# Megbízható numerikus számítások alapjai

Gergó Lajos, Huszárszky Szilvia
Lektorálta: G.-Tóth Boglárka

2013. február

# Tartalomjegyzék

# Bevezetés

Ez a jegyzet a programtervező informatikus mesterszak modellalkotó szakirányos hallgatói számára készült elsősorban, de szívesen ajánljuk minden olyan érdeklődőnek, aki szeretne megismerkedni a megbízható numerikus számítások alapjaival. A Numerikus analízis tárgy keretein belül egy félév alatt áttekintjük az intervallum aritmetikával kapcsolatos alapvető ismereteket, majd a numerikus módszerek néhány alapfeladatának az intervallum aritmetikai megoldását tárgyaljuk. Elsősorban a lineáris egyenletrendszerek intervallum alapú numerikus megoldásával foglalkozunk (Gauss-elimináció, a megoldásvektor különböző befoglalási módszerei) valamint a nemlineáris egyenletek különböző megoldási módszereit vizsgáljuk (Newton-iteráció, polinomok gyökeinek a szimultán meghatározása, interpolácis módszerek). Külön csemegének szánjuk a hetedik és kilencedik fejezetet, amelyekben az automatikus differenciálás keveset emlegetett módszere és globális optimumszámítási módszer kerül ismertetésre. A téma megértéséhez az alapszakos lineáris algebra, analízis és numerikus analízis ismeretek elegendőek.

A megbízható numerikus számítások lényege az, hogy olyan algoritmust kívánunk megadni, amely biztosítja azt, hogy az algoritmus befejeződésekor megad egy olyan intervallumot, amely tartalmazza a megoldást. Így garantált hibabecslést biztosít a lefutás végén. Nyilván akkor használható ez a módszer, ha az eredmény intervallum kellően kicsi átmérőjű. Mivel a hagyományos numerikus algoritmusok ezt nem tudják általában biztosítani, ha nagyon nagy szükség van igazán megbízható eredményre, akkor érdemes lehet több munkát fektetni a megoldásba és intervallum alapú, megbízható algoritmust felhasználni, ami garantált hibakorláttal rendelkező végeredményt képes produkálni.

Nézzünk néhány általános megjegyzést, elvet ezen módszerekkel

kapcsolatban!

A megbízható numerikus eredmények számítása két fő pillérre támaszkodik:

1. intervallum aritmetika elmélete,

2. alkalmas algoritmusok.

Megbízható numerikus eredményhez jutni legkönnyebben a megfelelő műveletek és változók intervallumos változatára való cseréjével lehet. Ezzel megbízható, ellenőrzött eredményhez jutunk, azonban a kapott befoglalások átmérője sokszor gyakorlatilag hasznosíthatatlanul szélesnek adódik. Szükségünk van tehát olyan módszerekre, amelyek hasznosítják az intervallum aritmetika előnyeit, és egyben, a már kiszámolt, de durva becslések finomításait adják.

Ilyen algoritmusok fejlesztése során nagyon óvatosnak kell lennünk, hogy *minek* is számoljuk a befoglalását. Például, ha egy közönséges differenciálegyenlet kezdeti érték problémájának megoldását Runge-Kutta módszerrel becslő programot készítünk, és az itt szereplő műveletekre intervallum műveletekkel való befoglalását számítanánk, akkor nem a differenciálegyenlet egy megoldásának befoglalását kapnánk, hanem a megfelelő Runge–Kutta módszer becslését! Ez a befoglalás a kerekítési hibákat igen, de a csonkolási hibákat nem tartalmazza. Egy megbízható algoritmusnak azonban az összes lehetséges hibaforrást le kell fednie, mint például a konverziós hibákat is, hogy tényleg megbízható bennfogalást kapjunk.

Az úgynevezett pont problémákra – azokra amelyekben a bemenő adat nem tartalmaz intervallumot – egy egyszerű megbízható megoldást kínál az iteratív finomítás módszere. Az első becslés lebegőpontos számolása után gépi intervallum számítással annak hibája le van fedve. Amennyiben ennek az átmérője kisebb a megkövetelt pontosságnál, akkor a megoldás egy ellenőrzött befoglalása a becslés és hibájának lefedése összegeként adódik. Máskülönben a becslést a hiba intervallum középpontjának hozzávételével megismételve egy finomabb becslés adódik. A megbízható numerikus algoritmusok gyakran fixpont tételek alkalmazásaira támaszkodnak, ebben az esetben az egyik lehetőség a Brouwer-féle fixpont tétel.

**Tétel.** (Brouwer fixpont tétele) Legyen $\mathbb{R}^n \to \mathbb{R}^n$ folytonos leképezés, $X \subseteq \mathbb{R}^n$ zárt, konvex és korlátos halmaz. Ha $f(X) \subseteq X$, akkor $f$ függvénynek van legalább egy $x^* \in X$ fixpontja.

Legyen $X = [x] \in \mathbb{R}^n$ egy gépi intervallum vektor (doboz az $n$-dimenziós térben). Ez kielégíti az előbbi tétel feltételeit. Tegyük fel, hogy találunk egy $[x]$ vektort úgy, hogy $f([x]) \subseteq [x]$. Ekkor $[x]$ biztosan tartalmazza legalább egy fixpontját az $f$ függvénynek. A tétel igaz marad, ha $f$ helyett annak $f_{[]}$ intervallum kiértékelését vesszük és arra biztosítjuk a tartalmazást, mivel $f([x]) \subseteq f_{[]}([x])$.

Ez a tétel egyfajta sablonként szolgálhat algoritmusainkhoz. Először keressünk egy $x = f(x)$ alakú, az eredetivel ekvivalens problémát, majd helyettesítjük a jobb oldali függvényt annak $f_{[]}$ intervallum kiértékelésével. Példaként a fixpont iterációs vagy más néven az egyszerű iterációs zérushely keresést tekintjük. Kezdjük valamely $[x]^{(0)}$ közelítő megoldással az alábbi iterációt

$$[x]^{(k+1)} = f_{[]}([x]^{(k)}) \quad k = 0, 1, 2, \ldots \tag{1}$$

Fejezzük be az iterálást, ha $[x]^{(k+1)} \subseteq [x]^{(k)}$ valamely $k \geq 0$ esetén. Ekkor matematikai értelemben beláttuk, hogy az eredeti problémának van legalább egy $x^*$ fixpontja $[x]^{(k)}$ intervallumban.

Megkülönböztetünk *a priori* és *a posteriori* módszereket a kezdő közelítésre. Az *a priori* eljárásban a kezdő közelítés már tartalmazza a fixpontot. Ekkor az (1) iterációt az alábbi módon alakítjuk át

$$[x]^{(k+1)} = f_{[]}([x]^{(k)}) \cap [x]^{(k)} \quad k = 0, 1, 2, \ldots$$

Az iteráció leáll, amennyiben elérte a maximális lépés számot, vagy két egymást követő eredmény azonos.

Az *a posteriori* módszer nem tartalmazza szükségszerűen a fixpontot. Itt az elvárás, hogy az iteráció során egyre közelebb kerüljünk a fixponthoz, és végül le is fedjük. Minél jobb a kezdő közelítés, annál gyorsabb a konvergencia. A gyakorlati tapasztalat az, hogy az iteráció közelít a fixponthoz, de csak ritka esetben tartalmazza azt. Egy egyszerű trükkel

segíthetünk ezen. Az új iterációs lépés előtt egy

$$[x] \bowtie \varepsilon := \begin{cases} [x] + [-\varepsilon, \varepsilon] \cdot d([x]) & \text{ha } d([x]) \neq 0 \\ [x] + [-x_{\min}, +x_{\min}] & \text{máskülönben} \end{cases}$$

$\varepsilon$-bővítéssel növeljük az aktuális intervallumot, ahol $x_{\min}$ a legkisebb pozitív gépi szám, $d([x])$ az $[x]$ intervallum szélessége, $\varepsilon > 0$. Ezután az *a posteriori* módszer iterációja a következő módon változik:

$$\left.\begin{array}{ll} [x]^{(k)} & = [x]^{(k)} \bowtie \varepsilon \\ [x]^{(k+1)} & = f_{[]}([x]^{(k)}) \end{array}\right\} \qquad k = 0, 1, 2, \ldots$$

Fixpont módszereink némelyike módosítható úgy, hogy a fixpont egyértelműsége is biztosított legyen.

# 1. fejezet

# Intervallum aritmetikai alapok

## 1.1. Valós intervallum aritmetika

A következő szakaszokban a valós számok halmazát $\mathbb{R}$ , elemeit kis betűk $(a, b, \ldots, x, y, z)$ jelölik. Az $\mathbb{R}$ alábbi részhalmazát

$$[a] := [\underline{a}, \overline{a}] := \{t \mid \underline{a} \leq t \leq \overline{a}, \underline{a}, \overline{a} \in \mathbb{R}\}$$

zárt, valós intervallumnak, vagy röviden intervallumnak nevezzük, ahol az intervallum alsó és felső korlátjára az

$$\underline{a}, \quad \overline{a}$$

jelölést használjuk. Ha $M$ egy tetszőleges halmaz, akkor $M^n$ és $M^{n\times m}$ jelöli az $n$ dimenziós vektorok, illetve az $(n \times m)$-es mátrixok halmazát, ahol a vektorok oszlopvektorként értendők. Az egységmátrix jele $I$. Mátrixok és vektorok maximum normájának jele $\|\cdot\|_\infty$. Az iteráció sorszámát a felső indexben jelöljük, pl: $x^{(k)}$. A zárt valós intervallumok halmazát $\mathbb{IR}$ jelöli, elemeit pedig a $[a], [b], \ldots, [x], [y], [z]$ szimbólumok. Ekkor az $x \in \mathbb{R}$ valós számok felfoghatók $\mathbb{IR}$ speciális elemeként: $[x, x]$, amiket pont-intervallumoknak nevezünk.

**1.1. Definíció.** *Az* $[a] = [\underline{a}, \overline{a}]$ *és* $[b] = [\underline{b}, \overline{b}]$ *intervallumok egyenlők,* $[a] = [b]$*, ha halmazelméleti értelemben egyenlők.*

Ebből közvetlenül következik, hogy

$$[a] = [b] \;\Leftrightarrow\; \underline{a} = \underline{b} \text{ és } \overline{a} = \overline{b}.$$

Az $=$ reláció $\mathbb{IR}$-ben reflexív, szimmetrikus, tranzitív.

A következőkben általánosítjuk a valós aritmetikát bevezetve az $\mathbb{IR}$-en értelmezett műveleteket.

**1.2. Definíció.** *Legyen* $\circ \in \{+, -, \cdot, :\}$ *egy bináris művelet a valós számokon értelmezve. Ha* $[a], [b] \in \mathbb{IR}$*, akkor*

$$[a] \circ [b] = \{z = a \circ b \mid a \in [a], b \in [b]\} \tag{1.1}$$

*definiálja a megfelelő* $\mathbb{IR}$*-beli műveletet.*

Az osztás esetén feltesszük, hogy $0 \notin [b]$, amit a továbbiakban nem említünk külön. Szintén megjegyezzük, hogy azonos szimbólumokat használunk az $\mathbb{R}$ illetve $\mathbb{IR}$-beli műveletekre.

Az $[a] = [\underline{a}, \overline{a}]$, $[b] = [\underline{b}, \overline{b}]$ intervallumokra vonatkozó műveletek explicit formája

$$\begin{aligned}
[a] + [b] &= [\underline{a} + \underline{b}, \overline{a} + \overline{b}], \\
[a] - [b] &= [\underline{a} - \overline{b}, \overline{a} - \underline{b}], \\
[a] \cdot [b] &= [\min\{\underline{ab}, \underline{a}\overline{b}, \overline{a}\underline{b}, \overline{ab}\}, \max\{\underline{ab}, \underline{a}\overline{b}, \overline{a}\underline{b}, \overline{ab}\}], \\
[a] : [b] &= [\underline{a}, \overline{a}] \cdot [1/\overline{b}, 1/\underline{b}].
\end{aligned} \tag{1.2}$$

Ez abból a tényből következik, hogy $z = f(x, y) = x \circ y, \circ \in \{+, -, \cdot, :\}$ kompakt halmazon vett, folytonos függvény, ennek okán felveszi legkisebb és legnagyobb, valamint az összes közbeeső értékét is, így $[a] \circ [b]$ szintén zárt valós intervallum. Az 1.2-beli képleteink ennek megfelelően $f(x, y)$ legkisebb, illetve legnagyobb elemét számítják ki. Az $\mathbb{IR}$ halmaz következésképp zárt a fenti műveletekre nézve, továbbá azonnal látszik, hogy a valós számok izomorfak a megfelelő pont-intervallumokkal, ezért egyszerűen használjuk az $[x, x] \circ [a] = x \circ [a]$ jelölést.

Mivel az intervallumok is halmazok – a halmazelméletben szokásos relációk, műveletek $(=, \in, \subseteq, \subset, \supseteq, \supset, \cap)$ az addigi értelemmel bírnak.

Bevezethetők újabb relációk is. Egy $[x]$ intervallumot tartalmazza $[y]$ pontosan akkor, ha $\underline{y} < \underline{x}$ és $\overline{x} < \overline{y}$. Ennek jele $[x] \overset{\circ}{\subset} [y]$ és belső tartalmazási relációnak is hívjuk. Néha használatos két intervallum burka:

$$[x] \underline{\cup} [y] := [\min\{\underline{x}, \underline{y}\}, \max\{\overline{x}, \overline{y}\}].$$

Az (1.1)-beli műveleteken túl gyakran használunk unáris intervallum műveleteket.

**1.3. Definíció.** *Ha $f(x)$ egy folytonos unáris művelet $\mathbb{R}$-en, akkor*

$$f([x]) = \left[\min_{x\in[x]} f(x), \max_{x\in[x]} f(x)\right]$$

*unáris művelet $\mathbb{IR}$-en.*

Példák ilyen unáris műveletekre $\mathbb{IR}$-en:

$$[x]^k (k \in \mathbb{R}), e^{[x]}, \ln[x], \sin[x], \cos[x], \ldots$$

A következő tételben összefoglaljuk az $\mathbb{IR}$-beli legfontosabb műveleti tulajdonságokat.

**1.4. Tétel.** *Legyen $[a], [b], [c] \in \mathbb{IR}$. Ekkor*

$$[a] + [b] = [b] + [a], \quad [a] \cdot [b] = [b] \cdot [a] \text{ (kommutativitás)}, \tag{1.3}$$

$$([a]+[b])+[c] = [a]+([b]+[c]), \quad ([a]\cdot[b])\cdot[c] = [a]\cdot([b]\cdot[c]) \text{ (asszociativitás)}, \tag{1.4}$$

$[0] = [0,0], [1] = [1,1]$ *egyértelműen meghatározott neutrális elemek az additív, illetve multiplikatív struktúrákban, azaz*

$$\begin{aligned} &[a] = [0] + [a] = [a] + [0] && \forall [a] \in \mathbb{IR} \Leftrightarrow [0] = [0,0], \\ &[a] = [1] \cdot [a] = [a] \cdot [1] && \forall [a] \in \mathbb{IR} \Leftrightarrow [1] = [1,1], \end{aligned} \tag{1.5}$$

$$\mathbb{IR} \text{ zérusosztó mentes}, \tag{1.6}$$

*az $[a] = [\underline{a}, \overline{a}] \in \mathbb{IR}, (\underline{a} \neq \overline{a})$, elemnek nincs sem additív, sem multiplikatív inverze, továbbá igaz, hogy*

$$0 \in [a] - [a] \quad \text{és } 1 \quad \in [a] : [a] \tag{1.7}$$

$$[a]([b]+[c]) \subseteq [a][b]+[a][c] \qquad \textit{(szubdisztributivitás)} \tag{1.8}$$
$$a([b]+[c]) = a[b]+a[c], \qquad a \in \mathbb{R}$$
$$[a]([b]+[c]) = [a][b]+[a][c], \qquad \textit{ha } bc \geq 0 \ \forall b \in [b], c \in [c].$$

**Bizonyítás:** Az (1.3) állítás belátása. Legyen $\circ \in \{+, \cdot\}$. Ekkor

$$\begin{aligned}[a] \circ [b] &= \{z = a \circ b \mid a \in [a], b \in [b]\} = \\ &= \{z = b \circ a \mid b \in [b], a \in [a]\} = [b] \circ [a].\end{aligned}$$

Az (1.4) állítás belátása. Legyen $\circ \in \{+, \cdot\}$. Ekkor

$$\begin{aligned}([a] \circ [b]) \circ [c] &= \{z = (a \circ b) \circ c \mid a \in [a], b \in [b], c \in [c]\} = \\ &= \{z = a \circ (b \circ c) \mid a \in [a], b \in [b], c \in [c]\} = [a] \circ ([b] \circ [c]).\end{aligned}$$

Az (1.5) állítás belátása. Tegyük fel, hogy $n$, $\widehat{n}$ két additív neutrális elem. Ekkor

$$n + \widehat{n} = \widehat{n} \text{ és } \widehat{n} + n = n.$$

A kommutativitás miatt $n = \widehat{n}$. Hasonlóan látható be a multiplikatív neutrális elem unicitása is.

Az (1.6) állítás belátása. Legyen $[a] \cdot [b] = 0$, azaz

$$[a] \cdot [b] = \{z = a \cdot b \mid a \in [a], b \in [b]\} = [0, 0].$$

Ebből következik, hogy $[a], [b] \in \mathbb{IR}$ legalább egyike $[0, 0]$.

Az (1.7) állítás belátása. Mindkét állítás egyenértékű az

$$\begin{aligned}[a] - [b] = [0, 0] &\Rightarrow [a] = [a, a] = [b], \\ [a] \cdot [b] = [1, 1] &\Rightarrow [a] = [a, a], [b] = [1/a, 1/a]\end{aligned}$$

állításokkal. Legyen

$$[a] - [b] = \{z = a - b \mid a \in [a], b \in [b]\} = [0, 0].$$

Következik, hogy $\forall a \in [a], b \in [b]$ esetén $z = a - b = 0$. Tetszőlegesen rögzítve $b \in [b]$ elemet, kapjuk, hogy $\forall a \in [a]$ esetén $a = b$, tehát $[a] = [b, b]$, vagy $a \in [a]$ elemet rögzítve $[b] = [a, a]$. A multiplikatív eset hasonlóan bizonyítható.

Mivel

$$0 = a - a \in \{z = x - y \mid x \in [a], y \in [a]\} \quad a \in [a],$$

következik, hogy $0 \in [a] - [a]$. Hasonlóan adódik, hogy $1 \in [a] : [a]$.

Az (1.8) állítás belátása.

$$\begin{aligned}[a]([b] + [c]) &= \{z = a \cdot (b + c) \mid a \in [a], b \in [b], c \in [c]\} \subseteq \\ &\subseteq \{z = ab + \widetilde{a}c \mid a, \widetilde{a} \in [a], b \in [b], c \in [c]\} = \\ &= [a][b] + [a][c].\end{aligned}$$

Egy ellenpélda elegendő az egyenlőség cáfolására.

$$\begin{gathered}[a] = [0, 1],\ [b] = [1, 1],\ [c] = [-1, -1] \\ [a]([b] + [c]) = [0, 0] \subset [-1, 1] = [a][b] + [a][c].\end{gathered}$$

Sőt, kapjuk, hogy $\forall a \in \mathbb{R}$ esetén

$$\begin{aligned}a([b] + [c]) &= \{z = a(b + c) | b \in [b], c \in [c]\} = \\ &= \{z = ab + ac | b \in [b], c \in [c]\} = \\ &= a[b] + a[c].\end{aligned}$$

Az utolsó állítás belátásához, az általánosság megszorítása nélkül feltehetjük, hogy $\underline{b} \geq 0$ és $\underline{c} \geq 0$. Ha $\underline{a} \geq 0$, akkor

$$[a]([b] + [c]) = [\underline{a}(\underline{b} + \underline{c}), \overline{a}(\overline{b} + \overline{c})]$$

és

$$[a][b] + [a][c] = [\underline{ab}, \overline{a}\overline{b}] + [\underline{ac}, \overline{ac}] = [\underline{a}(\underline{b} + \underline{c}), \overline{a}(\overline{b} + \overline{c})].$$

Ha $\overline{a} \leq 0$, akkor az előző esetre jutunk $-[a]$ helyettesítéssel. Amennyiben $\underline{a}\overline{a} \leq 0$, kapjuk, hogy

$$[a]([b] + [c]) = [\underline{a}(\overline{b} + \overline{c}), \overline{a}(\overline{b} + \overline{c})],$$

mint ahogy

$$[a][b] + [a][c] = [\underline{a}\overline{b}, \overline{a}\overline{b}] + [\underline{a}\overline{c}, \overline{ac}] = [\underline{a}(\overline{b} + \overline{c}), \overline{a}(\overline{b} + \overline{c})],$$

amiből az állítás adódik. □

Most ismertetjük, hogy mit mondhatunk az

$$[a][x] = [b] \quad [a] \neq [0,0], \quad [x] \in \mathbb{IR}$$

típusú intervallum-egyenlet megoldhatóságáról. A kérdés megválaszolásához szükségünk lesz a következő $\chi$ segéd függvényre

$$\chi[a] := \begin{cases} \underline{a}/\overline{a} & \text{ha } |\underline{a}| \leq |\overline{a}| \\ \overline{a}/\underline{a} & \text{különben.} \end{cases}$$

Ekkor igaz a következő: az $[a][x] = [b]$ egyenletet megoldja $[x] \in \mathbb{IR}$ pontosan akkor, ha

$$\chi[a] \geq \chi[b].$$

A megoldás pontosan akkor nem egyértelmű, ha

$$\chi[a] = \chi[b] \leq 0.$$

Tekintsünk egy példát. Legyen $[1,2][x] = [-1,3]$. Ennek egyetlen megoldása az $[x] = [-\frac{1}{2}, \frac{3}{2}]$, mivel

$$\chi[1,2] = 1/2 > \chi[-1,3] = -\frac{1}{3}.$$

Másrészt, tekintve az alábbi egyenlet megoldásait

$$ax = b \quad a \in [1,2] \quad b \in [-1,3],$$

amiből kapjuk, hogy

$$\left\{ x = \frac{b}{a} \mid a \in [1,2], b \in [-1,3] \right\} = \frac{[-1,3]}{[1,2]} = [-1,3] \supset [x].$$

Ez a megoldáshalmaz különbözik az $[x]$ intervallumtól, ezért az $[a][x] = [b]$ intervallum-egyenlet algebrai megoldásának nevezzük. Belátható, hogy általánosan is igaz a következő:

Legyen adott $[a][x] = [b], 0 \notin [a]$ és $[x] \in \mathbb{IR}$ egy megoldása. Ekkor

$$[x] \subseteq [b] : [a],$$

hiszen

$$x \in [x] \Rightarrow \exists\ a \in [a], b \in [b] :\ ax = b \Rightarrow x = b/a \in [b] : [a].$$

Megjegyzendő, hogy az $[a][x] = [b]$ egyenlet megoldható akkor is, ha $[b] : [a]$ nem definiált. Például

$$[-\frac{1}{3}, 1][x] = [-1, 2],$$

ahol $\chi[-\frac{1}{3}, 1] > \chi[-1, 2]$, így $[x] = [-1, 2]$ egyértelmű.

Az intervallum számítások egy alapvető tulajdonsága a befoglalásra vett monotonitás. Az alábbi tétel fogalmazza meg ezt a tulajdonságot.

**1.5. Tétel.** *Legyen* $[a], [b], [c], [d] \in \mathbb{IR}$ *és legyen*

$$[a] \subseteq [c],\ [b] \subseteq [d]$$

*Ekkor* $a \circ \in \{+, -, \cdot, :\}$ *műveletekre igaz, hogy*

$$[a] \circ [b] \subseteq [c] \circ [d]. \tag{1.9}$$

**Bizonyítás:** Mivel $[a] \subseteq [c],\ [b] \subseteq [d]$, következik, hogy

$$\begin{aligned} [a] \circ [b] &= \{z = x \circ y | x \in [a], y \in [b]\} \subseteq \\ &\subseteq \{w = u \circ v | u \in [c], v \in [d]\} = \\ &= [c] \circ [d]. \end{aligned}$$

□

Az 1.5. tétel egy speciális esete:

**1.6. Következmény.** *Legyen* $[a], [b] \in \mathbb{IR}$ *és* $a \in [a], b \in [b]$. *Ekkor*

$$a \circ b \in [a] \circ [b], \quad \circ \in \{+, -, \cdot, :\}.$$

Az 1.3. definíció műveleteire a megfelelő tulajdonságok:

$$\begin{aligned} [x] \subseteq [y] &\Rightarrow r([x]) \subseteq r([y]), \\ x \in [x] &\Rightarrow r(x) \subseteq r([x]). \end{aligned} \tag{1.10}$$

Ezen állítások közvetlen általánosításai intervallum kifejezésekre az 1.19. tételben találhatók.

## 1.2. További koncepciók, tulajdonságok

A következőkben bevezetjük az alapvető topológiai fogalmakat az intervallumok halmazán. Elsőként a távolság fogalmát definiáljuk $\mathbb{IR}$ halmazon.

**1.7. Definíció.** *Az* $[a] = [\underline{a}, \overline{a}]$ *és* $[b] = [\underline{b}, \overline{b}]$ *intervallumok távolsága*

$$q([a],[b]) = \max\{|\underline{a} - \underline{b}|, |\overline{a} - \overline{b}|\}.$$

A $q$ leképezés metrika $\mathbb{IR}$-ben, hiszen rendelkezik az alábbi tulajdonságokkal

$$q([a],[b]) \geq 0 \quad \text{és} \quad q([a],[b]) = 0 \Leftrightarrow [a] = [b],$$
$$q([a],[b]) \leq q([a],[c]) + q([b],[c]) \quad \text{(háromszög-egyenlőtlenség)}.$$

A háromszög-egyenlőtlenség belátható a következő módon:

$$\begin{aligned} q([a],[c]) + q([b],[c]) &= \max\{|\underline{a} - \underline{c}|, |\overline{a} - \overline{c}|\} + \max\{|\underline{c} - \underline{b}|, |\overline{c} - \overline{b}|\} \geq \\ &\geq \max\{|\underline{a} - \underline{c}| + |\underline{c} - \underline{b}|, |\overline{a} - \overline{c}| + |\overline{c} - \overline{b}|\} \geq \\ &\geq \max\{|\underline{a} - \underline{b}|, |\overline{a} - \overline{b}|\} = q([a],[b]). \end{aligned}$$

Ez a távolság fogalom redukálódik a szokásosra, amennyiben pont intervallumokra alkalmazzuk. Tehát

$$q([a,a],[b,b]) = |a - b|.$$

A fent bevezetett metrika az $\mathbb{IR}$ halmazon értelmezett Hausdorff metrika. Ez általánosítása a metrikus tér pontjai közt értelmezett távolságnak - jelen esetben $\mathbb{R}$ a $q(a,b) = |a-b|$ metrikával - ezen tér összes nem üres, kompakt részhalmazának halmazára. Ha $U, V$ ilyen halmazok, akkor a Hausdorff távolságuk

$$q(U,V) = \max\left\{\sup_{v\in V}\inf_{u\in U} q(u,v), \sup_{u\in U}\inf_{v\in V} q(u,v)\right\}$$

képlettel definiált.

Másfajta hasznos jellemzés is található a Hausdorff metrikára. Valós intervallumok $[a], [b]$ esetén könnyű meggyőződnünk arról, hogy az 1.7. definíció leírja a Hausdorff metrikát.

Az $\mathbb{IR}$ halmazon egy metrika bevezetésével nemcsak metrikus, de topologikus teret is kapunk. A továbbiakban a konvergencia és folytonosság fogalmai így a szokásos módon tárgyalhatók. Intervallumok egy sorozata $\left\{[a]^{(k)}\right\}_{k=0}^{\infty}$ konvergál az $[a]$ intervallumhoz pontosan akkor, ha a megfelelő intervallum korlátok konvergálnak $[a] = [\underline{a}, \overline{a}]$ korlátaihoz. Ekkor írhatjuk, hogy

$$\lim_{k\to\infty} [a]^{(k)} = [a] \Leftrightarrow \left(\lim_{k\to\infty} \underline{a}^{(k)} = \underline{a}\ ,\ \lim_{k\to\infty} \overline{a}^{(k)} = \overline{a}\right). \tag{1.11}$$

A bizonyítás következik az intervallumok távolság definíciójából, ezért az olvasóra bízzuk.

A fenti metrikára igaz a következő állítás, melynek bizonyítását az olvasóra bízzuk.

**1.8. Tétel.** $(\mathbb{IR}, q)$ *az 1.7. definíció szerinti metrikával teljes metrikus tér. (Intervallumok minden Cauchy sorozata konvergál valamely intervallumhoz.)*

Most az intervallum sorozatok egy fontos osztályának viselkedésére adunk jellemzést.

**1.9. Tétel.** *Legyen* $\left\{[a]^{(k)}\right\}_{k=0}^{\infty}$ *olyan intervallum-sorozat, melyre*

$$[a]^{(0)} \supseteq [a]^{(1)} \supseteq [a]^{(2)} \supseteq \cdots$$

*igaz. Ekkor* $\bigcap_{k=0}^{\infty} [a]^{(k)}$ *egy* $[a]$ *intervallumhoz konvergál.*

**Bizonyítás:** Legyen a korlátok sorozata

$$\underline{a}^{(0)} \leq \underline{a}^{(1)} \leq \underline{a}^{(2)} \leq \underline{a}^{(3)} \leq \cdots \leq \overline{a}^{(3)} \leq \overline{a}^{(2)} \leq \overline{a}^{(1)} \leq \overline{a}^{(0)}.$$

Az alsó korlátok sorozata így monoton növekvő számokból áll, amelyek felső korlátja $\overline{a}^{(0)}$. Egy ilyen sorozat konvergens és határértéke valamely $\underline{a}$ szám. Hasonlóan, a felső korlátok számsorozata monoton csökkenő és

alulról korlátos, ezért konvergens, az $\overline{a}$ határértékkel, ahol $\underline{a} \leq \overline{a}$. Az $[a] = \bigcap_{k=0}^{\infty} [a]^{(k)}$ egyenlőség ugyanilyen egyszerűen belátható. □

A bizonyítás azt is mutatja, hogy egy $\left\{[a]^{(k)}\right\}_{k=0}^{\infty}$, amelyre

$$[a]^{(0)} \supseteq [a]^{(1)} \supseteq [a]^{(2)} \supseteq \cdots \supseteq [b]$$

egy $[a] \supseteq [b]$ intervallumhoz konvergál.

Az intervallum műveletekről és a további műveletekről szól az alábbi állítás.

**1.10. Tétel.** *Az 1. fejezetben bevezetett* $+, -, \cdot, :$ *intervallum műveletek folytonosak.*

**Bizonyítás:** Csak az $+$ műveletére látjuk be az állítást, a többire hasonlóan elvégezhető. Legyen $\left\{[a]^{(k)}\right\}_{k=0}^{\infty}$ és $\left\{[b]^{(k)}\right\}_{k=0}^{\infty}$ két intervallum sorozat, amelyekre

$$\lim_{k\to\infty} [a]^{(k)} = [a] \ , \ \lim_{k\to\infty} [b]^{(k)} = [b].$$

Az összeg intervallumok sorozatára $\left\{[a]^{(k)} + [b]^{(k)}\right\}_{k=0}^{\infty}$ igaz, hogy

$$\begin{aligned}\lim_{k\to\infty} \left([a]^{(k)} + [b]^{(k)}\right) &= \lim_{k\to\infty} \left[\underline{a}^{(k)} + \underline{b}^{(k)}, \overline{a}^{(k)} + \overline{b}^{(k)}\right] = \\ &= \left[\lim_{k\to\infty} \left(\underline{a}^{(k)} + \underline{b}^{(k)}\right), \lim_{k\to\infty} \left(\overline{a}^{(k)} + \overline{b}^{(k)}\right)\right] = \\ &= [\underline{a} + \underline{b}, \overline{a} + \overline{b}] = [a] + [b]\end{aligned}$$

(1.11) miatt. □

Az 1.10. tétel kiterjesztése a (lásd 1.3. definíció)

**1.11. Következmény.** *Legyen* $f$ *egy folytonos függvény és*

$$f([x]) = [\min_{x\in[x]} f(x), \max_{x\in[x]} f(x)].$$

*Ekkor* $f([x])$ *folytonos intervallum kifejezés.*

A bizonyítás azonnal következik $f$ folytonosságából. Ez a következmény garantálja például az $[x]^k, \sin[x], e^{[x]}$ folytonosságát.

**1.12. Definíció.** *Az* $[a] = [\underline{a}, \overline{a}] \in \mathbb{IR}$ *abszolútértéke*

$$|[a]| = q([a], [0, 0]) = \max\{|\underline{a}|, |\overline{a}|\}.$$

Szokásos jelölése még

$$|[a]| = \max_{a \in [a]} \{|a|\}. \tag{1.12}$$

Ha $[a], [b] \in \mathbb{IR}$, akkor világos, hogy

$$[a] \subseteq [b] \Rightarrow |[a]| \leq |[b]|\,. \tag{1.13}$$

Definiálható továbbá az úgynevezett legkisebb abszolútérték

$$\langle [x] \rangle := \min \left\{ |x| \;\middle|\; x \in [x] \right\}.$$

Ekkor az 1.12. definíció a legnagyobb abszolútérték nevet is viselheti.

Most belátjuk az $\mathbb{IR}$-beli metrika néhány tulajdonságát.

**1.13. Tétel.** *Legyen* $[a] = [\underline{a}, \overline{a}], [b] = [\underline{b}, \overline{b}], [c] = [\underline{c}, \overline{c}], [d] = [\underline{d}, \overline{d}] \in \mathbb{IR}$. *Ekkor*

$$q([a] + [b], [a] + [c]) = q([b], [c]), \tag{1.14}$$

$$q([a] + [b], [c] + [d]) \leq q([a], [c]) + q([b], [d]), \tag{1.15}$$

$$q(\alpha[b], \alpha[c]) = |\alpha|\, q([b], [c]), \qquad \alpha \in \mathbb{R} \tag{1.16}$$

$$q([a][b], [a][c]) \leq |[a]|\, q([b], [c]). \tag{1.17}$$

**Bizonyítás:** (1.14) bizonyítása. A $q$ metrika definíciójából következik, hogy

$$\begin{aligned} q([a] + [b], [a] + [c]) &= \max\{|\underline{a} + \underline{b} - (\underline{a} + \underline{c})|, |\overline{a} + \overline{b} - (\overline{a} + \overline{c})|\} = \\ &= \max\{|\underline{b} - \underline{c}|, |\overline{b} - \overline{c}|\}. \end{aligned}$$

(1.15) bizonyítása. A háromszög-egyenlőtlenség, (1.14) valamint $q$ szimmetriája alapján

$$\begin{aligned} q([a] + [b], [c] + [d]) &\leq q([a] + [b], [b] + [c]) + q([c] + [d], [b] + [c]) = \\ &= q([a], [c]) + q([b], [d]). \end{aligned}$$

(1.16) bizonyítása.

$$q(\alpha[b], \alpha[c]) = \max\{|\alpha\underline{b} - \alpha\underline{c}|, |\alpha\overline{b} - \alpha\overline{c}|\} = |\alpha|\, q([b], [c]).$$

(1.17) bizonyítása. A bizonyítandó állítás felírható

$$q([a][b], [a][c]) = \max\{|\underline{[a][b]} - \underline{[a][c]}|, |\overline{[a][b]} - \overline{[a][c]}|\} \leq |[a]|q([b], [c])$$

alakban. Itt az egyenlőtlenséget csak az alsó korlátokra látjuk be:

$$|\underline{[a][b]} - \underline{[a][c]}| \leq |[a]|q([b], [c]).$$

Az

$$|\overline{[a][b]} - \overline{[a][c]}| \leq |[a]|q([b], [c])$$

egyenlőtlenség hasonlóan igazolható.

Legyen $a \in [a]$. Az (1.16) relációt felhasználva

$$\max\{|\underline{a[b]} - \underline{a[c]}|, |\overline{a[b]} - \overline{a[c]}|\} = |a|q([b], [c]).$$

Az általánosság korlátozása nélkül feltehetjük, hogy

$$\underline{[a][b]} \geq \underline{[a][c]}.$$

(Az $\underline{[a][b]} < \underline{[a][c]}$ eset hasonló.)

Mivel

$$[a][c] = \{ac \mid a \in [a], c \in [c]\},$$

ezért

$$\exists a \in [a] : \underline{[a][c]} = \underline{a[c]}.$$

A befoglalásra vett monotonitás miatt

$$a[b] \subseteq [a][b]$$

továbbá

$$\underline{a[b]} - \underline{a[c]} \geq \underline{[a][b]} - \underline{[a][c]} \geq 0.$$

Végül

$$\begin{aligned} |\underline{[a][b]} - \underline{[a][c]}| = \underline{[a][b]} - \underline{[a][c]} \leq \underline{a[b]} - \underline{a[c]} = \\ = |\underline{a[b]} - \underline{a[c]}| \leq |a|q([b], [c]) \leq \\ \leq |[a]|q([b], [c]). \end{aligned}$$

□

$|[a]| = q([a], 0)$ jelöléssel az abszolútérték könnyen igazolható tulajdonságai

$$\begin{aligned} &|[a]| \geq 0 \qquad \text{és} \qquad |[a]| = 0 \Leftrightarrow [a] = [0,0], \\ &|[a] + [b]| \leq |[a]| + |[b]|, \\ &|x[a]| = |x| \cdot |[a]|, \qquad\qquad x \in \mathbb{R}, \\ &|[a][b]| = |[a]| \cdot |[b]|. \end{aligned} \tag{1.18}$$

Az utolsó reláció igazolása:

$$\begin{aligned} |[a][b]| &= \max_{c \in [a][b]} |c| = \\ &= \max_{a \in [a], b \in [b]} |ab| = \\ &= \max_{a \in [a], b \in [b]} (|a| \cdot |b|) = \\ &= \max_{a \in [a]} |a| \max_{b \in [b]} |b| = \\ &= |[a]| \cdot |[b]|. \end{aligned}$$

A többi belátása hasonlóan történik.

**1.14. Definíció.** *Egy* $[a] = [\underline{a}, \overline{a}]$ *intervallum szélessége, átmérője*

$$d([a]) = \overline{a} - \underline{a} \geq 0.$$

A pont intervallumok ekkor írhatók

$$\{[a] \in \mathbb{IR} \mid d([a]) = 0\}$$

alakban.

Az intervallum sugara, középpontja is megadható az intervallum alsó, felső korlátjával

$$\begin{aligned} r([x]) := \operatorname{rad}([x]) &:= \frac{\overline{x} - \underline{x}}{2}, \\ m([x]) := \operatorname{mid}([x]) &:= \frac{\overline{x} + \underline{x}}{2}. \end{aligned}$$

Ekkor az $x \in [x]$ reláció $|x - m([x])| \leq r([x])$ alakba írható. Ha $x$ közelítéseként az $[x]$ intervallum középpontját választjuk, akkor ezen közelítés abszolút hibájának felső korlátja éppen $r([x])$.

Az $x$ valós számot tartalmazó $[x]$ intervallum minősítésére bevezetjük a relatív átmérő fogalmát

$$d_{\mathrm{rel}}([x]) := \begin{cases} \frac{d([x])}{\langle [x] \rangle} & \text{ha } 0 \notin [x], \\ d([x]) & \text{máskülönben.} \end{cases}$$

Azonnal adódnak az alábbi tulajdonságok

$$[a] \subseteq [b] \Rightarrow d([a]) \leq d([b]), \tag{1.19}$$

$$d([a] \pm [b]) = d([a]) + d([b]). \tag{1.20}$$

Az (1.19) bizonyítása triviális, azonnal adódik

$$d([a]) = \max_{a,b \in [a]} |a - b| \tag{1.21}$$

kifejezésből.

Az (1.20) állítás az $+$ műveletére igaz, mivel

$$\begin{aligned} d([a] + [b]) &= d([\underline{a} + \underline{b}, \overline{a} + \overline{b}]) = \\ &= \overline{a} + \overline{b} - (\underline{a} + \underline{b}) = \\ &= \overline{a} - \underline{a} + \overline{b} - \underline{b} = d([a]) + d([b]). \end{aligned}$$

Azonos gondolatmenetet követve $-$ műveletre is igaz (1.20).

**1.15. Tétel.** *Legyen* $[a], [b] \in \mathbb{IR}$. *Ekkor*

$$d([a][b]) \leq d([a]) \cdot |[b]| + |[a]| \cdot d([b]), \tag{1.22}$$

$$d([a][b]) \geq \max\{|[a]| \cdot d([b]), |[b]| \cdot d([a])\}, \tag{1.23}$$

$$d(\alpha[b]) = |\alpha| \cdot d([b]), \qquad \alpha \in \mathbb{R} \tag{1.24}$$

$$d([a]^n) \leq n|[a]|^{n-1} \cdot d([a]), \qquad n = 1, 2, \ldots, \tag{1.25}$$

$$\left( \textit{ahol } [a]^n := \prod_{i=1}^{n} [a] \right),$$

$$d(([x] - x)^n) \leq 2 \cdot d([x]^n), \qquad x \in [x], n = 1, 2, \ldots, \tag{1.26}$$

$$\left( \textit{ahol } ([x] - x)^n := \prod_{i=1}^{n} ([x] - x) \right).$$

*Egy* $0 \in [c] \in \mathbb{IR}$ *intervallumra igaz, hogy*

$$|[c]| \leq d([c]) \leq 2 \cdot |[c]|. \tag{1.27}$$

**Bizonyítás:** Az (1.22) állítás bizonyítása. Felhasználva (1.21) összefüggést

$$\begin{aligned}
d([a][b]) &= \max_{a,a^* \in [a], b, b^* \in [b]} |ab - a^*b^*| = \\
&= \max_{a,a^* \in [a], b, b^* \in [b]} |ab - ab^* + ab^* - a^*b^*| \leq \\
&\leq \max_{a,a^* \in [a], b, b^* \in [b]} \{|a(b - b^*)| + |(a - a^*)b^*|\} \leq \\
&\leq \max_{a \in [a], b, b^* \in [b]} |a| \cdot |b - b^*| + \max_{a,a^* \in [a], b^* \in [b]} |a - a^*| \cdot |b^*| = \\
&= \left( \max_{a \in [a]} |a| \right) \left( \max_{b, b^* \in [b]} |b - b^*| \right) + \\
&+ \left( \max_{a,a^* \in [a]} |a - a^*| \right) \left( \max_{b^* \in [b]} |b^*| \right) = \\
&= |[a]| \cdot d([b]) + d([a]) \cdot |[b]|.
\end{aligned}$$

Az (1.23) állítás bizonyítása. Először belátjuk, hogy

$$\begin{aligned}
d([a][b]) &= \max_{a,a^* \in [a], b, b^* \in [b]} |ab - a^*b^*| \geq \max_{a \in [a], b, b^* \in [b]} |ab - ab^*| = \\
&= \max_{a \in [a], b, b^* \in [b]} |a| \cdot |b - b^*| = |[a]| \cdot d([b]).
\end{aligned}$$

Hasonlóan

$$d([a][b]) \geq |[b]| \cdot d([a]),$$

így (1.23) azonnal adódik.

Az (1.24) állítás bizonyítása.

$$\begin{aligned}
d(\alpha[b]) &= \max_{b, b^* \in [b]} |\alpha b - \alpha b^*| = \max_{b, b^* \in [b]} \{|\alpha| \cdot |b - b^*|\} = \\
&= |\alpha| \max_{b, b^* \in [b]} |b - b^*| = |\alpha| \cdot d([b]).
\end{aligned}$$

Az (1.25) állítás bizonyítása. $n = 1$ esetén az állítás igaz. Ha egy $n \geq 1$ számra az egyenlőtlenség igaz, akkor felhasználva (1.22) összefüggést,

(1.18) utolsó relációját, kapjuk, hogy

$$\begin{aligned} d([a]^{n+1}) = d([a]^n[a]) &\le d([a]^n)\cdot|[a]| + |[a]|^n \cdot d([a]) \le \\ &\le n|[a]|^{n-1}\cdot d([a])\cdot|[a]| + |[a]|^n\cdot d([a]) = \\ &= (n+1)|[a]|^n \cdot d([a]). \end{aligned}$$

Az (1.26) állítás bizonyítása. Mivel $x \in [x]$, következik (1.19) és a befoglalásra vett monotonitás alapján, hogy

$$\begin{aligned} d(([x]-x)^n) \le d(([x]-[x])^n) &= d([-d([x]), d([x])]^n) = \\ &= d([(-d([x]))^n, (d([x]))^n]) = 2\cdot(d([x]))^n. \end{aligned}$$

Az (1.27) állítás bizonyítása. Minthogy $0 \in [c] = [\underline{c}, \overline{c}]$, ezért $\underline{c} \le 0 \le \overline{c}$, amiből

$$d([c]) = \overline{c} - \underline{c} = |\overline{c}| + |\underline{c}| \ge \max\{|\underline{c}|, |\overline{c}|\} = |[c]|,$$

továbbá

$$d([c]) = |\underline{c}| + |\overline{c}| \le 2\cdot\max\{|\underline{c}|, |\overline{c}|\} = 2|[c]|.$$

□

**1.16. Tétel.** *Legyen $[a], [b] \in \mathbb{IR}$, és tegyük fel, hogy $[a] = -[a]$, azaz $[a]$ szimmetrikus intervallum. Ekkor az alábbi tulajdonságok igazak*

$$[a][b] = |[b]|[a], \tag{1.28}$$

$$d([a][b]) = |[b]|\cdot d([a]). \tag{1.29}$$

A második tulajdonság igaz nem szimmetrikus esetben, ha $0 \in [a]$ és $\underline{b} \ge 0$ vagy $\overline{b} \le 0$.

