0	0
	14/336	0	0	35/336	162/336	125/336

Jelölje a továbbiakban EST a lokális hiba becsült értékének normáját. A szokásos adaptív Runge-Kutta sémát az alábbiakban írhatjuk le.

ADAPTÍV EGYLÉPÉSES MÓDSZEREK ALGORITMUSA:

Input t_0, y_0, b, tol; h_0 kiválasztása; $i = 0$.
while $t_i < b$
1. y_{i+1} kiszámítása és a lokális hiba (EST) becslése
2. **if** $EST \leq tol$
   $i = i + 1$
   else
   $h_i = h_{új}$
   goto 1
   end

end

Az új lépéshosszt ($h_{új}$) a $\|T\left(y\left(x_i\right),h_i\right)\|_\infty \approx c_3 h_i^{p+1} \approx EST$ becslésből és a $c_3 h_{új}^{p+1} \approx tol$ követelményből kaphatjuk meg. Eszerint $c_3 \approx EST/h_i^{p+1}$ és $h_{új}^{p+1} \approx tol/c_3 \approx \left(tol/EST\right) h_i^{p+1}$, ahonnan

$$h_{új} = h_i \left(\frac{tol}{EST}\right)^{1/(p+1)}. \tag{205}$$

Ha a becsült hiba lényegesen kisebb mint az előírt *tol* hibakorlát, akkor növelni lehet a lépéshosszt az előbbiek figyelembevételével. Bizonyos esetekben ez a stratégia optimális.

Példa. Oldjuk meg az alábbi "orbitális" differenciálegyenlet rendszert a $[0,20]$ intervallumon a MATLAB adaptív Ode23 programjával:

$$\begin{array}{ll} y_1' = y_3, & y_1\left(0\right) = 1-\lambda, \\ y_2' = y_4, & y_2\left(0\right) = 0, \\ y_3' = \frac{-y_1}{\left(y_1^2+y_2^2\right)^{3/2}}, & y_3\left(0\right) = 0, \\ y_4' = \frac{-y_2}{\left(y_1^2+y_2^2\right)^{3/2}}, & y_4\left(0\right) = \left[\left(1+\lambda\right)/\left(1-\lambda\right)\right]^{1/2}, \end{array}$$

ahol $\lambda = 0.3$. A program alapbeállításaival kapott megoldás komponensek (trajektóriák) képe a következő:

dtbpFU10.6097cm8.0001cm0ptMegoldás trajektóriákode.bmpAz ábrán látható közelítő megoldás trajektóriákat az Ode23 eljárás 114 lépésben kapta. A változó lépéshosszakat az alábbi ábra mutatja:

dtbpFU9.0896cm6.8688cm0ptA lépéshossz változásaode2.bmpA feladatot az explicit Euler-módszerrel is megoldva 4000 ekvidisztans ponton ($h = 0.005$) a következő közelítő trajektóriákat kaptuk:

dtbpFU10.5921cm8.0001cm0ptExplicit Euler-módszerrel kapott megoldás trajektóriákode3.bmpAz Ode23 programmal kapott trajektóriák jó közelítést adnak. Ugyanakkor az Euler-módszerrel kapottak még igen kis lépéshossz mellett is eléggé pontatlanok. Az összehasonlítás azt mutatja, hogy a magasabb rendű módszerek általában előnyösebbek.