**Bizonyítás:** Mivel $[a] = -[a]$, azaz $|\underline{a}| = |\overline{a}| = a$, ezért

$$\begin{aligned} [a][b] &= [\min\{a\underline{b}, a\overline{b}, -a\underline{b}, -a\overline{b}\}, \max\{a\underline{b}, a\overline{b}, -a\underline{b}, -a\overline{b}\}] = \\ &= [a\min\{\underline{b}, \overline{b}, -\underline{b}, -\overline{b}\}, a\max\{\underline{b}, \overline{b}, -\underline{b}, -\overline{b}\}] = \\ &= [a(-|[b]|), a|[b]|] = [-a, a]|[b]| = [a]|[b]|. \end{aligned}$$

Ebből következik (1.24) alapján (1.29). A többi eset analóg módon belátható. □

**1.17. Tétel.** *A következő tulajdonságok igazak az* $[a], [b] \in \mathbb{IR}$ *intervallumokra:*

$$d([a]) = |[a] - [a]|, \tag{1.30}$$

$$[a] \subseteq [b] \Rightarrow \frac{1}{2}(d([b]) - d([a])) \leq q([a], [b]) \leq d([b]) - d([a]). \tag{1.31}$$

**Bizonyítás:** Az (1.30) állítás bizonyítása.

$$d([a]) = \overline{a} - \underline{a} = |[a] - [a]|.$$

Az (1.31) állítás bizonyítása. Legyen $[a] \subseteq [b]$. Ekkor $\underline{b} \leq \underline{a} \leq \overline{a} \leq \overline{b}$, tehát

$$\begin{aligned} q([a], [b]) &= \max\{|\underline{a} - \underline{b}|, |\overline{a} - \overline{b}|\} = \max\{\underline{a} - \underline{b}, \overline{b} - \overline{a}\} \\ &\leq \overline{b} - \overline{a} + \underline{a} - \underline{b} = \overline{b} - \underline{b} - (\overline{a} - \underline{a}) = d([b]) - d([a]), \end{aligned}$$

továbbá

$$\begin{aligned} q([a], [b]) &= \max\{\underline{a} - \underline{b}, \overline{b} - \overline{a}\} \geq \frac{1}{2}(\underline{a} - \underline{b} + \overline{b} - \overline{a}) \\ &= \frac{1}{2}(d([b]) - d([a])). \end{aligned}$$

□

Most bevezetünk egy új bináris műveletet $\mathbb{IR}$ halmazon. Legyen $[a], [b] \in \mathbb{IR}$. Az

$$[a] \cap [b] = \{c | c \in [a], c \in [b]\} \tag{1.32}$$

összefüggés jelöli két halmaz metszetét a halmazelmélet szerint. E művelet eredménye pontosan akkor van $\mathbb{IR}$ halmazban, ha $[a] \cap [b]$ nem üreshalmaz. Ebben az esetben

$$[a] \cap [b] = [\max\{\underline{a}, \underline{b}\}, \min\{\overline{a}, \overline{b}\}]. \tag{1.33}$$

A metszet fontos tulajdonságait gyűjti össze az alábbi

**1.18. Következmény.** *Legyen* $[a], [b], [c], [d] \in \mathbb{IR}$. *Ekkor*

$$[a] \subseteq [c], [b] \subseteq [d] \Rightarrow [a] \cap [b] \subseteq [c] \cap [d]. \tag{1.34}$$

*(befoglalásra vett monotonitás)*

*A metszetképzés folytonos művelet, amennyiben elvégezhető* $\mathbb{IR}$ *halmazon.*

**Bizonyítás:** A befoglalásra vett monotonitás (1.34) következik az 1.32. definícióból. A folytonosság bizonyítása (1.33) segítségével elvégezhető. □

## 1.3. Intervallum kiértékelés, valós függvény értékkészlete

Ebben a fejezetben az $f$ valós, folytonos függvényekkel foglalkozunk. Az $f$ függvényhez tartozó $f(x)$ kifejezés jelenti azt a számítási eljárást, amellyel $f$ minden értelmezési tartománybeli eleméhez tartozó függvényértéket kiszámítjuk. Feltesszük, hogy a következőkben előforduló kifejezések véges sok műveletből állnak, amely műveletek az 1.2. és az 1.3. definícióval összhangban vannak. Ha egy $f$-hez tartozó kifejezés tartalmazza az $a^{(0)}, a^{(1)}, \ldots, a^{(m)}$ konstansokat, akkor ezt $f(x; a^{(0)}, a^{(1)}, \ldots, a^{(m)})$ módon jelöljük. Egyszerűsítés céljából feltesszük, hogy mindegyik konstans $a^{(k)} (0 \leq k \leq m)$ csak egyszer fordul elő az adott kifejezésben. Amennyiben többször is előfordulna valamelyik, akkor újabb indexet bevezetve a kívánt alakra hozható a kifejezés.

Például két kiszámítási szabálya ugyanannak a $g$ függvénynek lehet

$$g^{(1)}(x; a) = \frac{ax}{1-x}, \quad x \neq 1, \quad x \neq 0,$$

és

$$g^{(2)}(x; a) = \frac{a}{1/x - 1}, \quad x \neq 1, \quad x \neq 0.$$

Az alábbi

$$\begin{aligned} f([x]; [a]^{(0)}, \ldots, [a]^{(m)}) &= \\ &= \{f(x; a^{(0)}, \ldots, a^{(m)}) | x \in [x], a^{(k)} \in [a]^{(k)}, 0 \leq k \leq m\} = \\ &= \left[ \begin{array}{cc} \min\limits_{\substack{x \in [x] \\ a^{(k)} \in [a]^{(k)} \\ 0 \leq k \leq m}} f(x; a^{(0)}, \ldots, a^{(m)}), & \max\limits_{\substack{x \in [x] \\ a^{(k)} \in [a]^{(k)} \\ 0 \leq k \leq m}} f(x; a^{(0)}, \ldots, a^{(m)}) \end{array} \right] \end{aligned}$$

kifejezés jelöli a továbbiakban az $f$ függvény összes felvett értékének intervallumát (értékkészletét), amikor $x \in [x], a^{(k)} \in [a]^{(k)}, 0 \leq k \leq$

$m$ egymástól függetlenül felveszik lehetséges értékeiket. Ez a definíció független az $f$ függvénytől.

Például az előbbi $g$ függvényre és

$$[a] = [0,1], [x] = [2,3]$$

kapjuk, hogy

$$g([2,3];[0,1]) = \left\{ \frac{ax}{1-x} \;\middle|\; 2 \le x \le 3, 0 \le a \le 1 \right\} = [-2,0].$$

Az alábbiakban definiáljuk az $f$ függvény egy intervallum kiértékelését.

Legyen adva $f$ egy számítási szabálya. Cseréljük az összes változót intervallumokra, a műveleteket intervallum műveletekre. Az így kapott kifejezés $f_{[]}([x];[a]^{(0)},\ldots,[a]^{(m)})$. Ha az összes változó az 1.2. és az 1.3. definícióban foglalt műveletek értelmezési tartományába esik, akkor $f$ egy *intervallum kiértékelését* vagy *intervallum-aritmetikai kiértékelését* kapjuk.

A fenti átirat az általunk tárgyalt függvények esetén mindig lehetséges. A konstansok is intervallumokkal helyettesítendők. Az intervallum kiértékelés függ $f$ hozzárendelési szabályának konkrét alakjától. Később felhasználjuk ezt a tényt. Itt egy egyszerű példát adunk.

Legyen $g$ az előbbi példákból megismert függvénnyel azonos.

$$[a] = [0,1], [x] = [2,3]$$

mellett két különböző intervallum kiértékelést kapunk:

$$g^{(1)}([2,3];[0,1]) = \frac{[0,1][2,3]}{1-[2,3]} = [-3,0],$$
$$g^{(2)}([2,3];[0,1]) = \frac{[0,1]}{1/[2,3]-1} = [-2,0] \neq g^{(1)}([2,3],[0,1]).$$

A fenti jelölés többváltozós függvényekre is alkalmazható. Az $f(x^{(1)},\ldots,x^{(n)};a^{(0)},\ldots,a^{(m)})$ kifejezés értékkészlete $f([x]^{(1)},\ldots,[x]^{(n)};[a]^{(0)},\ldots,[a]^{(m)})$ értékekből áll, ahol $x^{(k)} \in [x]^{(k)}, 1 \le k \le n$, és $a^{(j)} \in [a]^{(j)}, 0 \le j \le m$ egymástól függetlenek. Az $f_{[]}([x]^{(1)},\ldots,[x]^{(n)};[a]^{(0)},\ldots,[a]^{(m)})$ intervallum kiértékelése hasonlóan értelmezhető.

Adunk egy példát olyan kifejezésre, amely értelmetlen intervallum kifejezésre vezet. Az

$$f(x) = \frac{1}{x^2 + \frac{1}{2}}$$

valós függvény értelmes $\mathbb{R}$ halmazon. Az $f$ függvény egy lehetséges hozzárendelési szabálya

$$\widetilde{f}(x) = \frac{1}{x \cdot x + \frac{1}{2}}.$$

A változót $[x] = [-1, 1]$ intervallumra cserélve ez részhalmaza az értelmezési tartománynak, a műveletek intervallum megfelelőit használva

$$\widetilde{f}_{[]}([-1, 1]) = \frac{1}{[-1, 1][-1, 1] + \frac{1}{2}} = \frac{1}{[-1, 1] + \frac{1}{2}} = \frac{1}{[-\frac{1}{2}, \frac{3}{2}]},$$

ami nincs értelmezve.

Az alábbi tétel a függvényérték intervallum kiértékelésének két fontos tulajdonságáról szól. Az 1.5. tétel és az 1.6. következmény alapján könnyen belátható, ezért a bizonyítástól eltekintünk.

**1.19. Tétel.** *Legyen $f$ az $x^{(1)}, \ldots, x^{(n)}$ változók folytonos függvénye és $f(x^{(1)}, \ldots, x^{(n)}; a^{(0)}, \ldots, a^{(m)})$ az $f$ egy kifejezése, továbbá tegyük fel, hogy az $f_{[]}([y]^{(1)}, \ldots, [y]^{(n)}; [b]^{(0)}, \ldots, [b]^{(m)})$ intervallum kiértékelés értelmes $[y]^{(1)}, \ldots, [y]^{(n)}, [b]^{(0)}, \ldots, [b]^{(m)}$ intervallumokra. Ekkor minden*

$$[x]^{(k)} \subseteq [y]^{(k)}, \quad [a]^{(j)} \subseteq [b]^{(j)}, \quad 1 \leq k \leq n, \quad 0 \leq j \leq m,$$

*esetén teljesül, hogy*

$$\begin{aligned} f([x]^{(1)}, \ldots, [x]^{(n)}; [a]^{(0)}, \ldots, [a]^{(m)}) \subseteq \\ \subseteq f_{[]}([x]^{(1)}, \ldots, [x]^{(n)}; [a]^{(0)}, \ldots, [a]^{(m)}) \\ \textit{(befoglalási tulajdonság)} \end{aligned} \tag{1.35}$$

*továbbá minden*

$$[x]^{(k)} \subseteq [z]^{(k)} \subseteq [y]^{(k)}, \quad [a]^{(j)} \subseteq [c]^{(j)} \subseteq [b]^{(j)}, \quad 1 \leq k \leq n, \quad 0 \leq j \leq m,$$

*esetén teljesül, hogy*

$$f_{[]}([x]^{(1)},\dots,[x]^{(n)};[a]^{(0)},\dots,[a]^{(m)}) \subseteq \tag{1.36}$$
$$\subseteq f_{[]}([z]^{(1)},\dots,[z]^{(n)};[c]^{(0)},\dots,[c]^{(m)})$$
*(befoglalásra vett monotonitás).*

Például, ha az $f$ függvény szabálya

$$f(x;a) = a - \frac{x}{1+x}, x \neq -1,$$

akkor

$$[x] = [-\tfrac{1}{2}, 1], \quad [z] = [-\tfrac{1}{2}, 2], \quad [a] = [c] = [2,3],$$

választással nyerjük, hogy

$$f([-\tfrac{1}{2},1],[2,3]) = [\tfrac{3}{2},4] \subset f_{[]}([-\tfrac{1}{2},1];[2,3]) = [0,4],$$
$$f_{[]}([-\tfrac{1}{2},1];[2,3]) = [0,4] \subset f_{[]}([-\tfrac{1}{2},2];[2,3]) = [-2,4].$$

Az (1.35) befoglalási tulajdonság kapcsolatot teremt a függvény értékkészlete és intervallum kiértékelése között. Ebben a szakaszban, többek között levezetünk képleteket az értékkészlet intervallum kiértékeléssel való becslésére.

Bizonyos esetekben az (1.35) relációban egyenlőség áll, például, ha $x^{(1)},\dots,x^{(n)},a^{(0)},\dots,a^{(m)}$ mennyiségek pontosan egyszer szerepelnek az $f(x^{(1)},\dots,x^{(n)},a^{(0)},\dots,a^{(m)})$ kifejezésben.

**1.20. Tétel.** *Legyen $p$ egy valós változós polinom a következő kifejezéssel definiálva*

$$p(x;a^{(0)},\dots,a^{(m)}) = (\cdots((a^{(m)}x + a^{(m-1)})^{n_{m-1}} + a^{(m-2)})^{n_{m-2}} +$$
$$+ \cdots + a^{(1)})^{n_1} + a^{(0)},$$

*ahol $n_\nu \geq 2, 1 \leq \nu \leq m-1$. Amennyiben a hatványokat az alábbi módon értékeljük ki*

$$[x]^k = \left[\min_{x\in[x]} x^k, \max_{x\in[x]} x^k\right]$$

*(lásd 1.3. definíciót), akkor*

$$p([x];a^{(0)},\dots,a^{(m)}) = p_{[]}([x];a^{(0)},\dots,a^{(m)}).$$

**Bizonyítás:** $m = 2$ esetben $p(x; a^{(0)}, a^{(1)}, a^{(2)}) = (a^{(2)}x + a^{(1)})^{n_1} + a^{(0)}$, ezért a bizonyítás triviális. A további esetek teljes indukcióval beláthatók. □

Egy polinomot azonban általában nem lehet az 1.20. tételben megkívánt alakra hozni. Egy másodfokú polinom

$$p(x; b^{(0)}, b^{(1)}) = x^2 + b^{(1)}x + b^{(0)}$$

viszont átalakítható

$$p(x; a^{(0)}, a^{(1)}) = (x + a^{(1)})^2 + a^{(0)}$$

alakra, ahol

$$a^{(1)} = b^{(1)}/2, \quad a^{(0)} = b^{(0)} - (b^{(1)})^2/4.$$

Az 1.19. tétel általánosan igaz állítása és a fentebb említett speciális esetekkel együtt az $f$ értékkészletének intervallum kiértékeléssel való becslésére ad kvalitatív állítást a következő tétel egyváltozós, valós függvény esetére. Mivel az állítás feltételei a következőkben több alkalommal is előfordúlnak, ezért külön jelölést vezetünk be rá.

**1.21. Definíció.** *Legyen $f$ valós egyváltozós függvény, $f(x; a^{(0)}, \ldots, a^{(m)})$ egy szabálya, ahol $a^{(i)}$-k konstansok. Az új $\widetilde{f}(x^{(1)}, \ldots, x^{(n)}; a^{(0)}, \ldots, a^{(m)})$ szabály jelentse az előbbi átiratát úgy, hogy $x$ változó minden előfordulásánál egy új $x^{(k)}, 1 \leq k \leq n$ változót vezetünk be. Ekkor azt mondjuk, hogy $f$ a rögzített $[y]$ intervallumon kielégíti a $(*)$ feltételt, ha értelmezve van az $[y], [a]^{(0)}, \ldots, [a]^{(m)} \in \mathbb{IR}$ intervallumokra $f$ intervallum kiértékelése $f_{[]}([y]; [a]^{(0)}, \ldots, [a]^{(m)})$, továbbá $\widetilde{f}(x^{(1)}, \ldots, x^{(n)}; a^{(0)}, \ldots, a^{(m)})$ kielégíti minden $x^{(k)}, 1 \leq k \leq n$ változóra az $[y]$ intervallumból a Lipschitz feltételt a $\gamma_k > 0$ Lipschitz konstanssal az $x^{(j)} \in [y], 1 \leq j \leq n, j \neq k$ változók alkalmas választása mellett.*

**1.22. Tétel.** *Legyen $f$ valós egyváltozós függvény, $f(x; a^{(0)}, \ldots, a^{(m)})$ egy szabálya. Tegyük fel, hogy $f$ kielégíti $[y]$ intervallumon a $(*)$ feltételt. Ekkor $[x] \subset [y]$ esetén $\exists \gamma > 0$, melyre*

$$q(f([x]; [a]^{(0)}, \ldots, [a]^{(m)}), f_{[]}([x]; [a]^{(0)}, \ldots, [a]^{(m)})) \leq \gamma d([x]), \qquad \gamma \geq 0. \tag{1.37}$$

**Bizonyítás:**

$$\widetilde{f}(x,\ldots,x;a^{(0)},\ldots,a^{(m)}) = f(x;a^{(0)},\ldots,a^{(m)}), \quad x \in [y].$$

Ekkor $f$ intervallum kiértékelése

$$f_{[]}([x];[a]^{(0)},\ldots,[a]^{(m)}) = \widetilde{f}([x],\ldots,[x];[a]^{(0)},\ldots,[a]^{(m)}), \quad [x] \subseteq [y].$$

Így a bizonyítandó állítás

$$q(f([x];[a]^{(0)},\ldots,[a]^{(m)}), \widetilde{f}([x],\ldots,[x];[a]^{(0)},\ldots,[a]^{(m)})) \leq \gamma d([x]),$$
$$[x] \subseteq [y].$$

Az $[x] \subseteq [y]$ esetben írhatjuk, hogy léteznek olyan

$$u, v \in [x], \quad a^{(j)}, b^{(j)} \in [a]^{(j)}, \quad 0 \leq j \leq m$$

értékek, amelyekre

$$f([x];[a]^{(0)},\ldots,[a]^{(m)}) = [f(u;a^{(0)},\ldots,a^{(m)}), f(v;b^{(0)},\ldots,b^{(m)})],$$

illetve léteznek olyan

$$x^{(k)}, y^{(k)} \in [x], \quad 1 \leq k \leq n, \quad c^{(j)}, e^{(j)} \in [a]^{(j)}, \quad 0 \leq j \leq m$$

értékek, amelyekre

$$\widetilde{f}([x],\ldots,[x];[a]^{(0)},\ldots,[a]^{(m)}) =$$
$$= [\widetilde{f}(x^{(1)},\ldots,x^{(n)};c^{(0)},\ldots,c^{(m)}), \widetilde{f}(y^{(1)},\ldots,y^{(n)};e^{(0)},\ldots,e^{(m)})],$$

és figyelembe véve a

$$f([x];[a]^{(0)},\ldots,[a]^{(m)}) \subseteq \widetilde{f}([x],\ldots,[x];[a]^{(0)},\ldots,[a]^{(m)})$$

relációt, az alsó korlátra kapjuk, hogy

$$
\begin{aligned}
&|f(u;a^{(0)},\dots,a^{(m)}) - \widetilde{f}(x^{(1)},\dots,x^{(n)};c^{(0)},\dots,c^{(m)})| = \\
&= f(u;a^{(0)},\dots,a^{(m)}) - \widetilde{f}(x^{(1)},\dots,x^{(n)};c^{(0)},\dots,c^{(m)}) \le \\
&\le f(u;c^{(0)},\dots,c^{(m)}) - \widetilde{f}(x^{(1)},\dots,x^{(n)};c^{(0)},\dots,c^{(m)}) = \\
&= \widetilde{f}(u,\dots,u;c^{(0)},\dots,c^{(m)}) - \widetilde{f}(x^{(1)},\dots,x^{(n)};c^{(0)},\dots,c^{(m)}) = \\
&= \widetilde{f}(u,\dots,u;c^{(0)},\dots,c^{(m)}) - \widetilde{f}(x^{(1)},u,\dots,u;c^{(0)},\dots,c^{(m)}) + \\
&+ \widetilde{f}(x^{(1)},u,\dots,u;c^{(0)},\dots,c^{(m)}) - \widetilde{f}(x^{(1)},x^{(2)},u,\dots,u;c^{(0)},\dots,c^{(m)}) + \\
&+ \widetilde{f}(x^{(1)},x^{(2)},u,\dots,u;c^{(0)},\dots,c^{(m)}) + \dots \\
&\cdots - \widetilde{f}(x^{(1)},\dots,x^{(n)};c^{(0)},\dots,c^{(m)}) \le \\
&\le \gamma_1|u-x^{(1)}| + \gamma_2|u-x^{(2)}| + \cdots + \gamma_n|u-x^{(n)}| \le \\
&\le \gamma \max_{1\le k\le n} |u-x^{(k)}| \le \gamma d([x]).
\end{aligned}
$$

A értékkészlet felső korlátainak különbsége hasonlóan becsülhető. E két becslés együtt bizonyítja az állítást. □

Az 1.22. tétel állításai, ahogy a bizonyításból is látszik, azonnal általánosíthatók $x^{(1)},\dots,x^{(n)}$ többváltozós függvényekre. Ekkor a következő mennyiségre jutunk

$$
\sum_{k=1}^{n} \gamma^{(k)} d([x]^{(k)}) \quad \left( \le \gamma \max_{1\le k\le n} d([x]^{(k)}) \right).
$$

Az alábbi példa bemutatja, hogy $f$ értékészletének intervallum kiértékeléssel való becslése függ $f$ értékeinek becslésére használt $f(x;a^{(0)},\dots,a^{(m)})$ szabály választásától.

Legyen $f(x) = x - x^2$ és $[x] = [0,1]$; Ekkor

$$
f([0,1]) = \left\{ x - x^2 \;\middle|\; 0 \le x \le 1 \right\} = [0, \tfrac{1}{4}].
$$

Az alábbi ekvivalens kifejezésekre más más eredmények adódnak:

$$\begin{aligned}
f^{(0)}(x) = x - x^2 &\Rightarrow f_{[]}^{(0)}([0,1]) = [0,1] - [0,1] = [-1,1],\\
f^{(1)}(x) = x(1-x) &\Rightarrow f_{[]}^{(1)}([0,1]) = [0,1](1-[0,1]) = [0,1],\\
f^{(2)}(x) = \tfrac{1}{4} - (x-\tfrac{1}{2})(x-\tfrac{1}{2}) &\Rightarrow\\
f_{[]}^{(2)}([0,1]) = \tfrac{1}{4} - ([0,1]-\tfrac{1}{2})([0,1]-\tfrac{1}{2}) &= [0,\tfrac{1}{2}],\\
f^{(3)}(x) = \tfrac{1}{4} - (x-\tfrac{1}{2})^2 &\Rightarrow\\
f_{[]}^{(3)}([0,1]) = \tfrac{1}{4} - ([0,1]-\tfrac{1}{2})^2 &= [0,\tfrac{1}{4}] = f([0,1]).
\end{aligned}$$

Az $f$ egy bizonyos alakú szabályára belátható az 1.22. tételnél élesebb állítás is. Ez az alak nem más, mint $f$ centralizált formája, ami egy $[x]$ halmazon kiértékelendő $f$ függvényhez tartozó speciális alak. Most koncentráljunk az egyváltozós valós esetre, válasszunk egy $z \in [x]$ pontot. Ekkor az $f(x)$ kifejezés előállítható

$$f(x) = f(z) + (x-z)h(x-z) \tag{1.38}$$

alakban, ahol a $h(x-z)$ tag az eltolt $\widetilde{z} = x - z$ változó függvénye. Az (1.38) alakot hívjuk $f(x)$ $z$ körüli centrális alakjának. Polinomok esetén (1.38) egyszerűen $f(x)$ $z$ körüli Taylor kifejtése $(x-z)$ alakra rendezve a nem konstans tagokat.

Legyen $f(x) = \frac{p(x)}{q(x)}$ racionális törtfüggvény, ekkor az alábbi centrális formára hozható. Legyen $n$ a $p(x), q(x)$ polinomok fokszámának maximuma. Ekkor $z \in [x]$ mellett értelmezzük az alábbi kifejezést

$$\gamma_\nu := p^{(\nu)}(z) - f(z)q^{(\nu)}(z), \quad 1 \le \nu \le n.$$

A

$$h(y) = \frac{\sum_{\nu=1}^{n} \gamma_\nu \frac{y^{\nu-1}}{\nu!}}{\sum_{\nu=0}^{s} \frac{y^\nu}{\nu!}}$$

függvény kielégíti az (1.38) függvényegyenletet.

**1.23. Tétel.** *Legyen $f$ a valós $x$ változó függvénye, és legyen*

$$f(x) = f(z) + (x-z)h(x-z)$$

*az $f$ centrális alakja. Tegyük fel, hogy létezik az $f_{[]}([y])$ intervallum kiértékelés valamely $[y] \in \mathbb{IR}$ halmazra és $h(x-z)$ kielégíti a $(*)$ feltételt az $[y]$ intervallumon. Ekkor tetszőleges $[x] \subseteq [y]$ esetén*

$$q(f([x]), f_{[]}([x])) \leq c \cdot (d([x]))^2, \quad c \geq 0. \tag{1.39}$$

**Bizonyítás:** Mivel

$$\widetilde{h}(x-z, \ldots, x-z) = h(x-z)$$

és

$$\widetilde{f}(x^{(0)}, \ldots, x^{(n)}) = f(z) + (x^{(0)} - z)\widetilde{h}(x^{(1)} - z, \ldots, x^{(n)} - z),$$

kapjuk, hogy

$$\begin{aligned} \widetilde{f}(x, \ldots, x) &= f(z) + (x-z)\widetilde{h}(x-z, \ldots, x-z) = \\ &= f(z) + (x-z)h(x-z) = f(x). \end{aligned}$$

$f$ centrális alakjának intervallum kiértékelése ekkor a következő alakban írható

$$f_{[]}([x]) = \widetilde{f}([x], \ldots, [x]),$$

így az állítás alakja

$$q(f([x]), \widetilde{f}([x], \ldots, [x])) \leq c \cdot (d([x]))^2, \quad c \geq 0.$$

Legyenek

$$x^{(k)}, y^{(k)} \in [x], 0 \leq k \leq n,$$

olyanok, hogy

$$\begin{aligned} \widetilde{f}([x], \ldots, [x]) = &[f(z) + (x^{(0)} - z)\widetilde{h}(x^{(1)} - z, \ldots, x^{(n)} - z), \\ &f(z) + (y^{(0)} - z)\widetilde{h}(y^{(1)} - z, \ldots, y^{(n)} - z)], \end{aligned}$$

és vegyük észre, hogy $f([x]) \subseteq \widetilde{f}([x], \ldots, [x])$. Használva az (1.31) relációt a következő becslés adódik

$$q(f([x]), \widetilde{f}([x], \ldots, [x])) \leq d(\widetilde{f}([x], \ldots, [x])) - d(f([x])).$$

Legyen $w$ olyan, hogy

$$\min_{x\in[x]} |h(x-z)| = |h(w-z)|.$$

Az

$$f(z) + ([x]-z)h(w-z) \subseteq f(z) + \{(x-z)h(x-z)|x \in [x]\} = f([x])$$

reláció könnyen igazolható $h(w-z)$ előjele miatt fellépő két eset vizsgálatával. Felhasználva az (1.19) és az (1.24) összefüggéseket, kapjuk, hogy

$$d(f([x])) \geq d(([x]-z)h(w-z)) = d([x])|h(w-z)|, \quad w \in [x].$$

Tovább becsülhetünk az alábbiak szerint

$$\begin{aligned}
&q(f([x]), \widetilde{f}([x],\ldots,[x])) \leq \\
&\leq (y^{(0)}-z)\widetilde{h}(y^{(1)}-z,\ldots,y^{(n)}-z) - (x^{(0)}-z)\widetilde{h}(x^{(1)}-z,\ldots,x^{(n)}-z) - \\
&- d([x])\cdot|h(w-z)| = \\
&= (y^{(0)}-z)\widetilde{h}(y^{(1)}-z,\ldots,y^{(n)}-z) - (y^{(0)}-z)\widetilde{h}(x^{(1)}-z,\ldots,x^{(n)}-z) + \\
&+ (y^{(0)}-z)\widetilde{h}(x^{(1)}-z,\ldots,x^{(n)}-z) - (x^{(0)}-z)\widetilde{h}(x^{(1)}-z,\ldots,x^{(n)}-z) - \\
&- d([x])\cdot|h(w-z)| = \\
&= (y^{(0)}-z)\left(\widetilde{h}(y^{(1)}-z,\ldots,y^{(n)}-z) - \widetilde{h}(x^{(1)}-z,\ldots,x^{(n)}-z)\right) + \\
&+ (y^{(0)}-x^{(0)})\widetilde{h}(x^{(1)}-z,\ldots,x^{(n)}-z) - d([x])\cdot|\widetilde{h}(w-z,\ldots,w-z)| \leq \\
&\leq |y^{(0)}-z|\cdot|\widetilde{h}(y^{(1)}-z,\ldots,y^{(n)}-z) - \widetilde{h}(x^{(1)}-z,\ldots,x^{(n)}-z)| + \\
&+ |y^{(0)}-x^{(0)}|\cdot|\widetilde{h}(x^{(1)}-z,\ldots,x^{(n)}-z)| - d([x])\cdot|\widetilde{h}(w-z,\ldots,w-z)| \leq \\
&\leq d([x])\cdot(|\widetilde{h}(y^{(1)}-z,\ldots,y^{(n)}-z) - \widetilde{h}(x^{(1)}-z,\ldots,x^{(n)}-z)| + \\
&+ \left||\widetilde{h}(x^{(1)}-z,\ldots,x^{(n)}-z)| - |\widetilde{h}(w-z,\ldots,w-z)|\right|) \leq \\
&\leq d([x])\cdot\left(c^{(1)}\max_{1\leq k\leq n}|y^{(k)}-x^{(k)}| + c^{(2)}\max_{1\leq k\leq n}|x^{(k)}-w|\right) \leq \\
&\leq d([x])\cdot(c^{(1)}+c^{(2)})\cdot d([x]) = c\cdot(d([x]))^2.
\end{aligned}$$

Itt felhasználtuk $\widetilde{h}, |\widetilde{h}|$ kifejezésekre a kapcsolódó Lipschitz feltételeket. □

Az előbbi bizonyítás többváltozós függvények esetére is átvihető.

Az 1.22. tétel következményeként becslést adunk az intervallum kiértékelés átmérőjére.

**1.24. Tétel.** *Legyen $f$ az $x$ valós változó függvénye, és $f(x)$ annak egy kiértékelési szabálya. Tegyük fel, hogy $f$-re $[y]$-on teljesül a $(*)$ feltétel. Ekkor*

$$d(f_{[]}([x])) \leq c \cdot d([x]), \quad c \geq 0, \tag{1.40}$$

*állítás igaz, ha $[x] \subseteq [y]$.*

**Bizonyítás:**

$$\begin{aligned} d(f_{[]}([x])) &\leq 2q(f_{[]}([x]), f([x])) + d(f([x])) \leq \\ &\leq 2c^{(1)} d([x]) + d(f([x])), \quad c^{(1)} \geq 0. \end{aligned}$$

Mivel a függvény eleget tesz a Lipschitz feltételnek, adódik, hogy

$$d(f([x])) = f(x) - f(y) \leq c^{(2)}|x - y|, \quad x, y \in [x], \quad c^{(2)} \geq 0,$$

amiből a

$$d(f_{[]}([x])) \leq 2c^{(1)} d([x]) + c^{(2)} d([x]) = c \cdot d([x])$$

állítás következik. □

Többváltozós esetben az állítás alakja

$$\begin{aligned} d(f_{[]}([x]^{(1)}, \ldots, [x]^{(n)})) &\leq \sum_{k=1}^{n} c^{(k)} d([x]^{(k)}) \leq \\ &\leq c \max_{1 \leq k \leq n} d([x]^{(k)}). \end{aligned} \tag{1.41}$$

A középérték tétel segítségével be szeretnénk látni az (1.35) típusú befoglalási tulajdonságot.

**1.25. Tétel.** *Legyen $f$ valós változós függvény, differenciálható $[x] = [\underline{x}, \overline{x}]$ intervallumban, továbbá $f'(x)$ legyen $f$ deriváltjának egy, az $[x]$ intervallumon kiértékelhető szabálya. Ekkor, ha $f'$ függvényre $[x]$ intervallumon teljesül a $(*)$ feltétel, akkor $y \in [x]$ esetén*

$$f([x]) \subseteq f(y) + f'_{[]}([x])([x] - y), \tag{1.42}$$

$$q(f([x]), f(y) + f'_{[]}([x])([x] - y)) \leq \widetilde{c} \cdot (d([x]))^2, \quad \widetilde{c} \geq 0. \tag{1.43}$$

**Bizonyítás:** Az (1.42) állítás bizonyítása. A középérték tételből tudjuk, hogy valamely $x, y \in [x]$ elemekre

$$f(x) = f(y) + f'(y + \theta(x - y))(x - y), \quad 0 < \theta < 1.$$

Az

$$y + \theta(x - y) \in y + [0, 1]([x] - y) = [x]$$

összefüggésből a befoglalásra vett monotonitás miatt következik, hogy

$$f(x) \in f(y) + f'_{[]}([x])([x] - y).$$

Ezzel (1.42) állítás bizonyított.

Az (1.43) állítás bizonyítása. Tekintsük az

$$f([x]) = [f(u), f(v)], \quad u, v \in [x]$$

kifejezést. A középérték tételből következik, hogy

$$\begin{aligned} d(f([x])) = f(v) - f(u) = |f(v) - f(u)| \geq \\ \geq |f(\underline{x}) - f(\overline{x})| = |f'(\xi)| d([x]), \quad \xi \in [x]. \end{aligned}$$

Mivel $\xi \in [x]$ és $f'(\xi) \in f'_{[]}([x])$, az (1.22), (1.13), (1.30) összefüggésekből kapjuk, hogy

$$q(f'_{[]}([x]), f'(\xi)) \leq d(f'_{[]}([x])).$$

Felhasználva az alábbi

$$|f'_{[]}([x])| - |f'(\xi)| \leq q(f'_{[]}([x]), f'(\xi))$$

egyenlőtlenséget, ami az (1.14), (1.15) és az 1.12. definíció alapján belátható, valamint (1.42) összefüggést (1.31) és az 1.24. tételt $f'_{[]}([x])$

kifejezésre alkalmazva kapjuk, hogy

$$\begin{aligned}
&q(f([x]), f(y) + f'_{[]}([x])([x] - y)) \leq \\
&\quad \leq d(f(y) + f'_{[]}([x])([x] - y)) - d(f([x])) \leq \\
&\quad \leq d(f'_{[]}([x]))d([x]) + (|f'_{[]}(x)| - |f'(\xi)|)d([x]) \leq \\
&\quad \leq d(f'_{[]}([x]))d([x]) + q(f'_{[]}([x]), f'(\xi))d([x]) \leq \\
&\quad \leq 2c \cdot (d([x]))^2 = \widetilde{c} \cdot (d([x]))^2.
\end{aligned}$$

□

Az 1.25. tételben a centrális formára kapott kvalitatív eredmény megkapható az 1.23. tételből

$$f_{[],z}([x]) := f(z) + f'_{[]}([x])([x] - z), \qquad z, x \in [x].$$

kifejezés felhasználásával, amit a szakirodalom standard centrális alaknak is nevez. Amennyiben $z = m([x])$, $f_{[],m}([x])$ kifejezést $f$ középérték alakjának nevezzük. Egy centrális alak általában sajnos nem rendelkezik a befoglalásra vett monotonitás tulajdonságával, csak a középérték alak.

A fenti állítás fontos tény, mivel már polinomok esetén is a teljes Horner séma szükségeltetik a centrális alak előállításához.

Az 1.25. tétel is általánosítható többváltozós függvényekre, de ezzel itt nem foglalkozunk.

Tekintsük az $f(x) = p(x)/q(x)$ racionális törtfüggvényeket. A $p(x) = \sum_{\nu=0}^{r} a_\nu x^\nu$ és $q(x) = \sum_{\nu=0}^{s} b_\nu x^\nu$ polinomokhoz bizonyos körülmények között léteznek a centrális alaknál, vagy az 1.25. tételbeli középérték alaknál egyszerűbb alakok, amelyek még teljesítik a

$$q(f([x]), f_{[]}([x])) \leq c \cdot (d([x]))^2, \quad c \geq 0, \tag{1.44}$$

feltételt.

Legyen $c = m([x])$ $[x]$ középpontja, és legyen adva a két polinom Taylor polinomja $p(x) = \sum_{\nu=0}^{r} a'_\nu (x - c)^\nu, q(x) = \sum_{\nu=0}^{s} b'_\nu (x - c)^\nu$. Az általánosság megszorítása nélkül feltehető, hogy $b'_0 = 1$ és $0 \notin q_{[]}([x]) := 1 + \sum_{\nu=1}^{s} b'_\nu ([x] - c)^\nu$. Ha most

$$\operatorname{sgn}(a'_1) \operatorname{sgn}(b'_1 a'_0) \leq 0, \tag{1.45}$$

akkor az

$$f_{[]}([x]) = \frac{\sum_{\nu=0}^{r} a'_\nu([x]-c)^\nu}{1+\sum_{\nu=1}^{s} b'_\nu([x]-c)^\nu}$$

intervallum kifejezés (1.44) tulajdonságú, amennyiben $p_{[]}([x]), q_{[]}([x])$ teljesíti a $d(p_{[]}([x])) \leq c_1 d([x]), d(q_{[]}([x])) \leq c_2 d([x])$ megkötéseket. Ezek a megkötések állnak a fenti két kifejezésre, akár $[x]-c$ hatványait, akár a Horner elrendezést használjuk. Ha most vesszük $p$ és $q$ centrális alakjait, ahol

$$0 \notin 1 + ([x]-c)q'_{[]}([x]),$$

akkor (1.45) teljesülése esetén

$$f_{[]}([x]) = \frac{a'_0 + ([x]-c)p'_{[]}([x])}{1+([x]-c)q'_{[]}([x])}$$

szintén kielégíti (1.44) feltételt. Itt $p'_{[]}([x])$ $p(x)$ deriváltjának egy intervallum kiértékelése $d(p'_{[]}([x])) \leq \alpha d([x])$ tulajdonsággal. Hasonlóan $q'_{[]}([x])$ $q(x)$ deriváltjának egy intervallum kiértékelése $d(q'_{[]}([x])) \leq \beta d([x])$ tulajdonsággal.

A 8. fejezetben a függvény meredekségének befoglalásait használjuk függvény zérushelyeinek befoglalásaihoz. A következőkben a különbségi hányados véges sok lehetséges befoglalását adjuk. Ezek részben rendezettek lesznek. Kiderül, hogy az optimális befoglalás egyszerűen és szisztematikusan megadható, és a megfelelő iterációval való számítás valamint a derivált intervallum kiértékelésének számolási igénye azonos.

Legyen adott az alábbi polinom

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

Az alábbi két egyenlőség algebrai átalakításokkal belátható:

$$\begin{aligned} p(x) - p(y) &= \sum_{i=0}^{n} a_i (x^i - y^i) = \\ &= \left( \sum_{i=1}^{n} a_i \sum_{j=1}^{i} x^{i-j} y^{j-1} \right) (x - y) = \\ &= \left( \sum_{i=1}^{n} \left( \sum_{j=i}^{n} a_j y^{j-i} \right) x^{i-1} \right) (x - y), \end{aligned} \tag{1.46}$$

$$\begin{aligned} p(x) - p(y) &= \sum_{i=0}^{n} a_i (x^i - y^i) = \\ &= \left( \sum_{i=1}^{n} a_i \sum_{j=1}^{i} y^{i-j} x^{j-1} \right) (x - y) = \\ &= \left( \sum_{i=1}^{n} \left( \sum_{j=i}^{n} a_j x^{j-i} \right) y^{i-1} \right) (x - y). \end{aligned} \tag{1.47}$$

Rögzített $y$ és tetszőleges $x \in [x]$ mellett (1.46) és a befoglalásra vonatkozó monotonitás alapján kapjuk, hogy

$$\frac{p(x) - p(y)}{x - y} \in \left( \sum_{i=1}^{n} c_{i-1} [x]^{i-1} \right)_H =: [j_1] \tag{1.48}$$

$$\subseteq [j_2] := \sum_{i=1}^{n} c_{i-1} [x]^{i-1}, \tag{1.49}$$

ahol

$$c_{i-1} = \sum_{j=i}^{n} a_j y^{j-i}, \quad 1 \leq i \leq n.$$

$H$ jelöli a Horner elrendezés szerinti kiértékelést. $[j_2]$ kifejezésben az $[x]^r$ hatványt $[x]^0 := 1$ és $[x]^r = [x]^{r-1}[x]$, $r \geq 1$ definiálja. A szubdisztributivitás miatt $[j_1] \subseteq [j_2]$. Viszont minden valós számra és

$[a]_j, 0 \leq j \leq n-1$ intervallumra

$$\sum_{j=1}^{n} [a]_{j-1} y^{j-1} = \left( \sum_{j=1}^{n} [a]_{j-1} y^{j-1} \right)_H .$$

Felhasználva a szubdisztributivitást és ezt az egyenlőséget, rögzített $y$ és tetszőleges $x \in [x], x \neq y$ mellett (1.47) miatt

$$\frac{p(x)-p(y)}{x-y} \in \sum_{i=1}^{n} ([c]_{i-1})_H \, y^{i-1} = \left( \sum_{i=1}^{n} ([c]_{i-1})_H \, y^{i-1} \right)_H =: [j_3] \quad (1.50)$$

$$\subseteq [j_4] := \sum_{i=1}^{n} [c]_{i-1} y^{i-1} = \left( \sum_{i=1}^{n} [c]_{i-1} y^{i-1} \right)_H , \quad (1.51)$$

ahol

$$([c]_{i-1})_H = \left( \sum_{j=i}^{n} a_j [x]^{j-i} \right)_H , \qquad 1 \leq i \leq n,$$

és

$$[c]_{i-1} = \sum_{j=i}^{n} a_j [x]^{j-i}, \qquad 1 \leq i \leq n.$$

**1.26. Tétel.** *A fenti kifejezések kielégítik az alábbi feltételeket:*

$$[j_1] \subseteq [j_2] \subseteq [j_4], \quad (1.52)$$

$$[j_1] \subseteq [j_3] \subseteq [j_4], \quad (1.53)$$

$$[j_4] \subseteq p'_{[]}([x]) \subseteq \sum_{i=1}^{n} i a_i [x]^{i-1}. \quad (1.54)$$

**Bizonyítás:** Az érthetőség kedvéért az $n = 4$ negyedrendű polinomok esetére korlátozzuk bizonyításunk. Az általános eset teljesen analóg módon látható be. Az (1.52) és az (1.54) állítások bizonyításához csak azt kell belátnunk, hogy $[j_2] \subseteq [j_4] \subseteq p'_{[]}([x])$. A befoglalásra vett mono-

tonitás és (1.8) alapján kapjuk, hogy

$$\begin{aligned}
[j_2] &= \sum_{i=1}^{4} c_{i-1}[x]^{i-1} = \\
&= (a_1 + a_2 y + a_3 y^2 + a_4 y^3)[x]^0 + (a_2 + a_3 y + a_4 y^2)[x] + \\
&\quad + (a_3 + a_4 y)[x]^2 + a_4[x]^3 \subseteq \\
&\subseteq a_1 + a_2[x] + a_3[x]^2 + a_4[x]^3 + a_2 y + a_3 y[x] + a_4 y[x]^2 + \\
&\quad + a_3 y^2 + a_4 y^2[x] + a_4 y^3 = \\
&= a_1 + a_2[x] + a_3[x]^2 + a_4[x]^3 + (a_2 + a_3[x] + a_4[x]^2)y + \\
&\quad + (a_3 + a_4[x])y^2 + a_4 y^3 = [j_4] \subseteq \\
&\subseteq a_1 + a_2[x] + a_3[x]^2 + a_4[x]^3 + a_2[x] + a_3[x]^2 + a_4[x]^3 + \\
&\quad + a_3[x]^2 + a_4[x]^3 + a_4[x]^3 = p'_{[]}([x]).
\end{aligned}$$

Az (1.53) állítás bizonyításához csak azt kell belátnunk, hogy $[j_1] \subseteq [j_3]$.

$$\begin{aligned}
[j_1] &= ((c_3[x] + c_2)[x] + c_1)[x] + c_0 = \\
&= ((a_4[x] + (a_3 + a_4 y))[x] + a_2 + a_3 y + a_4 y^2)[x] + \\
&\quad + a_1 + a_2 y + a_3 y^2 + a_4 y^3 \subseteq \\
&\subseteq ((a_4[x] + a_3)[x] + a_4 y[x] + a_2 + a_3 y + a_4 y^2)[x] + \\
&\quad + a_1 + a_2 y + a_3 y^2 + a_4 y^3 = \\
&= (((a_4[x] + a_3)[x] + a_2) + a_4 y[x] + a_3 y + a_4 y^2)[x] + \\
&\quad + a_1 + a_2 y + a_3 y^2 + a_4 y^3 = \\
&= (((a_4[x] + a_3)[x] + a_2) + (a_4[x] + a_3)y + a_4 y^2)[x] + \\
&\quad + a_1 + a_2 y + a_3 y^2 + a_4 y^3 \subseteq \\
&\subseteq ((a_4[x] + a_3)[x] + a_2)[x] + (a_4[x] + a_3)y[x] + a_4 y^2[x] + \\
&\quad + a_1 + a_2 y + a_3 y^2 + a_4 y^3 = \\
&= (((a_4[x] + a_3)[x] + a_2)[x] + a_1)y^0 + ((a_4[x] + a_3)[x] + a_2)y + \\
&\quad + (a_4[x] + a_3)y^2 + a_4 y^3 = [j_3].
\end{aligned}$$

Ezzel a tételt bizonyítottuk. □

Nincs általános szabály arra, hogy $[j_2]$ vagy $[j_3]$ adja a legjobb befoglalást. $[j_2] \subseteq [j_3]$ vagy $[j_3] \subseteq [j_2]$ is feltehető. Például legyen

$$p(x) = x^3 - x^2, \quad [x] = [-1, 2], \quad y = 1.$$

Ekkor

$$[j_2] = (a_1 + a_2y + a_3y^2)[x]^0 + (a_2 + a_3y)[x] + a_3[x]^2 = [x]^2 = [-2, 4]$$

és

$$\begin{aligned}[j_3] &= ((a_3[x] + a_2)[x] + a_1)y^0 + (a_3[x] + a_2)y + a_3y^2 = \\ &= ([x] - 1)[x] + ([x] - 1) + 1 = [-5, 4],\end{aligned}$$

tehát $[j_2] \subset [j_3]$.

Másfelől, ha $y = 0$ és így $c_{i-1} = a_i, 1 \leq i \leq n$, akkor

$$[j_2] = \sum_{i=1}^{n} a_i[x]^{i-1}, \qquad [j_3] = \left(\sum_{i=1}^{n} a_i[x]^{i-1}\right)_H,$$

ahol $[j_3] \subseteq [j_2]$. Tekintsük most az előbbi példát $y = 0, [x] = [0, 2]$ értékekkel. Ekkor

$$[j_2] = [x]^2 - [x] = [-2, 4] \qquad [j_3] = ([x] - 1)[x] = [-2, 2],$$

így $[j_3] \subset [j_2]$.

Az 1.26. tétel alapján a $[j_1], [j_2]$ intervallumok kiszámítása ismertnek feltételezi $c_{i-1} = \sum_{j=i}^{n} a_j y^{j-i}, \quad 1 \leq i \leq n$ értékeit. Amennyiben a $p(x)$ polinom $y$ helyen vett értéke is adott, mint például a 8. fejezet iterációs eljárásainál, akkor $c_{i-1}$ számítása nem igényel további aritmetikai műveleteket, ezek ugyanis kiszámításra kerülnek $p(y)$ számításakor. Legyen adott

$$p(x) = \sum_{i=1}^{n} a_i x^i,$$

mint feljebb. Ekkor a

$$p_n := a_n, \quad \text{és} \quad p_{i-1} := p_i y + a_{i-1}, \qquad i = n, \ldots, 1$$

Horner elrendezés szerint számolva kapjuk $p_0 = p(y)$ értékét. A definícióból

$$\begin{aligned} c_{n-1} &= a_n && (= p_n), \\ c_{n-2} &= a_n y + a_{n-1} && (= p_{n-1}), \\ &\vdots && \vdots \\ c_0 &= c_1 y + a_1 && (= p_1), \end{aligned}$$

ezzel $c_{i-1} = p_i,\ 1 \le i \le n$.

**Példák:**

a) $p(x) = x^4 - 1, \qquad [x] = [0.5, 3.5], \qquad y = 2.$

$$[j_1] = [j_2] = [j_3] = [j_4] = [10.625, 89.375], \tag{1.55}$$
$$p'_{[]}([x]) = \left(p'_{[]}([x])\right)_H = [0.5, 171.5]. \tag{1.56}$$

b) $p(x) = x^3 + 4x - 16, \qquad [x] = [-1, 3] \qquad y = 1.$

$$[j_1] = [j_2] = [j_3] = [j_4] = [1, 17], \tag{1.57}$$
$$p'_{[]}([x]) = \left(p'_{[]}([x])\right)_H = [-5, 31], \tag{1.58}$$

c) $p(x) = \sum_{i=0}^{n} a_i x^i, \qquad 0 \in [x] \qquad y = 0$. Ekkor

$$c_0 = a_1, \qquad c_1 = a_2, \ldots, c_{n-1} = a_n$$

és

$$[j_1] = \left(\sum_{i=1}^{n} c_{i-1}[x]^{i-1}\right)_H = \left(\sum_{i=1}^{n} a_i[x]^{i-1}\right)_H.$$

d) $p(x) = x^3 - x^2, \qquad [x] = [1, 3] \qquad y = 2.$

$$[j_1] = [j_2] = [j_3] = [4, 14] \subset [2, 16] = [j_4] \subset$$
$$\subset \left(p'_{[]}([x])\right)_H = [1, 21] \subset [-3, 25] = p'_{[]}([x]).$$

e) Legyen $x_0 \in [x]$ és $f \in C^{n+1}([x])$. A Taylor kifejtéssel adódik, hogy

$$f(x) = p(x) + \phi(x),$$

ahol

$$p(x) = \sum_{k=0}^{n} \frac{(x - x_0)^k}{k!} f^{(k)}(x_0)$$

és

$$\phi(x) = \int_{x_0}^{x} \frac{(x - t)^n}{n!} f^{(n+1)}(t)dt.$$

$\phi$ differenciálható és

$$\phi'(x) = \int_{x_0}^{x} \frac{(x-t)^{n-1}}{(n-1)!} f^{(n+1)}(t)dt.$$

A integrálokra vonatkozó középérték tétel miatt

$$\phi'(x) = f^{(n+1)}(\eta) \int_{x_0}^{x} \frac{(x-t)^{n-1}}{(n-1)!} dt = \frac{(x-x_0)^n}{n!} f^{(n+1)}(\eta)$$

valamely $x \leq \eta \leq x_0$ számra. A középérték tételt $\phi$ függvényre alkalmazva adódik

$$\begin{aligned} f(x) - f(y) &= p(x) - p(y) + \phi(x) - \phi(y) \\ &= \left\{ \sum_{k=1}^{n} c_{k-1}(x-x_0)^{k-1} + \phi'(\xi) \right\} (x-y), \end{aligned}$$

ahol

$$c_{k-1} = \sum_{j=k}^{n} (y-x_0)^{j-k} \frac{f^{(k)}(x_0)}{k!}, \qquad 1 \leq k \leq n$$

és

$$\phi'(\xi) = \frac{(\xi - x_0)^n}{n!} f^{(n+1)}(\eta),$$

ahol $x \leq \xi \leq y$ és $x_0 \leq \eta \leq \xi$. $y = x_0$ választással adódik

$$c_0 = f'(x_0)/1!, \quad \ldots, \quad c_{n-1} = f^{(n)}(x_0)/n!.$$

Ha az $(n+1)$-edik deriváltnak létezik kiszámítható intervallum szabálya, akkor $y = x_0$ esetére

$$\frac{f(x) - f(x_0)}{x - x_0} \in \sum_{k=1}^{n} \frac{f^{(k)}(x_0)}{k!} ([x] - x_0)^{k-1} + f^{(n+1)}([x]) \frac{([x] - x_0)^n}{n!},$$

mivel $\eta, \xi \in [x]$.

f) $p(x) = x^7 + 3x^6 - 4x^5 - 12x^4 - x^3 - 3x^2 + 4x + 12$, $[x] = [1.8, 3]$, $y = 2$. Kapjuk, hogy

$$\begin{aligned} [j_1] &= [173.2362, 2400], & [j_2] &= [161.4762, 2411.76] \\ [j_3] &= [24.72, 2400], & [j_4] &= [-870.2933, 3443.5296] \\ (p'_{[]}([x]))_H &= [71.79808, 6520], & p'_{[]}([x]) &= [-2378.791292, 8970.592]. \end{aligned}$$

Ezek a gondolatok többváltozós esetben is végig vihetők.

## 1.4. Gépi intervallum aritmetika

Rátérünk az intervallumműveletek gépi megvalósítására. Mint jól ismert, a számítógépek véges számhalmazzal dolgoznak, amelyet gyakran szemilogaritmikus alakban írnak le fix hosszúságú, lebegőpontos számokkal:

$$x = m \cdot b^e,$$

ahol $m$ a mantissza, $b$ a hatványalap, $e$ a karisztika. A számok belső gépi ábrázolása rendszerint $b = 2$ alappal és a mantissza normalizált $(1/2 \leq |m| < 1)$ formájával történik. A kitevő korlátok közé esik $e_{\min} \leq e \leq e_{\max}$.

A gépi számok fenti típusú halmazát $\mathcal{R}$ jelöli és feltesszük, hogy a további meggondolásoknál $\mathcal{R}$ szimmetrikus, azaz

$$\mathcal{R} = -\mathcal{R}.$$

A $[\min_{y \in \mathcal{R}} y, \max_{y \in \mathcal{R}} y]$ intervallumba tartozó valós számok hatékonyan közelíthetők $\widetilde{x} \in \mathcal{R}$ gépi számokkal, az alábbi leképezés segítségével

$$\bigcirc : \mathbb{R} \ni x \mapsto \widetilde{x} = \bigcirc(x) \in \mathcal{R}. \tag{1.59}$$

Ezt a leképezést *kerekítés*nek nevezzük, amennyiben teljesül

$$x \leq y \Rightarrow \bigcirc(x) \leq \bigcirc(y) \quad \text{(monotonitás)}. \tag{1.60}$$

Az

$$x \in \mathcal{R} \Rightarrow \bigcirc(x) = x \tag{1.61}$$

tulajdonságú kerekítéseket optimális kerekítéseknek nevezzük. Különösen érdekesek az irányított kerekítések, tehát azok, amelyek mindig fel, vagy le kerekítenek. Ha $\bigtriangledown$ kerekítésre igaz, hogy

$$x \in \mathbb{R} \Rightarrow \bigtriangledown x \leq x, \tag{1.62}$$

akkor lefelé irányított kerekítésről beszélünk. Felhasználva a

$$\bigtriangleup x := -(\bigtriangledown(-x)), \quad x \in \mathbb{R} \tag{1.63}$$

definíciót, felfelé irányított kerekítéshez jutunk; a fel- és lefelé irányított kerekítésre kézenfekvő példa rendre a felső, ill. alsó egészrész.

A valós számok gépi számokkal való ábrázolásával azonos módon ábrázolhatók a valós intervallumok gépi intervallumokkal. A feladat egy

$$[x] \in \mathbb{IR},\ [x] \subseteq \left[\min_{y\in\mathcal{R}} y, \max_{y\in\mathcal{R}} y\right]$$

intervallum ábrázolása alkalmas gépi intervallummal az alábbi halmazból

$$I\mathcal{R} = \left\{[x_1, x_2] \;\middle|\; x_1, x_2 \in \mathcal{R},\ x_1 \leq x_2\right\} \subset \mathbb{IR}.$$

Az

$$\Diamond : \mathbb{IR} \ni [x] \to \Diamond[x] \in I\mathcal{R}$$

intervallum kerekítésnek rendelkeznie kell az alábbi tulajdonságokkal

$$[x] \in \mathbb{IR} \Rightarrow [x] \subseteq \Diamond[x] \tag{1.64}$$

és

$$[x], [y] \in \mathbb{IR}, \quad [x] \subseteq [y] \Rightarrow \Diamond[x] \subseteq \Diamond[y], \tag{1.65}$$

hogy az intervallumműveletek alapvető tulajdonságait gépi intervallum műveletekre átvihessük. Amennyiben egy $[x] = [x_1, x_2]$ intervallum és annak $\widetilde{[x]} = [\widetilde{x}_1, \widetilde{x}_2]$ gépi ábrázolása közti átmenetet tekintjük, (1.65) szerint ezt a megfelelő korlátok kerekítésével, (1.64) szerint pedig ezeket a kerekítéseket a megfelelő irányítással kell megvalósítanunk, amiből következik, hogy minden intervallumkerekítés előáll az alábbi alakban

$$\Diamond[x] = \Diamond[x_1, x_2] = [\bigtriangledown x_1, \bigtriangleup x_2]. \tag{1.66}$$

A fentiekből következik, hogy elegendő egy lefelé irányított kerekítés az intervallum kerekítés megvalósításához, azonban nem szükségszerű, hogy az (1.63) összefüggéssel kapcsolódjon $\bigtriangledown$ és $\bigtriangleup$.

Ha két $x, y \in \mathcal{R}$ gépi számmal végzünk $\circ \in \{+, -, \cdot, :\}$ műveletet, az eredmény is egy $z \in \mathcal{R}$ gépi szám. Ha nem lépünk ki $\mathcal{R}$ értékei közül (alul-, túlcsordulás), akkor az eredmény

$$z = \bigcirc(x \circ y) \tag{1.67}$$

alakban előállítható egy alkalmas $\bigcirc$ kerekítéssel. Ezúton a gépi műveletek eredményére adható az alábbi

**1.27. Definíció.** *Legyen* $[a], [b] \in I\mathcal{R}, \circ \in \{+, -, \cdot, :\}$, *és legyen adott egy intervallum kerekítés. Ekkor az* $[a], [b]$ *elemekre alkalmazott* $\circ$ *művelet* $\Diamond$ *intervallum kerekítéssel kapott eredménye*

$$[c] = \Diamond([a] \circ [b]) \in I\mathcal{R}. \tag{1.68}$$

Belátjuk, hogy az intervallum aritmetika alapvető tulajdonságai továbbra is állnak ezen definíció alkalmazásával.

**1.28. Tétel.** *Az 1.27. definícióban értelmezett gépi műveletekre igaz a következő állítás*

$$\begin{aligned} &[a]^{(k)}, [b]^{(k)} \in I\mathcal{R}, \circ \in \{+, -, \cdot, :\}, [a]^{(k)} \subseteq [b]^{(k)}, k = 1, 2 \\ &\quad \Rightarrow [c]^{(1)} = \Diamond([a]^{(1)} \circ [a]^{(2)}) \subseteq [c]^{(2)} = \Diamond([b]^{(1)} \circ [b]^{(2)}) \end{aligned} \tag{1.69}$$

A bizonyítás azonnal adódik (1.65) alapján. (1.69) nem más mint a bennfoglalásra vett monotonitás (1.9) tulajdonsága gépi intervallum műveletekre. Az alábbi tulajdonságok a kerekítés hibabecslésénél válnak érdekessé.

**1.29. Tétel.** *Legyen* $\Diamond$ *az (1.66) alapján értelmezett,* $\bigtriangledown, \bigtriangleup$ *kerekítésekre támaszkodó intervallum kerekítés, és legyen* $\circ \in \{+, -, \cdot, :\}$. *Ekkor*

$$\begin{aligned} [a], [b] \in I\mathcal{R} &\Rightarrow [a] \circ [b] \subseteq [c] = \Diamond([a] \circ [b]) \in I\mathcal{R}, \\ a \in [a], b \in [b] &\Rightarrow a \circ b \in [c] = \Diamond([a] \circ [b]) \in I\mathcal{R}. \end{aligned} \tag{1.70}$$

*Ha az* $\bigcirc$ *kerekítésre áll*

$$\bigtriangledown x \leq \bigcirc(x) \leq \bigtriangleup x, \quad x \in \mathbb{R}, \tag{1.71}$$

*akkor* $x, y, z \in \mathcal{R}$ *esetén következik, hogy*

$$z = \bigcirc(x \circ y) \in [z] = \Diamond([x, x] \circ [y, y]) \in I\mathcal{R}.$$

Az (1.70) és (1.71) tulajdonságok elemi bizonyítása azonnal adódik a megfelelő definíciókból, így elhagyjuk. A fenti eredmények összefoglalását adjuk.

Egy függvényszabály 1.27. definícióra támaszkodó intervallum műveletek segítségével történő gépi intervallum kiértékelése bennfoglalja a függvényszabály intervallum kiértékelését. Ezek egyben tartalmazzák a függvény értékkészletére vonatkozó becsléseket is, továbbá kielégítik a bennfoglalásra vett monotonitás tulajdonságát is.

A gépi intervallum műveletek praktikus megvalósítása a megfelelő gépi műveletek segítségével történik. Ezek a műveletek vagy egy magasabb szintű programozási nyelv részei, vagy megvalósíthatók például ALGOL nyelven írt szubrutinokkal. Tekintsük át az utóbbi esetet röviden. Szubrutinok egy ilyen halmaza gyakran rendelkezik egy $\bigtriangledown$ lefele irányított kerekítést generáló művelettel. Ez például a LOW eljárással megvalósítható. Ezt az eljárást használva az ADD, SUB, MUL, DIV műveleteket definiáljuk a standard intervallum aritmetikai műveletek ábrázolására. Az 1.3. definíció unáris műveletei, az úgynevezett elemi függvények hasonló módon értelmezhetők.

Most a valós számok halmazán működő algoritmusokat tekintjük. Például a Horner elrendezést, Gauss algoritmust. Amennyiben ezeket az algoritmusokat gépi aritmetika segítségével számítógépeken futtatjuk, általában még a bemenő adatot sem tudjuk pontosan ábrázolni. Ez a probléma orvosolható gépi intervallum aritmetika használatával. A bemenő adat egyszerűen egy – gépi számokkal, mint korlátokkal megadott – intervallumba esik. Ha az algoritmust a kerekítési hibák figyelmen kívül hagyásával futtatjuk, akkor az eredmény, általában, továbbra is az eredeti adattal nem összekapcsolható mértékű kiszélesedéssel jár, mint azt az 1.3. fejezetben láttuk. Ezt a jelenséget vesszük nagyító alá, amikor a kerekítési hibákat is figyelembe vesszük. Ezért megvizsgáljuk, hogy

mekkora pontosság növekedést érhetünk el, amennyiben $t_1$ jegyű után $t_2 > t_1$ jegyű mantisszával rendelkező gépi intervallum aritmetikával futtatjuk algoritmusaink. Feltesszük, hogy eközben a karakterisztika nem változik. Ekkor minden $t_1$ jegyű szám egyben $t_2$ jeggyel is ábrázolható.

Legyen $x \in \mathbb{R}$, $x \neq 0$, és

$$x = \left( \sum_{i=1}^{\infty} a_i b^{-i} \right) b^e, \quad 1 \leq a_1 \leq b-1, \quad 0 \leq a_i \leq b-1, \quad i \geq 2.$$

Az egyértelműség garantálásához feltesszük, hogy $a_i \neq b-1$, $i \leq i_0$, egy rögzített $i_0$ esetén, továbbá $x$ nem pontosan ábrázolható $t_1$ jegyű mantisszából álló lebegő pontos rendszerben.(Ha az lenne, a következő meggondolás biztosan túlcsordulna.) Feltesszük még, hogy az (1.66) intervallum kerekítést a korlátok optimális kerekítésével hajtjuk végre. Az $x > 0$ esetben, (1.66) figyelembe vételével, kapjuk, hogy

$$\Diamond x = \Diamond[x, x] = [\bigtriangledown x, \bigtriangleup x],$$

ahol

$$\bigtriangledown x = \left( \sum_{i=1}^{t_1} a_i b^{-i} \right) b^e, \quad \bigtriangleup x = \left( \sum_{i=1}^{t_1} a_i b^{-i} \right) b^e + b^{-t_1+e}.$$

Világos, hogy $\Diamond x$ átmérője

$$d(\Diamond x) = b^{-t_1+e}.$$

Ez az eredmény adódik $x < 0$ esetben is.

Annak érdekében, hogy észrevegyük az eredmény mantissza hossztól való függését, a továbbiakban $\bigcirc_1(x)$ és $\bigcirc_2(x)$ jelölést használjuk. $\bigcirc$ egy valós szám (később valós intervallum) intervallum kerekítését jelöli. A fenti reláció ezzel a következő alakot ölti

$$d(\bigcirc_1(x)) = b^{-t_1+e}.$$

Analóg módon

$$d(\bigcirc_2(x)) \leq b^{-t_1+e-l}$$

adódik $t_2 = t_1 + l$ jegyű mantisszára. A szigorú egyenlőtlenség abban az esetben áll, ha $x$ pontosan ábrázolható $t_2$ jegyű mantisszával. Az előzőekből adódik, hogy

$$d(\bigcirc_2(x)) \leq b^{-l} d(\bigcirc_1(x)). \tag{1.72}$$

Az intervallum kerekítésre tett megszorításokból adódik az $[a], [b]$ gépi intervallumokra, hogy

$$\Diamond([a] \circ [b]) = \bigcirc_1([a] \circ [b]) = [(1-\varepsilon_1)(\underline{[a] \circ [b]}), (1+\varepsilon_2)(\overline{[a] \circ [b]})].$$

Itt $([a] \circ [b])_1, ([a] \circ [b])_2$ a pontos eredmény korlátjait számolja, így

$$-\varepsilon_1(\underline{[a] \circ [b]}) \leq 0, \quad \varepsilon_2(\overline{[a] \circ [b]}) \geq 0,$$

szintúgy, mint

$$|\varepsilon_1|, |\varepsilon_2| \leq b^{1-t_1}.$$

Írható, hogy

$$\bigcirc_1([a] \circ [b]) = [a] \circ [b] + [\varepsilon_1(\underline{[a] \circ [b]}), \varepsilon_2(\overline{[a] \circ [b]})]. \tag{1.73}$$

Az eredmény átmérőjére pedig

$$d(\bigcirc_1([a] \circ [b])) \leq d([a] \circ [b]) + 2b^{1-t_1}|[a] \circ [b]|. \tag{1.74}$$

Ez a közelítés mutatja, hogy a pontos intervallum eredmény abszolútértéke felelős a $d([a] \circ [b])$ intervallum átmérő növekedéséért fix mantissza hossz mellett.

Legyen $\widetilde{x} \in [x] \in \mathbb{IR}$. Ekkor javasolt egy $x \in [x]$ elemet választani $\widetilde{x}$ közelítésére. Az abszolút hiba

$$|x - \widetilde{x}| \leq d([x]) =: \delta([x]), \tag{1.75}$$

és, ha $0 \notin [x]$, $\widetilde{x} \neq 0$, a relatív hiba

$$\left|\frac{x-\widetilde{x}}{\widetilde{x}}\right| \leq \frac{d([x])}{\min\{|x| \,\big|\, x \in [x]\}} =: \rho([x]). \tag{1.76}$$

**1.30. Tétel.** *Legyenek $[a], [b], [a'], [b']$ valós gépi intervallumok, amelyekre*

$$[a] \subseteq [a'], \quad [b] \subseteq [b'] \tag{1.77}$$

$$\begin{gathered} d([a']) \leq s_1, \quad d([b']) \leq s_2 \\ d([a]) \leq b^{-1}s_1, \quad d([b]) \leq b^{-1}s_2. \end{gathered} \tag{1.78}$$

*Jelölje $\circ$ a valós intervallum műveletek valamelyikét. Ekkor egy $\Delta(\bigcirc_1([a'] \circ [b'])), \rho(\bigcirc_1([a'] \circ [b'])$ korlátjainál $b^{-1}$ faktorral kisebb korlátokat kapunk $\Delta(\bigcirc_2([a] \circ [b])), \rho(\bigcirc_2([a] \circ [b]))$ kifejezésekre, ha $0 \notin \bigcirc_1([a'] \circ [b'])$.*

**Bizonyítás:** Felhasználva (1.74),(1.20),(1.22) relációkat,

$$d(1/[x]) \leq |1/[x]|^2 d([x]) \quad (0 \notin [x])$$

és (1.78) első sorát, a következő egyenlőtlenségre jutunk

$$\begin{aligned} &d(\bigcirc_1([a'] \circ [b'])) \leq d([a'] \circ [b']) + 2b^{1-t_1}|[a'] \circ [b']| \leq \\ &\leq \left\{ \begin{array}{ll} s_1 + s_2, & \circ = +, - \\ |[a']|s_2 + s_1|[b']|, & \circ = \cdot \\ |[a']||1/[b']|^2 s_2 + |1/[b']|s_1, & \circ = : \end{array} \right\} + 2b^{1-t_1}|[a'] \circ [b']|. \end{aligned}$$

Felhasználva (1.77), (1.78) állításokat analóg módon igazolható, hogy

$$d([a] \circ [b]) \leq b^{-1} \left\{ \begin{array}{ll} s_1 + s_2, & \circ = +, - \\ |[a']|s_2 + s_1|[b']|, & \circ = \cdot \\ |[a']||1/[b']|^2 s_2 + |1/[b']|s_1, & \circ = : \end{array} \right\}. \quad (1.79)$$

(1.77) miatt, az 1.28. tételből a bennfoglalásra

$$\bigcirc_2([a] \circ [b]) \subseteq \bigcirc_2([a'] \circ [b']) \subseteq \bigcirc_1([a'] \circ [b']),$$

mivel feltettük, hogy a korlátok optimális kerekítésével számoljuk az intervallum kerekítést. Ezért adódik, hogy

$$\min\{|x| \,\big|\, x \in \bigcirc_2([a] \circ [b])\} \geq \min\{|x| \,\big|\, x \in \bigcirc_1([a'] \circ [b'])\}. \quad (1.80)$$

Végül (1.74), (1.79) és $|[a] \circ [b]| \leq |[a'] \circ [b']|$ miatt következik, hogy

$$\begin{aligned} &d(\bigcirc_2([a] \circ [b])) \leq d([a] \circ [b]) + 2b^{1-t_1-l}|[a'] \circ [b']| \leq \\ &\leq b^{-1} \left\{ \begin{array}{ll} s_1 + s_2, & \circ = +, - \\ |[a']|s_2 + s_1|[b']|, & \circ = \cdot \\ |[a']||1/[b']|^2 s_2 + |1/[b']|s_1, & \circ = : \end{array} \right\} + 2b^{1-t_1-l}|[a'] \circ [b']|. \end{aligned}$$

Ezzel az abszolút hiba felső korlátjára vonatkozó állítást beláttuk. (1.80) miatt azonnal kapjuk a relatív hiba felső korlátjára az eredményt. □

Egy elemi, de annál fontosabb következménye ennek a tételnek az alábbi

**1.31. Tétel.** *Az előbbi, a gépi intervallum aritmetikára vonatkozó feltételezésekkel itt is élünk. Most a valós számokra készített algoritmusok számítógépen való futtatásához gépi intervallum aritmetikát használunk $t_1$ jegyű mantisszával. Ha ezután $t_2 = t_1 + \ell$ jegyű ($\ell \geq 0$) mantisszájú gépi intervallum aritmetikával futtatjuk az algoritmust, akkor mind az abszolút, mind a relatív hibakorlátokat redukáljuk egy $b^{-\ell}$ faktorral. (Egy algoritmus itt egy egyértelműen meghatározott aritmetikai műveletsorozatot jelent adott bemenő adatokkal.)*

**Bizonyítás:** (1.72) alapján a bemenő adat intervallumkerekítése kielégíti az 1.30. tétel (1.78) feltételezését. Az intervallum aritmetika tulajdonságai megerősítik (1.77) állítást. A bizonyítás ezek után adódik az 1.30. tételből teljes indukcióval. □

Az 1.31. tétel alapján utalást kapunk arra, hogyan számoljuk a kimenetet előre adott abszolút, illetve relatív pontossággal. Legyen például $d_1$ a keletkező maximális intervallumhossz $t_1$ jegyű mantisszával számolva, és legyen az elvárt pontosság $\varepsilon$. Ha $d_1 \leq \varepsilon$, akkor végeztünk. Máskülönben $l$ jeggyel növeljük a mantissza jegyeinek számát úgy, hogy

$$b^{-l} d_1 \leq \varepsilon.$$

(Ezzel a választással az abszolút hiba $b^{-l}$ faktorral való redukciója nem biztosított. Az 1.31. tételnek megfelelően ez csak az abszolút hiba felső korlátjára igaz.)

| | mantissza jegyeinek száma | | | |
|---|---|---|---|---|
| i | 15 | 20 | 25 | 30 |
| 1 | $> 1^{a}$ | $0.11 \times 10^{-3}$ | $0.11 \times 10^{-8}$ | $0.11 \times 10^{-13}$ |
| 2 | $0.34 \times 10^{0}$ | $0.29 \times 10^{-5}$ | $0.29 \times 10^{-10}$ | $0.29 \times 10^{-15}$ |
| 3 | $0.18 \times 10^{-1}$ | $0.17 \times 10^{-6}$ | $0.17 \times 10^{-11}$ | $0.17 \times 10^{-16}$ |
| 4 | $0.16 \times 10^{-2}$ | $0.16 \times 10^{-7}$ | $0.16 \times 10^{-12}$ | $0.16 \times 10^{-17}$ |
| 5 | $0.26 \times 10^{-3}$ | $0.25 \times 10^{-8}$ | $0.25 \times 10^{-13}$ | $0.25 \times 10^{-18}$ |
| 6 | $0.64 \times 10^{-4}$ | $0.64 \times 10^{-9}$ | $0.64 \times 10^{-14}$ | $0.64 \times 10^{-19}$ |
| 7 | $0.58 \times 10^{-4}$ | $0.58 \times 10^{-9}$ | $0.58 \times 10^{-14}$ | $0.58 \times 10^{-19}$ |

1.1. táblázat. A Gauss algoritmus relatív hibájának $\rho([x]_i)$ felső korlátja

Az 1.31. tételben tárgyalt és bizonyított tényekre konkrét példaként egy egyenletrendszert választottunk, amit egy $7 \times 7$ Hilbert mátrix, jobb

oldalon pedig $(1,\ldots,1)^T$ határoz meg. A Gauss algoritmusnál gépi intervallum aritmetikát használtunk 15, 20, 25, 30, 35 decimális jeggyel a mantisszában. Az eredményeket az 1.4. táblázat tartalmazza, ahol csak a relatív hiba $\rho([x]_i)$ felső korlátját adtuk meg a megoldásvektor komponenseire.

Tekintsük a következő problémát: legyenek adva gépi intervallumok (olyan valós intervallumok, amelyek végpontjai gépi számok), mondjuk

$$[c]_0, [a]_0, [b]_0, [d]_0, [a]_1, [b]_1, [d]_1, \ldots, [a]_{n-1}, [b]_{n-1}, [d]_{n-1}$$

és egy $a_0$ gépi szám. Az

$$[r]_n = \frac{1}{a_n}\{[c]_0 - [a]_0([b]_0 - [d]_0) - [a]_1([b]_1 - [d]_1) - \ldots - [a]_{n-1}([b]_{n-1} - [d]_{n-1})\}$$

kifejezést szeretnénk kiszámolni.

Elméletileg használhatjuk a következő algoritmust:

$$\begin{aligned} &[s]_0 := [c]_0 \\ &[s]_i := [s]_{i-1} - [a]_{i-1}([b]_{i-1} - [d]_{i-1}), \quad 1 \le i \le n, \\ &[r]_n := [s]_n / a_n. \end{aligned} \tag{s}$$

Gyakorlatban azonban a következő műveleteket végezzük el:

$$\begin{aligned} &\widehat{[s]}_0 := [s]_0 := [c]_0 \\ &\widehat{[s]}_i := \bigcirc(\widehat{[s]}_{i-1} - \bigcirc([a]_{i-1}(\bigcirc([b]_{i-1} - [d]_{i-1})))), \quad 1 \le i \le n, \\ &\widehat{[r]}_n := \bigcirc([s]_n / a_n). \end{aligned} \tag{$\widehat{[s]}$}$$

Kezdjük (1.73) egyenlettel, ahol rögzítjük $\varepsilon := \frac{1}{2}b^{1-t}$ értékét, majd általános $[a], [b]$ intervallumokra kapjuk, hogy

$$\bigcirc([a] \circ [b]) \subseteq [a] \circ [b] + [-\varepsilon, \varepsilon]([a] \circ [b]), \tag{1.81}$$

ahol $\max\{|\varepsilon_1|, |\varepsilon_2|\} \le 2\varepsilon$ igaz.

Tegyük fel egy pillanatra, hogy már kiszámoltuk az

$$\widehat{[s]}_0 = [s]_0 = [c]_0, \widehat{[s]}_1, \ldots, \widehat{[s]}_{n-1}$$

[a] $\rho([x]_1) > 1$ jelentése, hogy $0 \in [x]_1$.

értékeket. Ekkor (1.81) miatt

$$\begin{aligned}
&\bigcirc([b]_{n-1}-[d]_{n-1}) \subseteq [b]_{n-1}-[d]_{n-1}+|[b]_{n-1}-[d]_{n-1}|[-\varepsilon,\varepsilon],\\
&\bigcirc([a]_{n-1}\bigcirc([b]_{n-1}-[d]_{n-1})) \subseteq\\
&\subseteq [a]_{n-1}([b]_{n-1}-[d]_{n-1}+|[b]_{n-1}-[d]_{n-1}|[-\varepsilon,\varepsilon])+\\
&+|[a]_{n-1}([b]_{n-1}-[d]_{n-1}+|[b]_{n-1}-[d]_{n-1}|[-\varepsilon,\varepsilon])|[-\varepsilon,\varepsilon] \subseteq\\
&\subseteq [a]_{n-1}([b]_{n-1}-[d]_{n-1})+|[a]_{n-1}||[b]_{n-1}-[d]_{n-1}|[-2\varepsilon-\varepsilon^2,+2\varepsilon+\varepsilon^2],
\end{aligned}$$

ezért

$$\begin{aligned}
\widehat{[s]}_n \subseteq &\widehat{[s]}_{n-1}-[a]_{n-1}([b]_{n-1}-[d]_{n-1})-\\
&-|[a]_{n-1}||[b]_{n-1}-[d]_{n-1}|[-2\varepsilon-\varepsilon^2,+2\varepsilon+\varepsilon^2]+\\
&+|\widehat{[s]}_{n-1}-[a]_{n-1}([b]_{n-1}-[d]_{n-1})-\\
&-|[a]_{n-1}||[b]_{n-1}-[d]_{n-1}|[-2\varepsilon-\varepsilon^2,+2\varepsilon+\varepsilon^2]|[-\varepsilon,\varepsilon] \subseteq\\
\subseteq &\widehat{[s]}_{n-1}-[a]_{n-1}([b]_{n-1}-[d]_{n-1})+|\widehat{[s]}_{n-1}|[-\varepsilon,\varepsilon]+\\
&+|[a]_{n-1}||[b]_{n-1}-[d]_{n-1}|[-3\varepsilon-3\varepsilon^2-\varepsilon^3,3\varepsilon+3\varepsilon^2+\varepsilon^3].
\end{aligned} \tag{1.82}$$

Teljes indukcióval belátjuk, hogy igaz

$$\begin{aligned}
\widehat{[s]}_n \subseteq&\\
\subseteq\quad &[s]_n+[-\varepsilon,\varepsilon]\sum_{i=0}^{n-1}|\widehat{[s]}_i|+\\
&+[-3\varepsilon-3\varepsilon^2-\varepsilon^3,3\varepsilon+3\varepsilon^2+\varepsilon^3]\sum_{i=0}^{n-1}|[a]_i||[b]_i-[d]_i|.
\end{aligned} \tag{1.83}$$

$n=1$ esetén $\widehat{[s]}_0=[s]_0=[c]_0$ felhasználásával (1.82) alapján

$$\begin{aligned}
\widehat{[s]}_1 \subseteq&\widehat{[s]}_0-[a]_0([b]_0-[d]_0)+|\widehat{[s]}_0|[-\varepsilon,\varepsilon]\\
&+|[a]_0||[b]_0-[d]_0|[-3\varepsilon-3\varepsilon^2-\varepsilon^3,3\varepsilon+3\varepsilon^2+\varepsilon^3]\\
=&[s]_1+[-\varepsilon,\varepsilon]|\widehat{[s]}_0|+[-3\varepsilon-3\varepsilon^2-\varepsilon^3,3\varepsilon+3\varepsilon^2+\varepsilon^3]|[a]_0||[b]_0-[d]_0|,
\end{aligned}$$

így az állítás igaz $n = 1$ esetén. Ha (1.83) igaz valamely $n \geq 1$ esetre, akkor $n$ helyett $(n+1)$-et helyettesítve (1.82) kifejezésbe és felhasználva (s) összefüggést, adódik, hogy

$$\begin{aligned}\widehat{[s]}_{n+1} \subseteq & \widehat{[s]}_n - [a]_n([b]_n - [d]_n) + |\widehat{[s]}_n|[-\varepsilon,\varepsilon] + \\ & + |[a]_n||[b]_n - [d]_n|[-3\varepsilon - 3\varepsilon^2 - \varepsilon^3, 3\varepsilon + 3\varepsilon^2 + \varepsilon^3] \subseteq \\ \subseteq & [s]_{n+1} + [-\varepsilon,\varepsilon]\sum_{i=0}^{n} |\widehat{[s]}_i| + \\ & + [-3\varepsilon - 3\varepsilon^2 - \varepsilon^3, 3\varepsilon + 3\varepsilon^2 + \varepsilon^3]\sum_{i=0}^{n} |[a]_i||[b]_i - [d]_i|,\end{aligned}$$

ami éppen (1.83) a változócserével. Alkalmazva még egyszer (1.81) kifejezést, adódik az eredmény

$$\widehat{[r]}_n \subseteq \widehat{[s]}_n/a_n + (|\widehat{[s]}_n|/|a_n|)[-\varepsilon,\varepsilon]. \tag{1.84}$$

Ez azt mutatja, hogy a gépi intervallum aritmetikával kiszámolt formula relatív hibája $[-\varepsilon,\varepsilon]$ intervallum, vagyis a formulát stabilan számoltuk ki.

# 2. fejezet

# Komplex intervallum aritmetika

Ebben a fejezetben szeretnénk definiálni és használni egy úgynevezett komplex intervallum aritmetikát. Megmutatjuk, hogy a valós esetnél tárgyalt legtöbb tulajdonság átvihető a komplex esetre is. Ennek érdekében definiálnunk kell a komplex számok egy olyan halmazát, amely éppen a komplex intervallumot alkotja. Két ésszerű választást tekintünk az alábbiakban:

## 2.1. Téglalapok, mint komplex intervallumok

**2.1. Definíció.** *Legyen* $[a_{re}], [a_{im}] \in \mathbb{IR}$. *Ekkor*

$$[a] = \left\{ a = a_{re} + ia_{im} \;\middle|\; a_{re} \in [a_{re}], a_{im} \in [a_{im}] \right\}$$

*komplex számhalmazt komplex intervallumnak nevezzük.*

A 2.1. definícióban értelmezett komplex számhalmaz a koordinátatengelyekkel párhuzamos oldalú téglalapnak felel meg a komplex síkon, jele $R\mathbb{C}$. Az $R\mathbb{C}$ halmaz elemeit $[a], [b], [c], \ldots, [x], [y], [z] \in R\mathbb{C}$ jelöli, így $[a] = [a_{re}] + i[a_{im}]$ írható, ahol $[a_{re}], [a_{im}] \in \mathbb{IR}$. Egy $a = a_{re} + ia_{im}$ komplex szám ekkor

$$[a] = [a_{re}, a_{re}] + i[a_{im}, a_{im}] \in R\mathbb{C}$$

komplex pont intervallumnak is tekinthető. Minden $[a] \in \mathbb{IR}$ elem $[a] = [a_{re}]+i[0,0] \in R\mathbb{C}$ elemnek is gondolható, amiből világos, hogy $\mathbb{IR} \subset R\mathbb{C}$.

**2.2. Definíció.** *Legyen* $[a] = [a_{re}] + i[a_{im}], [b] = [b_{re}] + i[b_{im}] \in R\mathbb{C}$. *Ekkor* $[a] = [b]$ *pontosan akkor, ha*

$$[a_{re}] = [b_{re}] \qquad \text{és} \qquad [a_{im}] = [b_{im}].$$

Az előbb definiált = reláció reflexív, szimmetrikus, tranzitív.

Általánosítsuk a komplex aritmetikát $R\mathbb{C}$-beli komplex intervallum aritmetikára.

**2.3. Definíció.** *Legyen* $\circ \in \{+, -, \cdot, :\}$ *bináris művelet* $\mathbb{IR}$ *elemein. Ekkor* $[a] = [a_{re}] + i[a_{im}], [b] = [b_{re}] + i[b_{im}] \in R\mathbb{C}$ *mellett*

$$[a] \pm [b] = [a_{re}] \pm [b_{re}] + i([a_{im}] \pm [b_{im}]), \tag{2.1}$$

$$[a] \cdot [b] = [a_{re}][b_{re}] - [a_{im}][b_{im}] + i([a_{re}][b_{im}] + [a_{im}][b_{re}]), \tag{2.2}$$

$$\begin{aligned}[a] : [b] = {} & ([a_{re}][b_{re}] + [a_{im}][b_{im}]) : ([b_{re}]^2 + [b_{im}]^2) + \\ & + i([a_{im}][b_{re}] - [a_{re}][b_{im}]) : ([b_{re}]^2 + [b_{im}]^2).\end{aligned} \tag{2.3}$$

Természetesen most is feltesszük, hogy $0 \notin ([b_{re}]^2 + [b_{im}]^2)$ osztáskor. Azonban most $0 \notin [b_{re}]+i[b_{im}]$ nem elegendő feltétel, ahogy azt az alábbi példával illusztráljuk is.

Legyen

$$[b] = [-1, 1] + i[1, 3].$$

Ekkor

$$0 \in [0, 10] = [-1, 1] + [1, 9] = [b_{re}][b_{re}] + [b_{im}][b_{im}].$$

Ha azonban a 2.3. definícióbeli osztásnál a $[b_{re}]^2 + [b_{im}]^2$ kifejezést

$$[b_{re}]^2 + [b_{im}]^2 = \left\{ b_{re}^2 \;\middle|\; b_{re} \in [b_{re}] \right\} + \left\{ b_{im}^2 \;\middle|\; b_{im} \in [b_{im}] \right\}$$

módon számoljuk, akkor a fenti példát ezúton számolva

$$[b_{re}]^2 + [b_{im}]^2 = [0, 1] + [1, 9] = [1, 10].$$

Vegyük közelebbről szemügyre a fent bevezetett komplex intervallum aritmetika tulajdonságait.

Nyilvánvaló, hogyha $[a], [b] \in R\mathbb{C}$, akkor

$$[a] \pm [b] = \left\{ a \pm b \;\middle|\; a \in [a], b \in [b] \right\}$$

igaz $R\mathbb{C}$ halmazon. Általánosságban ez nem igaz a szorzásra és osztásra, mint az alábbi példa mutatja.

Legyen

$$[a] = [2,4] + i[0,0], \qquad [b] = [1,1] + i[1,1].$$

A 2.3. definícióból

$$[a][b] = [2,4] + i[2,4].$$

Másfelől

$$\left\{ ab \;\middle|\; a \in [a], b \in [b] \right\} = \left\{ s(1+i) \;\middle|\; s \in \mathbb{R}, 2 \le s \le 4 \right\} \subset [a][b].$$

Az alábbi tétel azonban érvényes.

**2.4. Tétel.** *(Tartalmazási tétel) A 2.3. definíció műveleteire*

$$\left\{ a \circ b \;\middle|\; a \in [a], b \in [b] \right\} \subseteq [a] \circ [b].$$

*Az összeadás és a kivonás esetén egyenlőség is teljesül. A szorzásra*

$$[a][b] = \inf \left\{ [x] \in R\mathbb{C} \;\middle|\; \left\{ a \cdot b \;\middle|\; a \in [a], b \in [b] \right\} \subseteq [x] \right\},$$

*ahol az infimumot $R\mathbb{C}$ halmazon a halmazelméleti bennfoglalás által definiált részben rendezés szerint vesszük. Ez azt jelenti, hogy ez az a legszűkebb intervallum, ami tartalmazza az $[a]$ és $[b]$ intervallumok komplexusszorzatát.*

**Bizonyítás:** Az összeadás, kivonás esetét már feljebb tárgyaltuk. Legyen $a \in [a], b \in [b]$. A valós intervallumokra vonatkozó bennfoglalásra vett monotonitást felhasználva $a = a_{re} + ia_{im}, b = b_{re} + ib_{im}$ mellett kapjuk, hogy

$$\begin{aligned} ab &= a_{re}b_{re} - a_{im}b_{im} + i(a_{re}b_{im} + a_{im}b_{re}) \\ &\in [a_{re}][b_{re}] - [a_{im}][b_{im}] + i([a_{re}][b_{im}] + [a_{im}][b_{re}]) = [a][b]. \end{aligned}$$

Mivel $a_{re}b_{re}-a_{im}b_{im}$ kifejezésben minden változó pontosan egyszer fordul elő kapjuk, hogy

$$\left\{a_{re}b_{re} - a_{im}b_{im} \;\middle|\; a \in [a], b \in [b]\right\} = [a_{re}][b_{re}] - [a_{im}][b_{im}].$$

Ugyanezen alapon

$$\left\{a_{re}b_{im} - a_{im}b_{re} \;\middle|\; a \in [a], b \in [b]\right\} = [a_{re}][b_{im}] + [a_{im}][b_{re}].$$

Ez utóbbi kettőből látszik, hogy minden

$$\begin{aligned} &c_{re} = a_{re}b_{re} - a_{im}b_{im} \in [a_{re}][b_{re}] - [a_{im}][b_{im}], \\ &a_k \in [a]_k, \quad b_k \in [b]_k, \quad k = 1, 2, \end{aligned}$$

valós számhoz található olyan

$$\begin{aligned} &c_{im} = a_{im}b_{re} + a_{re}b_{im} \in [a_{im}][b_{re}] - [a_{re}][b_{im}], \\ &a_k \in [a]_k, \quad b_k \in [b]_k, \quad k = 1, 2, \end{aligned}$$

valós szám, hogy $c = c_{re} + ic_{im} \in [a][b]$, amit meg kellett mutatnunk. A tétel osztásra vonatkozó állítása következik a bennfoglalásra vett monotonitásból. □

A 2.4. tétel szorzásra adott eredménye általában nem igaz az osztásra.

## 2.2. Körlapok, mint komplex intervallumok

**2.5. Definíció.** *Legyen $a \in \mathbb{C}$, $r \geq 0$. Azt mondjuk, hogy*

$$[z] = \left\{z \in \mathbb{C} \;\middle|\; |z - a| \leq r\right\}$$

*egy körlap, körszerű intervallum, vagy egyszerűen egy komplex intervallum, ha nem keverhető a téglalap alakú komplex intervallumokkal.*

Ezen körlapok halmazát $K\mathbb{C}$ jelöli, elemeit $[a], [b], [c], \ldots, [x], [y], [z]$. Az $a$ középpontú $r$ sugarú körlapokat

$$[z] = \langle a, r\rangle$$

alakban is írjuk. A komplex számokat ekkor $K\mathbb{C}$ $\langle a, 0\rangle$ alakú elemeinek tekinthetjük, amiből világos, hogy $\mathbb{C} \subset K\mathbb{C}$.

**2.6. Definíció.** *Két körlap,* $[a] = \langle a, r_a \rangle$ *és* $[b] = \langle b, r_b \rangle$ *pontosan akkor egyenlő, ha halmazelméleti értelemben azok. Ekkor* $a = b$ *és* $r_a = r_b$*.*

Ez a reláció ismét ekvivalencia reláció.

$K\mathbb{C}$ halmazra a következő módon általánosítjuk a valós számokon szokásos műveleteket.

**2.7. Definíció.** *Legyen* $\circ \in \{+, -, \cdot, :\}$ *a komplex számokon értelmezett bináris művelet. Ekkor* $[a] = \langle a, r_a \rangle$ *és* $[b] = \langle b, r_b \rangle$ *mellett*

$$
\begin{aligned}
[a] \pm [b] &= \langle a \pm b, r_a \pm r_b \rangle \,, \\
[a] \cdot [b] &= \langle ab,\ |a| r_b + |b| r_a + r_a r_b \rangle \,, \\
\frac{1}{[b]} &= \left\langle \frac{\overline{b}}{b\overline{b} - r_b^2}, \frac{r_b}{b\overline{b} - r_b^2} \right\rangle && 0 \notin [b], \\
[a] : [b] &= [a] \cdot \frac{1}{[b]} && 0 \notin [b].
\end{aligned}
$$

*Itt* $|a| = \sqrt{a_1^2 + a_2^2}$ *az a komplex szám euklideszi normáját,* $\overline{b} = b_1 - ib_2$ *pedig a b komplex szám konjugáltját jelöli.*

Körlapok összeadására és szorzására világos, hogy teljesül

$$[a] \pm [b] = \left\{ a \pm b \;\middle|\; a \in [a], b \in [b] \right\}.$$

Ez áll a körlap inverzére is, ugyanis ha alkalmazzuk a konform leképezések elméletét a nullát nem tartalmazó körlapok leképezésére, akkor a $w = 1/z$ leképezéssel újabb körlapot kapunk:

$$1/[b] = \left\{ 1/b \;\middle|\; b \in [b] \right\}.$$

Elemi számolással ellenőrizhető, hogy a 2.7. definíció $1/[b]$ kifejezésre vonatkozó képlete helyes.

A 2.7. definícióbeli szorzásra (és így az osztásra is) általában csak

$$\left\{ z_1 z_2 \;\middle|\; z_1 \in [a], z_2 \in [b] \right\} \subseteq [a][b]$$

igaz. Ez az alábbi egyenlőtlenségekből következik

$$\begin{aligned}|z_1z_2 - ab| &= |a(z_2-b) + b(z_1-a) + (z_1-a)(z_2-b)| \\ &\le |a||z_2-b| + |b||z_1-a| + |b||z_1-a||z_2-b| \\ &\le |a|r_b + |b|r_a + r_ar_b.\end{aligned}$$

Az 1.4. tételnek megfelelően összegyűjtjük a az $R\mathbb{C}$-beli műveleti tulajdonságokat most $K\mathbb{C}$ halmazra való tekintettel. Hacsak másképp nem mondjuk, $\mathrm{I}\mathbb{C}$ legyen $R\mathbb{C}$ a 2.3. vagy $K\mathbb{C}$ a 2.7. definícióbeli műveletekkel.

**2.8. Tétel.** *Legyen* $[a],[b],[c] \in \mathrm{I}\mathbb{C}$ *és* $[d],[e],[f] \in K\mathbb{C}$. *Ekkor*

$$[a]+[b] = [b]+[a], \quad [a][b] = [b][a] \qquad \textit{(kommutativitás)}, \tag{2.4}$$

$$\begin{aligned}&([a]+[b])+[c] = [a]+([b]+[c]), \\ &([d][e])[f] = [d]([e][f]), \quad \textit{(asszociativitás)},\end{aligned} \tag{2.5}$$

$$[0,1] + i[0,0] \in R\mathbb{C}, \quad \textit{illetve} \quad \langle 0,0\rangle \in K\mathbb{C}, \tag{2.6}$$

*és*

$$[1,1] + i[0,0] \in R\mathbb{C}, \quad \textit{illetve} \quad \langle 1,0\rangle \in K\mathbb{C},$$

*az egyértelműen meghatározott additív illetve multiplikatív neutrális elemek.*

$$\mathrm{I}\mathbb{C} \textit{ nullosztómentes.} \tag{2.7}$$

*Egy* $[z] \in \mathrm{I}\mathbb{C}$ *elemnek pontosan akkor létezik additív és multiplikatív inverze, ha* $[z] \in \mathbb{C}$ *és szorzás esetén* $[z] \neq 0$. *Mindenesetre igaz, hogy* $0 \in [a]-[a]$ *és* $1 \in [a]:[a]$.

(2.8)

$$[a]([b]+[c]) \subseteq [a][b]+[a][c] \quad \textit{(szubdisztributivitás)}, \tag{2.9}$$

$$a([b]+[c]) = a[b]+a[c] \quad a \in \mathbb{C}.$$

**Bizonyítás:** A bizonyítások következnek a 2.3. és a 2.7. definíciókból. Példaként bemutatjuk (2.9) bizonyítását $K\mathbb{C}$ esetre. Ha $[a] = \langle a, r_a\rangle, [b] = \langle b, r_b\rangle, [c] = \langle c, r_c\rangle \in K\mathbb{C}$, akkor

$$\begin{aligned}[a]([b]+[c]) &= \langle a, r_a\rangle \langle b+c, r_b+r_c\rangle \\ &= \langle a(b+c), |a|(r_b+r_c) + |b+c|r_a + r_a(r_b+r_c)\rangle \\ &\subseteq \langle ab+ac, |a|r_b + |a|r_c + |b|r_a + |c|r_a + r_ar_b + r_ar_c\rangle \\ &= \langle ab, |a|r_b + |b|r_a + r_ar_b\rangle + \langle ac, |a|r_c + |c|r_a + r_ar_c\rangle \\ &= [a][b]+[a][c].\end{aligned}$$

Az $[a] = \langle a, 0 \rangle$, azaz $r_a = 0$ esetben a bizonyításból látszik, hogy

$$a([b] + [c]) = a[b] + a[c].$$

□

Lényeges kiemelni, hogy a (2.5) asszociatív törvény általában nem teljesül $R\mathbb{C}$ elemeire. Például

$$\begin{gathered}[a] = [2,4] + i[0,0], \quad [b] = [1,1] + i[1,1], \quad [c] = [1,1] + i[1,1], \\ ([a][b])[c] = ([2,4] + i[2,4])([1,1] + i[1,1]) = [-2,2] + i[4,8], \\ [a]([b][c]) = ([2,4] + i[0,0])([0,0] + i[2,2]) = [0,0] + i[4,8].\end{gathered}$$

A bennfoglalásra vett monotonitás igaz $\mathrm{I}\mathbb{C}$ halmazon is.

**2.9. Tétel.** *Legyen* $[a]^{(k)}, [b]^{(k)} \in \mathrm{I}\mathbb{C}$, $k = 1, 2$ *úgy, hogy*

$$[a]^{(k)} \subseteq [b]^{(k)}, \quad k = 1, 2.$$

*Ekkor*

$$[a]^{(1)} \circ [a]^{(2)} \subseteq [b]^{(1)} \circ [b]^{(2)}$$

*teljesül* $\circ \in \{+, -, \cdot, :\}$ *műveletekre.*

**Bizonyítás:** Az állítás igaz $R\mathbb{C}$ esetén, mivel a bennfoglalásra vett monotonitás teljesül $\mathbb{IR}$ elemeire (lásd az 1.5. tételt). $K\mathbb{C}$-beli összeadás és kivonás esetén

$$\begin{aligned}[a]^{(1)} \pm [a]^{(2)} &= \left\{ z = x \pm y \;\middle|\; x \in [a]^{(1)}, y \in [a]^{(2)} \right\} \\ &\subseteq \left\{ w = u \pm v \;\middle|\; u \in [b]^{(1)}, v \in [b]^{(2)} \right\} = [b]^{(1)} \pm [b]^{(2)}.\end{aligned}$$

Tekintsük a szorzást $K\mathbb{C}$ esetén és legyen

$$[a]^{(k)} = \left\langle a^{(k)}, r^{(k)} \right\rangle, \quad [b]^{(k)} = \left\langle b^{(k)}, s^{(k)} \right\rangle, \quad k = 1, 2.$$

Ekkor az $[a]^{(k)} \subseteq [b]^{(k)}, \quad k = 1, 2$ ekvivalens azzal, hogy

$$|a^{(k)} - b^{(k)}| \leq s^{(k)} - r^{(k)}, \quad k = 1, 2,$$

továbbá

$$[a]^{(1)}[a]^{(2)} = \left\langle a^{(1)}a^{(2)}, |a^{(1)}|r^{(2)} + |a^{(2)}|r^{(1)} + r^{(1)}r^{(2)} \right\rangle,$$
$$[b]^{(1)}[b]^{(2)} = \left\langle b^{(1)}b^{(2)}, |b^{(1)}|s^{(2)} + |b^{(2)}|s^{(1)} + s^{(1)}s^{(2)} \right\rangle.$$

Bizonyítandó, hogy

$$|a^{(1)}a^{(2)} - b^{(1)}b^{(2)}| \leq$$
$$\leq |b^{(1)}|s^{(2)} + |b^{(2)}|s^{(1)} + s^{(1)}s^{(2)} - |a^{(1)}|r^{(2)} + |a^{(2)}|r^{(1)} + r^{(1)}r^{(2)}.$$

A háromszög egyenlőtlenségből kapjuk, hogy

$$-|b^{(k)}| \leq -|a^{(k)}| + |a^{(k)} - b^{(k)}|, \quad k = 1, 2$$

és mivel

$$|a^{(k)} - b^{(k)}| \leq s^{(k)} - r^{(k)}, \quad k = 1, 2,$$

kapjuk, hogy

$$\begin{aligned} -|b^{(2)}|r^{(1)} &\leq -|a^{(2)}|r^{(1)} + r^{(1)}(s^{(2)} - r^{(2)}) = \\ &= -|a^{(2)}|r^{(1)} + r^{(1)}s^{(2)} - r^{(1)}r^{(2)}, \\ -|b^{(1)}|r^{(2)} &\leq -|a^{(1)}|r^{(2)} + r^{(2)}(s^{(1)} - r^{(1)}) = \\ &= -|a^{(1)}|r^{(2)} + r^{(2)}s^{(1)} - r^{(2)}r^{(1)}. \end{aligned}$$

Ebből adódik, hogy

$$\begin{aligned} &|a^{(1)}a^{(2)} - b^{(1)}b^{(2)}| \leq \\ &\leq |b^{(2)}||a^{(1)} - b^{(1)}| + |b^{(1)}||a^{(2)} - b^{(2)}| + |a^{(1)} - b^{(1)}||a^{(2)} - b^{(2)}| \leq \\ &\leq |b^{(2)}|(s^{(1)} - r^{(1)}) + |b^{(1)}|(s^{(2)} - r^{(2)}) + (s^{(1)} - r^{(1)})(s^{(2)} - r^{(2)}) \leq \\ &\leq |b^{(2)}|s^{(1)} + |b^{(1)}|s^{(2)} + s^{(1)}s^{(2)} - (|a^{(2)}|r^{(1)} + |a^{(1)}|r^{(2)} + r^{(1)}r^{(2)}), \end{aligned}$$

ami a szorzásra vonatkozó állítást bizonyítja.

$$\begin{aligned} 1/([a]^{(2)}) &= \left\{ z = 1/x \;\middle|\; x \in [a]^{(2)} \right\} \\ &\subseteq \left\{ w = 1/u \;\middle|\; u \in [b]^{(2)} \right\} = 1/([b]^{(2)}) \end{aligned}$$

miatt igaz, hogy

$$[a]^{(1)} : [a]^{(2)} = [a]^{(1)} \cdot \frac{1}{[a]^{(2)}} \subseteq [b]^{(1)} \cdot \frac{1}{[b]^{(2)}} = [b]^{(1)} : [b]^{(2)}.$$

□

A 2.9. tétel speciális eseteként adódik az alábbi

**2.10. Következmény.** *Legyen* $[a], [b] \in \mathrm{I}\mathbb{C}$ *és* $a \in [a], b \in [b]$. *Ekkor*

$$a \circ b \in [a] \circ [b].$$

**2.11. Megjegyzés.** *Az* $R\mathbb{C}$*-beli aritmetika gépi megvalósítása nem okoz problémát, mivel azt* $\mathbb{IR}$*-beli műveletekkel definiáltuk, amire már bemutattunk egy - a legfontosabb aritmetikai tulajdonságokat megőrző - lehetséges gépi megvalósítást az 1.4. fejezetben. Eszerint* $\mathbb{IR}$*-beli becsléseink* $R\mathbb{C}$*-re is átvihetők.*

## 2.3. Metrika, abszolútérték és szélesség $\mathrm{I}\mathbb{C}$-ben

Ebben a fejezetben $q$ az 1.7 definícióban bevezetett $\mathbb{IR}$-beli metrikát jelöli. Az alábbiakban egy metrikát definiálunk $R\mathbb{C}$-n.

**2.12. Definíció.** *Legyen* $[a] = [a_{re}] + i[a_{im}],\ [b] = [b_{re}] + i[b_{im}] \in R\mathbb{C}$. *Ekkor az* $[a]$ *és* $[b]$ *elemek távolsága definíció szerint legyen:*

$$p([a], [b]) = q([a_{re}], [b_{re}]) + q([a_{im}], [b_{im}])$$

Leszűkítve $p$-t $\mathbb{IR}$-re ugyanazt az eredményt kapjuk, mint a az 1.7-es definícióban. Ezért a továbbiakban jelöljük $R\mathbb{C}$-ben a távolságot $q$-val és így

$$q([a], [b]) = q([a_{re}], [b_{re}]) + q([a_{im}], [b_{im}]).$$

Felhasználva, hogy $q$ metrika $\mathbb{IR}$-ben, könnyen igazolható, hogy $q$ metrika $R\mathbb{C}$-ben. A $q$ metrika bevezetésével $R\mathbb{C}$ egy topológikus térré

válik. Ha most a metrikus terekben szokásos módon bevezetjük a konvergencia fogalmát, akkor azt mondhatjuk, hogy egy $\left\{[a^{(k)}]\right\}_{k=0}^{\infty}$ $R\mathbb{C}$-beli sorozat (ahol $[a^{(k)}] = [a_{re}^{(k)}] + i[a_{im}^{(k)}]$,) akkor és csak akkor tart egy $[a] = [a_{re}] + i[a_{im}] \in R\mathbb{C}$ elemhez, ha

$$\lim_{k\to\infty}[a_{re}^{(k)}] = [a_{re}] \text{ és } \lim_{k\to\infty}[a_{im}^{(k)}] = [a_{im}]. \tag{2.10}$$

Felhasználva, hogy $(\mathbb{IR}, q)$ metrikus tér teljes, (2.10) alapján következik, hogy $R\mathbb{C}$ a $q$ metrikával szintén teljes metrikus tér.

**2.13. Definíció.** *Legyen* $[a] = [a_{re}] + i[a_{im}] \in R\mathbb{C}$. *Ekkor*

$$|[a]| = q([a], 0) = |[a_{re}]| + |[a_{im}]| = q([a_{re}], 0) + q([a_{im}], 0)$$

*az* $[a]$ *abszolútértéke.*

Ha $[a] = [a_{re}, a_{re}] + i[a_{im}, a_{im}] = a_{re} + ia_{im} = a$, akkor a következőt kapjuk:

$$|[a]| = |a| = |a_{re}| + |a_{im}|\,. \tag{2.11}$$

Egy $[a] \in R\mathbb{C}$ elem abszolútértéke tehát nem számolható át a komplex számok euklideszi abszolútértékére. A továbbiakban a szövegkörnyezetből nyilvánvaló lesz, mikor használjuk az euklideszi abszolútértéket és mikor a 2.13 definícióbeli abszolútértéket. Végül megemlítenénk, hogy a (2.11) használatával igaz marad az

$$|[a]| = \max_{a\in[a]} |a|$$

reláció.

Jelölje $d$ egy valós intervallum szélességét, úgy ahogy azt az 1.14 definícióban bevezettük. Ekkor a következőt kapjuk:

**2.14. Definíció.** *Legyen* $[a] = [a_{re}] + i[a_{im}] \in R\mathbb{C}$. *Ekkor a*

$$d([a]) = d([a_{re}]) + d([a_{im}])$$

*mennyiséget az* $[a]$ *szélességének nevezzük.*

Most bevezetjük a megfelelő fogalmakat $K\mathbb{C}$-ben.

**2.15. Definíció.** *Legyen* $[a] = \langle a, r_a \rangle$, $[b] = \langle b, r_b \rangle \in K\mathbb{C}$. *Ekkor*

*(a)* $q([a],[b]) = |a-b| + |r_a - r_b|$ *az* $[a]$ *és a* $[b]$ *elemek távolsága,*

*(b)* $|[a]| = |a| + r_a$ *az* $[a]$ *abszolútértéke, és*

*(c)* $d([a]) = 2r_a$ *az* $[a]$ *szélessége.*

Az előző definícióban a komplex-sík két kör-intervallumának távolságát az euklideszi metrika segítségével definiáltuk. A kör-intervallum abszolútértéke az euklideszi abszolútértékre vezet, ha a komplex számok halmazára szűkítjük le. Megjegyeznénk, hogy az

$$|[a]| = \max_{a \in [a]} |a|$$

reláció itt is igaz marad.

A $K\mathbb{C}$ tér teljessége a $q$ metrikával könnyen igazolható, ha a $K\mathbb{C}$-beli sorozatok konvergenciáját a $q$ metrikában a szokásos módon definiáljuk. Ezzel a definícióval a következőt kapjuk

$$\lim_{k\to\infty} [a^{(k)}] = [a] \;\Leftrightarrow\; \lim_{k\to\infty} a^{(k)} = a, \text{ és } \lim_{k\to\infty} r^{(k)} = r, \tag{2.12}$$

ahol

$$\left\{[a^{(k)}]\right\}_{k=0}^{\infty} = \left\{\left\langle a^{(k)}, r^{(k)}\right\rangle\right\}_{k=0}^{\infty} \quad \text{és} \quad [a] = \langle a, r \rangle .$$

Most pedig összegyűjtjük a metrika, az abszolútérték és a szélesség legfontosabb tulajdonságait az $R\mathbb{C}$ és a $K\mathbb{C}$ halmazokon.

**2.16. Tétel.** *Legyenek* $[a],[b],[c],[d] \in I\mathbb{C}$, *ekkor igazak a következők:*

$$q([a]+[b],[a]+[c]) = q([b],[c]), \tag{2.13}$$

$$q([a]+[b],[c]+[d]) \le q([a],[c]) + q([b],[d]), \tag{2.14}$$

$$q(a[b],a[c]) \le |a|\, q([b],[c]),\ a \in \mathbb{C}. \tag{2.15}$$

*A (2.15)-ban mindig fennáll az egyenlőség, ha* $[b],[c] \in K\mathbb{C}$.

$$q([a][b],[a][c]) \le |[a]|\, q([b],[c]), \tag{2.16}$$

$$|[a]| \ge 0,\ |[a]| = 0 \Leftrightarrow [a] = 0, \tag{2.17}$$

$$|[a]+[b]| \le |[a]| + |[b]|, \tag{2.18}$$

$$|a[b]| \le |a|\,|[b]|,\ \forall a \in \mathbb{C}. \tag{2.19}$$

*A (2.19)-ban mindig fenáll az egyenlőség, ha* $[b] \in K\mathbb{C}$.

$$|[a][b]| \leq |[a]|\,|[b]|\,, \tag{2.20}$$
$$d(a[b]) = |a|\,d([b]),\ a \in \mathbb{C}, \tag{2.21}$$
$$d([a][b]) \leq |[a]|\,d([b]) + |[b]|\,d([a]), \tag{2.22}$$
$$d([a]) = |[a] - [a]|\,, \tag{2.23}$$
$$d([a][b]) \geq |[a]|\,d([b]), \tag{2.24}$$
$$d([a] \pm [b]) = d([a]) + d([b]), \tag{2.25}$$
$$[a] \subseteq [b] \Rightarrow \frac{1}{2}\,(d([b]) - d([a])) \leq q([a],[b]) \leq d([b]) - d([a]). \tag{2.26}$$

**Bizonyítás:** A fenti tulajdonságokat először $R\mathbb{C}$-re bizonyítjuk.

A (2.13)-(2.16) tulajdonságok egyszerűen a valós intervallumokra vonatkozó az 1.13 tétel megfelelő állításaiból igazolhatóak. Legyen ezért

$$[a] = [a_{re}] + i[a_{im}], \quad [b] = [b_{re}] + i[b_{im}],$$
$$[c] = [c_{re}] + i[c_{im}], \quad [d] = [d_{re}] + i[d_{im}] \in R\mathbb{C}.$$

(2.13) bizonyításához tekintsük:

$$\begin{aligned}
&q([a]+[b],[a]+[c]) = \\
&= q\left([a_{re}]+[b_{re}]+i([a_{im}]+[b_{im}]),[a_{re}]+[c_{re}]+i([a_{im}]+[c_{im}])\right) = \\
&= q\left([a_{re}]+[b_{re}],[a_{re}]+[c_{re}]\right) + q\left([a_{im}]+[b_{im}],[a_{im}]+[c_{im}]\right) = \\
&= q([b_{re}],[c_{re}]) + q([b_{im}],[c_{im}]) = q([b],[c]).
\end{aligned}$$

(2.14) bizonyításához tekintsük:

$$\begin{aligned}
&q([a]+[b],[c]+[d]) = \\
&= q\left([a_{re}]+[b_{re}],[c_{re}]+[d_{re}]\right) + q\left([a_{im}]+[b_{im}],[c_{im}]+[d_{im}]\right)) \leq \\
&\leq q\left([a_{re}],[c_{re}]\right) + q\left([b_{re}],[d_{re}]\right) + q\left([a_{im}],[c_{im}]\right) + q\left([b_{im}],[d_{im}]\right) = \\
&= q([a],[c]) + q([b],[d]).
\end{aligned}$$

(2.15) és (2.16) bizonyítását egyszerre végezzük, ugyanis (2.15) speciális esete (2.16)-nak $[a] = [a, a]$ választással.

$$\begin{aligned}
&q([a][b],[a][c]) = \\
&= q\left([a_{re}][b_{re}] - [a_{im}][b_{im}], [a_{re}][c_{re}] - [a_{im}][c_{im}]\right) + \\
&+ q\left([a_{re}][b_{im}] + [a_{im}][b_{re}], [a_{re}][c_{im}] + [a_{im}][c_{re}]\right) \leq \\
&\leq |[a_{re}]|\, q\left([b_{re}],[c_{re}]\right) + |[a_{im}]|\, q\left([b_{im}],[c_{im}]\right) + \\
&+ |[a_{re}]|\, q\left([b_{im}],[c_{im}]\right) + |[a_{im}]|\, q\left([b_{re}],[c_{re}]\right) = \\
&= \left(|[a_{re}]| + |[a_{im}]|\right) q([b],[c]) = |[a]|\, q([b],[c]).
\end{aligned}$$

A (2.17)-(2.20) eredmények $|[a]|$ definíciójának felhasználásával igazolhatók.

(2.17) bizonyítása:

$$\begin{aligned}
&|[a]| = q\left([a],0\right) = q\left([a_{re}],0\right) + q\left([a_{im}],0\right) = |[a_{re}]| + |[a_{im}]| \geq 0, \\
&|[a]| = 0 \Leftrightarrow |[a_{re}]| = |[a_{im}]| = 0 \Leftrightarrow [a] = 0.
\end{aligned}$$

(2.18) bizonyítása, (2.14)-et felhasználva:

$$|[a] + [b]| = q\left([a] + [b], 0\right) \leq q\left([a],0\right) + q\left([b],0\right) = |[a]| + |[b]|\,.$$

(2.19) és (2.20) bizonyítása, felhasználva (2.15)-öt és (2.16)-ot:

$$|[a][b]| = q\left([a][b], 0\right) = q\left([a][b], [a]\cdot 0\right) \leq |[a]|\, q\left([b],0\right) = |[a]|\,|[b]|\,.$$

(2.21) bizonyítása:
Legyen $a = a_{re} + i a_{im} \in \mathbb{C}$. A 2.3 Definíció alapján kapjuk:

$$a[b] = a_{re}[b_{re}] - a_{im}[b_{im}] + i\left(a_{re}[b_{im}] + a_{im}[b_{re}]\right)$$

felhasználva (2.11)-t kaphatjuk, hogy:

$$\begin{aligned}
d(a[b]) &= d\left(a_{re}[b_{re}] - a_{im}[b_{im}]\right) + d\left(a_{re}[b_{im}] + a_{im}[b_{re}]\right) = \\
&= d\left(a_{re}[b_{re}]\right) + d\left(a_{im}[b_{im}]\right) + d\left(a_{re}[b_{im}]\right) + d\left(a_{im}[b_{re}]\right) = \\
&= |a_{re}|\, d\left([b_{re}]\right) + |a_{im}|\, d\left([b_{im}]\right) + |a_{re}|\, d\left([b_{im}]\right) + |a_{im}|\, d\left([b_{re}]\right) = \\
&= \left(|a_{re}| + |a_{im}|\right)\left(d\left([b_{re}]\right) + d\left([b_{im}]\right)\right) = |a|\, d\left([b]\right).
\end{aligned}$$

(2.22) bizonyítása:

$$\begin{aligned}
&d([a][b]) = d([a_{re}][b_{re}] - [a_{im}][b_{im}]) + d([a_{re}][b_{im}] + [a_{im}][b_{re}]) = \\
&= d([a_{re}][b_{re}]) + d([a_{im}][b_{im}]) + d([a_{re}][b_{im}]) + d([a_{im}][b_{re}]) \leq \\
&\leq |[a_{re}]|\, d([b_{re}]) + |[b_{re}]|\, d([a_{re}]) + |[a_{im}]|\, d([b_{im}]) + |[b_{im}]|\, d([a_{im}]) + \\
&\quad + |[a_{re}]|\, d([b_{im}]) + |[b_{im}]|\, d([a_{re}]) + |[a_{im}]|\, d([b_{re}]) + |[b_{re}]|\, d([a_{im}]) = \\
&= (|[a_{re}]| + |[a_{im}]|)\,(d([b_{re}]) + d([b_{im}])) + \\
&\quad + (|[b_{re}]| + |[b_{im}]|)\,(d([a_{re}]) + d([a_{im}])) = \\
&= |[a]|\, d([b]) + |[b]|\, d([a])\,.
\end{aligned}$$

(2.23) bizonyítása:

$$d([a]) = d([a_{re}]) + d([a_{im}]) = |[a_{re}] - [a_{re}]| + |[a_{im}] - [a_{im}]| = |[a] - [a]|\,.$$

(2.24) bizonyítása:

$$\begin{aligned}
&d([a][b]) = d([a_{re}][b_{re}] - [a_{im}][b_{im}]) + d([a_{re}][b_{im}] + [a_{im}][b_{re}]) \geq \\
&\geq |[a_{re}]|\, d([b_{re}]) + |[a_{im}]|\, d([b_{im}]) + |[a_{re}]|\, d([b_{im}]) + |[a_{im}]|\, d([b_{re}]) = \\
&= (|[a_{re}]| + |[a_{im}]|)\,(d([b_{re}]) + d([b_{im}])) = |[a]|\, d([b])\,.
\end{aligned}$$

(2.25) bizonyítása:

$$\begin{aligned}
d([a] \pm [b]) &= d([a_{re}] \pm [b_{re}]) + d([a_{im}] \pm [b_{im}]) = \\
&= d([a_{re}]) + d([a_{im}]) + d([b_{re}]) + d([b_{im}]) = \\
&= d([a]) + d([b])\,.
\end{aligned}$$

(2.26) egyenes következménye (1.31)-nek.

$K\mathbb{C}$ esetén a bizonyítások a következők.

$$\begin{aligned}
&[a] = \langle a, r_a \rangle\,, \quad [b] = \langle b, r_b \rangle\,, \\
&[c] = \langle c, r_c \rangle\,, \quad [d] = \langle d, r_d \rangle \in K((\mathbb{C})).
\end{aligned}$$

(2.13):

$$\begin{aligned}
q([a] + [b], [a] + [c]) &= |a + b - (a + c)| + |r_a + r_b - (r_a + r_c)| = \\
&= |b - c| + |r_b - r_c| = q([b], [c])\,.
\end{aligned}$$

(2.14):

$$\begin{aligned} q([a]+[b],[c]+[d]) &= |a+b-(c+d)|+|r_a+r_b-(r_c+r_d)| \leq \\ &\leq |a-c|+|r_a-r_c|+|b-d|+|r_b-r_d| = \\ &= q([a],[c])+q([b],[d]). \end{aligned}$$

(2.15):

$$\begin{aligned} q(a[b],a[c]) &= |ab-ac|+||a|r_b-|a|r_c| = \\ &= |a|\{|b-c|+|r_b-r_c|\} = |a|\,q([b],[c]). \end{aligned}$$

(2.16):

$$\begin{aligned} &q([a][b],[a][c]) = \\ &= |ab-ac|+||a|r_b+|b|r_a+r_ar_b-(|a|r_c+|c|r_a+r_ar_c)| \leq \\ &\leq |a|\,|b-c|+|a|\,|r_b-r_c|+r_a\,||b|-|c||+r_a\,|r_b-r_c| \leq \\ &\leq (|a|+r_a)\,(|b-c|+|r_b-r_c|) = |[a]|\,q([b],[c]). \end{aligned}$$

(2.17):

$$|[a]| = |a|+r_a \geq 0, \quad |[a]| = 0 \Leftrightarrow (a=0, r_a=0).$$

(2.18):

$$|[a]+[b]| = |a+b|+|r_a+r_b| \leq |a|+r_a+|b|+r_b = |[a]|+|[b]|.$$

(2.19):

$$|a[b]| = |ab|+|a|\,r_b = |a|\,|[b]|.$$

(2.20) bizonyítása (2.16) felhasználásával:

$$|[a][b]| = q([a][b],0) = q([a][b],[a]\cdot 0) \leq |[a]|\,q([b],0) = |[a]|\,|[b]|.$$

(2.21):

$$d(a[b]) = 2\,|a|\,r_b = |a|\,d([b]).$$

(2.22):

$$\begin{aligned} d([a][b]) &= 2\{|a|\,r_b+|b|\,r_a+r_ar_b\} = \\ &= 2\{(|a|+r_a)\,r_b+|b|\,r_a\} \leq \\ &\leq 2\{(|a|+r_a)\,r_b+(|b|+r_b)\,r_a\} = \\ &= |[a]|\,d([b])+|[b]|\,d([a]). \end{aligned}$$

(2.23):

$$d([a]) = 2r_a = |\langle 0, 2r_a\rangle| = |[a] - [a]|\,.$$

(2.24):

$$\begin{aligned} d([a][b]) &= 2\{|a|\,r_b + |b|\,r_a + r_a r_b\} = \\ &= 2\{(|a| + r_a)\,r_b + |b|\,r_a\} \geq \\ &\geq 2\,(|a| + r_a)\,r_b = |[a]|\,d([b])\,. \end{aligned}$$

(2.25):

$$d([a] \pm [b]) = d(\langle a \pm b, r_a + r_b\rangle) = 2\,(r_a + r_b) = d([a]) + d([b])\,.$$

(2.26): $[a] \subseteq [b]$ akkor és csak akkor, ha $|a - b| \leq r_b - r_a$. Ezért

$$\begin{aligned} \frac{1}{2}\,(d([b]) - d([a])) &= |r_b| - |r_a| \leq |r_b - r_a| \leq |a - b| + |r_a - r_b| = \\ &= q([a], [b]) \leq r_b - r_a + |r_b - r_a| = d([b]) - d([a])\,. \end{aligned}$$

□

**2.17. Tétel.** *Az R$\mathbb{C}$-n és a K$\mathbb{C}$-n definiált $\{+, -, \cdot, :\}$ műveletek folytonos leképezések.*

**Bizonyítás:** Legyenek $\left\{[a^{(k)}]\right\}_{k=0}^{\infty}, \left\{[b^{(k)}]\right\}_{k=0}^{\infty}$ sorozatok, melyekre

$$[a^{(k)}] = [a_{re}^{(k)}] + i[a_{im}^{(k)}], \quad [b^{(k)}] = [b_{re}^{(k)}] + i[b_{im}^{(k)}] \in R\mathbb{C}$$

és legyenek

$$\lim_{k\to\infty}[a^{(k)}] = A = [a_{re}] + i[a_{im}], \quad \lim_{k\to\infty}[b^{(k)}] = [b] = [b_{re}] + i[b_{im}].$$

Megmutatjuk, hogy a szorzás folytonos művelet. Ezért elvégezzük az alábbi számítást:

$$\begin{aligned} &\lim_{k\to\infty}[a^{(k)}][b^{(k)}] = \\ &= \lim_{k\to\infty}\left\{[a_{re}^{(k)}][b_{re}^{(k)}] - [a_{im}^{(k)}][b_{im}^{(k)}] + i\left([a_{re}^{(k)}][b_{im}^{(k)}]\right) + [a_{im}^{(k)}][b_{re}^{(k)}]\right\} = \\ &= \lim_{k\to\infty}\left([a_{re}^{(k)}][b_{re}^{(k)}] - [a_{im}^{(k)}][b_{im}^{(k)}]\right) + i\lim_{k\to\infty}\left([a_{re}^{(k)}][b_{im}^{(k)}] - [a_{im}^{(k)}][b_{re}^{(k)}]\right) = \\ &= [a_{re}][b_{re}] - [a_{im}][b_{im}] + i\,([a_{re}][b_{im}] + [a_{im}][b_{re}]) = [a][b], \end{aligned}$$

mivel a komplex számok valós és imaginárius részekre bontása folytonos művelet $\mathbb{IR}$-n. Hasonló bizonyítás végezhető el a többi műveletre $R\mathbb{C}$-n és az összes műveletre $K\mathbb{C}$-n. □

A valós esethez hasonlóan új kétváltozós műveleteket vezetünk be $R\mathbb{C}$-ben.
Legyen $[a], [b] \in R\mathbb{C}$ két intervallum ezek halmazelméleti metszetének nevezzük $[a]$ és $[b]$ metszetét:

$$[a] \cap [b] = \{c | c \in [a], c \in [b]\} . \tag{2.27}$$

$[a]$ és $[b]$ elemek metszete $R\mathbb{C}$-beli, ha a halmazelméleti metszet nem üres. Ha $[a] = [a_{re}] + i[a_{im}]$, $[b] = [b_{re}] + i[b_{im}]$, akkor

$$[a] \cap [b] = [a_{re}] \cap [b_{re}] + i \left( [a_{im}] \cap [b_{im}] \right) , \tag{2.28}$$

ahol $[a_i] \cap [b_i]$-t a az (1.33)-nak megfelelően kell kialakítani.

Az 1.18 következmény megfelelője:

**2.18. Következmény.** *Legyen* $[a], [b], [c], [d] \in R\mathbb{C}$. *Ekkor*

$$[a] \subseteq [c], \ [b] \subseteq [d] \ \Rightarrow \ [a] \cap [b] \subseteq [c] \cap [d] \tag{2.29}$$

*tartalmazási monotonitás, továbbá a metszet művelet folytonos művelet, ha az eredmény* $R\mathbb{C}$*-beli.*

A fenti következmény a az 1.18 következmény valós illetve képzetes részekre való alkalmazásával igazolható.

# 3. fejezet

# Intervallum-együtthatós lineáris egyenletrendszerek

## 3.1. Intervallummátrixok

A következő részben az intervallummátrixok legfontosabb tulajdonságait foglaljuk össze bizonyítás nélkül. Megjegyezzük, hogy az 1. fejezetben tárgyalt intervallumokra vonatkozó tulajdonságok itt is igazak.

Az $m \times n$-es valós mátrixok halmazát a szokásos $\mathbb{R}^{m\times n}$, az egy oszlopból álló mátrixokat, azaz az oszlopvektorokat $\mathbb{R}^n$ jelöli. Jelölje $\mathbb{IR}^{m\times n}$ az olyan $m\times n$-es mátrixok halmazát, melyek komponensei intervallumok, az intervallumvektorokat pedig $\mathbb{IR}^n$.

**3.1. Definíció.** $\mathbf{A} = ([a]_{ij}) \in \mathbb{IR}^{m\times n}$ *és* $\mathbf{B} = ([b]_{ij}) \in \mathbb{IR}^{m\times n}$ *egyenlők, azaz* $\mathbf{A} = \mathbf{B}$ *pontosan akkor, ha minden komponensük egyenlő, azaz* $[a]_{ij} = [b]_{ij}$, $1 \le i \le m$, $1 \le j \le n$.

Definiálunk egy részbenrendezést $\mathbb{IR}^{m\times n}$-en.

**3.2. Definíció.** *Legyen* $\mathbf{A} = ([a]_{ij})$ *és* $\mathbf{B} = ([b]_{ij}) \in \mathbb{IR}^{m\times n}$. *Ekkor azt mondjuk, hogy* $\mathbf{A} \subseteq \mathbf{B}$, *ha* $[a]_{ij} \subseteq [b]_{ij}$ $1 \le i \le m$, $1 \le j \le n$.

**3.3. Megjegyzés.** *Ha* $A$ *pontmátrix, azaz* $A \in \mathbb{R}^{m\times n}$, *akkor az* $A \in \mathbf{B}$ *jelölést használjuk.*

**3.4. Definíció.** *1. Ha* $\mathbf{A} = ([a]_{ij})$ *és* $\mathbf{B} = ([b]_{ij}) \in \mathbb{IR}^{m\times n}$*, akkor*

$$\mathbf{A} \pm \mathbf{B} := ([a]_{ij} \pm [b]_{ij}).$$

*2. Ha* $\mathbf{A} = ([a]_{ij}) \in \mathbb{IR}^{m\times r}$ *és* $\mathbf{B} = ([b]_{ij}) \in \mathbb{IR}^{r\times n}$*, akkor*

$$\mathbf{AB} := \left( \sum_{k=1}^{r} [a]_{ik}[b]_{kj} \right).$$

*Speciálisan, ha* $\mathbf{u} = ([u]_i) \in \mathbb{IR}^n$*, akkor*

$$\mathbf{Au} = \left( \sum_{k=1}^{r} [a]_{ik}[u]_k \right).$$

*3. Ha* $\mathbf{A} = ([a]_{ij}) \in \mathbb{IR}^{m\times n}$ *és* $[x] \in \mathbb{IR}$*, akkor*

$$[x]\mathbf{A} = \mathbf{A}[x] := ([x][a]_{ij}).$$

**3.5. Állítás.** *Legyen* $\mathbf{A} \in \mathbb{IR}^{m\times r}$ *és* $\mathbf{B} \in \mathbb{IR}^{r\times n}$*. Ekkor*

$$\{AB : A \in \mathbf{A}, B \in \mathbf{B}\} \subseteq \{C : C \in \mathbf{AB}\}.$$

Egyenlőség általában nem igazolható.

**3.6. Állítás.** *Legyen* $\mathbf{A}, \mathbf{B} \in \mathbb{IR}^{m\times n}$ *és* $c \in \mathbb{R}^n$*. Ekkor*

*1.* $\{A + B : A \in \mathbf{A}, B \in \mathbf{B}\} = \mathbf{A} + \mathbf{B}$*, és*

*2.* $\{Ac : A \in \mathbf{A}\} = \mathbf{A}c$*.*

Tehát az intervallummátrixok halmaza zárt az előző definícióban bevezetett műveletekre.

**3.7. Tétel.** *Legyenek* $\mathbf{A}, \mathbf{B}$ *és* $\mathbf{C}$ *olyan méretű intervallummátrixok, amelyekre az adott műveletek értelmehetők. Ekkor*

*1.* $\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$*.*

*2.* $\mathbf{A} + (\mathbf{B} + \mathbf{C}) = (\mathbf{A} + \mathbf{B}) + \mathbf{C}$*,*

3. $\mathbf{A} + 0 = 0 + \mathbf{A} = \mathbf{A}$, *ahol* $0$ *a megfelelő méretű nullmátrix.*

4. $\mathbf{A}I = I\mathbf{A} = \mathbf{A}$, *ahol* $I$ *a megfelelő méretű egységmátrix.*

5. $(\mathbf{A} + \mathbf{B})\mathbf{C} \subseteq \mathbf{AC} + \mathbf{BC}$ *és* $\mathbf{C}(\mathbf{A} + \mathbf{B}) \subseteq \mathbf{CA} + \mathbf{CB}$.

6. $(\mathbf{A}+\mathbf{B})C = \mathbf{A}C+\mathbf{B}C$ *és* $C(\mathbf{A}+\mathbf{B}) = C\mathbf{A}+C\mathbf{B}$, *ahol* $C \in \mathbb{R}^{k\times m}$.

7. $\mathbf{A}(BC) \subseteq (\mathbf{A}B)C$, *ahol* $B$ *és* $C$ *valós mátrixok.*

8. $(A\mathbf{B})\mathbf{C} \subseteq A(\mathbf{BC})$, *ha* $\mathbf{C} = -\mathbf{C}$, *és* $A \in \mathbb{R}^{k\times m}$.

9. $A(\mathbf{B}C) = (A\mathbf{B})C$, *ahol* $C \in \mathbb{R}^{n\times k}$.

10. $\mathbf{A}(\mathbf{BC}) = (\mathbf{AB})\mathbf{C}$, *ha* $\mathbf{B} = -\mathbf{B}$ *és* $\mathbf{C} = -\mathbf{C}$.

**3.8. Tétel.** *Legyenek* $\mathbf{A}^{(k)}, \mathbf{B}^{(k)}$, $k = 1, 2$ *intervallummátrixok és* $[x], [y]$ *intervallumok. Továbbá tegyük fel, hogy* $\mathbf{A}^{(k)} \subseteq \mathbf{B}^{(k)}$, $k = 1, 2$ *és* $[x] \subseteq [y]$. *Ekkor*

1. $\mathbf{A}^{(1)} * \mathbf{A}^{(2)} \subseteq \mathbf{B}^{(1)} * \mathbf{B}^{(2)}$, *ahol* $* = \{+, -, \cdot\}$, *és*

2. $[x]\mathbf{A}^{(1)} \subseteq [y]\mathbf{B}^{(1)}$.

**3.9. Megjegyzés.** *Ha speciálisan* $A \in \mathbf{A}$, $B \in \mathbf{B}$ *és* $x \in [x]$, *akkor*

1. $A * B \in \mathbf{A} * \mathbf{B}$, *ahol* $* = \{+, -, \cdot\}$, *és*

2. $xA \in [x]\mathbf{A}$.

Az intervallumokhoz hasonlóan a következőkben definiáljuk az intervallummátrixok szélességét és abszolútértékét.

**3.10. Definíció.** *Legyen* $\mathbf{A} = ([a]_{ij}) \in \mathbb{IR}^{m\times n}$. *Ekkor*

$$d(\mathbf{A}) := (d([a]_{ij}))$$

*az* $\mathbf{A}$ *szélességmátrixa.*

**3.11. Definíció.** *Legyen* $\mathbf{A} = ([a]_{ij}) \in \mathbb{IR}^{m\times n}$. *Ekkor*

$$|\mathbf{A}| := (|[a]_{ij}|)$$

*az* $\mathbf{A}$ *abszolútérték-mátrixa.*

**3.12. Definíció.** *Legyen* $X = (x_{ij}), Y = (y_{ij}) \in \mathbb{R}^{m\times n}$. *Ekkor azt mondjuk, hogy* $X \leq Y$, *ha* $x_{ij} \leq y_{ij}$ $\forall 1 \leq i \leq m$ *és* $1 \leq j \leq n$ *esetén.*

**3.13. Állítás.** *Legyen* $\mathbf{A}$ *és* $\mathbf{B}$ *intervallummátrix, ekkor a következők teljesülnek.*

1. *Ha* $\mathbf{A} \subseteq \mathbf{B}$, *akkor* $d(\mathbf{A}) \leq d(\mathbf{B})$.
2. $d(\mathbf{A} \pm \mathbf{B}) = d(\mathbf{A}) \pm d(\mathbf{B})$.
3. $d(\mathbf{A}) = \sup_{A,A'\in\mathbf{A}} |A - A'|$.
4. $\mathbf{A} \subseteq \mathbf{B}$ *esetén* $|\mathbf{A}| \leq |\mathbf{B}|$.
5. $|\mathbf{A}| = \sup_{A\in\mathbf{A}} |A|$.
6. - $|\mathbf{A}| \geq 0$ *és* $|\mathbf{A}| = 0 \Leftrightarrow \mathbf{A} = 0$,
   - $|\mathbf{A} + \mathbf{B}| \leq |\mathbf{A}| + |\mathbf{B}|$,
   - $|x\mathbf{A}| = |\mathbf{A}x| = |x||\mathbf{A}|$ $\forall x \in \mathbb{R}$ *és*
   - $|\mathbf{AB}| \leq |\mathbf{A}||\mathbf{B}|$.
7. $d(\mathbf{AB}) \leq d(\mathbf{A})|\mathbf{B}| + |\mathbf{A}|d(\mathbf{B})$.
8. $d(\mathbf{AB}) \geq |\mathbf{A}|d(\mathbf{B})$ *és* $d(\mathbf{AB}) \geq d(\mathbf{A})|\mathbf{B}|$.
9. - $d(a\mathbf{B}) = |a|d(\mathbf{B})$ $\forall a \in \mathbb{R}$ *esetén,*
   - $d(A\mathbf{B}) = |A|d(\mathbf{B})$, *ha* $A$ *megfelelő méretű valós mátrix.*
   - $d(\mathbf{B}A) = d(\mathbf{B})|A|$, *ha* $A$ *megfelelő méretű valós mátrix.*
10. *Ha* $0$ *a nullmátrixot jelöli, akkor* $0 \in \mathbf{A}$ *esetén* $|\mathbf{A}| \leq d(\mathbf{A}) \leq 2|\mathbf{A}|$.
11. *Ha* $\mathbf{A} = -\mathbf{A}$, *akkor* $\mathbf{AB} = \mathbf{A}|\mathbf{B}|$.

12. *Legyen* $\mathbf{B} = ([b]_{ij})$ *és tegyük fel, hogy* $0 \in \mathbf{A}$ *és* $0 \notin [b]_{ij}$. *Ekkor* $d(\mathbf{AB}) = d(\mathbf{A})|\mathbf{B}|$.

**3.14. Definíció.** *Legyen* $\mathbf{A} = ([a]_{ij})$ *és* $\mathbf{B} = ([b]_{ij}) \in \mathbb{IR}^{m\times n}$. *Ekkor az* $\mathbf{A}$ *és* $\mathbf{B}$ *intervallummátrixok távolsága*

$$q(\mathbf{A}, \mathbf{B}) := (q([a]_{ij}, [b]_{ij})).$$

**3.15. Állítás.** *Legyenek* $\mathbf{A}, \mathbf{B}, \mathbf{C}$ *és* $\mathbf{D}$ *olyan méretű intervallummátrixok, amelyekre az adott műveletek értelmezhetők. Ekkor*

1. $q(\mathbf{A}, \mathbf{B}) = 0 \Leftrightarrow \mathbf{A} = \mathbf{B}$,
2. $q(\mathbf{A}, \mathbf{B}) \leq q(\mathbf{A}, \mathbf{C}) + q(\mathbf{B}, \mathbf{C})$,
3. $q(\mathbf{A} + \mathbf{C}, \mathbf{B} + \mathbf{C}) = q(\mathbf{A}, \mathbf{B})$,
4. $q(\mathbf{A} + \mathbf{B}, \mathbf{C} + \mathbf{D}) = q(\mathbf{A}, \mathbf{C}) + q(\mathbf{B}, \mathbf{D})$,
5. $q(\mathbf{AB}, \mathbf{AC}) \leq |\mathbf{A}|q(\mathbf{B}, \mathbf{C})$.

A fent definiált távolságfogalommal és egy tetszőleges monoton mátrixnormával metrikát kapunk $\mathbb{IR}^{m\times n}$-en. Mivel $\mathbb{IR}^{m\times n}$ felfogható úgy, hogy $\mathbb{IR} \times \mathbb{IR} \times ... \times \mathbb{IR}$ ($nm$ db) és $\mathbb{IR}$ teljes metrikus tér, ezért $\mathbb{IR}^{m\times n}$ is az. A konvergencia a pontonkénti konvergencia, azaz

$$\lim_{k\to\infty} \mathbf{A}^{(k)} = \mathbf{A} \Leftrightarrow \lim_{k\to\infty} [a]_{ij}^{(k)} = [a]_{ij},$$

$1 \leq i \leq m$, $1 \leq j \leq n$.

**3.16. Következmény.** *Legyen* $\{\mathbf{A}^{(k)}\}_{k=0}^{\infty}$ *olyan intervallummátrixsorozat, melyre* $\mathbf{A}^{(0)} \supseteq \mathbf{A}^{(1)} \supseteq ....$ *Ekkor* $\{\mathbf{A}^{(k)}\}_{k=0}^{\infty}$ *konvergens, és*

$$\lim_{k\to\infty} \mathbf{A}^{(k)} = \mathbf{A} = ([a]_{ij}),$$

*ahol*

$$[a]_{ij} = \bigcap_{k=0}^{\infty} [a]_{ij}^{(k)}.$$

**3.17. Következmény.** *Az $\mathbb{IR}^{m\times n}$-en definiált műveletek $(+,-,\cdot)$ folytonosak.*

**3.18. Állítás.** *Legyen $\mathbf{X}\subseteq\mathbf{Y}\in\mathbb{IR}^{m\times n}$. Ekkor*

$$\frac{1}{2}(d(\mathbf{Y})-d(\mathbf{X}))\leq q(\mathbf{X},\mathbf{Y})\leq d(\mathbf{Y})-d(\mathbf{X}).$$

**3.19. Definíció.** *Legyen $\mathbf{A},\mathbf{B}\in\mathbb{IR}^{m\times n}$. Ekkor*

$$\mathbf{A}\cap\mathbf{B}:=\{C:C\in\mathbf{A},C\in\mathbf{B}\},$$

*azaz a halmazelméleti metszete a két mátrixnak.*

**3.20. Állítás.** *Legyen $\mathbf{A}=([a]_{ij})$ és $\mathbf{B}=([b]_{ij})\in\mathbb{IR}^{m\times n}$. Ekkor $\mathbf{A}\cap\mathbf{B}$ pontosan akkor $\mathbb{IR}^{m\times n}$-beli, ha nem üres. Ebben az esetben*

$$\mathbf{A}\cap\mathbf{B}=([a]_{ij}\cap[b]_{ij}),$$

$1\leq i\leq m$, $1\leq j\leq n$.

**3.21. Következmény.** *(Tartalmazási monotonitás) Legyenek $\mathbf{A},\mathbf{B},\mathbf{C},\mathbf{D}$ intervallummátrixok. Továbbá tegyük fel, hogy $\mathbf{A}\subseteq\mathbf{C}$ és $\mathbf{B}\subseteq\mathbf{D}$. Ekkor*

$$\mathbf{A}\cap\mathbf{B}\subseteq\mathbf{C}\cap\mathbf{D}.$$

A következőkben olyan $\mathbf{A}x=\mathbf{b}$ lineáris egyenletrendszerekkel fogunk foglalkozni, melyek $\mathbf{A}$ mátrixa intervallummátrix és a jobb oldal intervallumvektor.

## 3.2. Intervallum-együtthatós lineáris egyenletrenszerek megoldása

Ebben a részben az intervallum-együtthatós lineáris egyenletrendszerek megoldhatóságának kérdését tárgyaljuk általános esetben. Legyen

$$\mathbf{A}=[\underline{A},\overline{A}]\in\mathbb{IR}^{m\times n},\qquad\mathbf{b}=[\underline{b},\overline{b}]\in\mathbb{IR}^{m}.$$

**3.22. Definíció.** *Egy*

$$\mathbf{A}x = \mathbf{b}$$

*intervallum-együtthatós lineáris egyenletrendszert megoldhatónak nevezünk, ha*

$$Ax = b$$

*megoldható minden* $A \in \mathbf{A}$ *és* $b \in \mathbf{b}$ *esetén.*

A következő jelöléseket fogjuk használni a továbbiakban. Legyen

$$A_c := \frac{1}{2}(\underline{A} + \overline{A})$$

az $\mathbf{A}$ intervallummátrix középmátrixa,

$$\Delta := \frac{1}{2}(\overline{A} - \underline{A})$$

a sugármátrix. Ekkor

$$\mathbf{A} = [A_c - \Delta, A_c + \Delta].$$

Ugyanígy a jobb oldali $\mathbf{b}$ vektorra

$$b_c := \frac{1}{2}(\underline{b} + \overline{b})$$

és

$$\delta := \frac{1}{2}(\overline{b} - \underline{b}),$$

esetén

$$\mathbf{b} = [b_c - \delta, b_c + \delta].$$

Továbbá legyen

$$Y_m := \{y \in \mathbb{R}^m : y_j \in \{-1, 1\} \forall j\},$$

azaz $Y_m$ tartalmazza az összes $m$-dimenziós $\pm 1$ vektort. $Y_m$ elemszáma $2^m$. Végül $\forall y \in Y_m$ vektor esetén jelölje

$$T_y = diag(y_1, ..., y_m).$$

Már most megjegyezzük, hogy $\forall y \in Y_m$ esetén

$$A_c - T_y\Delta \in \mathbf{A}, \quad A_c + T_y\Delta \in \mathbf{A}, \quad b_c + T_y\delta \in \mathbf{b}.$$

Most kimondjuk azt a két állítást, amit a megoldhatóságról szóló tétel bizonyításánál használni fogunk. Az első a jól ismert Farkas-lemma.

**3.23. Lemma.** *(Farkas) Legyen $A \in \mathbb{R}^{m\times n}$ és $b \in \mathbb{R}^m$. Ekkor az*

$$Ax = b,$$

$$x \geq 0$$

*rendszernek akkor és csak akkor létezik megoldása, ha $\forall p \in \mathbb{R}^m$ esetén, melyre*

$$A^T p \geq 0,$$

*igaz, hogy*

$$b^T p \geq 0.$$

**3.24. Tétel.** *(Oettli-Prager) Legyen*

$$X = \{x : |A_c x - b_c| \leq \Delta|x| + \delta\}.$$

*Ekkor minden $x \in X$ esetén létezik $A \in \mathbf{A}$ és $b \in \mathbf{b}$, melyre $Ax = b$.*

Attól az esettől eltekintve, amikor $\underline{A} = \overline{A}$ és $\underline{b} = \overline{b}$ az $\mathbf{A}x = \mathbf{b}$ intervallum-együtthatós lineáris egyenletrenszer végtelen sok lineáris egyenletrendszert tartalmaz. A most következő tétel, ami egyébként ennek a fejezetnek a legfontosabb állítása, azt mondja ki, hogy az $\mathbf{A}x = \mathbf{b}$ megoldása karakterizálható véges sok nemnegatív megoldással. Persze ezek száma általában exponenciális a mátrix méretében.

**3.25. Tétel.** *Az $\mathbf{A}x = \mathbf{b}$ intervallum-együtthatós lineáris egyenletrendszer akkor és csak akkor megoldható, ha $\forall y \in Y_m$ esetén az*

$$\begin{aligned}(A_c - T_y\Delta)x^{(1)} - (A_c + T_y\Delta)x^{(2)} &= b_c + T_y\delta,\\ x^{(1)} \geq 0, \quad x^{(2)} &\geq 0,\end{aligned} \tag{3.1}$$

*rendszernek létezik $x_y^{(1)}, x_y^{(2)}$ megoldása. Továbbá ebben az esetben $\forall A \in \mathbf{A}$, $b \in \mathbf{b}$ esetén az $Ax = b$ egyenletrendszernek létezik megoldása a*

$$Conv\{x_y^{(1)} - x_y^{(2)} : y \in Y_m\}$$

*halmazban.*

**Bizonyítás:** Először nézzük a szükségességet. Tegyük fel, hogy az $\mathbf{A}x = \mathbf{b}$ intervallum-együtthatós lineáris egyenletrendszer megoldható, és indirekt tegyük fel, hogy (3.1) rendszernek nem létezik megoldása. Ekkor a Farkas-lemma szerint $\exists p \in \mathbb{R}^m$, melyre

$$(A_c - T_y\Delta)^T p \geq 0, \tag{3.2}$$

$$(A_c + T_y\Delta)^T p \leq 0, \tag{3.3}$$

$$(b_c + T_y\delta)^T p < 0. \tag{3.4}$$

Ekkor (3.2) és (3.3) szerint

$$\Delta^T T_y p \leq A_c^T p \leq -\Delta^T T_y p,$$

így

$$|A_c^T p| \leq -\Delta^T T_y p = |-\Delta^T T_y p| \leq \Delta^T |p|.$$

Mivel

$$p \in \{x : |A_c^T x| \leq \Delta^T |x|\},$$

ezért az Oettli-Prager-tételt az

$$[A_c^T - \Delta^T, A_c^T + \Delta^T]z = [0, 0]$$

intervallum-együtthatós lineáris egyenletrendszerre alkalmazva azt kapjuk, hogy $\exists A \in \mathbf{A}$, melyre

$$A^T p = 0. \tag{3.5}$$

Tehát $\exists p \in \mathbb{R}^m$, melyre (3.4) és (3.5) teljesül. Ha erre alkalmazzuk a Farkas-lemmát, akkor azt kapjuk, hogy $\nexists x \in \mathbb{R}^n$, melyre

$$Ax = b_c + T_y\delta.$$

Ez ellentmond annak a feltételnek, miszerint az $\mathbf{A}x = \mathbf{b}$ intervallum-együtthatós lineáris egyenletrendszer megoldható, ugyanis $A \in \mathbf{A}$ és $b_c + T_y\delta \in \mathbf{b}$.
Most nézzük az elégségesség bizonyítását. Tegyük fel, hogy $\forall y \in Y_m$ esetén (3.1) rendszernek létezik megoldása: $x_y^{(1)}$, $x_y^{(2)}$. Legyen $A \in \mathbf{A}$ és $b \in \mathbf{b}$ tetszőleges. Azt kell megmutatni, hogy ekkor az $Ax = b$ lineáris

egyenletrendszernek létezik megoldása. Ehhez először azt mutatjuk meg, hogy $\forall y \in Y_m$ esetén

$$T_y A x_y \geq T_y b, \tag{3.6}$$

ahol $x_y = x_y^{(1)} - x_y^{(2)}$. Tehát legyen $y \in Y_m$ tetszőleges. Ekkor

$$T_y(Ax_y - b) = T_y(A_c x_y - b_c) + T_y(A - A_c)x_y + T_y(b_c - b).$$

Mivel

$$|T_y(A - A_c)x_y| \leq \Delta |x_y|,$$

ezért

$$T_y(A - A_c)x_y \geq -\Delta |x_y|,$$

és ugyanígy, mivel

$$|T_y(b_c - b)| \leq \delta,$$

ezért

$$T_y(b_c - b) \geq -\delta,$$

és így

$$\begin{aligned} T_y(Ax_y - b) &\geq T_y(A_c x_y - b_c) - \Delta |x_y| - \delta = \\ &= T_y(A_c(x_y^{(1)} - x_y^{(2)}) - b_c) - \Delta |x_y^{(1)} - x_y^{(2)}| - \delta \geq \\ &\geq T_y(A_c(x_y^{(1)} - x_y^{(2)}) - b_c) - \Delta(x_y^{(1)} + x_y^{(2)}) - \delta. \end{aligned}$$

Ha felbontjuk a zárójeleket és kiemeljük $x_y^{(1)}$-t és $x_y^{(2)}$-t, akkor azt kapjuk, hogy

$$T_y(Ax_y - b) \geq T_y((A_c - T_y\Delta)x_y^{(1)} - (A_c + T_y\Delta)x_y^{(2)} - (b_c + T_y\delta)).$$

Mivel $x_y^{(1)}, x_y^{(2)}$ megoldása a (3.1) renszernek, ezért

$$T_y(Ax_y - b) \geq T_y((A_c - T_y\Delta)x_y^{(1)} - (A_c + T_y\Delta)x_y^{(2)} - (b_c + T_y\delta)) = 0,$$

ami igazolja (3.6)-ot.
Ezt felhasználva megmutatjuk, hogy ha $\lambda_y \geq 0$ és $y \in Y_m$, akkor a

$$\begin{aligned} \sum_{y \in Y_m} \lambda_y A x_y &= b, \\ \sum_{y \in Y_m} \lambda_y &= 1 \end{aligned} \tag{3.7}$$

lineáris egyenletrendszernek létezik megoldása. A Farkas-lemma szerint elég azt megmutatni, hogy $\forall p \in \mathbb{R}^m$, $p_0 \in \mathbb{R}$ esetén ha

$$p^T A x_y + p_0 \geq 0 \quad \forall y \in Y_m, \tag{3.8}$$

akkor

$$p^T b + p_0 \geq 0. \tag{3.9}$$

Tegyük fel tehát, hogy $p \in \mathbb{R}^m$ és $p_0 \in \mathbb{R}$ kielégíti (3.8)-at. Definiáljuk $y \in Y_m$-t a következő módon

$$y_i = \begin{cases} -1, & \text{ha } p_i \geq 0, \\ 1 & \text{különben,} \end{cases}$$

$(i = 1, 2, ..., m)$. Mivel $p = -T_y|p|$ és $T_y = T_y^T$, ezért

$$p^T b + p_0 = -|p|^T T_y b + p_0.$$

(3.6) miatt

$$p^T b + p_0 \geq -|p|^T T_y A x_y + p_0 = p^T A x_y + p_0.$$

Végül (3.8) miatt

$$p^T b + p_0 \geq p^T A x_y + p_0 \geq 0,$$

ami igazolja (3.9)-et. Így ha $\lambda_y \geq 0$ és $y \in Y_m$, akkor a (3.7) egyenletrendszernek létezik megoldása.
Legyen

$$x = \sum_{y \in Y_m} \lambda_y x_y,$$

ekkor (3.7) miatt $Ax = b$ és

$$x \in Conv\{x_y : y \in Y_m\} = Conv\{x_y^{(1)} - x_y^{(2)} : y \in Y_m\},$$

és ezzel a tétel bizonyítása teljes. □

A következőkben megnézzük, hogy mit is mond valójában az imént belátott tétel. Ha $y_i = 1$, akkor az $A_c - T_y\Delta$ és az $A_c + T_y\Delta$ $i$-edik sora

megegyezik $\underline{A}$ és $\overline{A}$ $i$-edik sorával, és $(b_c + T_y\delta)_i = \overline{b}_i$. Ez azt jelenti, hogy ebben az esetben (3.1) $i$-edik egyenlete a következő

$$(\underline{A}x^{(1)} - \overline{A}x^{(2)})_i = \overline{b}_i. \tag{3.10}$$

Ugyanígy, ha $y_i = -1$, akkor

$$(\overline{A}x^{(1)} - \underline{A}x^{(2)})_i = \underline{b}_i. \tag{3.11}$$

Tehát $\forall y \in Y_m$-re a (3.1) rendszerek családja megegyezik az olyan rendszerek családjával, ahol az $i$-edik egyenlet vagy a (3.10), vagy a (3.11) alakban van, $i = 1, ..., m$. A különböző ilyen rendszerek száma pontosan $2^q$, ahol a $q$ a $(\Delta, \delta)$ mátrix nemnulla sorainak számát jelöli. Így a megoldandó rendszerek száma exponenciális, ezért az előző tétel a gyakorlatban csak akkor használható, ha $q$ viszonylag kicsi.

Most megmutatjuk, hogy hogyan lehet konstruálni tetszőleges $A \in \mathbf{A}$ és $b \in \mathbf{b}$ esetén az $Ax = b$ azon megoldását, amelyik a $Conv\{x_y^{(1)} - x_y^{(2)} : y \in Y_m\}$ halmazban van. Ehhez az $Y_m$ elemeinek egy speciális sorrendjére lesz szükség, amit indukcióval definiálunk a következőképpen.

1. Az $Y_1$ elemeinek sorrendje legyen a következő: $(-1)$, $(1)$.

2. Ha az $Y_j$ sorrendje $y^{(1)}, y^{(2)}, ..., y^{(2^j)}$, akkor az $Y_{j+1}$ sorrendje legyen

$$\begin{pmatrix} y^{(1)} \\ -1 \end{pmatrix}, \ldots, \begin{pmatrix} y^{(2^j)} \\ -1 \end{pmatrix}, \begin{pmatrix} y^{(1)} \\ 1 \end{pmatrix}, \ldots, \begin{pmatrix} y^{(2^j)} \\ 1 \end{pmatrix}.$$

Továbbá egy $z^{(1)}, z^{(2)}, ..., z^{(2h)}$ páros elemszámú sorozatban a $z^{(j)}$, $z^{(j+h)}$ $j \leq h$ párokat konjugált pároknak nevezzük. Legyen minden $y \in Y_m$ esetén $x_y^{(1)}$, $x_y^{(2)}$ a (3.1) rendszer megoldása. Ekkor az algoritmus a következő.

1. Válasszunk egy tetszőleges $A \in \mathbf{A}$-t és $b \in \mathbf{b}$-t.

2. Az $((x_{-y}^{(1)} - x_{-y}^{(2)})^T, (A(x_{-y}^{(1)} - x_{-y}^{(2)}) - b)^T)^T$ vektorokat tegyük a nekik megfelelő $y$-ok $Y_m$-beli sorrendjébe.

3. Az aktuális sorban minden $x,\ x'$ konjugált párhoz legyen

$$\lambda = \begin{cases} \frac{x'_k}{x'_k - x_k}, & \text{ha } x'_k \neq x_k, \\ 1 & \text{különben,} \end{cases}$$

ahol $k$ az aktuális utolsó komponens indexe. Legyen

$$x := \lambda x + (1 - \lambda)x'.$$

4. Töröljük a sorozat második felét, majd a megmaradó részben töröljük a vektorok utolsó koordinátáját.

5. Ha egyetlen $x$ vektor maradt, akkor $x$ megoldása $Ax = b$-nek és

$$x \in Conv\{x_y^{(1)} - x_y^{(2)} : y \in Y_m\}.$$

Ellenkező esetben menjünk vissza a 3. lépésre.

Az algoritmus $2^m$ db $n+m$ hosszú vektorral indul, és minden lépésben megfelezi a vektorok számát, illetve eggyel csökkenti a dimenzióját. Így a végére egyetlen $x \in \mathbb{R}^n$ vektor marad.

A megoldhatóság ellenőrzését szolgáló rendszerek, azaz (3.1) száma általában exponenciális az **A** intervallummátrix sorában. Ez az eredmény valószínűleg lényegesen nem javítható a következő tétel miatt.

**3.26. Tétel.** *Az intervallum-együtthatós lineáris egyenletrendszerek megoldhatóságának ellenőrzése NP-nehéz feladat.*

Az állítás abból a tényből következik, hogy egy intervallummátrix regularitásának ellenőrzése NP-teljes. Ez nyilvánvalóan polinom időben visszavezethető az intervallum-együtthatós lineáris egyenletrendszerek megoldhatóságának kérdésére, ami így NP-nehéz.

# 4. fejezet

# Gauss-elimináció

## 4.1. Gauss-elimináció algoritmusa intervallummátrixokra

Legyen $\mathbf{A} = ([a]_{ij})$ intervallummátrix, $\mathbf{b} = ([b]_i)$ intervallumvektor. Feltesszük, hogy $A^{-1}$ létezik minden $A \in \mathbf{A}$ esetén. Keressük a

$$\Sigma = \{x : Ax = b, A \in \mathbf{A}, b \in \mathbf{b}\}$$

halmazt. Mivel ez a halmaz általában túl bonyolúlt, ezért ehelyett egy olyan intervallumvektort keresünk, ami ezt tartalmazza. A Gauss-eliminációt fogjuk alkalmazni az intervallum-együtthatós rendszerre. A kezdőtáblázatunk a következő:

$$\left[\begin{array}{ccc|c} [a]_{11} & \cdots & [a]_{1n} & [b]_1 \\ \vdots & & \vdots & \vdots \\ [a]_{n1} & \cdots & [a]_{nn} & [b]_n \end{array}\right].$$

Ha feltesszük, hogy $0 \notin [a]_{11}$, akkor az első eliminációs lépés után a következő táblázatot kapjuk:

$$\left[\begin{array}{cccc|c} [a]'_{11} & [a]'_{12} & \cdots & [a]'_{1n} & [b]'_1 \\ 0 & [a]'_{22} & \cdots & [a]'_{2n} & [b]'_2 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & [a]'_{n2} & \cdots & [a]'_{nn} & [b]'_n \end{array}\right],$$

ahol az első sor ugyanaz, mint az előző táblázat első sora, és az $i$-edik sort úgy kapjuk, hogy az előző tábla $i$-edik sorából kivonjuk az első sor $[a]_{i1}/[a]_{11}$-szeresét $2 \leq i \leq n$, azaz

$$\begin{array}{cc} [a]'_{1j} = [a]_{1j} & 1 \leq j \leq n, \\ [b]'_1 = [b]_1 & \\ [a]'_{ij} = [a]_{ij} - [a]_{1j}([a]_{i1}/[a]_{11}) & 2 \leq i, j \leq n, \\ [b]'_i = [b]_i - [b]_1([a]_{i1}/[a]_{11}) & 2 \leq i \leq n, \\ [a]'_{i1} = 0 & 2 \leq i \leq n. \end{array}$$

**4.1. Állítás.** *Az eredeti rendszer megoldáshalmaza része az új rendszer megoldáshalmazának, azaz*

$$\{x : Ax = b, A \in \mathbf{A}, b \in \mathbf{b}\} \subseteq \{y : A'y = b', A' \in \mathbf{A}', b' \in \mathbf{b}'\}.$$

**Bizonyítás:** Legyen $A = (a_{ij}) \in \mathbf{A}$ és $b = (b_i) \in \mathbf{b}$, és tekintsük az alábbi lineáris egyenletrenszert:

$$Ax = b.$$

Legyen $A' := (a'_{ij})$ és $b' := (b'_i)$, ahol

$$\begin{array}{cc} a'_{1j} = a_{1j} & 1 \leq j \leq n, \\ b'_1 = b_1 & \\ a'_{ij} = a_{ij} - a_{1j}(a_{i1}/a_{11}) & 2 \leq i, j \leq n, \\ b'_i = b_i - b_1(a_{i1}/a_{11}) & 2 \leq i \leq n, \\ a'_{i1} = 0 & 2 \leq i \leq n. \end{array}$$

Ismert, hogy az $A'y = b'$ lineáris egyenletrendszer megoldása ugyan az, mint az $Ax = b$ rendszeré. A tartalmazási monotonitás miatt $A' \in \mathbf{A}'$ és $b' \in \mathbf{b}'$, ami bizonyítja az állítást. □

Ha ezt a lépést $n-1$-szer elvégezzük, akkor az erdeti táblából egy felső háromszög alakút kapunk:

$$\left(\begin{array}{cccc|c} \widetilde{[a]}_{11} & \widetilde{[a]}_{12} & \cdots & \widetilde{[a]}_{1n} & \widetilde{[b]}_1 \\ & \widetilde{[a]}_{22} & \cdots & \widetilde{[a]}_{2n} & \widetilde{[b]}_2 \\ & & \ddots & \vdots & \vdots \\ & & & \widetilde{[a]}_{nn} & \widetilde{[b]}_n \end{array}\right),$$

melyre igaz, hogy

$$\{x : Ax = b, A \in \mathbf{A}, b \in \mathbf{b}\} \subseteq \{\widetilde{x} : \widetilde{A}\widetilde{x} = \widetilde{b}, \widetilde{A} \in \widetilde{\mathbf{A}}, \widetilde{b} \in \widetilde{\mathbf{b}}\}.$$

Legyen

$$[x]_n := \frac{\widetilde{[b]}_n}{\widetilde{[a]}_{nn}},$$

$$[x]_i := \frac{\widetilde{[b]}_i - \sum_{j=i+1}^{n} \widetilde{[a]}_{ij}[x]_j}{\widetilde{[a]}_{ii}}, \qquad 1 \leq i \leq n-1.$$

Ekkor $\mathbf{x} := ([x]_i)$ intervallumvektor esetén

$$\Sigma = \{x : Ax = b, A \in \mathbf{A}, b \in \mathbf{b}\} \subseteq \mathbf{x}.$$

A következőkben a Gauss-eliminációval kapott intervallumvektor néhány tulajdonságával foglalkozunk, majd megnézzük, hogy milyen feltételek mellett hajtható végre. Azt már most megjegyezzük, hogy ha speciálisan $A = (a_{ij})$ reguláris pontmátrix, akkor a Gauss-elimináció a részleges főelemkiválasztással minden jobb oldali intervallumvektor esetén végrehajtható.
Legyen

$$g : \mathbb{R}^{n\times n} \times \mathbb{R}^n \to \mathbb{R}^n$$

olyan leképezés, ami egy reguláris $A$ mátrixhoz és egy tetszőleges $b$ vektorhoz az $Ax = b$ lineáris egyenletrendszer részleges főelemkiválasztásos Gauss-eliminációval kapott megoldását rendeli, azaz

$$x = g(A, b).$$

A $g$ leképezés egyértelmű, de több kifejezése is lehet. Például teljes főelemkiválastás esetén ugyanazt az értéket kapjuk, mint részleges főelemkiválasztásnál, de a kifejezéps más pivotelemet választ. Tehát $g$ kifejezése függ attól is, hogy a Gauss-elimináció során hogy választjuk a pivotelemeket. A következő állításban szereplő tulajdonságok függetlenek a pivotelemek választásától.

**4.2. Állítás.** *Legyen* $g(\mathbf{A}, \mathbf{b})$ *a fent definiált leképezés intervallumkiértékelése. Az* $\mathbf{x} = g(\mathbf{A}, \mathbf{b})$ *intervallumvektor a fent leírt módon, Gauss-eliminációval kiszámítható.*

1. *Legyen* $\mathbf{A}, \mathbf{B} \in \mathbb{IR}^{n\times n}$ *és* $\mathbf{a}, \mathbf{b} \in \mathbb{IR}^n$. *Továbbá tegyük fel, hogy* $\mathbf{A} \subseteq \mathbf{B}$ *és* $\mathbf{a} \subseteq \mathbf{b}$. *Ekkor*
$$g(\mathbf{A}, \mathbf{a}) \subseteq g(\mathbf{B}, \mathbf{b}).$$
2. *Legyen* $A \in \mathbb{R}^{n\times n}$ *és* $\mathbf{b} = \mathbf{u} + \mathbf{v} \in \mathbb{IR}^n$. *Ekkor*
$$g(A, \mathbf{b}) = g(A, \mathbf{u}) + g(A, \mathbf{v}).$$
3. *Legyen* $A \in \mathbb{R}^{n\times n}$ *és* $\mathbf{b} \in \mathbb{IR}^n$. *Ekkor*
$$A^{-1}\mathbf{b} \subseteq g(A, \mathbf{b}).$$
4. *Legyen* $A \in \mathbb{R}^{n\times n}$ *és* $\mathbf{a}, \mathbf{b} \in \mathbb{IR}^n$. *Továbbá tegyük fel, hogy létezik* $\alpha \geq 0$, *hogy* $d(\mathbf{a}) \leq \alpha d(\mathbf{b})$. *Ekkor*
$$d(g(A, \mathbf{a})) \leq \alpha d(g(A, \mathbf{b})).$$

**Bizonyítás:**

1. A tartalmazási monotonitás miatt triviális.
2. Mivel $A \in \mathbb{R}^{n\times n}$ és tudjuk, hogy $a([b] + [c]) = a[b] + a[c]$ $\forall a \in \mathbb{R}, [b], [c] \in \mathbb{IR}$, ezért ha ezt a Gauss-elimináció képleteibe beírjuk, akkor megkapjuk az állítást.
3. Ismeretes, hogy ha $f_1$ és $f_2$ az $f$ függvény két kifejezése, melyekre $f_1$-ben a változó pontosan egyszer fordul elő, míg $f_2$-ben $m$-szer, akkor $f_1([x]) \subseteq f_2([x])$. Ez igaz többváltozós függvényekre is. Tekintsük az $i$-edik ($1 \leq i \leq n$) komponensét $A^{-1}\mathbf{b}$-nek és $g(A, \mathbf{b})$-nek. A Gauss-elimináció képleteiben a $\mathbf{b}$ intervallumvektor komponensei többször is előfordulnak, míg $A^{-1}\mathbf{b}$ $i$-edik komponensének kiszámítása során csak egyszer.
4. Ismeretes, hogy $d([a] \pm [b]) = d([a]) + d([b])$ és $d(a[b]) = |a|d([b])$ minden $a \in \mathbb{R}$ és $[a], [b] \in \mathbb{IR}$ esetén. Valamint feltettük, hogy létezik $\alpha \geq 0$, amelyre $d(\mathbf{a}) \leq \alpha d(\mathbf{b})$. Ezeket a Gauss-elimináció algoritmusában használva rögtön megkapjuk az állítást. □

## 4.2. Gauss-elimináció elvégezhetősége

Most térjünk rá a Gauss-elimináció elvégezhetőségének kérdésére. A következő tétel az 1 illetve a 2-dimenziós esetről szól.

**4.3. Tétel.** *Legyen $1 \leq n \leq 2$, és tegyük fel, hogy $\mathbf{A} = ([a]_{ij}) \in \mathbb{IR}^{n\times n}$ nem tartalmaz szinguláris mátrixot. Ekkor a Gauss-elimináció algoritmusa elvégezhető.*

**Bizonyítás:**

1. $n = 1$ eset: Ebben az esetben $\mathbf{A} = [a]_{11}$ és a tétel feltétele ekvivalens azzal, hogy $0 \notin [a]_{11}$, ami bizonyítja az állítást.

2. $n = 2$ eset: Az egyenletrendszerünk a következő:

$$\begin{pmatrix} [a]_{11} & [a]_{12} \\ [a]_{21} & [a]_{22} \end{pmatrix} \begin{pmatrix} [x]_1 \\ [x]_2 \end{pmatrix} = \begin{pmatrix} [b]_1 \\ [b]_2 \end{pmatrix}.$$

Ekkor $[a]_{11}$ és $[a]_{21}$ közül legalább az egyik nem tartalmazza a 0-t, mert ellenkező esetben létezne $A \in \mathbf{A}$, ami szinguláris. Esetleges sorcserével elérhetjük, hogy $0 \notin [a]_{11}$. A Gauss-elimináció szerint

$$[a]'_{22} = [a]_{22} - (1/[a]_{11})[a]_{21}[a]_{12}.$$

Tekinthetjük $[a]'_{22}$-t egy $f$ függvény intervallumaritmetikai kiértékelésének, ahol az $f$ változói $a_{11}, a_{12}, a_{21}$ és $a_{22}$,

$$f(a_{11}, a_{12}, a_{21}, a_{22}) = a_{22} - (1/a_{11})a_{21}a_{12}. \tag{4.1}$$

Mivel feltettük, hogy minden $A \in \mathbf{A}$-ra

$$\det(A) = a_{11}a_{22} - a_{21}a_{12} \neq 0,$$

ezért

$$f(a_{11}, a_{12}, a_{21}, a_{22}) = (1/a_{11}) \det(A) \neq 0.$$

Az intervallumkiértékelés a pontos értéket adja, ha $a_{11}$-et $[a]_{11}$-gyel, $a_{12}$-t $[a]_{12}$-vel, $a_{21}$-et $[a]_{21}$-gyel és $a_{22}$-t $[a]_{22}$-vel helyettesítjük, mivel minden változó pontosan egyszer fordul elő a (4.1) kifejezésben. Tehát $0 \notin [a]'_{22}$, ami azt jelenti, hogy a Gauss-elimináció elvégezhető. □

A fenti bizonyítás $n \geq 3$ esetre nem általánosítható. A fejezet további részében szeretnénk megkapni az intervallummátrixok egy olyan osztályát, amelyre a Gauss-elimináció esetleges sorcserékkel mindig elvégezhető. Mostantól az intervallumokat nem a kezdő és végpontjukkal adjuk meg, hanem a középpontjával és a sugarával, vagy más néven a félszélességével. Azaz $[a] = [\underline{a}, \overline{a}]$ a következő alakban is felírható:

$$[a] = [a - r, a + r] =: \langle a, r \rangle,$$

ahol

$$a = \frac{1}{2}(\underline{a} + \overline{a}), \qquad r = \frac{1}{2}d([a]) = \frac{1}{2}(\overline{a} - \underline{a}).$$

Könnyen igazolható, hogy ha $[a] = \langle a, r \rangle, [b] = \langle b, s \rangle \in \mathbb{IR}$, akkor

$$[a] \pm [b] = \langle a \pm b, r + s \rangle.$$

A szorzás esetében csak a következő egyenlőségre lesz szükségünk:

$$[-r, r][-s, s] = \langle 0, r \rangle \langle 0, s \rangle = \langle 0, rs \rangle.$$

Tegyük fel, hogy $0 \notin [a] = \langle a, r \rangle$. Mivel

$$\frac{1}{[a]} = \left[\frac{1}{a + r}, \frac{1}{a - r}\right] = \left[\frac{a}{a^2 - r^2} - \frac{r}{a^2 - r^2}, \frac{a}{a^2 - r^2} + \frac{r}{a^2 - r^2}\right],$$

ezért

$$\frac{1}{[a]} = \left\langle \frac{a}{a^2 - r^2}, \frac{r}{a^2 - r^2} \right\rangle.$$

Az $[a]$ abszolútértékét a következőképpen számolhatjuk:

$$|[a]| = \max\{\underline{a}, \overline{a}\} = |a| + r.$$

Továbbá az is igaz, hogy

$$0 \notin [a] \Leftrightarrow |a| - r > 0.$$

És végül

$$[a] = \langle a, r \rangle \subseteq \langle 0, |[a]| \rangle = \langle 0, |a| + r \rangle.$$

**4.4. Lemma.** *Legyenek* $[a] = \langle a, r_a\rangle, [b] = \langle b, r_b\rangle, [c] = \langle c, r_c\rangle$ *és* $[d] = \langle d, r_d\rangle$ *valós intervallumok. Továbbá tegyük fel, hogy* $0 \notin [d]$. *Ekkor*

$$[z] = \langle z, r_z\rangle = [a] - \frac{1}{[d]}[b][c]$$

*esetén*

$$|a| - r_a - \frac{|[b]||[c]|}{|d| - r_d} \leq |z| - r_z.$$

**Bizonyítás:** A tartalmazási monotonitás miatt

$$[z] = \langle z, r_z\rangle = [a] - [b][c]\frac{1}{[d]} \subseteq [a] - \langle 0, |[b]|\rangle\langle 0, |[c]|\rangle \left\langle \frac{d}{d^2 - r_d^2}, \frac{r_d}{d^2 - r_d^2} \right\rangle \subseteq$$

$$\subseteq \langle a, r_a\rangle - \left\langle 0, |[b]||[c]|\frac{|d|}{d^2 - r_d^2} + |[b]||[c]|\frac{r_d}{d^2 - r_d^2} \right\rangle =$$

$$= \langle a, r_a\rangle - \left\langle 0, |[b]||[c]|\frac{1}{|d| - r_d} \right\rangle =$$

$$= \left\langle a, r_a + |[b]||[c]|\frac{1}{|d| - r_d} \right\rangle =: \langle a, r_6\rangle.$$

Mivel $[z] \subseteq \langle a, r_6\rangle$, ezért

$$|a| - |z| \leq |a - z| \leq r_6 - r_z.$$

ezt átrendezve

$$|z| - r_z \geq |a| - r_6 = |a| - r_a - |[b]||[c]|\frac{1}{|d| - r_d},$$

és ez volt az állítás. □

**4.5. Definíció.** *Legyen* $B = (b_{ij}) \in \mathbb{R}^{n\times n}$. *Ekkor* $B$ *egy M-mátrix, ha*

1. $b_{ij} \leq 0$, *ha* $i \neq j$ *és*
2. $B^{-1} \geq 0$.

Ismeretes, hogy a definíció második feltétele ekvivalens azzal, hogy $\exists u = (u_i) \in \mathbb{R}^n$, melyre $u_i > 0$, $1 \leq i \leq n$ és

$$Bu > 0.$$

Továbbá azt is tudjuk, hogy egy M-mátrix diagonális elemei mindig pozitívak.

**4.6. Tétel.** *Legyen* $\mathbf{A} = ([a]_{ij}) \in \mathbb{IR}^{n\times n}$ *és* $[a]_{ij} = \langle a_{ij}, r_{ij}\rangle$, $1 \leq i, j \leq n$. *Továbbá legyen* $B = (b_{ij}) \in \mathbb{R}^{n\times n}$, *melyre*

$$b_{ij} := \begin{cases} |a_{ii}| - r_{ii}, & \text{ha } i = j \\ -|[a]_{ij}| & \text{különben.} \end{cases}$$

*Ha B M-mátrix, akkor a Gauss-elimináció elvégezhető* $\mathbf{A}$ *intervallummátrixra sor- és oszlopcserék nélkül.*

**Bizonyítás:** Mivel $B$ M-mátrix, ezért $\exists u = (u_i) \in \mathbb{R}^n$, melyre $u_i > 0$, $1 \leq i \leq n$ és $Bu > 0$. Ez azt jelenti, hogy

$$(|a_{ii}| - r_{ii})u_i > \sum_{j=1, j\neq i}^{n} |[a]_{ij}|u_j,$$

$1 \leq i \leq n$. Mivel a jobb oldal nemnegatív és $u_i > 0$, ezért $i = 1$-re $|a_{11}| - r_{11} > 0$, amiből az következik, hogy $0 \notin [a]_{11}$. Tehát a Gauss-elimináció első lépését el lehet végezni, és így megkapjuk az $\mathbf{A}' = ([a]'_{ij})$ intervallummátrixot. Ha megmutatjuk, hogy a tétel feltételei fennállnak az $\widetilde{\mathbf{A}}' = (\widetilde{[a]}'_{ij}) \in \mathbb{IR}^{(n-1)\times(n-1)}$-re, melyre

$$\widetilde{[a]}'_{ij} = [a]'_{ij} = \langle a'_{ij}, r'_{ij}\rangle, \qquad 2 \leq i, j \leq n,$$

akkor teljes indukcióval beláttuk az állítást.
Legyen $i \geq 2$, ekkor

$$\sum_{j=2, j\neq i}^{n} |[a]'_{ij}|u_j = \sum_{j=2, j\neq i}^{n} \left| [a]_{ij} - [a]_{1j}\frac{[a]_{i1}}{[a]_{11}} \right| u_j \leq$$

$$\leq \sum_{j=2,j\neq i}^{n} |[a]_{ij}|u_j + |[a]_{i1}| \left|\frac{1}{[a]_{11}}\right| \sum_{j=2,j\neq i}^{n} |[a]_{1j}|u_j.$$

Vegyük észre, hogy a fenti képletben $j$ index kettőtől megy $n$-ig. Tekintsük ismét a bizonyítás elején szereplő egyenlőtlenséget $i = 1$-re és a szumma $k$-adik tagját, ahol $k \geq 2$, vigyük át a másik oldalra. Ekkor

$$\sum_{j=2,j\neq k}^{n} |[a]_{1j}|u_j < (|a_{11}| - r_{11})u_1 - |[a]_{1k}|u_k. \tag{4.2}$$

Továbbá

$$\left|\frac{1}{[a]_{11}}\right| = \left|\left\langle \frac{a_{11}}{a_{11}^2 - r_{11}^2}, \frac{r_{11}}{a_{11}^2 - r_{11}^2} \right\rangle\right| = \frac{|a_{11}|}{a_{11}^2 - r_{11}^2} + \frac{r_{11}}{a_{11}^2 - r_{11}^2} = \frac{1}{|a_{11}| - r_{11}}.$$

A legutóbbi összefüggést és (4.2)-t $k = i$ helyettesítéssel felhasználva kapjuk, hogy

$$\sum_{j=2,j\neq k}^{n} |[a]'_{ij}|u_j \leq \sum_{j=2,j\neq i}^{n} |[a]_{ij}|u_j + |[a]_{i1}| \frac{1}{|a_{11}| - r_{11}} ((|a_{11}| - r_{11})u_1 - |[a]_{1i}|u_i).$$

Ha a zárójelet felbontjuk, és az első tagját egyszerűsítjük $|a_{11}| - r_{11}$-gyel, akkor azt be tudjuk vinni a szummába, és az alábbi becslést kapjuk

$$\sum_{j=2,j\neq i}^{n} |[a]'_{ij}|u_j \leq \sum_{j=1,j\neq i}^{n} |[a]_{ij}|u_j - \frac{|[a]_{i1}||[a]_{1i}|}{|a_{11}| - r_{11}} u_i.$$

Erre megint alkalmazhatjuk az első egyenlőtlenséget, ekkor

$$\sum_{j=2,j\neq i}^{n} |[a]'_{ij}|u_j < u_i \left( |a_{ii}| - r_{ii} - \frac{|[a]_{i1}||[a]_{1i}|}{|a_{11}| - r_{11}} \right).$$

Végül ha az előző lemmát az $[a] = [a]_{ii}$, $[b] = [a]_{i1}$, $[c] = [a]_{1i}$ és $[d] = [a]_{11}$ intervallumokra alkalmazzuk, akkor

$$[z] = [a] - \frac{1}{[d]}[b][c] = [a]_{ii} - \frac{1}{[a]_{11}}[a]_{i1}[a]_{1i} = [a]'_{ii},$$

és így

$$|a_{ii}| - r_{ii} - \frac{|[a]_{i1}||[a]_{1i}|}{|a_{11}| - r_{11}} \leq |a'_{ii}| - r'_{ii}.$$

Ezzel tovább tudunk becsülni, és a következőre jutunk:

$$\sum_{j=1, j\neq i}^{n} |[a]'_{ij}|u_j < (|a'_{ii}| - r'_{ii})u_i,$$

és ezt kellett belátnunk. □

Az intervallummátrixok egy igen fontos osztálya teljesíti az előző tétel feltételeit.

**4.7. Definíció.** *Legyen* $\mathbf{A} = ([a]_{ij}) \in \mathbb{IR}^{n\times n}$ *és* $[a]_{ij} = \langle a_{ij}, r_{ij}\rangle$, $1 \leq i, j \leq n$. *Az* $\mathbf{A}$ *intervallummátrix szigorúan diagonálisan domináns, ha*

$$|a_{ii}| - r_{ii} > \sum_{j=1, j\neq i}^{n} |[a]_{ij}|, \qquad 1 \leq i \leq n.$$

A definícióból rögtön következik, hogy egy szigorúan diagonálisan domináns $\mathbf{A}$ intervallummátrix diagonális elemei nem tartalmazhatják a 0-t. Továbbá az is látszik, hogy minden valós $\widehat{A} = (\widehat{a}_{ij}) \in \mathbf{A}$ mátrix esetén

$$|\widehat{a}_{ii}| > \sum_{j=1, j\neq i}^{n} |\widehat{a}_{ij}|, \qquad 1 \leq i \leq n.$$

Azaz minden valós $\widehat{A} \in \mathbf{A}$ mátrix szigorúan diagonálisan domináns a hagyományos értelemben, ezáltal nemszinguláris.
Egy szigorúan diagonálisan domináns $\mathbf{A}$ intervallummátrix teljesíti az előző tétel feltételét is, azaz a megfelelő $B$ mátrix egy M-mátrix az $u = (u_i)$, $u_i = 1$, $1 \leq i \leq n$ választással. Tehát kimondhatjuk a következő következményt.

**4.8. Következmény.** *Legyen* $\mathbf{A}$ *szigorúan diagonálisan domináns intervallummátrix. Ekkor a Gauss-elimináció elvégezhető az* $\mathbf{A}$ *intervallummátrixra sor- és oszlopcserék nélkül.*

## 4.3. Gauss-elimináció tridiagonális intervallummátrixokra

Legyen

$$\mathbf{A} = \begin{pmatrix} [a]_1 & [c]_1 & & & \\ [b]_2 & [a]_2 & [c]_2 & & 0 \\ & \ddots & \ddots & \ddots & \\ 0 & & \ddots & \ddots & [c]_{n-1} \\ & & & [b]_n & [a]_n \end{pmatrix}.$$

**4.9. Tétel.** *Legyen az* $\mathbf{A}$ *intervallummátrix tridiagonális, és tegyük fel, hogy*

$$\begin{aligned} [a]_i &= \langle a_i, r_i \rangle, && 1 \le i \le n, \\ [b]_i &= \langle b_i, s_i \rangle \neq 0, && 2 \le i \le n, \\ [c]_i &= \langle c_i, t_i \rangle \neq 0, && 1 \le i \le n-1. \end{aligned}$$

*Továbbá tegyük fel, hogy*

$$\begin{aligned} |a_1| - r_1 &> |[c]_1|, \\ |a_i| - r_i &\ge |[b]_i| + |[c]_i|, \quad 2 \le i \le n-1, \ . \\ |a_n| - r_n &> |[b]_n|. \end{aligned}$$

*Ekkor a Gauss-elimináció elvégezhető* $\mathbf{A}$ *intervallummátrixra sor- és oszlopcserék nélkül.*

**Bizonyítás:** Írjuk fel az előző tételbeli $B$ mátrixot ebben az esetben.

$$B = \begin{pmatrix} |a_1| - r_1 & -|[c]_1| & & & \\ -|[b]_2| & |a_2| - r_2 & -|[c]_2| & & 0 \\ & \ddots & \ddots & \ddots & \\ 0 & & \ddots & \ddots & -|[c]_{n-1}| \\ & & & -|[b]_n| & |a_n| - r_n \end{pmatrix}.$$

Tehát $B$ olyan diagonálisan domináns tridiagonális valós mátrix, melyre teljesül, hogy az első és az utolsó sorban szigorú egyenlőtlenség van, azaz M-mátrix és az előző tétel alkalmazható $\mathbf{A}$-ra. □

## 4.4. Gauss-elimináció nem diagonálisan domináns mátrixokra

Ebben a fejezetben megnézzük, hogy mit lehet tenni abban az esetben, ha a lineáris egyenletrendszer $\mathbf{A}$ mátrixa nem szigorúan diagonálisan domináns. Az ötlet az, hogy alkalmazunk egy olyan transzformációt a rendszerre, ami szigorúan diagonálisan dominánssá transzformálja az $\mathbf{A}$ mátrixot.
Legyen $\mathbf{A} = ([a]_{ij}) \in \mathbb{IR}^{n\times n}$ és $[a]_{ij} = \langle a_{ij}, r_{ij}\rangle$, $1 \le i, j \le n$. Tegyük fel továbbá, hogy minden $A \in \mathbf{A}$ valós mátrix esetén létezik $A^{-1}$. Legyen $A_c := (a_{ij}) \in \mathbb{R}^{n\times n}$. Ez invertálható, hiszen $A_c \in \mathbf{A}$. Szorozzuk be az egyenlet mindkét oldalát $A_c^{-1}$-zel, ekkor az

$$\widetilde{\mathbf{A}} := A_c^{-1}\mathbf{A}$$

és

$$\widetilde{\mathbf{b}} := A_c^{-1}\mathbf{b}$$

jelöléseket használva az új egyenletrendszerünk

$$\widetilde{\mathbf{A}}x = \widetilde{\mathbf{b}}.$$

Ekkor

$$\{x : Ax = b, A \in \mathbf{A}, b \in \mathbf{b}\} \subseteq \{y : \widetilde{A}y = \widetilde{b}, \widetilde{A} \in \widetilde{\mathbf{A}}, \widetilde{b} \in \widetilde{\mathbf{B}}\}.$$

Ugyanis legyen az $x$ egy eleme a baloldali halmaznak, azaz létezik $A \in \mathbf{A}$ és $b \in \mathbf{b}$, hogy

$$Ax = b.$$

Ekkor

$$A_c^{-1}Ax = A_c^{-1}b,$$

és mivel

$$A_c^{-1}A \in \widetilde{\mathbf{A}}, \quad A_c^{-1}b \in \widetilde{\mathbf{b}},$$

az állítást beláttuk.
Ha az $\mathbf{A}$ mátrix elemei nem túl szélesek, akkor az $\widetilde{\mathbf{A}}$ erősen diagonálisan domináns és a Gauss-elimináció elvégezhető. Ha ugyanis $d(\mathbf{A}) = 0$, akkor $\widetilde{\mathbf{A}} = I$ és ekkor $\widetilde{\mathbf{A}}$ persze erősen diagonálisan domináns. Ha az $\mathbf{A}$

elemeinek szélessége nem túl nagy, akkor $\widetilde{\mathbf{A}}$ nem sokban fog eltérni az egységmátrixtól.
Azonban az $\widetilde{\mathbf{A}}$ intervallummátrix erősen diagonális dominanciája nem csak az $\mathbf{A}$ mátrix elemeinek szélességétől függ. Legyen

$$[a]_{ij} = \langle a_{ij}, r_{ij} \rangle, \quad 1 \le i, j \le n,$$

$$\mathbf{D} = ([d]_{ij}), \quad [d]_{ij} = \langle 0, r_{ij} \rangle, \quad 1 \le i, j \le n.$$

Ekkor

$$\widetilde{\mathbf{A}} = A_c^{-1}\mathbf{A} = A_c^{-1}(A_c + \mathbf{D}) =$$
$$= I + A_c^{-1}\mathbf{D} = I + \mathbf{H},$$

ahol $\mathbf{H} = A_c^{-1}\mathbf{D}$. Mivel

$$\|\mathbf{H}\| \le \|A_c^{-1}\| \cdot \|\mathbf{D}\| = \frac{1}{2}\|A_c^{-1}\| \cdot \|d(\mathbf{A})\| =$$

$$= \frac{1}{2}\|A_c^{-1}\| \cdot \|A_c\| \cdot \frac{\|d(\mathbf{A})\|}{\|A_c\|} = \frac{1}{2} cond(A_c) \frac{\|d(\mathbf{A})\|}{\|A_c\|},$$

ezért az $\widetilde{\mathbf{A}}$ annál inkább diagonálisan domináns, minél kisebb az $A_c$ kondíciószáma.

**Példa:** Legyen

$$\mathbf{A} := \begin{pmatrix} \left[\frac{29}{30}, \frac{31}{30}\right] & \left[\frac{14}{30}, \frac{16}{30}\right] \\ \left[\frac{14}{30}, \frac{16}{30}\right] & \left[\frac{9}{30}, \frac{11}{30}\right] \end{pmatrix},$$

ekkor az $\mathbf{A}$ elemeinek szélessége $\frac{1}{15}$ és a középmátrixa

$$A_c = \begin{pmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{3} \end{pmatrix},$$

így $cond_1(A_c) = 27$ és $\|A_c\|_1 = 1.5$, ezért a fenti becslés alapján

$$\|\mathbf{H}_\mathbf{A}\| \le \frac{6}{5}.$$

Ugyanakkor legyen

$$\mathbf{B} := \begin{pmatrix} \left[\frac{59}{30}, \frac{61}{30}\right] & \left[-\frac{31}{30}, -\frac{29}{30}\right] \\ \left[-\frac{31}{30}, -\frac{29}{30}\right] & \left[\frac{59}{30}, \frac{61}{30}\right] \end{pmatrix}.$$

Vegyük észre, hogy a $\mathbf{B}$ intervallummátrix csak a közepében tér el az $\mathbf{A}$ intervallummátrixtól, a szélessége ugyanannyi.

$$B_c = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix},$$

így $cond_1(B_c) = 3$ és $\|A_c\|_1 = 3$, ezért a fenti becslés alapján

$$\|\mathbf{H_B}\| \leq \frac{1}{15}.$$

# 5. fejezet

# Megoldáshalmaz behatárolása reguláris esetben

Mint azt az előző fejezetben láttuk, a Gauss-eliminációt olyan intervallummátrixok esetén lehet jól használni, melyekben az elemek viszonylag keskenyek. Ebben a fejezetben két olyan eljárást ismertetünk, ami abban az esetben hatékony, amikor ezek az intervallumok viszonylag szélesek. Viszont a hátrányuk az, hogy több számolással járnak, mint a Gauss-elimináció. Először E. R. Hansen eredményét közöljük. Itt a bizonyításokra nem térünk ki, mivel a második módszer, melyet J. Rohn közölt, lényegében ugyanarra az eredményre jut, mint a Hansen-féle, de $2n$ db lineáris egyenletrendszer megoldása helyett csak egy mátrix invertálása szükséges.

## 5.1. E. R. Hansen módszere

Legyen $\mathbf{A} \in \mathbb{IR}^{n\times n}$ és $\mathbf{b} \in \mathbb{IR}^n$.

**5.1. Definíció.** *Egy intervallumot, intervallumvektort illetve intervallummátrixot centráltnak nevezünk, ha a centruma a* 0 *szám, vektor illetve mátrix. Egy intervallummátrixot az identitás körül centráltnak nevezünk, ha a centruma az identitásmátrix.*

Tegyük fel, hogy $\forall A \in \mathbf{A}$ reguláris. Ekkor az $\mathbf{A}x = \mathbf{b}$ intervallum-együtthatós lineáris egyenletrendszer megoldáshalmaza a követ-

kezőképpen adható meg:

$$\Sigma = \{x = A^{-1}b : A \in \mathbf{A}, b \in \mathbf{b}\}.$$

A pontos megoldáshalmaz helyett most is a legszűkebb olyan intervallumvektort keressük, ami azt tartalmazza.

Ha elvégezzük az intervallum-együtthatós lineáris egyenletrendszeren az előző fejezetben ismertetett transzformációt, akkor — mint azt láttuk — ha $\mathbf{A}$ és $\mathbf{b}$ elemei viszonylag szűkek, akkor csak kis mértékben növeli a megoldáshalmazt, ha viszont szélesek, akkor nagyon megnövelheti azt. Enélkül viszont az intervallumok szélessége általában nagyon gyorsan nő a megoldás során és a végső eredmény kevéssé használható lesz. Így az eredeti intervallum-együtthatós lineáris egyenletrendszer helyett tekinsük az

$$\widetilde{\mathbf{A}}x = \widetilde{\mathbf{b}}$$

intervallum-együtthatós lineáris egyenletrendszert, ahol

$$\widetilde{\mathbf{A}} := A_c^{-1}\mathbf{A}$$

és

$$\widetilde{\mathbf{b}} := A_c^{-1}\mathbf{b}.$$

Láttuk, hogy $\widetilde{\mathbf{A}}$ az identitás körül centrált, így

$$\widetilde{\mathbf{A}} = [I - \Delta, I + \Delta], \qquad \widetilde{\mathbf{b}} = [b_c - \delta, b_c + \delta].$$

**5.2. Tétel.** *Tegyük fel, hogy $\forall A \in \widetilde{\mathbf{A}}$ mátrix szigorúan diagonálisan domináns. (Ekkor $\widetilde{\mathbf{A}}$ nem tartalmaz szinguláris mátrixot.) Ekkor az alábbiak teljesülnek a megoldáshalmazt tartalmazó $\mathbf{x}$ intervallumvektorra.*

1. *$\mathbf{x}_i$ maximális értéke, ha az nemnegatív:*

$$\overline{\mathbf{x}}_i = e_i^T(I - \Delta)^{-1}s^{(i)},$$

*ahol*

$$s_j^{(i)} = \begin{cases} \overline{\widetilde{b}}_j, & ha\ j = i \\ |\widetilde{b}_j|, & ha\ j \neq i. \end{cases}$$

*2.* $\mathbf{x}_i$ *minimális értéke, ha az nemnegatív:*

$$\underline{\mathbf{x}}_i = \frac{1}{2((I-\Delta)^{-1})_{ii}-1} e_i^T (I-\Delta)^{-1} t^{(i)},$$

*ahol*

$$t_j^{(i)} = \begin{cases} \widetilde{\underline{b}}_j, & ha\ j = i \\ -|b_j|, & ha\ j \neq i. \end{cases}$$

*3.* $\mathbf{x}_i$ *maximális értéke, ha az negatív:*

$$\overline{\mathbf{x}}_i = \frac{1}{2((I-\Delta)^{-1})_{ii}-1} e_i^T (I-\Delta)^{-1} s^{(i)}.$$

*4.* $\mathbf{x}_i$ *minimális értéke, ha az negatív:*

$$\underline{\mathbf{x}}_i = e_i^T (I-\Delta)^{-1} t^{(i)}.$$

Megjegyezzük, hogy $\mathbf{x}_i$ maximális értéke csak úgy lehet negatív, ha $(b_c+\delta)_i < 0$, ugyanis az $\widetilde{\mathbf{A}}x = \widetilde{\mathbf{b}}$ intervallum-együtthatós lineáris egyenletrendszer megoldáshalmaza tartalmazza a $\widetilde{\mathbf{b}}$ intervallumvektort, mivel $I \in \widetilde{\mathbf{A}}$. Ezért ha $(b_c+\delta)_i \geq 0$, akkor $\overline{\mathbf{x}}_i \geq 0$.

Azt is megjegyezzük, hogy $s^{(i)}$ és $t^{(i)}$ kiszámítható elágazás nélkül, ugyanis ha $b_c > 0$, akkor

$$\max\{-(b_c-\delta)_j, (b_c+\delta)_j\} = (b_c)_j + \delta_j,$$

és

$$\min\{(b_c-\delta)_j, -(b_c+\delta)_j\} = -\max\{-(b_c-\delta)_j, (b_c+\delta)_j\} = -(b_c)_j + \delta_j.$$

## 5.2. J. Rohn módszere

Ugyanazokat a jelöléseket használjuk, mint az előző részben. Az előző tételben a szigorúan diagonális dominancia volt a regularitás elégséges feltétele. Az $[I-\Delta, I+\Delta]$ intervallummátrix akkor és csak akkor reguláris, ha

$$\varrho(\Delta) < 1,$$

ahol $\varrho(\Delta)$ a $\Delta$ spektrálsugara. Ebből az következik, hogy az

$$M = (I - \Delta)^{-1} = (m_{ij})$$

mátrix létezik és nemnegatív. Legyen

$$\underline{x}_i := \min_{x \in X} x_i,$$

$$\overline{x}_i := \max_{x \in X} x_i,$$

ahol $X$ az

$$[I - \Delta, I + \Delta]x = [b_c - \delta, b_c + \delta]$$

intervallum-együtthatós lineáris egyenletrendszer megoldáshalmaza.

**5.3. Tétel.** *Tegyük fel, hogy* $\varrho(\Delta) < 1$. *Ekkor* $\forall i = 1, 2, ..., n$*-re*

$$\underline{x}_i =$$
$$\min \left\{ m_{ii}(b_c + |b_c|)_i - (M(|b_c| + \delta))_i, -\frac{(M(|b_c| + \delta))_i + m_{ii}(b_c + |b_c|)_i}{2m_{ii} - 1} \right\},$$

$$\overline{x}_i =$$
$$\max \left\{ (M(|b_c| + \delta))_i + m_{ii}(b_c - |b_c|)_i, \frac{(M(|b_c| + \delta))_i + m_{ii}(b_c - |b_c|)_i}{2m_{ii} - 1} \right\}.$$

**Bizonyítás:** A tétel bizonyítása három részből áll.

1. Belátjuk, hogy minden $x \in X$ esetén

   $$x_i \leq \max\{\widetilde{x}_i, \nu_i \widetilde{x}_i\},$$

   ahol

   $$\widetilde{x}_i = (M(|b_c| + \delta))_i + m_{ii}(b_c - |b_c|)_i$$

   és

   $$\nu_i = \frac{1}{2m_{ii} - 1}.$$

2. Megmutatjuk, hogy $\widetilde{x}_i = x'_i$ és $\nu_i \widetilde{x}_i = x''_i$ valamely $x', x'' \in X$-re. Ebből az $\overline{x}_i$-ra vonatkozó állítás következik.

3. Megmutatjuk az $\underline{x}_i$-ra vonatkozó állítást.

1. Először lássuk be, hogy

$$M\Delta = \Delta M = M - I. \tag{5.1}$$

Ugyanis

$$M - I = (I-\Delta)^{-1} - (I-\Delta)^{-1}(I-\Delta) = (I-\Delta)^{-1}(I-(I-\Delta)) =$$
$$= (I-\Delta)^{-1}(I - I + \Delta) = M\Delta,$$

és

$$M - I = (I-\Delta)^{-1} - (I-\Delta)(I-\Delta)^{-1} = (I-(I-\Delta))(I-\Delta)^{-1} =$$
$$= (I - I + \Delta)(I-\Delta)^{-1} = \Delta M.$$

$\nu_i \in (0,1]$, ugyanis $m_{ii} \geq 1$, ezért $2m_{ii} - 1 \geq 1$, és így

$$\frac{1}{2m_{ii}-1} = \nu_i \in (0,1]. \tag{5.2}$$

Legyen $D$ diagonális mátrix, $j = 1, 2, ..., n$,

$$D_{jj} := \begin{cases} 1, & \text{ha } j \neq i \text{ és } (b_c)_j \geq 0, \\ -1, & \text{ha } j \neq i \text{ és } (b_c)_j < 0, \\ 1, & \text{ha } j = i. \end{cases}$$

És legyen

$$\widehat{b} := Db_c + \delta = \begin{pmatrix} |(b_c)_1| \\ \vdots \\ |(b_c)_{i-1}| \\ (b_c)_i \\ |(b_c)_{i+1}| \\ \vdots \\ |(b_c)_n| \end{pmatrix} + \delta.$$

Azaz $\widehat{b}$ a $|b_c| + \delta$ vektortól csak az $i$. koordinátájában tér el, ahol $(\overline{b}_c)_i$ lesz. Ekkor

$$\widetilde{x}_i = (M(|b_c| + \delta))_i + m_{ii}(b_c - |b_c|)_i = (M\widehat{b})_i. \tag{5.3}$$

Legyen $x \in X$ tetszőleges, azaz $\exists A \in [I - \Delta, I + \Delta]$ és $b \in [b_c - \delta, b_c + \delta]$, hogy $Ax = b$. Továbbá legyen

$$x' = Dx = \begin{pmatrix} |x_1| \\ \vdots \\ |x_{i-1}| \\ x_i \\ |x_{i+1}| \\ \vdots \\ |x_n| \end{pmatrix}.$$

Ekkor belátható, hogy

$$M(x' - |x|) + |x| \leq M\widehat{b}, \tag{5.4}$$

ugyanis

$$\begin{aligned} x'_i = x_i = b_i + ((I - A)x)_i \leq \\ \leq (b_c + \delta)_i + (\Delta|x|)_i = (\widehat{b} + \Delta|x|)_i, \end{aligned} \tag{5.5}$$

és $j \neq i$ esetén

$$\begin{aligned} x'_j = |x_j| \leq |b_j| + |((I - A)x)_j| \leq \\ \leq |b_c|_j + \delta_j + (\Delta|x|)_j = (\widehat{b} + \Delta|x|)_j. \end{aligned} \tag{5.6}$$

(5.5) és (5.6) alapján

$$x' \leq \widehat{b} + \Delta|x|.$$

Az egyenlet mindkét oldalát balról $M$-mel szorozva

$$Mx' \leq M\widehat{b} + M\Delta|x|.$$

Mivel $M\Delta = M - I$,

$$Mx' \leq M\widehat{b} + (M - I)|x|.$$

Ha $(M - I)|x|$-t átvisszük a másik oldalra megkapjuk (5.4)-t. Két eset van.

- Ha $x_i \geq 0$, akkor $x' = |x|$, és így (5.4) miatt

$$x_i = |x_i| \leq (M\widehat{b})_i = \widetilde{x}_i.$$

- Ha $x_i < 0$, akkor $x'_i = x_i$ és $|x_i| = -x_i$. (5.4) miatt

$$\begin{aligned}(M(x' - |x|))_i + |x_i| = 2m_{ii}x_i - x_i = \\ = (2m_{ii} - 1)x_i \leq (M\widehat{b})_i = \widetilde{x}_i.\end{aligned}$$

Ezért $x_i \leq \nu_i\widetilde{x}_i$, ami bizonyítja az első részt.

2. Legyen $x' := DM\widehat{b}$ és $x'' := DM(\widehat{b} - 2\nu_i\widetilde{x}_i\Delta e_i)$. Megmutatjuk, hogy $x', x'' \in X$ és, hogy $x'_i = \widetilde{x}_i$ és $x''_i = \nu_i\widetilde{x}_i$.

Először nézzük $x'$-t. Mivel $M\Delta = M - I$,

$$\begin{aligned}(I - D\Delta D)x' &= (I - D\Delta D)DM\widehat{b} = \\ &= DM\widehat{b} - D\Delta M\widehat{b} = \\ &= DM\widehat{b} - D(M - I)\widehat{b} = \\ &= DM\widehat{b} - DM\widehat{b} + D\widehat{b} = \\ &= D\widehat{b} = D(Db_c + \delta) = b_c + D\delta.\end{aligned}$$

Azaz

$$(I - D\Delta D)x' = b_c + D\delta. \tag{5.7}$$

Mivel

- $I - D\Delta D \in [I - \Delta, I + \Delta]$ és
- $b_c + D\delta \in [b_c - \delta, b_c + \delta]$,

ezért (5.7) miatt $x' \in X$ teljesül.

Most nézzük $x''$-t. Legyen $D'$ diagonális mátrix, ahol $D'_{ii} = -1$ és $D'_{jj} = D_{jj}$.

$$\begin{aligned}(I - D\Delta D')DM &= DM - D\Delta D'DM = \\ &= DM - D\Delta(I - 2e_ie_i^T)M = \\ &= DM - D\Delta M + D\Delta 2e_ie_i^TM = \\ &= DM - D(M - I) + D\Delta 2e_ie_i^TM = \\ &= DM - DM + D + 2D\Delta e_ie_i^TM = \\ &= D + 2D\Delta e_ie_i^TM.\end{aligned}$$

Ezt felhasználva, és hogy $\widetilde{x}_i = (M\widehat{b})_i = e_i^T M\widehat{b}$

$$\begin{aligned}
&(I - D\Delta D')x'' = (I - D\Delta D')DM(\widehat{b} - 2\nu_i\widetilde{x}_i\Delta e_i) = \\
&= (D + 2D\Delta e_i e_i^T M)(\widehat{b} - 2\nu_i\widetilde{x}_i\Delta e_i) = \\
&= D\widehat{b} - 2\nu_i\widetilde{x}_i D\Delta e_i + 2D\Delta e_i e_i^T M\widehat{b} - 4\nu_i\widetilde{x}_i D\Delta e_i e_i^T M\Delta e_i = \\
&= D\widehat{b} - 2\nu_i\widetilde{x}_i D\Delta e_i + 2D\Delta e_i\widetilde{x}_i - 4\nu_i\widetilde{x}_i D\Delta e_i e_i^T(M - I)e_i = \\
&= D\widehat{b} + 2\widetilde{x}_i D\Delta e_i(-\nu_i + 1 - 2\nu_i e_i^T(M - I)e_i) = \\
&= D\widehat{b} + 2\widetilde{x}_i D\Delta e_i\left(-\frac{1}{2m_{ii} - 1} + 1 - \frac{2(m_{ii} - 1)}{2m_{ii} - 1}\right) = \\
&= D\widehat{b} = b_c + D\delta.
\end{aligned}$$

Azaz

$$(I - D\Delta D')x'' = b_c + D\delta. \tag{5.8}$$

Mivel

- $I - D\Delta D' \in [I - \Delta, I + \Delta]$ és
- $b_c + D\delta \in [b_c - \delta, b_c + \delta]$,

ezért (5.8) miatt $x'' \in X$ teljesül.

A második pont igazolásához még azt kell belátni, hogy $x_i' = \widetilde{x}_i$ és $x_i'' = \nu_i\widetilde{x}_i$.

- Mivel $e_i^T D = e_i^T$, ezért

$$x_i' = e_i^T DM\widehat{b} = e_i^T M\widehat{b} = (M\widehat{b})_i = \widetilde{x}_i.$$

- $e_i^T D = e_i^T$ és (5.1) miatt

$$\begin{aligned}
x_i'' &= (DM\widehat{b})_i - (2\nu_i\widetilde{x}_i DM\Delta e_i)_i = \\
&= \widetilde{x}_i - 2\nu_i\widetilde{x}_i e_i^T D(M - I)e_i = \\
&= \widetilde{x}_i - 2\nu_i\widetilde{x}_i(m_{ii} - 1) = \\
&= \widetilde{x}_i - \frac{2\widetilde{x}_i(m_{ii} - 1)}{2m_{ii} - 1} = \nu_i\widetilde{x}_i.
\end{aligned}$$

Ezzel beláttuk a tétel maximumra vonatkozó állítását.

3. Tekintsük az $[I - \Delta, I + \Delta]x = [-b_c - \delta, -b_c + \delta]$ intervallum-együtthatós lineáris egyenletrendszer $X_0 = -X$ megoldáshalmazát. Ha az imént belátottakat erre alkalmazzuk, akkor megkapjuk a minimumra vonatkozó állítást. □

# 6. fejezet

# Megoldáshalmaz behatárolása általános esetben

## 6.1. Elméleti háttér

Egy általános módszert írunk le, mely megadja egy tetszőleges intervallum-együtthatós lineáris egyenletrendszer megoldáshalmazát tartalmazó legszűkebb intervallumvektort, vagy ad egy szinguláris mátrixot, mely eleme a rendszer baloldali mátrixának. Az alábbi meggondolások és az algoritmus ismét J. Rohn nevéhez fűződnek. Az alábbi állítások bizonyításai [8], [9], [10], [11] cikkekben találhatók. Tehát most az

$$\mathbf{A} = [A_c - \Delta, A_c + \Delta] \in \mathbb{IR}^{n \times n}$$

és a

$$\mathbf{b} = [b_c - \delta, b_c + \delta] \in \mathbb{IR}^n$$

intervallummátrixról és vektorról nem teszünk fel semmit.

A következőkben az alábbi jelöléseket használjuk.

**6.1. Definíció.** *Legyen $x \in \mathbb{R}^n$ tetszőleges vektor, ekkor*

$$(sgn(x))_i := \begin{cases} 1, & ha \quad x_i \geq 0, \\ -1, & ha \quad x_i < 0 \end{cases} \qquad (i = 1, ..., n).$$

**6.2. Definíció.** *Jelölje $\mathbb{R}^n_z$ azt az ortánst, amire*

$$\mathbb{R}^n_z := \{x \in \mathbb{R}^n : T_z x \geq 0\},$$

*ahol $T_z = diag(z_1, ..., z_n)$ és $z \in Y_n$ előre rögzített vektor.*

**6.3. Definíció.** *Legyen $z, z' \in Y_n$. Ekkor azt mondjuk, hogy $z$ és $z'$ szomszédosak, ha pontosan egy koordinátájukban különböznek.*

Az $\mathbf{A}x = \mathbf{b}$ intervallum-együtthatós lineáris egyenletrendszer megoldáshalmazát továbbra is $\Sigma$-val jelöljük, azaz

$$\Sigma = \{x : \exists A \in \mathbf{A} \wedge \exists b \in \mathbf{b}, Ax = b\}.$$

Az Oettli-Prager-tétel szerint ez a megoldáshalmaz a következőképpen írható le:

$$\Sigma = \{x : |A_c x - b_c| \leq \Delta|x| + \delta\}.$$

Ismeretes, hogy ha $\mathbf{A}$ reguláris, akkor $\Sigma$ kompakt és összefüggő halmaz, ellenkező esetben pedig $\Sigma$ minden komponense (azaz nemüres összefüggő részhalmaza, ami a tartalmazásra nézve maximális) nemkorlátos. A megoldáshalmaz általában egy bonyolult nemkonvex struktúra, ezért most is az őt tartalmazó legszűkebb intervallumvektort keressük, melyet $\mathbf{x}(\mathbf{A}, \mathbf{b})$-vel jelölünk. Azaz

$$\mathbf{x}(\mathbf{A}, \mathbf{b}) = [\underline{x}, \overline{x}],$$

ahol

$$\underline{x}_i = \min\{x_i : x \in \Sigma\},$$
$$\overline{x}_i = \max\{x_i : x \in \Sigma\},$$

$(i = 1, ..., n)$. Ha $\mathbf{A}$ szinguláris, akkor $\Sigma$ vagy üres, vagy nemkorlátos, ezért ebben az esetben $\mathbf{x}(\mathbf{A}, \mathbf{b})$-t nem definiáljuk.

A megoldáshalmazt tartalmazó legszűkebb intervallumvektor megadásáról szóló fő tétel előtt kimondjuk az ezt megalapzó három egymásra épülő tételt.

**6.4. Tétel.** *Legyen $\mathbf{A} \in \mathbb{IR}^{n \times n}$ és $\mathbf{b} \in \mathbb{IR}^n$, és legyen $Z \subseteq Y_n$ melyre a következők teljesülnek:*

1. *$sgn(x) \in Z$ valamely $x \in \Sigma$ esetén,*
2. *$\Sigma \cap \mathbb{R}^n_z$ korlátos halmaz minden $z \in Z$ esetén,*
3. *ha $z \in Z$ és $y \in Y_n$ szomszédosak és $\Sigma \cap \mathbb{R}^n_z \cap \mathbb{R}^n_y \neq 0$, akkor $y \in Z$.*

*Ekkor* **A** *reguláris és*

$$\Sigma \subseteq \bigcup_{z \in Z} \mathbb{R}^n_z.$$

Tehát a tétel ad egy szükséges feltételt az **A** intervallummátrix regularitására, és a megoldáshalmazba tartozó vektorok előjeleit korlátozza a $Z$ halmazra. A következő tételben kicsit változtatunk a $Z$ halmaz tulajdonságain, és így egy $\Sigma$-t tartalmazó intervallumvektort tudunk adni, ami persze még nem biztos, hogy a legszűkebb.

**6.5. Tétel.** *Legyen $\mathbf{A} \in \mathbb{IR}^{n \times n}$ és $\mathbf{b} \in \mathbb{IR}^n$, és legyen $Z \subseteq Y_n$ melyre a következők teljesülnek:*

1. *$sgn(x) \in Z$ valamely $x \in \Sigma$ esetén,*
2. *minden $z \in Z$-re, melyre $\Sigma \cap \mathbb{R}^n_z \neq 0$, létezik egy $[\underline{x}_z, \overline{x}_z]$ intervallumvektor, melyre $\Sigma \cap \mathbb{R}^n_z \subseteq [\underline{x}_z, \overline{x}_z]$,*
3. *ha $z \in Z$, $\Sigma \cap \mathbb{R}^n_z \neq 0$ és $(\underline{x}_z)_j (\overline{x}_z)_j \leq 0$ valamely $j$ esetén, akkor $z - 2z_j e_j \in Z$.*

*Ekkor* **A** *reguláris és*

$$\Sigma \subseteq \bigcup_{z \in Z_0} [\underline{x}_z, \overline{x}_z],$$

*ahol*

$$Z_0 = \{z \in Z : \Sigma \cap \mathbb{R}^n_z \neq 0\}.$$

A következő tételben egy abszolútértékes egyenlőtlenségrendszer megoldására vezetjük vissza a problémát, melynek megoldására később még visszatérünk. Ismét változtatunk a $Z$ halmaz tulajdonságain, amivel az előzőnél egy jobban használható eredményre jutunk.

**6.6. Tétel.** *Legyen $\mathbf{A} = [A_c - \Delta, A_c + \Delta] \in \mathbb{IR}^{n \times n}$ és $\mathbf{b} = [b_c - \delta, b_c + \delta] \in \mathbb{IR}^n$, és legyen $Z \subseteq Y_n$ melyre a következők teljesülnek:*

1. *$sgn(x) \in Z$ valamely $x \in \Sigma$ esetén,*

2. *minden $z \in Z$-re az alábbi egyenlőtlenségeknek*

$$(QA_c - I)T_z \geq |Q|\Delta \tag{6.1}$$

$$(QA_c - I)T_{-z} \geq |Q|\Delta \tag{6.2}$$

*létezik $Q_z$ és $Q_{-z}$ megoldása,*

3. *ha $z \in Z$, $Q_{-z}b_c - |Q_{-z}|\delta \leq Q_z b_c + |Q_z|\delta$ és $(Q_{-z}b_c - |Q_{-z}|\delta)_j (Q_z b_c + |Q_z|\delta)_j \leq 0$ valamely $j$ esetén, akkor $z - 2z_j e_j \in Z$.*

*Ekkor* **A** *reguláris és*

$$\Sigma \subseteq \bigcup_{z \in Z_1} [Q_{-z}b_c - |Q_{-z}|\delta, Q_z b_c + |Q_z|\delta] \subseteq$$

$$\subseteq [\min_{z \in Z_1}(Q_{-z}b_c - |Q_{-z}|\delta), \max_{z \in Z_1}(Q_z b_c + |Q_z|\delta)],$$

*ahol*

$$Z_1 = \{z \in Z : Q_{-z}b_c - |Q_{-z}|\delta \leq Q_z b_c + |Q_z|\delta\}.$$

Legyen mostantól

$$\overline{x}_z := Q_z b_c + |Q_z|\delta,$$
$$\underline{x}_z := Q_{-z}b_c - |Q_{-z}|\delta.$$

Tehát ha a tétel feltételei teljesülnek, akkor

$$\Sigma \subseteq [\min_{z \in Z_1} \underline{x}_z, \max_{z \in Z_1} \overline{x}_z] \tag{6.3}$$

A következő tétel azt mondja ki, hogy ha az (6.1), (6.2) abszolútértékes egyenlőtlenségeket egyenlőséggel oldjuk meg, akkor az (6.3)-béli tartalmazó intervallum legszűkebb tartalmazó intervallummá válik.

**6.7. Tétel.** *Legyen $\mathbf{A} = [A_c - \Delta, A_c + \Delta] \in \mathbb{IR}^{n \times n}$ és $\mathbf{b} = [b_c - \delta, b_c + \delta] \in \mathbb{IR}^n$, és legyen $Z \subseteq Y_n$ melyre a következők teljesülnek:*

1. *$sgn(x) \in Z$ valamely $x \in \Sigma$ esetén,*

2. *minden $z \in Z$-re az alábbi egyenlőségeknek*

$$QA_c - |Q|\Delta T_z = I \tag{6.4}$$

$$QA_c - |Q|\Delta T_{-z} = I \tag{6.5}$$

*létezik $Q_z$ és $Q_{-z}$ megoldása,*

3. *ha $z \in Z$, $\underline{x}_z \leq \overline{x}_z$ és $(\underline{x}_z)_j(\overline{x}_z)_j \leq 0$ valamely $j$ esetén, akkor $z - 2z_j e_j \in Z$.*

*Ekkor* **A** *reguláris és*

$$\mathbf{x}(\mathbf{A}, \mathbf{b}) = [\min_{z \in Z_1} \underline{x}_z, \max_{z \in Z_1} \overline{x}_z], \tag{6.6}$$

*ahol*

$$Z_1 = \{z \in Z : \underline{x}_z \leq \overline{x}_z\}.$$

Tehát a fenti tétel segítségével meg tudjuk adni egy tetszőleges intervallum egyenletrendszer megoldáshalmazát tartalmazó legszűkebb intervallumvektort, ha van ilyen.

Most térjünk rá az abszolútértékes egyenlet (6.4), (6.5) megoldására. Legyen

$$x^T = Q_{i.} \quad i \in \{1, 2, ..., n\},$$

ahol $Q_{i.}$ jelöli a $Q$ mátrix $i.$ sorát. Ekkor $x$ vektor az

$$x^T A_c - |x|^T \Delta T_z = e_i^T \tag{6.7}$$

megoldása, és így

$$A_c^T x - T_z \Delta^T |x| = e_i, \tag{6.8}$$

ami

$$Ax + B|x| = b \tag{6.9}$$

alakban van, ahol $A, B \in \mathbb{R}^{n \times n}, b \in \mathbb{R}^n$. A megoldás minket abban az esetben érdekel, ha nem létezik olyan $S$ szinguláris mátrix, melyre

$$|S - A| \leq |B|, \tag{6.10}$$

hiszen ha létezik ilyen $S$, akkor ez eleme az **A** intervallummátrixnak, és így az szinguláris.

A következőkben felsoroljuk azokat az állításokat, melyeket (6.9) egyenletrendszer megoldása során felhasználunk.

Először egy intervallum-mátrix szingularitásának ekvivalens megfogalmazását adjuk meg.

**6.8. Állítás.** *Legyen* $\mathbf{A} = [A - |B|, A + |B|] \in \mathbb{IR}^{n\times n}$. $\mathbf{A}$ *akkor és csak akkor szinguláris, ha* $|Ax| \leq |B||x|$ *egyenlőtlenségnek létezik nemtriviális megoldása.*

A következő állítás egy szükséges feltételt ad a probléma megoldására.

**6.9. Állítás.** *Legyen* $\mathbf{A} = [A - |B|, A + |B|] \in \mathbb{IR}^{n\times n}$ *reguláris és*

$$(A + BT_{z'})x' = (A + BT_{z''})x''$$

*valamely* $z', z'' \in Y_n$, $x' \neq x''$ *esetén. Ekkor létezik olyan* $j$ *index, melyre* $z'_j z''_j = -1$ *és* $x'_j x''_j > 0$.

Az alábbi állítás elégséges feltételt ad arra, hogy az intervallumos egyenletrendszerünk mátrixa mikor tartalmaz szinguláris mátrixot.

**6.10. Állítás.** *Legyen*

$$(A + BT_{z'})x' = (A + BT_{z''})x''$$

*valamely* $z', z'' \in Y_n$ *esetén és* $x' \neq x''$ *olyan, hogy minden* $l$ *indexre, amire* $z'_l z''_l = -1$, *igaz, hogy* $x'_l x''_l \leq 0$. *Továbbá legyen* $x = x' - x''$,

$$y_j = \begin{cases} (Ax)_j/(|B||x|)_j, & ha (|B||x|)_j > 0 \\ 1, & ha (|B||x|)_j = 0 \end{cases} \qquad (j = 1, ..., n) \tag{6.11}$$

*és*

$$z = sgn(x). \tag{6.12}$$

*Ekkor*

$$S = A - T_y|B|T_z \tag{6.13}$$

*szinguláris mátrix, melyre* $|S - A| \leq |B|$ *és* $Sx = 0$.

A fenti állítások képezik a magját a következő részben leírt algoritmusoknak.

## 6.2. Algoritmusok

Először azt az algoritmust írjuk le, amely vagy megoldja az (6.9) abszolútértékes egyenletrendszert, vagy ad egy $S$ szinguláris mátrixot, melyre $|S - A| \leq |B|$.

**6.11. Algoritmus.** *A lépések a következők:*

1. *Ha $A$ szinguláris, akkor $S = A$ és kész vagyunk.*
2. *Legyen $z = sgn(A^{-1}b)$.*
3. *Ha $A + BT_z$ szinguláris, akkor $S = A + BT_z$ és kész vagyunk.*
4. *Legyen $x = (A + BT_z)^{-1}b$ és $C = -(A + BT_z)^{-1}B$.*
5. *Legyen $i = 0$, $r = 0 \in \mathbb{R}^n$, $X = 0 \in \mathbb{R}^{n\times n}$.*
6. *Amíg $z_j x_j < 0$ valamely $j$-re*
   (a) *Legyen $i = i + 1$ és $k = \min\{j : z_j x_j < 0\}$.*
   (b) *Ha $1 + 2z_k C_{kk} \leq 0$, akkor $S = A + B(T_z + (1/C_{kk})e_k e_k^T)$ és kész vagyunk.*
   (c) *Ha ($k < n$ és $r_k > \max_{j>k} r_j$) vagy ($k = n$ és $r_n > 0$), akkor*
      i. *$x = x - X_{.k}$, ahol $X_{.k}$ az $X$ mátrix $k$. oszlopát jelöli.*
      ii. *Ha $(|B||x|)_j > 0$, akkor legyen $y_j = (Ax)_j/(|B||x|)_j$ egyébként legyen $y_j = 1$ $(j = 1, 2, ..., n)$.*
      iii. *Legyen $z = sgn(x)$ és $S = A - T_y|B|T_z$ és kész vagyunk.*
   (d) *Legyen $r_k = i$, $X_{.k} = x$, $z_k = -z_k$ és $\alpha = 2z_k/(1 - 2z_k C_{kk})$.*
   (e) *Legyen $x = x + \alpha x_k C_{.k}$ és $C = C + \alpha C_{.k} C_{k.}$.*

Tehát a fenti algoritmussal $A = A_c^T$, $B = -T_z \Delta^T$, $b = e_i$, $(i = 1, 2, ..., n)$ választással $Q_z$ illetve $Q_{-z}$ sorait ki tudjuk számítani.

Most térjünk rá arra az algoritmusra, amely egy intervallum-együtthatós lineáris egyenletrendszerhez megadja a megoldáshalmazát tartalmazó legszűkebb intervallumvektort, ha ilyen létezik. Ellenkező esetben megad egy olyan szinguláris $S$ mátrixot, ami benne van az egyenletrendszer együttható intervallummátrixában.

**6.12. Algoritmus.** *A lépések a következők:*

1. *Ha $A_c$ szinguláris, akkor $S = A_c$, és kész vagyunk.*
2. *Legyen $x_c = A_c^{-1}b_c$, $z = sgn(x_c)$, $\underline{x} = \overline{x} = x_c$, $Z = \{z\}$ és $D = \emptyset$.*
3. *Amíg $Z \neq \emptyset$:*
   (a) *Választunk egy $z \in Z$-t, $Z = Z \backslash \{z\}$ és $D = D \cup \{z\}$.*
   (b) *A 6.11 algoritmussal kiszámítjuk $Q_z$-t és $Q_{-z}$-t, ha léteznek. Ha valamelyik nem létezik, akkor az algoritmus ad egy $S$ szinguláris mátrixot, és kész vagyunk.*
   (c) *Legyen $\overline{x}_z = Q_z b_c + |Q_z|\delta$ és $\underline{x}_z = Q_{-z}b_c - |Q_{-z}|\delta$.*
   (d) *Ha $\underline{x}_z \leq \overline{x}_z$, akkor*
      i. *Legyen $\underline{x} = \min\{\underline{x}, \underline{x}_z\}$ és $\overline{x} = \max\{\overline{x}, \overline{x}_z\}$.*
      ii. *Válasszunk egy tetszőleges $z$-vel szomszédos $z'$-t, és legyen $j$ az az index, amelyre $z'_j = -z_j$. Ha $(\underline{x})_j(\overline{x})_j \leq 0$ és $z' \notin Z \cup D$, akkor legyen $Z = Z \cup \{z'\}$. Ezt addig ismételjük, amíg $z$ összes szomszédját meg nem vizsgáltuk.*
4. *$\mathbf{x}(\mathbf{A}, \mathbf{b}) = [\underline{x}, \overline{x}]$.*

Megjegyezzük, hogy a fenti algoritmusok alapvetően lineáris algebrai műveleteket tartalmaznak, ezért például MATLAB környezetben könnyen megvalósíthatóak.

# 7. fejezet

# Automatikus Differenciálás

A gyakorlatban előforduló numerikus számítások többségében szükséges, hogy meghatározzuk a függvények különböző deriváltjait. Egyszerű példa ilyen alkalmazásra a nemlineáris függvények zérushely keresése, vagy szélsőértékeinek meghatározása. A deriváltak kiszámítására háromféle módszer alkalmazható: numerikus differenciálás, szimbolikus differenciálás és automatikus differenciálás.

A numerikus differenciálás módszere (véges) differenciákkal közelíti a derivált értékeit. A szimbolikus differenciálás a deriválás szabályai alapján explicit meghatározza a derivált függvény alakját. Ezeket a megfelelő pontokban még ki kell értékelni, hogy megkapjuk a derivált értékét. Az automatikus differenciálás szintén a jól ismert deriválási szabályokon alapszik, de felhasználja a tényleges numerikus értékeket is. Ez egyesíti a szimbolikus és a numerikus módszer előnyeit, mivel a szimbolikus kifejezések helyett elegendő számokkal dolgozni, és a feldolgozás után rögtön megkapjuk a derivált numerikus értékét is. A legfőbb előny, hogy a deriválandó függvénynek elegendő egy kiszámítási szabályát ismerni, nem szükséges a deriváltak explicit alakjának ismerete.

Ebben a fejezetben az automatikus deriválás módszereit terjesztjük ki az intervallum aritmetika használatával, hogy a függvény deriváltjának értékét garantáltan befoglaló intervallumot kapjunk.

Az automatikus differenciálás alapvető építőköve a megbízható numerikus algoritmusoknak, hiszen a legtöbb intervallum algoritmus számára szükséges a magasabbrendű derivált értékének befoglalása, hogy a nume-

rikus hiba korlátja kiszámítható legyen.

Megjegyezzük, hogy az automatikus differenciálásnak létezik egy úgynevezett visszafelé haladó változata, de itt most erre nem térünk ki.

## 7.1. Elméleti háttér

Az automatikus differenciálás módszerében algoritmussal, vagy formulával megadott függvények deriváltjainak értékét számítjuk ki differenciál aritmetika segítségével, amelyet a következőkben definiálunk.

### 7.1.1. Elsőrendű deriváltak rendezett párokkal

Az egydimenziós, elsőrendű esetben a differenciál aritmetika építőkövei az

$$U = (u, u'), u, u' \in \mathbb{R}$$

alakú rendezett párok. Az U első komponense tartalmazza $u(x)$-et, azaz az $u : \mathbb{R} \to \mathbb{R}$ függvény értékét az $x \in \mathbb{R}$ helyen. A második komponens tartalmazza a derivált értékét, azaz $u'(x)$-et. A négy alapműveletre a következő differenciál aritmetikai szabályok érvényesek:

$$\begin{aligned} U + V &= (u, u') + (v, v') = (u + v, u' + v') \\ U - V &= (u, u') + (v, v') = (u - v, u' - v') \\ U \cdot V &= (u, u') \cdot (v, v') = (u \cdot v, u \cdot v' + u' \cdot v) \\ U/V &= (u, u')/(v, v') = (u/v, (u' - u/v \cdot v')/v), v \neq 0 \end{aligned}$$

A második komponens kiszámításánál az analízisből jól ismert deriválási szabályokat alkalmaztuk. A zárójeleken belüli kifejezésekben valós számokon végett műveleteket találunk. A differenciál aritmetika kiértékelése során bármely x független változó helyén az $X = (x, 1)$, c tetszőleges konstans helyén pedig a $C = (c, 0)$ rendezett pár helyettesíthető be, hiszen $\frac{dx}{dx} = 1$, illetve $\frac{dc}{dx} = 0$.

Legyen $x$ az $f : \mathbb{R} \to \mathbb{R}$ függvény független változója. Helyettesítsük az összes előfordulását $X = (x, 1)$-el, és az összes formulabeli c konstanst a megfelelő $C = (c, 0)$ elemmel. Ekkor az f függvény differenciál aritmetikai kiértékelése megadja a következő

$$f(X) = f((x, 1)) = (f(x), f'(x))$$

rendezett párt.

**Példa:** Számítsuk ki az $f(x) = x \cdot (4+x)/(3-x)$ függvény deriváltjának értékét az $x = 1$ pontban!

$$\begin{aligned} f(X) = (f, f') &= (x,1) \cdot ((4,0) + (x,1))/((3,0) - (x,1)) \\ &= (1,1) \cdot ((4,0) + (1,1))/((3,0) - (1,1)) \\ &= (1,1) \cdot (5,1)/(2,-1) \\ &= (5,6)/(2,-1) \\ &= (2.5, 4.25) \end{aligned}$$

Látható, hogy $f(1) = 2.5$ és $f'(1) = 4.25$.

Az $s : \mathbb{R} \to \mathbb{R}$ elemi függvények esetén a deriválás lánc szabályának megfelelő

$$s(U) = s((u, u')) = (s(u), u' \cdot s'(u))$$

szabály alkalmazható a derivált értékének kiszámítására.

Például a szinusz függvény esetén:

$$sin\ U = sin(u, u') = (sin\ u, u' \cdot cos\ u).$$

### 7.1.2. Másodrendű deriváltak rendezett hármasokkal

A másodrendű differenciál-aritmetikában a következő szám-hármasokat használjuk

$$U = (u, u', u''), \quad \text{ahol} \quad u, u', u'' \in \mathbb{R}$$

Itt $u, u', u''$ jelöli rendre a függvény-, az első derivált- és a második derivált értékét az $x \in \mathbb{R}$ pontban. Az $u(x) = c$ konstans függvény helyettesítése $C = (c, 0, 0)$. Az $u(x) = x$ függvényé pedig $U = (x, 1, 0)$. A négy alapműveletre korábban definiált differenciál aritmetikai szabályokat kiterjesztjük a harmadik komponens számításához $U = (u, u', u'')$ és $V = (v, v', v'')$ jelölések mellett:

$$\begin{aligned} W = U + V &\Rightarrow w'' = u'' + v'' \\ W = U - V &\Rightarrow w'' = u'' - v'' \\ W = U \cdot V &\Rightarrow w'' = u \cdot v'' + 2 \cdot v' \cdot u' + u'' \cdot v \\ W = U/V &\Rightarrow w'' = (u'' - 2 \cdot w' \cdot v' - w \cdot v'')/v, \ v \neq 0 \end{aligned}$$

Az elemi $s : \mathbb{R} \to \mathbb{R}$ függvények esetére a lánc szabály a következőképpen módosul, $U = (u, u', u'')$ jelölés mellett:

$$s(U) = (s(u), s'(u) \cdot u', s'(u) \cdot u'' + s''(u) \cdot (u')^2).$$

Itt feltesszük, hogy léteznek $s$ első- és második deriváltjai: $s' : \mathbb{R} \to \mathbb{R}$ és az $s'' : \mathbb{R} \to \mathbb{R}$.

A függvényértékek és a derivált értékek befoglalásait egy intervallum aritmetikára épített differenciál aritmetika segítségével fogjuk kiszámítani. Az $u, u', u''$ értékeit helyettesítjük a megfelelő intervallum értékekkel, és a valós aritmetikai és függvény kiértékeléseket helyettesítjük a nekik megfelelő intervallum aritmetikai kiértékelésekkel. Így az $f : \mathbb{R} \to \mathbb{R}$ függvény intervallumos differenciál aritmetikai

$$f(X) = f(([x], 1, 0)) = ([f], [f'], [f''])$$

kiértékelésére teljesülnek a következők:

$$f([x]) \subseteq [f],\ f'([x]) \subseteq [f'],\ f''([x]) \subseteq [f''].$$

**Példa:** Egy nemlineáris $f : \mathbb{R} \to \mathbb{R}$ függvény zérushelyének Newton módszerrel történő meghatározásához szükséges $f'$ ismerete. Vannak olyan módszerek is, amelyek másodrendű, vagy magasabbrendű deriváltakat alkalmaznak. A Halley módszer az első- és másodrendű deriváltakon alapszik. Kiindulva egy $x^{(0)} \in \mathbb{R}$ elemből, a következő iteráció alkalmazható:

$$a^{(k)} := -\frac{f(x^{(k)})}{f'(x^{(k)})} \tag{7.1}$$

$$b^{(k)} := a^{(k)} \cdot \frac{f''(x^{(k)})}{f'(x^{(k)})} \tag{7.2}$$

$$x^{(k+1)} := x^{(k)} + \frac{a^{(k)}}{1 + \frac{b^{(k)}}{2}} \tag{7.3}$$

$k = 0, 1, 2, \ldots$

## 7.2. Gradiens, Jacobi- és Hesse-mátrix számítása

Az előző részben az egyváltozós automatikus differenciálással foglalkoztunk, de számos olyan numerikus módszer is előfordul az alkalmazásokban, ahol többdimenziós függvények deriváltértékeit kell kiszámolnunk. Ebben a részben kiterjesztjük az automatikus differenciálás eszközeit a többdimenziós esetre. Alkalmazzuk a jól ismert deriválási szabályokat a gradiens, Jacobi- és Hesse-mátrixok kiszámítására. Hasonlóan az egydimenziós esethez, itt is elegendő a függvény kiszámítási algoritmusát, vagy formuláját ismerni. Nincs szükség explicit formulákra a gradiens, Jacobi- és Hesse-mátrixok számításához. Módszert adunk a gradiens, a Jacobi- és a Hesse-mátrixok garantált befoglalására.

### 7.2.1. Elméleti háttér

Legyen $f : \mathbb{R}^n \to \mathbb{R}$ egy skalárértékű, kétszer folytonosan differenciálható függvény. Egyrészt az $f$ függvény gradiensét szeretnénk kiszámolni:

$$\nabla f(x) = \begin{pmatrix} \frac{\partial f}{\partial x_1}(x) \\ \frac{\partial f}{\partial x_2}(x) \\ \vdots \\ \frac{\partial f}{\partial x_n}(x) \end{pmatrix}$$

Másrészt az $f$ függvény Hesse-mátrixát:

$$\nabla^2 f(x) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2}(x) & \frac{\partial^2 f}{\partial x_1 \partial x_2}(x) & \dots & \frac{\partial^2 f}{\partial x_1 \partial x_n}(x) \\ \frac{\partial^2 f}{\partial x_2 \partial x_1}(x) & \frac{\partial^2 f}{\partial^2 x_2}(x) & \dots & \frac{\partial^2 f}{\partial x_2 \partial x_n}(x) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1}(x) & \frac{\partial^2 f}{\partial x_n \partial x_2}(x) & \dots & \frac{\partial^2 f}{\partial^2 x_n}(x) \end{pmatrix}$$

Az egyváltozós függvények esetére ismertetett eljárásban úgy jártunk el, hogy a differenciálandó függvényt egy elemi függvényekből és aritmetikai műveletekből álló véges kód-listába konvertáltuk, amelyet aztán a

differenciál aritmetika segítségével értelmeztünk. A többváltozós esetben is hasonló sémát követünk. Itt csak az első- és másodrendű deriváltak kiszámításával foglalkozunk, de a módszerek általánosíthatóak magasabbrendű deriváltak kiszámításához is.

A Gradiens /Hesse aritmetika alap építőköve a következő rendezett hármas:

$$U = (u_f, u_g, u_h), \quad \text{ahol} \quad u_f \in \mathbb{R}, u_g \in \mathbb{R}^n, u_h \in \mathbb{R}^{n \times n}$$

ahol az $u_f$ skalár jelöli a kétszer differenciálható $u : \mathbb{R}^n \to \mathbb{R}$ függvény $u(x)$ értékét az $x \in \mathbb{R}^n$ pontban. Hasonlóan $u_g$, illetve $u_h$ jelöli a $\nabla u(x)$ gradienst és a $\nabla^2 u(x)$ Hesse-mátrixot a megadott x pontban. A konstans $u(x) = c$ függvény esetén a behelyettesítendő rendezett hármas az $U = (u_f, u_g, u_h) = (c, 0, 0)$. Az $u(x) = x_k$, $(k \in \{1, 2, \ldots, n\})$ függvények esetén a behelyettesítés pedig $U = (u_f, u_g, u_h) = (x_k, e_k, 0)$, ahol $e_k \in \mathbb{R}^n$ a k.-ik egységvektor. A 0 jelöli a nullvektort és a nullmátrixot a megfelelő dimenziókban. A többdimenziós differenciál aritmetika kiszámítási szabályai a következőek:

$$W = U + V \Rightarrow \begin{cases} w_f = u_f + v_f \\ w_g = u_g + v_g \\ w_h = u_h + v_h \end{cases}$$

$$W = U - V \Rightarrow \begin{cases} w_f = u_f - v_f \\ w_g = u_g - v_g \\ w_h = u_h - v_h \end{cases}$$

$$W = U \cdot V \Rightarrow \begin{cases} w_f = u_f \cdot v_f \\ w_g = u_f \cdot v_g + v_f \cdot u_g \\ w_h = v_f \cdot u_h + u_g \cdot v_g^T + v_g \cdot u_g^T + u_f \cdot v_h \end{cases}$$

$$W = U/V \Rightarrow \begin{cases} w_f = u_f / v_f \\ w_g = (u_g - w_f \cdot v_g)/v_f \\ w_h = (u_h - w_g \cdot v_g^T - v_g \cdot w_g^T - w_f \cdot v_h)/v_f \end{cases}$$

ahol látható, hogy a második és harmadik komponensben a többdimenziós deriválási szabályokat alkalmaztuk. Fel kell ezen kívül még

tennük, hogy az osztás esetén $v_f \neq 0$. A $w_f, w_g, w_h$ változókon csak valós számokon, vektorokon és mátrixokon végzett alapműveleteket hajtunk végre.

Kiindulunk az $f : \mathbb{R}^n \to \mathbb{R}$ függvényből, és annak összes független $x_i$ változóját helyettesítjük az $X_i = (x_i, e_i, 0)$ értékkel, összes $c_k$ konstansát pedig a megfelelő $(c_k, 0, 0)$ értékkel. Ekkor kiszámítható az $f$ differenciál aritmetikai kiértékelése:

$$f(X) = f\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} = f\begin{pmatrix} (x_1, e^{(1)}, 0) \\ (x_2, e^{(2)}, 0) \\ \vdots \\ (x_n, e^{(n)}, 0) \end{pmatrix} = (f(x), \nabla f(x), \nabla^2 f(x))$$

**Példa:** Számítsuk ki az $f(x) = x_1 \cdot (4 + x_2)$ függvény értékét a gradiens és a Hesse-mátrixszal együtt az $x = (1, 2)^T$ pontban! A differenciál aritmetikai számítások alapján kapjuk, hogy:

$$\begin{aligned}
f(X) &= (f_f, f_g, f_h) \\
&= \left(x_1, \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}\right) \cdot \left((4,0,0) + \left(x_2, \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}\right)\right) \\
&= \left(1, \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}\right) \cdot \left((4,0,0) + \left(2, \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}\right)\right) \\
&= \left(1, \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}\right) \cdot \left(6, \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}\right) \\
&= \left(6, \begin{pmatrix} 6 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\right)
\end{aligned}$$

Ebből következően $f(x) = 6$, $\nabla f(x) = \begin{pmatrix} 6 \\ 1 \end{pmatrix}$, és $\nabla^2 f(x) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ az $x = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$ értékre.

Az elemi $s : \mathbb{R} \to \mathbb{R}$ függvény és $U = (u_f, u_g, u_h)$ esetén

$$W = s(U) \Rightarrow \begin{cases} w_f = s(u_f) \\ w_g = s'(u_f) \cdot u_g \\ w_h = s''(u_f) \cdot u_g \cdot u_g^T + s'(u_f) \cdot u_h \end{cases}$$

Itt feltesszük, hogy létezik $s$ első deriváltja $s' : \mathbb{R} \to \mathbb{R}$, és második deriváltja $s'' : \mathbb{R} \to \mathbb{R}$.

### 7.2.2. Intervallum aritmetika alapú differenciál aritmetika

Az eddig bemutatott szabályok a pontos értékeket tartalmazták. Most bevezetjük az intervallum alapú differenciál-aritmetikát a függvény gradiensének, és Hesse-mátrixának kiszámításához. Az $u_f$, $u_g$ és $u_h$ komponenseket intervallumokra cseréljük, a differenciálaritmetikában szereplő alapműveleteket pedig az intervallumos megfelelőjükre cseréljük. Ennek eredményeképp az $f : \mathbb{R}^n \to \mathbb{R}$ függvény egy adott $\mathbf{x} \in \mathbb{IR}^n$ argumentummal történő intervallumos differenciál aritmetikai kiértékelése után

$$f(X) = ([f_f], [f_g], [f_h])$$

rendelkezni fog a következő tulajdonságokkal:

$$f(\mathbf{x}) \subset [f_f],\ \nabla f(\mathbf{x}) \subset [f_g], \nabla^2 f(\mathbf{x}) \subset [f_h]$$

Legyen $f : \mathbb{R}^n \to \mathbb{R}^n$ egy vektorértékű, differenciálható függvény, és számítsuk ki a Jacobi mátrixot:

$$J_f(x) = \begin{pmatrix} \frac{\delta f_1}{\delta x_1}(x) & \frac{\delta f_1}{\delta x_2}(x) & \dots & \frac{\delta f_1}{\delta x_n}(x) \\ \frac{\delta f_2}{\delta x_1}(x) & \frac{\delta f_2}{\delta x_2}(x) & \dots & \frac{\delta f_2}{\delta x_n}(x) \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\delta f_n}{\delta x_1}(x) & \frac{\delta f_n}{\delta x_2}(x) & \dots & \frac{\delta f_n}{\delta x_n}(x) \end{pmatrix}.$$

Ezt megtehetjük úgy, hogy a gradiens differenciál-aritmetikát alkalmazzuk az intervallum aritmetikai műveletek segítségével minden $f_i$, $i = 1, 2, \dots, n$ függvénykomponensre. Ebben az esetben nem szükséges a Hesse komponensek kiszámítása a differenciál aritmetikai szabályokban.

### 7.2.3. Algoritmikus leírás

Ebben a szekcióban bemutatjuk az elemi operátorok $(+, -, \cdot, /)$ és az elemi függvények ($s \in \{$ sqr, sqrt, power, exp, ln, sin, cos, tan, cot,

arcsin, arccos, arctan, arccot, sinh, cosh, tanh, coth, arsinh, arcosh, artanh, arcoth }) intervallum aritmetika alapú differenciál aritmetikai szabályaihoz tartozó algoritmikus lépéseket, amelyekkel egy $f : \mathbb{R}^n \to \mathbb{R}$ kétszer folytonosan differenciálható függvény gradiensének, és Hesse-mátrixának befoglalása kiszámítható. Legyen $U := ([u_f], \mathbf{u_g}, \mathbf{U_h})$, $[u_f] \in \mathbb{IR}$, $\mathbf{u_g} \in \mathbb{IR}^n$, és $\mathbf{U_h} \in \mathbb{IR}^{n\times n}$. Definiáljuk a következő intervallum-mátrix osztályt:

$$\mathbb{IR}^{\hat{n}\times\hat{n}} := \left\{ \mathbf{A} \in \mathbb{IR}^{(n+1)\times(n+1)} | \mathbf{A} = ([a]_{ij})_{i,j\in[0,\ldots,n]} \right\} \tag{7.4}$$

Tegyük meg a következő megfeleltetéseket $U$ és egy $[U] \in \mathbb{IR}^{\hat{n}\times\hat{n}}$ mátrix között:

$$[u_f] = [u]_{00}, \tag{7.5}$$

$$\mathbf{u_g} = ([u]_{01}, [u]_{02}, \ldots, [u]_{0n})^T, \tag{7.6}$$

$$\mathbf{U_h} = \begin{pmatrix} [u]_{11} & [u]_{12} & \ldots & [u]_{1n} \\ [u]_{21} & [u]_{22} & \ldots & [u]_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ [u]_{n1} & [u]_{n2} & \ldots & [u]_{nn} \end{pmatrix}. \tag{7.7}$$

Ez a jelölés megadja $[U] \in \mathbb{IR}^{\hat{n}\times\hat{n}}$ particionálását a következő alakban:

$$[U] = \left( \begin{array}{c|c} [u_f] & \mathbf{u_g}^T \\ \hline & \mathbf{U_h} \end{array} \right). \tag{7.8}$$

A Hesse-mátrix szimmetriája miatt elegendő az $i = 0, \ldots, n$ és a $j = 1, \ldots, i$ indexű $[u]_{ij}$ komponenseket kiszámítani.

**7.1. Algoritmus.** $+([U], [V])$ *operátor*

*1.* $[w]_{00} := [u]_{00} + [v]_{00}$*;* { *függvényérték* }

*2.* **for** $i := 1$ **to** $n$ **do**

*(a)* $[w]_{0i} := [u]_{0i} + [v]_{0i}$*;* { *gradiens komponensek* }

*(b)* **for** $j := 1$ **to** $i$ **do**

$[w]_{ij} := [u]_{ij} + [v]_{ij}$;

*3.* **return** $[W]$;

**7.2. Algoritmus.** $-([U],[V])$ *operátor*

*1.* $[w]_{00} := [u]_{00} - [v]_{00}$; { *függvényérték* }

*2.* **for** $i := 1$ **to** $n$ **do**

*(a)* $[w]_{0i} := [u]_{0i} - [v]_{0i}$; { *gradiens komponensek* }

*(b)* **for** $j := 1$ **to** $i$ **do**

$[w]_{ij} := [u]_{ij} - [v]_{ij}$;

*3.* **return** $[W]$;

**7.3. Algoritmus.** $\cdot([U],[V])$ *operátor*

*1.* $[w]_{00} := [u]_{00} \cdot [v]_{00}$; { *függvényérték* }

*2.* **for** $i := 1$ **to** $n$ **do**

*(a)* $[w]_{0i} := [v]_{00} \cdot [u]_{0i} + [u]_{00} \cdot [v]_{0i}$; { *gradiens komponensek* }

*(b)* **for** $j := 1$ **to** $i$ **do**

$[w]_{ij} := [v]_{00} \cdot [u]_{ij} + [u]_{0i} \cdot [v]_{0j} + [v]_{0i} \cdot [u]_{0j} + [u]_{00} \cdot [v]_{ij}$;

*3.* **return** $[W]$;

Az osztás implementálásakor nem vesszük figyelembe a $0 \in [v]_{00}$ esetet, mivel nincs értelme folytatni a számításokat, ha ez az eset felmerül.

**7.4. Algoritmus.** $/([U],[V])$ *operátor*

*1.* $[w]_{00} := [u]_{00}/[v]_{00}$;

*2.* **for** $i := 1$ **to** $n$ **do**

*a)* $[w]_{0i} := ([u]_{0i} - [w]_{00} \cdot [v]_{0i})/[v]_{00}$;

**for** $j := 1$ **to** $i$ **do**

$$[w]_{ij} := ([u]_{ij} - [w]_{0i} \cdot [v]_{0j} - [v]_{0i} \cdot [w]_{0j} - [w]_{00} \cdot [v]_{ij})/[v]_{00};$$

*3.* **return** $[W]$*;*

A következő algoritmus az elemi függvényekkel történő kompozíció deriváltját számítja ki. A nullával való osztás esetéhez hasonlóan járunk el itt is abban az esetben, ha a komponálandó elemi függvény értelmezési tartománya szűkebb, mint a megadott $[u]_{00}$ intervallum. A hibakezelésnek ilyenkor az algoritmus első két pontjában fellépő hibákat kell lekezelnie.

**7.5. Algoritmus.** $s([U])$

*1.* $[w]_{00} := s([u]_{00})$

*2.* $[h_1] := s'([u]_{00}); [h_2] := s''([u]_{00});$

*3.* **for** $i := 1$ **to** $n$ **do**

*(a)* $[w]_{0i} := [h_1] \cdot [u]_{0i}$*;*

*(b)* **for** $j := 1$ **to** $i$ **do**

$$[w]_{ij} := [h_2] \cdot [u]_{0i} \cdot [u]_{0j} + [h_1] \cdot [u]_{ij};$$

*4.* **return** $s := [W]$*;*

# 8. fejezet

# Valós egyváltozós függvény zérushelyének befoglalása

Ebben a fejezetben eljárásokat vizsgálunk, amelyek alkalmasak egy valós függvény zérushelyeinek befoglalására. Az eljárások lehetővé teszik, hogy találjunk egy intervallum-halmazt, a lehető legkisebb szélességgel, amelynek minden eleme tartalmazza az $f$ függvény egy, vagy több zérushelyét kiindulva egy adott $[x^{(0)}] \in \mathbb{IR}$ intervallumból. Az eljárásokhoz szükséges feltételek igen bő függvényosztályra teljesülnek. Másrészről, gyököket tartalmazó intervallumokat kapunk, ha az eljárást számítógéppel hajtjuk végre, ahol a hagyományos intervallum aritmetika helyett az 1.4. fejezetben bemutatott gépi intervallum aritmetikát használjuk.

Egyszerű megvalósítását adják ezeknek az eljárásoknak az úgynevezett *felosztási algoritmusok* (subdivision methods). Ezek az intervallumos megfelelői a bináris keresésnek és egyéb keresési algoritmusoknak. Egy rövid magyarázatot adunk ezekhez az algoritmusokhoz. Ehhez csak az $f$ függvény egy intervallumkiértékelésére van szükség az $[x^{(0)}]$ intervallumban (lásd 1.3. fejezet). Hogy pontosítsuk a gyököket tartalmazó intervallumokat, felosztjuk $[x^{(0)}]$-t az

$$m([x^{(0)}]) = \frac{1}{2}(\underline{x}^{(0)} + \overline{x}^{(0)})$$

ponttal egy $[u^{(0)}]$ és egy $[v^{(0)}]$ intervallumra, melyekre

$$[u^{(0)}] = [\underline{x}^{(0)}, m([x^{(0)}])], \text{ és } [v^{(0)}] = [m([x^{(0)}]), \overline{x}^{(0)}].$$

Világos, hogy

$$[x^{(0)}] = [\underline{x}^{(0)}, m([x^{(0)}])] \cup [m([x^{(0)}]), \overline{x}^{(0)}] = [u^{(0)}] \cup [v^{(0)}].$$

Ha $0 \in f_{[]}([u^{(0)}])$, akkor lehetséges, hogy az $f$ egy gyökét az $[u^{(0)}]$ tartalmazza és ezért az eljárást megismételjük $[u^{(0)}]$-ra. Ha $0 \in f_{[]}([v^{(0)}])$, akkor hasonlóan megismételjük az eljárást a $[v^{(0)}]$ intervallumra. Másrészről viszont ha azt kapjuk, hogy $0 \notin f_{[]}([u^{(0)}])$ vagy $0 \notin f_{[]}([v^{(0)}])$, akkor a megfelelő intervallumot elhagyhatjuk, mivel a befoglalási tulajdonság miatt nem tartalmazhatja $f$ egyik gyökét sem. Ez az intervallum tehát elhagyható a további számításokból. Ez az iteráció az $[x^{(0)}]$ részintervallumainak egy olyan sorozatát generálja, amely tartalmazhatja $f$ egy gyökét. Ezen intervallumok szélessége tart 0-hoz, mivel a szélesség minden lépésben feleződik. Ezek a lépésről lépésre számolt intervallumok szükségszerűen konvergálnak $f$ $[x^{(0)}]$-beli gyökeihez, ha (1.40) igaz.

Hogy megakadályozzuk a vizsgálandó intervallumok számának túl nagyra növését, vezessük be a következő módosítást. Minden lépésben a keletkező két részintervallum közül csak a jobb (vagy csak a bal) oldali intervallumot vizsgáljuk. Ha valamelyik lépésben azt kapjuk, hogy $0 \notin f([y])$ a vizsgált félintervallumra ($[y]$), akkor az eljárást újraindítjuk az $[\underline{x}^{(0)}, \underline{y}] \subset [x^{(0)}]$ (illetve $[\overline{y}, \overline{x}^{(0)}] \subset [x^{(0)}]$) intervallumra. Ezzel a módszerrel meghatározhatjuk az egyes gyököket jobbról balra (illetve balról jobbra) haladva sorban. Így elkerülhetjük a nagy számú vizsgálandó intervallum eltárolásának problémáját.

## 8.1. Newton-szerű eljárás

Ebben és a következő szakaszban a Newton-módszer intervallumos megfelelőit vizsgáljuk. Ezért tekintsünk egy folytonos $f$ függvényt, amelynek az adott $[x^{(0)}] = [\underline{x}^{(0)}, \overline{x}^{(0)}]$ intervallumban van zérushelye, azaz

$$f(\xi) = 0$$

valamely $\xi \in [x^{(0)}]$-ra. Legyen

$$f(\underline{x}^{(0)}) < 0 \text{ és } f(\overline{x}^{(0)}) > 0 \tag{8.1}$$

az $[x^{(0)}]$ végpontjaiban. Továbbá legyenek $\underline{m}$ és $\overline{m}$ az osztott differenciák korlátai, azaz

$$0 < \underline{m} \leq \frac{f(x) - f(\xi)}{x - \xi} = \frac{f(x)}{x - \xi} \leq \overline{m} < \infty, \quad \xi \neq x \in [x^{(0)}]. \tag{8.2}$$

Ezek a határok egy $[m] = [\underline{m}, \overline{m}] \in \mathbb{IR}$ intervallumot határoznak meg. (Hasonló értelmezés írható fel $f(\underline{x}^{(0)}) > 0$ és $f(\overline{x}^{(0)}) < 0$ esetén is.) A fenti feltételek mellett nyilvánvaló, hogy $f$-nek $[x^{(0)}]$-ban nincs másik gyöke.

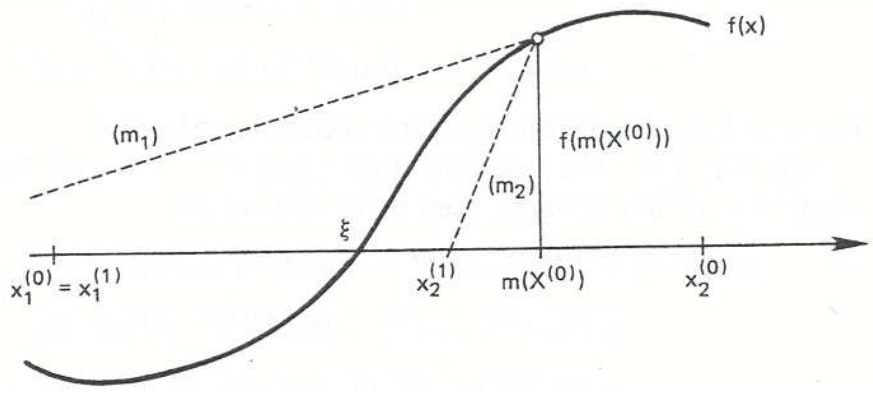
Az $[x^{(0)}] \ni \xi$ kiindulási intervallumból indulva számoljuk az új $[x^{(k)}]$, $k \geq 1$ intervallumokat, ismétlődően a következő eljárásnak megfelelően:

$$[x^{(k+1)}] = \left\{ m([x^{(k)}]) - \frac{f(m([x^{(k)}]))}{[m]} \right\} \cap [x^{(k)}], \ k \geq 0, \tag{8.3}$$

$$\text{ahol } m([x^{(k)}]) \in [x^{(k)}].$$

Általában $m([x^{(k)}]) \in [x^{(k)}]$ választása tetszőleges, viszont tipikusan az intervallum középpontjára esik választásunk, melyet korábban szintén így jelöltünk.

A 8.1 ábra tisztázza az iteráció első lépését.




8.1. ábra.

A (8.3) iteráció intervallum műveletek nélkül is felírható:

$$\left\{\begin{array}{rcl} \underline{x}^{(k+1)} & = & \left\{\begin{array}{ll} \max\left\{\underline{x}^{(k)}, m([x^{(k)}]) - \frac{f(m([x^{(k)}]))}{\underline{m}}\right\} & \text{ha } f(m([x^{(k)}])) \geq 0 \\ m([x^{(k)}]) - \frac{f(m([x^{(k)}]))}{\overline{m}} & \text{ha } f(m([x^{(k)}])) \leq 0 \end{array}\right. \\ \overline{x}^{(k+1)} & = & \left\{\begin{array}{ll} m([x^{(k)}]) - \frac{f(m([x^{(k)}]))}{\overline{m}} & \text{ha } f(m([x^{(k)}])) \geq 0 \\ \min\left\{\overline{x}^{(k)}, m([x^{(k)}]) - \frac{f(m([x^{(k)}]))}{\underline{m}}\right\} & \text{ha } f(m([x^{(k)}])) \leq 0. \end{array}\right. \end{array}\right. \tag{8.3'}$$

Mind a (8.3), mind pedig a (8.3') formulában használt

$$m : \mathbb{IR} \ni [x] \mapsto m([x]) \in \mathbb{R}$$

helyettesítés magában foglal egy kiválasztási eljárást melynek során kiválasztunk egy intervallumból egy valós $m$ számot. Gyakran használt választás a középpont:

$$m([x]) = \frac{1}{2}(\underline{x} + \overline{x}) \tag{8.4}$$

Összegyűjtjük az iteráció során generált $\{[x^{(k)}]\}_{k=0}^{\infty}$ sorozat legfontosabb tulajdonságait.

**8.1. Tétel.** *Legyen $f$ egy folytonos függvény és $\xi$ pedig $f$ egy gyöke az $[x^{(0)}]$ intervallumban. (8.1) és (8.2) teljesüljön az $[m] = [\underline{m}, \overline{m}]$, $\underline{m} > 0$ intervallum esetén. Ekkor a (8.3) alapján számolt $\{[x^{(k)}]\}_{k=0}^{\infty}$ sorozat az alábbi tulajdonságokkal rendelkezik:*

$$\xi \in [x^{(k)}],\ k \geq 0, \tag{8.5}$$

$$[x^{(0)}] \supset [x^{(1)}] \supset [x^{(2)}] \supset \cdots \quad , \textit{ ahol } \lim_{k\to\infty} [x^{(k)}] = \xi, \tag{8.6}$$

*vagy a sorozat véges sok lépésben lecseng és megáll a $[\xi, \xi]$ pontban. Továbbá az intervallumok hosszáról elmondható, hogy*

$$d([x^{(k+1)}]) \leq \left(1 - \frac{\underline{m}}{\overline{m}}\right) d([x^{(k)}]). \tag{8.7}$$

**Bizonyítás:** (8.5) bizonyítása:
(8.2)-ből és az 1.6 következményből kapjuk, hogy

$$\begin{aligned}\xi &= m([x^{(0)}]) - \frac{f(m([x^{(0)}]))}{\frac{f(m([x^{(0)}]))}{m([x^{(0)}])-\xi}} \in \\ &\in \left\{ m([x^{(0)}]) - \frac{f(m([x^{(0)}]))}{[m]} \right\} \cap [x^{(0)}] = [x^{(1)}].\end{aligned}$$

$k > 1$ esetén a bizonyítás teljes indukcióval történik.

(8.6) és (8.7) bizonyítása:
Tegyük fel, hogy $f(m([x^{(k)}])) > 0$. Ha most

$$f(m([x^{(k)}])) \geq (m([x^{(k)}]) - \underline{x}^{(k)})\underline{m}$$

teljesül, akkor (8.3')-t felhasználva kapjuk, hogy

$$\begin{aligned}d([x^{(k+1)}]) &= \overline{x}^{(k+1)} - \underline{x}^{(k+1)} = m([x^{(k)}]) - \frac{f(m([x^{(k)}]))}{\overline{m}} - \underline{x}^{(k)} \leq \\ &\leq (m(x^{(k)}) - \underline{x}^{(k)}) - \frac{(m([x^{(k)}]) - \underline{x}^{(k)})\underline{m}}{\overline{m}} = \\ &= (m([x^{(k)}]) - \underline{x}^{(k)})(1 - \underline{m}/\overline{m}) \leq d([x^{(k)}])(1 - \underline{m}/\overline{m}).\end{aligned}$$

Ha most $f(m([x^{(k)}])) \leq (m([x^{(k)}]) - \underline{x}^{(k)})\underline{m}$, akkor (8.3')-t felhasználva kapjuk, hogy

$$\begin{aligned}d([x^{(k+1)}]) &= \overline{x}^{(k+1)} - \underline{x}^{(k+1)} = \\ &= m([x^{(k)}]) - \frac{f(m([x^{(k)}]))}{\overline{m}} - m([x^{(k)}]) + \frac{f(m([x^{(k)}]))}{\underline{m}} = \\ &= f(m([x^{(k)}])) \left( \frac{1}{\underline{m}} + \frac{1}{\overline{m}} \right) = \frac{f(m([x^{(k)}]))}{\underline{m}}(1 - \underline{m}/\overline{m}) \leq \\ &\leq (m([x^{(k)}]) - \underline{x}^{(k)})(1 - \underline{m}/\overline{m}) \leq \\ &\leq d([x^{(k)}])(1 - \underline{m}/\overline{m}).\end{aligned}$$

Az $f(m([x^{(k)}])) < 0$ eset hasonló módon bizonyítható.

Ha azonban $f(m([x^{(k)}])) = 0$, akkor $m([x^{(k)}]) = \xi$ és ezért $d([x^{(k+1)}]) = 0$ és $[x^{(k+i)}] = \xi$, $i \geq 1$. Ez bizonyítja (8.7)-t. Mivel $\underline{m} \leq \overline{m}$ kapjuk, hogy

$$d([x^{(k+1)}]) \leq \gamma^{k+1} d([x^{(0)}]) \quad 0 \leq \gamma = (1 - \underline{m}/\overline{m}) < 1,$$

így

$$\lim_{k\to\infty} d([x^{(k+1)}]) = 0.$$

Mivel (8.5) miatt $\xi \in [x^{(k)}],\ k \geq 0$, ezért $\lim_{k\to\infty}[x^{(k)}] = \xi$, kivéve, ha $[x^{(k_0+i)}] = \xi,\ i \geq 1$ már teljesül valamely $k_0$-ra. A (8.6) tulajdonság a (8.3) eljárás közvetlen következménye. □

Tehát a 8.1. tétel garantálja, hogy a megadott feltételek mellett az $[x^{(k)}],\ k \geq 0$ iteráció az $f$ függvény $\xi$ gyökéhez konvergáljon. Ekkor minden, az iterációban szereplő intervallum tartalmazza a kívánt gyököt. Másrészt viszont, ha a (8.3) eljárást egy olyan $[x^{(0)}]$ intervallumra alkalmazzuk, amelyre $\xi \notin [x^{(0)}]$, akkor van olyan $k_0$ index, amelyre a (8.3)-ban felírt metszet üres. Ugyanis (8.7) felhasználásával ellentmondásra jutunk kiindulva abból a feltételből, hogy a metszet nem üres.

A (8.3) iteráció két módosítását vizsgáljuk, melyek az $m$ pont választásából származnak. Először $m$ választását rögzítjük, így a következőhöz jutunk:

**8.2. Következmény.** *Legyenek a feltételek és a jelölések azonosak a 8.1. tétel feltételeivel illetve jelöléseivel. Kiegészítésként válasszuk minden lépésben az intervallum középpontját*

$$m([x^{(k)}]) = \frac{1}{2}(\underline{x}^{(k)} + \overline{x}^{(k)}),\ k \geq 0.$$

*Ekkor a*

$$d([x^{(k+1)}]) \leq \frac{1}{2}(1 - \underline{m}/\overline{m})d([x^{(k)}]), \tag{8.8}$$

*egyenlőtlenség igaz az* $\{[x^{(k)}]\}_{k=0}^{\infty}$ *iterációs sorozatra, amely a (8.7) becslés javítása.*

**Bizonyítás:** A 8.1. tétel (8.7) állításának bizonyításában $m([x^{(k)}])$ választásából

$$m([x^{(k)}]) - \underline{x}^{(k)} = \frac{1}{2}d([x^{(k)}])$$

adódik, amiből (8.8) kapható. □

Tehát ha a középpontot választjuk $m([x^{(k)}])$-nak, akkor garantált, hogy a tartalmazó intervallum szélessége minden lépésben legalább feleződik.

Más lehetőségeket is vizsgáltak $m([x^{(k)}])$ választására, például

$$m([x^{(k)}]) = m([x^{(k-1)}]) - f(m([x^{(k-1)}]))/m_0, \text{ ahol } m_0 \in [m],$$

illetve

$$m([x^{(k)}]) \in \{\underline{x}^{(k)}, \overline{x}^{(k)}\}, \text{ ha } m([x^{(k)}]) \notin [x^{(k)}],\ k \geq 0.$$

Az, hogy az $[m]$ intervallum határai az osztott differenciák korlátjai (lásd (8.2)), mind a 8.1. tételhez, mind a 8.2. következményhez fontosak. Ha az $f$ folytonosan differenciálható, és $f'(x) \neq 0,\ x \in [x^{(0)}]$, akkor választható

$$[m] = \left[ \inf_{y \in [x^{(0)}]} f'(y),\ \sup_{y \in [x^{(0)}]} f'(y) \right],$$

felhasználva a középérték tételt. Általában ez az egyetlen lehetséges becslés olyan halmazra amely, tartalmazza ezt az intervallumot. Becslést például az $f'$ intervallum kiértékelésén keresztül nyerhetünk, vagyis

$$[m] = f'([x^{(0)}]).$$

Az $\underline{m} > 0$ feltétel biztosítható, ha az $\inf_{y \in [x^{(0)}]} f'(y)$-nak egy alsó becslését vesszük.

## 8.2. Optimális eljárás meghatározása

Az előző szakaszban tekintett (8.3) iterációnál egy bizonyos mértékű szabadsággal rendelkezünk $m([x^{(k)}]) \in [x^{(k)}]$ választásában. Attól függően, hogy $[x^{(k)}]$ melyik elemét választjuk $m([x^{(k)}])$-nak más és más $\{[x^{(k)}]\}_{k=0}^{\infty}$ tartalmazó intervallum sorozatot kapunk. Ezek a sorozatok általában nem hasonlíthatók össze elemről elemre tartalmazás tekintetében. Nyilvánvaló cél tehát, az eljárás számára olyan $m([x^{(k)}]) \in [x^{(k)}]$ választása, amely olyan $\{[x^{(k)}]\}_{k=0}^{\infty}$ sorozatot generál, melyben az egyes elemek szélessége a lehető legkisebb. Szeretnénk ezt világosabban definiálni, ezért jelöljük $\phi[x]$-szel azon $f$ függvények osztályát, melyekre teljesülnek a következők:

1. $f(\underline{x}) < 0$ és $f(\overline{x}) > 0$.

2. Az $[m] = [\underline{m}, \overline{m}]$ intervallumra, amelyre $\underline{m} > 0$ teljesül, igaz hogy

$$\underline{m} \leq \frac{f(x) - f(y)}{x - y} \leq \overline{m}, \text{ ha } x \neq y,\ x, y \in [x].$$

Nyilvánvaló, hogy minden $f \in \phi[x]$ függvénynek egy és csak egy $\xi$ gyöke van az $[x]$ intervallumban. Minden feltétel teljesül, amely a (8.3) iterációhoz szükséges, és a 8.1. tétel összes állítása igaz.

Hogy meghatározzuk az alkalmas $m([x^{(k)}]) \in [x^{(k)}]$ elemet egy lépegetős módszert (stepwise manner) használunk. Jelöljük a (8.3) iterációhoz tartozó sorozatot $\{[x^{(k)}]\}_{k=0}^{\infty}$-val. Az iteráció $[x^{(k+1)}]$ új lépésének kiszámításához szükségünk van az $m([x^{(k)}])$ és az $f(m([x^{(k)}]))$ mennyiségekre. Ha $m([x^{(k)}]) = x \in [x^{(k)}]$-t rögzítjük, akkor $[x^{(k+1)}]$ csak $f(m([x^{(k)}]))$-tól függ. Ez a függvényérték bárhogy változhat, de csak bizonyos $\underline{y}^{(k)}$ és $\overline{y}^{(k)}$ korlátok között, mivel $f \in \phi[x]$ és mivel $f(m([x^{(i)}]))$, $0 \leq i \leq k$ rögzített. Ez lehetővé teszi, hogy meghatározhassuk a lehető legnagyobb szélességet

$$\max\{d([x^{(k+1)}]) \mid m([x^{(k)}]) = x,\ \underline{y}^{(k)} \leq f(m([x^{(k)}])) \leq \overline{y}^{(k)}\}.$$

Ez a lehető legrosszabb eset, amely $f \in \phi[x]$ függvény mellett történhet.

Most meghatározzuk azt az $\tilde{x} = m([x^{(k)}]) \in [x^{(k)}]$ amely esetén a legnagyobb szélesség minimális. Vagyis kiszámítva

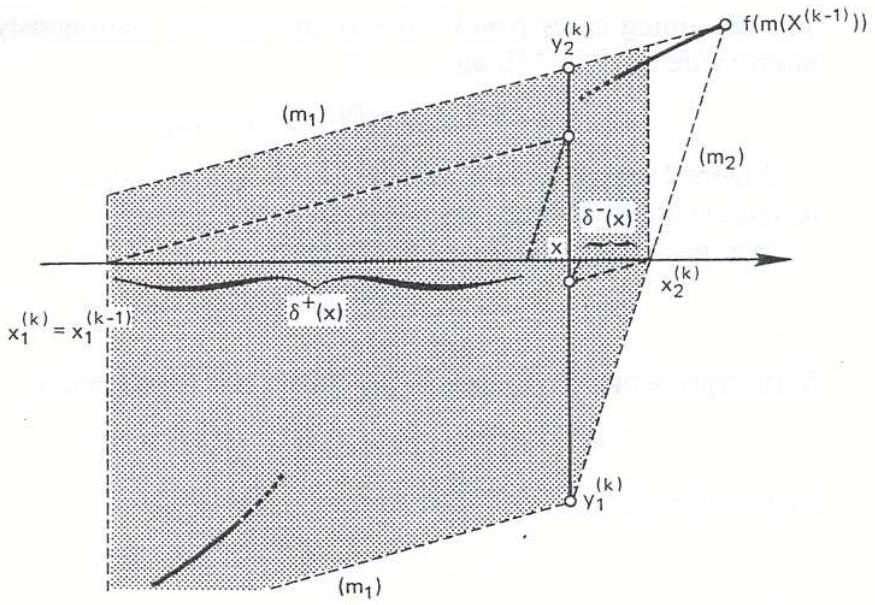
$$\min_{x \in [x^{(k)}]} \max\{d([x^{(k+1)}]) \mid m([x^{(k)}]) = x,\ \underline{y}^{(k)} \leq f(m([x^{(k)}])) \leq \overline{y}^{(k)}\}.$$

értéket és a megfelelő $\tilde{x}$ értéket $m([x^{(k)}])$-nak választjuk. Az $m([x^{(k)}])$ meghatározása tehát a legrosszabb eset minimalizálásával történik.

Megadjuk a fenti eljárás részletes leírását. Az általánosság megszorítása nélkül tekintsük azt az esetet, amikor $f(m([x^{(k)}])) > 0$. A 8.2 ábrán a besatírozott terület mutatja az $f(m([x^{(k)}]))$ függvényértékek lehetséges tartományát, ha $f \in \phi[x]$ és $f(m([x^{(k-1)}])) > 0$ feltételek teljesülnek.

$d([x^{(k+1)}])$ lehetséges értékeit felírjuk, ha $m([x^{(k)}]) = x \in [x^{(k)}]$ meghatározott. Legyen először $f(m([x^{(k)}])) \geq 0$. Az összes

$$0 \leq f(x) \leq (x - \underline{x}^{(k)})\underline{m},$$




8.2. ábra.

értékre (8.3') alapján kapjuk, hogy

$$d([x^{(k+1)}]) = x - \frac{f(x)}{\overline{m}} - x + \frac{f(x)}{\underline{m}} = f(x)\left(\frac{1}{\underline{m}} - \frac{1}{\overline{m}}\right).$$

Hasonlóan az összes

$$(x - \underline{x}^{(k)})\underline{m} \leq f(x) \leq \overline{y}^{(k)},$$

értékre

$$d([x^{(k+1)}]) = x - f(x)/\overline{m} - \underline{x}^{(k)}.$$

Jegyezzük meg, hogy mivel

$$\underline{x}^{(k)} = \max\left\{\underline{x}^{(k-1)}, m([x^{(k-1)}]) - \frac{f(m([x^{(k-1)}]))}{\underline{m}}\right\},$$

így mindig igaz, hogy $\overline{y}^{(k)} \geq (x - \underline{x}^{(k)})\underline{m}$.
Az első esetben $d([x^{(k+1)}])$ egy monoton növő, a második esetben egy

monoton csökkenő függvénye $f(x)$-nek. $f(x) = (x - \underline{x}^{(k)})\underline{m}$ esetén a maximum

$$\delta^+(x) = (x - \underline{x}^k)(1 - \underline{m}/\overline{m}).$$

A fennmaradó $f(m([x^{(k)}])) \leq 0$ eseteket hasonlóan kezelve adható meg $d([x^{(k+1)}])$ maximuma,

$$\delta^-(x) = (\overline{x}^k - x)(1 - \underline{m}/\overline{m}).$$

A 8.2 ábra megmutatja a két lehetőséget $[x^{(k+1)}]$ kiértékelésére, amelyek a $\delta^+(x)$ (illetve $\delta^-(x)$) maximális szélességekhez vezetnek.

Figyeljük meg, hogy $\delta^+(x)$ és $\delta^-(x)$ lineáris függvények.

Most meghatározzuk a minimumot:

$$\min_{x \in [x^{(k)}]} \max\{\delta^+(x), \delta^-(x)\}.$$

A $\delta^+(x)$ és a $\delta^-(x)$ kifejezés teljesíti a

$$\delta^+\left(\frac{1}{2}(\underline{x}^{(k)} + \overline{x}^{(k)}) - t\right) = \delta^-\left(\frac{1}{2}(\underline{x}^{(k)} + \overline{x}^{(k)}) + t\right)$$

követelményt, ha $|t| \leq \frac{1}{2}(\overline{x}^{(k)} - \underline{x}^{(k)})$. A minimum tehát az

$$\tilde{x} = \frac{1}{2}(\underline{x}^{(k)} + \overline{x}^{(k)}).$$

pontban van és értéke

$$d([x^{(k+1)}]) = \frac{1}{2} d([x^{(k)}])(1 - \underline{m}/\overline{m}),$$

Vessük össze ezt az eredményt a 8.2 következménnyel.

Szeretnénk az $[x^{(k+1)}]$ kiszámításánál használt optimalizáció alapelvét kiterjeszteni $m([x^i])$, $0 \leq i \leq k$ értékének meghatározására. Ugyanolyan módon próbáljuk meghatározni $m([x^{(0)}]) = x^{(0)}, \ldots, m([x^{(k)}]) = x^{(k)}$ értékeket, ahogy a

$$\min_{x^{(0)} \in [x^{(0)}]} \max_{\underline{y}^{(0)} \leq f(x^{(0)}) \leq \overline{y}^{(0)}} \ldots \min_{x^{(k)} \in [x^{(k)}]} \max_{\underline{y}^{(k)} \leq f(x^{(k)}) \leq \overline{y}^{(k)}} d([x^{(k+1)}])$$

értéket kaptuk. Ez könnyen előállítható, mivel $d([x^{(k+1)}])$ optimális értéke rögzített $m([x^{(k-1)}])$ esetén arányos $d([x^{(k)}])$-val. Az $f(m([x^{(k-1)}]))$ függvényértékek megengedett tartománya kizárólag $f(m([x^{(k-2)}]))$ felhasználásával meghatározható. Ezért a fenti gondolatmenet végigvihető $m([x^{(k-1)}])$-re, ahogy $m([x^{(k)}])$-ra, kapjuk az

$$m([x^{(k-1)}]) = \frac{1}{2}(\underline{x}^{(k-1)} + \overline{x}^{(k-1)}).$$

optimális értéket. Hasonlóan kapjuk az

$$m([x^{(i)}]),\ i = k-2, k-3, \ldots, 0$$

értékeket a jelölt sorrendben.

**8.3. Tétel.** *Alkalmazzuk a (8.3) iterációt $f \in \phi[x]$ függvényekre. Ha az*

$$m([x^{(k)}]) = \frac{1}{2}(\underline{x}^{(k)} + \overline{x}^{(k)}), \quad 0 \le k \le i,\ i \ge 0,$$

*szabályt használjuk, akkor a $d([x^{(i+1)}])$ maximális szélesség $f \in \phi[x]$ függvényekre kisebb, mint bármely más $m([x^{(k)}])$ választás mellett. Ha $f \in \phi[x]$, akkor*

$$d([x^{(i+1)}]) \le \frac{1}{2^{i+1}}(1 - \underline{m}/\overline{m})^{i+1} d([x^{(0)}]).$$

*Továbbá létezik egy $g \in \phi[x]$ függvény, amelyre a fenti relációban az egyenlőség áll fent.*

A fent tárgyaltak során bizonyítottuk ezt a tételt. Ami a létezést illeti, kihangsúlyoznánk, hogy a $g \in \phi[x]$ függvény választható egy szakaszonként lineáris függvénynek az $(m([x^{(k)}]), f(m([x^{(k)}]))),\ 0 \le k \le i$ pontokon át.

## 8.3. Négyzetesen konvergáló eljárások

Ahhoz, hogy a (8.3) eljárást használjuk szükségünk van az $f$ osztott-differenciáinak rögzített $\underline{m}$ illetve $\overline{m}$ korlátjára. Ez az eljárás megfelel

az egyszerűsített Newton-iteráció egy intervallumos verziójának. Ha feltesszük, hogy az $f$ folytonosan differenciálható és az $f'$ deriváltnak létezik $f'([x])$ intervallumkiértékelése (lásd: 1.3. fejezet), akkor definiálhatjuk a szokásos Newton-iteráció intervallumos megfelelőjét is. Az új eljárás a (8.3) iteráció módosításával kapható, úgy, hogy minden iterációs lépésben kiértékeljük az $[m]$ intervallumot:

$$[m^{(k)}] = f'([x^{(k)}]). \tag{8.9}$$

Ha ismerünk valamilyen a priori becslést

$$0 < \underline{l} \leq f'(x) \leq \overline{l}, \quad x \in [x^{(0)}],$$

akkor garantálhatjuk, hogy $\underline{m} > 0$ és használhatjuk az

$$[m^{(k)}] = [\underline{m}^{(k)}, \overline{m}^{(k)}] = f'([x^{(k)}]) \cap [l], \quad [l] = [\underline{l}, \overline{l}] \tag{8.10}$$

kifejezést. Így az alábbi formulát kapjuk

$$[x^{(k+1)}] = \{m([x^{(k)}]) - f(m([x^{(k)}]))/[m^{(k)}]\} \cap [x^{(k)}], \tag{8.11}$$
$$k \geq 0,\ m([x^{(k)}]) \in [x^{(k)}].$$

A (8.11) iterációt használva egy $\{[x^{(k)}]\}_{k=0}^{\infty}$ intervallum sorozatot kapunk, amelyre a 8.1. tételhez hasonló állítást bizonyítunk.

**8.4. Tétel.** *Legyen $f$ egy folytonosan differenciálható függvény és teljesítse $f'$ az $[x^{(0)}]$ intervallumon az 1.3. fejezet 1.24. tételének feltételeit. Továbbá teljesüljön a (8.1) reláció az $[x^{(0)}]$ intervallumon. Jelölje $\xi$ az $f$ függvény $[x^{(0)}]$-beli gyökét, és az $[m^{(k)}]$ intervallumokat definiálják a (8.9) és a (8.10) kifejezések. Ekkor a 8.1 tétel szerint az $\{[x^{(k)}]\}_{k=0}^{\infty}$ intervallumsorozat teljesíti az alábbiakat:*

$$\xi \in [x^{(k)}],\ k \geq 0,$$
$$[x^{(0)}] \supset [x^{(1)}] \supset [x^{(2)}] \supset \cdots \quad , \textit{ahol} \quad \lim_{k\to\infty} [x^{(k)}] = \xi,$$

*vagy a sorozat véges sok lépésben lecseng és megáll a $[\xi, \xi]$ pontban. Továbbá az intervallumok hosszáról elmondható, hogy*

$$d([x^{(k+1)}]) \leq (1 - \underline{m}^{(k)}/\overline{m}^{(k)}) d([x^{(k)}]) \leq \beta (d([x^{(k)}]))^2, \quad \beta \geq 0, \tag{8.12}$$

*azaz a (8.11) iteráció legalább másodrendben konvergál.*

**Bizonyítás:** $x \in [x^{(k)}]$ esetén teljesül, hogy

$$\frac{f(x)}{x-\xi} = \frac{f(x)-f(\xi)}{x-\xi} = f'(\eta) \in [m^{(k)}], \quad \eta = x + \theta(\xi - x), \quad 0 < \theta < 1.$$

Tehát az $[m^{(k)}]$ intervallumokra egy hasonló következtetés bizonyítható, mint a 8.1. tételben.

A (8.12) állítás igazolása maradt vissza. Ugyanúgy, mint a 8.1. tétel bizonyítása során kapjuk:

$$d([x^{(k+1)}]) \leq \left(1 - \frac{\underline{m}^{(k)}}{\overline{m}^{(k)}}\right) d([x^{(k)}]) = \frac{\overline{m}^{(k)} - \underline{m}^{(k)}}{\overline{m}^{(k)}} d([x^{(k)}])$$

és ezért, felhasználva az (1.19) összefüggést és az 1.3. fejezet 1.24 tételét

$$\begin{aligned} d([x^{(k+1)}]) &\leq \frac{d([m^{(k)}])}{\underline{m}^{(0)}} d([x^{(k)}]) \leq \frac{d(f'([x^{(k)}]))}{\underline{m}^{(0)}} d([x^{(k)}]) \leq \\ &\leq (c/\underline{m}^0)(d([x^{(k)}]))^2, \quad c/\underline{m}^{(0)} \geq 0. \end{aligned}$$

□

Módosítsunk egy kicsit az iteráción. Ehhez jegyezzük meg, hogy attól függően, hogy $f(m([x^{(k)}])) > 0$, vagy $f(m([x^{(k)}])) < 0$, a keresett $\xi$ gyök az $[\underline{x}^{(k)}, m([x^{(k)}])]$ intervallumban, illetve $[m([x^{(k)}]), \overline{x}^{(k)}]$ intervallumban lesz. Ha $f(m([x^{(k)}])) = 0$, akkor $m([x^{(k)}]) = \xi$ és az iteráció megáll. Ezért a (8.11)-ben elegendő az

$$[m^{(k)}] = f'([y^{(k)}]) \cap [l]$$

intervallummal számolni, ahol $[l]$ a (8.10)-ben bevezetett intervallum és

$$[y^{(k)}] = \begin{cases} [\underline{x}^{(k)}, m([x^{(k)}])] & \text{, ha } f(m([x^{(k)}])) > 0 \\ \left[m([x^{(k)}], \overline{x}^{(k)}\right] & \text{, ha } f(m([x^{(k)}])) < 0 \\ [x^{(k)}] & \text{egyébként.} \end{cases} \tag{8.13}$$

Ekkor $f'([y^{(k)}]) \subseteq f'([x^{(k)}])$ és $d([y^{(k)}]) \leq d([x^{(k)}])$ igaz és az $\underline{m}^{(k)} > 0$ feltétel ezen a módon lényegesen könnyebben kielégíthető. A 8.4. tétel szintén igaz a (8.13) szerinti választással.

A (8.11) eljárás során $m([x^{(k)}]) \in [x^{(k)}]$ választásra vonatkozóan a 8.2. következményhez hasonló állítás tehető és a 8.2. fejezetben levezetett tárgyaláshoz hasonlóan vizsgálható. Most ezt nem részletezzük tovább.

Néhány numerikus példával világítjuk meg az intervallumos Newton iteráció működését.

**Példák:**

1. Az
$$f(x) = x^2 \left( \frac{1}{3} x^2 + \sqrt{2} \sin x \right) - \frac{\sqrt{3}}{19}$$
függvénynek van $\xi$ gyöke az $[x^{(0)}] = [0.1,\ 1]$ intervallumban. Az
$$f'(x) = x \left( \frac{4}{3} x^2 + \sqrt{2}(2 \sin x + x \cos x) \right)$$
derivált az $[x^{(0)}]$ intervallumon
$$\underline{l} = 0.00133 \leq f'(x) \leq \overline{l} = 5.57598, \quad x \in [x^{(0)}].$$
határokkal becsülhető. Az
$$\begin{array}{ll} [x^{(k)}], & k \geq 0 \quad \text{felhasználva (8.10)-t,} \\ [y^{(k)}], & k \geq 0 \quad \text{felhasználva (8.13)-t,} \end{array}$$
tartalmazó intervallumokat a (8.11) eljárás alapján számoltuk számítógéppel, egészen addig amikor már nem tapasztalható javulás. A 8.1. táblázatban szereplő értékeket kaptuk.

2. A
$$p(x) = x(x^9 - 1) - 1$$
polinomnak egyetlen $\xi$ gyöke van az $[x^{(0)}] = [1,\ 1.5]$ intervallumban. A
$$p'(x) = 10x^9 - 1$$

| k | $[x^{(k)}]$ | | |
|---|---|---|---|
| 0 | [0.09999999999999 | , | 1.0000000000000] |
| 1 | [0.09999999999999 | , | 0.4384388546433] |
| 2 | [0.3382030708107 | , | 0.4384388546433] |
| 3 | [0.3915056049954 | , | 0.3924484948316] |
| 4 | [0.3923789206719 | , | 0.3923799504692] |
| 5 | [0.3923795071350 | , | 0.3923795071378] |

| k | $[y^{(k)}]$ | | | $d([x^{(k)}])/d([y^{(k)}])$ |
|---|---|---|---|---|
| 1 | [0.09999999999990 | , | 0.5181776715881] | 0.809 |
| 2 | [0.3455588336928 | , | 0.5181776715881] | 0.581 |
| 3 | [0.3739864679691 | , | 0.4075613709040] | 0.028 |
| 4 | [0.3922481030413 | , | 0.392544130626] | 0.004 |
| 5 | [0.3923794945039 | , | 0.3923795211850] | 0.001 |
| 6 | [0.3923795071350 | , | 0.3923795071378] | - |

8.1. táblázat.

derivált $p'([x])$ kifejtése teljesíti a $0 \notin p'([x])$ feltételt minden $[x] \subseteq [x^{(0)}]$ intervallumra. Az

$[x^{(k)}]$, $k \geq 0$ felhasználva (8.13)-t, ahol $[l] = p'([x^{(0)}])$
$[y^{(k)}]$, $k \geq 0$ felhasználva (8.9)-t,

iterált tartalmazó intervallumokat a (8.11) eljárással számoltuk (8.4) választást használva. A 8.2. táblázatban szereplő értékeket kaptuk.

A (8.11) iteráció $\underline{m} > 0$ feltétele a gyakorlatban fennáll. Megmutat-

| k | $[x^{(k)}]$ | | |
|---|---|---|---|
| 0 | [1.000000000000 | , | 1.500000000000] |
| 1 | [1.000000000000 | , | 1.153909281002] |
| 2 | [1.074525733152 | , | 1.075772270022] |
| 3 | [1.075764355129 | , | 1.075767749943] |
| 4 | [1.075766066086 | , | 1.075766066088] |

| k | $[y^{(k)}]$ | | | $d([x^{(k)}])/d([y^{(k)}])$ |
|---|---|---|---|---|
| 1 | [1.000000000000 | , | 1.231579011696] | 0.665 |
| 2 | [1.018539065305 | , | 1.102153489956] | 0.015 |
| 3 | [1.0718097668336 | , | 1.084762444669] | $3 \cdot 10^{-4}$ |
| 4 | [1.075647094319 | , | 1.075931180877] | $6 \cdot 10^{-9}$ |
| 5 | [1.075766039501 | , | 1.075766097327] | ... |
| 6 | [1.075766066085 | , | 1.075766066090] | ... |
| 7 | [1.075766066085 | , | 1.075766066088] | ... |

8.2. táblázat.

tuk, hogy ez (8.10) miatt teljesíthető, felhasználva az $f'(x)$ egy ismert alsó korlátját az $[x^{(0)}]$ intervallumon. Ha nem ismert ilyen $\underline{l}$ alsó korlát és ha $0 \in f'([x^{(0)}])$, akkor a (8.11) eljárás nem indítható el. Ezért, hogy elindíthassuk az eljárásunkat, először lefuttathatjuk az intervallum felosztó eljárásunkat néhányszor, ahogy azt a szakasz bevezetőjében leírtuk. Így találhatunk egy $[y^{(0)}] \subset [x^{(0)}]$ intervallumot melyre a $0 \notin [y^{(0)}]$ feltétel teljesül.

Van egy másik módosítása az intervallumos Newton-módszernek, amely alkalmazható a fenti esetben, mikor $0 \in f'([x^{(0)}])$. Ez az eljárás akkor is alkalmazható, ha $f$-nek több gyöke is van az $[x^{(0)}]$ intervallum-

ban. Ezt fogjuk most körvonalazni. Ha $0 \notin f'([x^{(0)}])$, akkor ez az eljárás a (8.11) iterációval megegyezik. Tegyük fel tehát, hogy $0 \in f'([x^{(0)}])$. Tekintsük az $[x^{(0)}]$ intervallum

$$\begin{aligned} [u^{(1)}] &= \left[\underline{x}^{(0)}, m([x^{(0)}]) - \frac{|f(m([x^{(0)}]))|}{\overline{m}^{(0)}}\right], \\ [v^{(1)}] &= \left[m([x^{(0)}]) + \frac{|f(m([x^{(0)}]))|}{\overline{m}^{(0)}}, \overline{x}^{(0)}\right], \end{aligned}$$

részintervallumait, feltéve, hogy $f(m([x^{(0)}])) \neq 0$. Az $f$ összes $[x^{(0)}]$-beli gyökének az $[u^{(1)}] \cup [v^{(1)}]$-ben kell lennie. Ugyanis bármely $\xi \in [x^{(0)}]$ zérushelyre teljesülnie kell, hogy

$$\left|\frac{f(m([x^{(0)}]))}{\xi - m([x^{(0)}])}\right| \leq \overline{m}^{(0)},$$

ahonnan

$$\frac{|f(m([x^{(0)}]))|}{\overline{m}^{(0)}} \leq |\xi - m([x^{(0)}])|$$

és

$$\xi \geq m([x^{(0)}]) + \frac{|f(m([x^{(0)}]))|}{\overline{m}^{(0)}}, \quad \text{vagy} \quad \xi \leq m([x^{(0)}]) - \frac{|f(m([x^{(0)}]))|}{\overline{m}^{(0)}}$$

következik. Az utolsó egyenlőtlenségek magukban foglalják, hogy $\xi \in [u^{(1)}] \cup [v^{(1)}]$. Továbbá igaz, hogy

$$d([u^{(1)}]) + d([v^{(1)}]) = d([x^{(0)}]) - 2|f(m([x^{(0)}]))|/\overline{m}^{(0)} < d([x^{(0)}]),$$

amit az biztosít, hogy $f(m([x^{(0)}])) \neq 0$.

Ez az eljárás most megismételhető az $[u^{(1)}]$ és $[v^{(1)}]$ részintervallumokra és így tovább. Ezen intervallumok teljes szélessége tart a 0-hoz. Ha $f$-nek az $[x^{(0)}]$ intervallumban csupa egyszeres gyöke van, akkor az iteráció egy bizonyos lépése után ezek diszjunkt részintervallumokba kerülnek. Továbbá az eljárás egy bizonyos $k$ indexnél visszatér a (8.11) iterációhoz. Ennek az iterációnak a hatására

tehát a részintervallum vagy egy olyan intervallumba tart, amely egy gyököt tartalmaz, vagy valahol egy üres metszetet kapunk.

A (8.11) során a (8.3)-nak megfelelő $[m^{(k)}] := f'([x^{(k)}])$ helyett polinomok esetén használhatjuk az 1.3. fejezet 1.26. tételében bevezetett $[j_1], [j_2], [j_3]$ és $[j_4]$ intervallumokat, ahol a derivált behatárolásához $[y] := m([x^{(k)}])$ és $[x] := [x^{(k)}]$. A 8.4. tétel összes állítása továbbra is igaz. Mivel az 1.3. fejezet 1.26. tételében megmutattuk, hogy $[j_4]$ az optimális tartalmazó intervallum, ésszerű ezt választani a derivált tartalmazójának, hogy minden lépésben a legjobb tartalmazó intervallumot kapjuk a gyökökre.

Ennek megfelelően tekintsük a következő példát.

**Példa:** Legyen

$$p(x) = x^7 + 3x^6 - 4x^5 - 12x^4 - x^3 - 3x^2 + 4x + 12$$

egy polinom, melynek az $[x^{(0)}] = [1.8,\ 2.4]$ intervallumban van egy $\xi$ gyöke. A (8.11) iterációt használva számoljuk a gyököt tartalmazó intervallumokat a Horner elrendezés segítségével kiszámolva az $[m^{(k)}] := p'([x^{(k)}])$ intervallumot. A 8.3. táblázat tartalmazza a kiszámított intervallumokat.

Ha $p'([x^{(k)}])$ intervallumot $[j_1]$ intervallumra cseréljük, hasonló módon nyerjük a 8.4. táblázat adatait. A 8.5. táblázatban bemutatjuk a $d_1/d_2$ hányados értékét, amely az első iterált intervallum szélességének és a második iterált intervallum szélességének hányadosa minden egyes lépésben. Ezt a példát a Berlini Műszaki Egyetem Számítóközpontjának CDC 6500-as gépén 48 bites mantisszával számolták.

## 8.4. Magasabbrendű eljárások

Most magasabbrendű eljárásokat fogunk fejleszteni szigorúan monoton növő, vagy fogyó függvények $\xi$, $[x^{(0)}] = [\underline{x}^{(0)}, \overline{x}^{(0)}]$-beli gyökeinek megtalálására, ha a függvény elegendően magasrendű deriváltja folytonos. Ezek az eljárások mindig konvergensek. A konstrukció alapelvét Ehrmann fektette le. Intervallum analitikai eszközöket és az alapelvet használva olyan eljárásokat fejleszthetünk, amelyek mindig

| k | $[x^{(k)}]$ |
|---|---|
| 0 | [1.8, 2.4] |
| 1 | [1.8, 2.0727618077482] |
| 2 | [1.9742900052812, 2.0727618077842] |
| 3 | [1.9948757147483, 2.0059215482353] |
| 4 | [1.9999888234200, 2.0000115390070] |
| 5 | [1.9999999999894, 2.0000000000107] |
| 6 | [2.0, 2.0] |

8.3. táblázat.

| k | $[x^{(k)}]$ |
|---|---|
| 0 | [1.8, 2.4] |
| 1 | [1.9419538108826, 2.0566964050488] |
| 2 | [1.9999999975872, 2.0001112993369] |
| 3 | [1.9999999975872, 2.0000000029595] |
| 4 | [2.0, 2.0] |

8.4. táblázat.

szükségszerűen konvergálnak. Ahogy a korábbi szakaszokban itt is az általánosság megszorítása nélkül feltehetjük, hogy

$$f(\underline{x}^{(0)}) < 0 \text{ és } f(\overline{x}^{(0)}) > 0.$$

Legyenek $\underline{m}$ és $\overline{m}$ az osztott differenciák korlátjai, azaz

$$0 < \underline{m} \leq \frac{f(x) - f(\xi)}{x - \xi} = \frac{f(x)}{x - \xi} \leq \overline{m} < \infty, \quad \xi \neq x \in [x^{(0)}].$$

| k | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| $d_1^{(k)}/d_2^{(k)}$ | 1 | 2.37 | 492.35 | $3.7 \cdot 10^6$ |

8.5. táblázat.

Legyen $[m] = [\underline{m}, \overline{m}]$ az $\underline{m}$ és $\overline{m}$ korlátok által alkotott intervallum. Továbbá legyen az $f$ függvény $(p+1)$-szer folytonosan differenciálható és $[f_i] \in I\mathbb{R},\ 2 \le i \le p+1$ intervallumokra igaz

$$f^{(i)}(x) \in [f_i], \quad x \in [x^{(0)}]. \tag{8.14}$$

Az $[f_i]$ intervallumok például az $f$ deriváltjainak $[x^{(0)}]$ feletti kifejtéseiből számolhatók. Ha a deriváltakra vonatkozó intervallum-kifejezés nem értelmezett (például egy $[x]$ intervallummal kellene osztani, ahol $0 \in [x]$), akkor például részintervallumokra oszthatjuk $[x^{(0)}]$-t és az $[f_i]$ intervallumot az egyes részintervallumok kifejtésének uniójaként kaphatjuk.

Tekintsük a következő iterációt

$$\left\{\begin{array}{rcl} x^{(k)} & = & m([x^{(k)}]) \in [x^{(k)}], \\ [x^{(k+1,0)}] & = & \{x^{(k)} - f(x^{(k)})/[m]\} \cap [x^{(k)}], \\ [x^{(k+1,i)}] & = & \Big\{x^{(k)} - \frac{1}{f'(x^{(k)})}\Big[f(x^{(k)}) + \\ & & + \sum_{\nu=2}^{i} \frac{f^{(\nu)(x^{(k)})}}{\nu!}([x^{(k+1,i-1)}] - x^{(k)})^{\nu} \\ & & + \frac{1}{(i+1)!}[f_{i+1}]([x^{(k+1,i-1)}] - x^{(k)})^{i+1}\Big]\Big\} \cap [x^{(k+1,i-1)}] \\ [x^{(k+1)}] & = & [x^{(k+1,p)}], \end{array}\right. \tag{8.15}$$

$(1 \le i \le p,\ k \ge 0)$.

Ahogy a 8.1. szakaszban, jelentsen $m([x])$ egy tetszőlegesen választott valós számot az $[x]$ intervallumból. A fent megadott iterációhoz $f(x^{(k)}), f'(x^{(k)}), \ldots, f^{(p)}(x^{(k)})$ értékek kiszámítása szükséges minden lépésben, és az iteráció az alábbi tulajdonságokkal rendelkezik.

**8.5. Tétel.** *Legyen $f$ egy $(p+1)$-szer folytonosan differenciálható függvény, $p \geq 1$, és legyen az $[x^{(0)}]$ intervallumon igaz a (8.1) reláció. Legyen $\xi$ az $f$ függvény $[x^{(0)}]$-beli zérushelye és legyen az $[m] = [\underline{m}, \overline{m}]$ intervallum (8.2 alapján definiálva. Legyen igaz továbbá a (8.15) iterációra (8.14), ekkor*

$$\xi \in [x^{(k)}], \quad k \geq 0, \tag{8.16}$$

$$[x^{(0)}] \supset [x^{(1)}] \supset [x^{(2)}] \supset \cdots \quad \text{és} \quad \lim_{k\to\infty} [x^{(k)}] = \xi \tag{8.17}$$

*vagy a sorozat véges sok lépésben lecseng és megáll a $[\xi, \xi]$ pontban.*

$$d([x^{(k+1)}]) \leq \gamma (d([x^{(k)}]))^{p+1}, \tag{8.18}$$

*ahol $\gamma \geq 0$. Azaz a fent definiált iteráció legalább $p+1$-edrendben konvergál.*

**Bizonyítás:** (8.16) bizonyítása: Tegyük fel, hogy $\xi \in [x^{(k)}]$ valamely $k \geq 0$ esetén. A tétel feltételei miatt $k = 0$ esetén ez teljesül. Ahogy a 8.1. tételben, megmutatható, hogy

$$\xi \in [x^{(k+1,0)}].$$

Tegyük fel, hogy $\xi \in [x^{(k+1,i)}]$ valamely $i \geq 0$. Ez $i = 0$ esetén teljesül a fentiek alapján. Ekkor kapjuk, hogy

$$\xi - x^{(k)} \in [x^{(k+1,i)}] - x^{(k)}.$$

A Taylor-formulából kapjuk

$$\begin{aligned} 0 = f(\xi) \;=\; & f(x^{(k)}) + f'(x^{(k)})(\xi - x^{(k)}) + \cdots + \\ & + \frac{1}{(i+1)!} f^{(i+1)}(x^{(k)})(\xi - x^{(k)})^{i+1} + \\ & + \frac{1}{(i+2)!} f^{(i+2)}(\eta_{i+2})(\xi - x^{(k)})^{i+2}, \end{aligned}$$

valamely $\eta_{i+2}$ $x^{(k)}$ és $\xi$ közötti számra. A fenti egyenlőség jobb oldalának második tagjából $\xi$-t kifejezve, a tartalmazás monotonitása miatt kapjuk

az alábbi relációt:

$$\begin{aligned}\xi &= x^{(k)} - \frac{1}{f'(x^{(k)})}\left[f(x^{(k)}) + \sum_{\nu=2}^{i+1}\frac{f^{(v)}(x^{(k)})}{\nu!}(\xi - x^{(k)})^{\nu} + \right.\\ &\quad \left. + \frac{f^{(i+2)}(\eta_{i+2})}{(i+2)!}(\xi - x^{(k)})^{i+2}\right] \in \\ &\in \left\{x^{(k)} - \frac{1}{f'(x^{(k)})}\left[f(x^{(k)}) + \sum_{\nu=2}^{i+1}\frac{f^{(\nu)}(x^{(k)})}{\nu!}([x^{(k+1,i)}] - x^{(k)})^{\nu} + \right.\right.\\ &\quad \left.\left. + \frac{[f_{i+2}]}{(i+2)!}([x^{(k+1,i)} - x^{(k)}])^{i+2}\right]\right\} \cap [x^{(k+1,i)}] = \\ &= [x^{(k+1,i+1)}].\end{aligned}$$

Ezért igaz, hogy $\xi \in [x^{(k+1,i)}],\ 0 \le i \le p$, és $\xi \in [x^{(k+1)}] = [x^{(k+1,p)}]$.

(8.17) bizonyítása: A 8.1. tételben használt módon megmutatható, hogy $[x^{(k)}] \supset [x^{(k+1,0)}]$ és mivel a (8.15) eljárásban metszetet vettünk kapjuk $[x^{(k)}] \supset [x^{(k+1)}],\ k \ge 0$. Továbbá, ahogy a 8.1. tételben, itt is igaz, hogy

$$d([x^{(k+1,0)}]) \le (1 - \underline{m}/\overline{m})d([x^{(k)}]).$$

Mivel a (8.15) eljárásban metszetet vettünk kapjuk

$$d([x^{(k+1)}]) \le (1 - \underline{m}/\overline{m})d([x^{(k)}]), \quad k \ge 0.$$

Ahogy a 8.1. tételben is, kapjuk a konvergenciára vonatkozó állítást $\lim_{k\to\infty}[x^{(k)}] = \xi$. (8.17) fennmaradó állításai ugyanúgy igazolhatóak, mint az a 8.1. tételben.

(8.18) bizonyítása: $d([x^{(k+1,0)}]) \le d([x^{(k)}])$ és ezért

$$\begin{aligned}d([x^{(k+1,1)}]) &\le d\left(x^{(k)} - \frac{1}{f'(x^{(k)})}(f(x^{(k)})) + \frac{1}{2}[f_2]([x^{(k+1,0)}] - x^{(k)})^2\right) \le \\ &\le \frac{1}{2}d\left(\frac{[f_2]}{f'(x^{(k)})}([x^{(k)}] - [x^{(k)}])^2\right) \le \\ &\le \frac{1}{2}d\left(\frac{[f_2]}{[m]}[-(d([x^{(k)}])^2, (d([x^{(k)}]))^2]\right).\end{aligned}$$

Alkalmazva (1.29)-et kapjuk, hogy

$$d([x^{(k+1,1)}]) \le |[f_2]/[m]|(d([x^{(k)}]))^2 = \gamma_1(d([x^{(k)}]))^2,$$

ahol $\gamma_1 = |[f_2]/[m]|$ $k$-tól független konstans.

Tegyük fel, hogy valamely $i \geq 1$ esetén

$$d([x^{(k+1,i)}]) \leq \gamma_i (d([x^{(k)}]))^{i+1},$$

ahol $\gamma_1$ független $k$-tól. Ezt $i = 1$ esetén fent bizonyítottuk. $i > 1$ esetén a (8.15) iterációból felhasználva az 1.2. fejezet szélességre vonatkozó szabályát, kapjuk:

$$\begin{aligned} d([x^{(k+1,i+1)}]) \leq \; & d\left(\textstyle\sum_{\nu=2}^{i+1} \frac{f^{(\nu)}(x^{(k)})}{\nu! f'(x^{(k)})} ([x^{(k+1,i)}] - x^{(k)})^{\nu} + \right. \\ & \left. + \tfrac{1}{(i+2)!} \tfrac{[f_{i+2}]}{f'(x^{(k)})} ([x^{(k+1,i)}] - x^{(k)})^{i+2} \right) \leq \end{aligned}$$

$$\begin{aligned} \leq \; & \textstyle\sum_{\nu=2}^{i+1} \frac{1}{\nu!} \left| \frac{f^{(\nu)}(x^{(k)})}{f'(x^{(k)})} \right| d(([x^{(k+1,i)}] - x^{(k)})^{\nu}) + \\ & + \tfrac{1}{(i+2)!} d\left( \tfrac{[f_{i+2}]}{f'(x^{(k)})} ([x^{(k+1,i)}] - x^{(k)})^{i+2} \right) \leq \end{aligned}$$

$$\begin{aligned} \leq \; & \textstyle\sum_{\nu=2}^{i+1} \frac{1}{\nu!} \left| \frac{[f_\nu]}{[m]} \right| \nu |[x^{(k+1,i)}] - x^{(k)}|^{\nu-1} d([x^{(k+1,i)}] - x^{(k)}) + \\ & + \tfrac{1}{(i+2)!} d\left( \tfrac{[f_{i+2}]}{f'(x^{(k)})} ([x^{(k+1,i)}] - x^{(k)})^{i+2} \right) \leq \end{aligned}$$

$$\begin{aligned} \leq \; & \textstyle\sum_{\nu=2}^{i+1} \frac{1}{(\nu-1)!} \left| \frac{[f_\nu]}{[m]} \right| |[x^{(k)}] - [x^{(k)}]|^{\nu-1} d([x^{(k+1,i)}]) + \\ & + \tfrac{1}{(i+2)!} d\left( \tfrac{[f_{i+2}]}{[m]} ([x^{(k)}] - [x^{(k)}])^{i+2} \right) \leq \end{aligned}$$

$$\begin{aligned} \leq \; & \textstyle\sum_{\nu=2}^{i+1} \frac{1}{(\nu-1)!} \left| \frac{[f_\nu]}{[m]} \right| (d([x^{(k)}]))^{\nu-1} \gamma_i (d([x^{(k)}]))^{i+1} + \\ & + \tfrac{1}{(i+2)!} d\left( \tfrac{[f_{i+2}]}{[m]} [-(d([x^{(k)}]))^{i+2}, (d([x^{(k)}]))^{i+2}] \right) = \end{aligned}$$

$$\begin{aligned} = \; & (d([x^{(k)}]))^{i+2} \textstyle\sum_{\nu=2}^{i+1} \frac{1}{(\nu-1)!} \left| \frac{[f_\nu]}{[m]} \right| \gamma_i (d([x^{(k)}]))^{\nu-2} + \\ & + \tfrac{2}{(i+2)!} \left| \tfrac{[f_{i+2}]}{[m]} \right| (d([x^{(k)}]))^{i+2} \leq \end{aligned}$$

$$\leq \underbrace{\left(\sum_{\nu=2}^{i+1} \frac{1}{(\nu-1)!}\left|\frac{[f_\nu]}{[m]}\right| \gamma_i (d([x^{(0)}]))^{\nu-2} + \frac{2}{(i+2)!}\left|\frac{[f_{i+2}]}{[m]}\right|\right)}_{\gamma_{i+1}} \cdot$$

$$\cdot (d([x^{(k)}]))^{i+2} =$$

$$= \gamma_{i+1}(d([x^{(k)}]))^{i+2}$$

ahol $\gamma_{i+1}$ egy $k$-tól független konstans. Ezért a

$$d([x^{(k+1,i)}]) \leq \gamma_i (d([x^{(k)}]))^{(p+1)}$$

reláció igaz, ha $1 \leq i \leq p$. Így

$$d([x^{(k+1)}]) = d([x^{(k+,1,p)}]) \leq \gamma_p (d([x^{(k)}]))^{p+1}$$

ahol $\gamma_p$ független $k$-tól. Ez pedig megegyezik a (8.18) állításával $\gamma = \gamma_p$-vel és így a tételt igazoltuk. □

Most a $p = 1$ esetben szeretnénk megvizsgálni néhány további részletet, azaz amikor az $f$ függvény kétszer folytonosan differenciálható. Ekkor a (8.15) iteráció

$$\left\{\begin{array}{rcl} x^{(k)} & = & m([x^{(k)}]) \in [x^{(k)}], \\ [x^{(k+1,0)}] & = & \left\{x^{(k)} - f(x^{(k)})/[m]\right\} \cap [x^{(k)}], \\ [x^{(k+1,1)}] & = & \left\{x^{(k)} - (1/f'(x^{(k)}))(f(x^{(k)})) + \right. \\ & & \left. + \frac{1}{2}[f_2]([x^{(k+1,0)}] - x^{(k)})^2\right\} \cap [x^{(k+1,0)}], \\ [x^{(k+1)}] & = & [x^{(k+1,1)}], \quad k \geq 0 \end{array}\right.$$

alakban írható. Az eljárás ugyanazokkal a tulajdonságokkal rendelkezik, mint a 8.3. szakaszban tárgyalt módszerek. Eltekintve néhány járulékos aritmetikai művelettől ehhez kevesebb munkára van szükség, hiszen mind a függvényértékeket, mind a derivált értékeit az $x^{(k)}$ pontban kell számolni. Az ezt megelőző eljárások esetében a deriváltat ki kellett értékelni az $[x^{(k)}]$ intervallumot felhasználva. Ez általában több számítási műveletet igényel, mint az $x^{(k)}$ pontban való kiértékelés. Ha az $[f_2]$ intervallum egyszerűen számolható, akkor a (8.15) eljárás $p = 1$ esetben jobban alkalmazható, mint az előző szakaszban tárgyalt

eljárások. Ezek az eredmények csak elméletileg igazak, amikor pontos számításokat feltételezünk. Ha számítógépes számítás során szeretnénk egy a gyököt tartalmazó intervallumot garantálni, akkor a kerekítési hibákat is számításba kell vennünk. Ez úgy tehető meg, ha minden műveletet gépi intervallum műveletként végzünk el. Különösen fontos $f'(x^{(k)})$ értékét gépi intervallum aritmetikát használva számolni. Ebben az esetben a (8.15) eljárás, eltekintve néhány aritmetikai művelettől, összességében ugyanannyi műveletet igényel, mint a 8.3. szakaszban tárgyalt módszerek. Mivel az $[f_2]$ intervallumot szintén számolni kell, az előző szakaszban leírt eljárást érdemesebb választani, ha a kerekítési hibákkal is számolni kell.

Ezen a ponton szeretnénk megemlíteni azt is, hogy Krawzcyk az alábbi eljárást vizsgálta:

$$\left\{\begin{array}{rcl} x^{(k)} & = & m([x^{(k)}]) \in [x^{(k)}], \\ [x^{(k+1)}] & = & \left\{x^{(k)} - (1/f'(x^{(k)}))(f(x^{(k)}) + \right. \\ & & \left. \frac{1}{2}f''([x^{(k)}])([x^{(k)}] - x^{(k)})^2)\right\} \cap [x^{(k)}], \quad k \geq 0. \end{array}\right.$$

amellett a feltétel mellett, hogy $f$ kétszer differenciálható. Igaz, hogy $\xi \in [x^{(k)}]$, $k \geq 0$. A $\lim_{k\to\infty}[x^{(k)}] = \xi$ konvergencia feltételei nem adottak. Ha az eljárás konvergens, akkor az iterációs intervallumok szélességeinek sorozata négyzetesen tart 0-hoz, ha $f'(\xi) \neq 0$. Összehasonlítva (8.15) eljárással, ahol a $p = 1$ esetet vizsgáltuk, most ki kell értékelni a második deriváltat is az $[x^{(k)}]$ intervallumon minden lépésben. Ez csökkenti a konvergencia konstansát, de nem javítja a konvergencia rendjét. (Ugyanez igaz a (8.15) iterációra, a $p = 1$ esetben, ha az $[f_2]$ konstans intervallumot minden lépésben kicseréljük $f''([x^{(k)}])$-ra.) Ennek az eljárásnak a gyakorlati alkalmazása során, ha a kerekítési hibákat figyelembe vesszük, háromszor annyi műveletre van szükség. Mivel a konvergencia nem biztosított, az eljárás sokkal kevésbé vonzó.

Amikor a (8.15) eljárást használjuk el kell határoznunk magunkat egy bizonyos rendre. Szintén megjegyezzük, hogy a szokásos feltételek mellett azt az eredményt kaphatjuk, hogy a (8.15) eljárás $p = 2$ esetén optimális, ami egy harmadrendű eljárás.

## 8.5. Polinomok valós zérushelyeinek szimultán meghatározása

Ebben a fejezetben olyan Newton-szerű intervallum eljárásokat vizsgálunk, melyekkel befoglalhatjuk egy valós polinom összes valós gyökét. Először azt az esetet vizsgáljuk, amikor a polinom összes gyöke valós. A komplex gyököket a következő részben vizsgáljuk. Ha a polinom összes gyöke valós és egyszeres, akkor egy egylépéses eljárást konstruálhatunk, amely négyzetesnél gyorsabban konvergál. Egy alkalmazásként ezzel az eljárással meghatározhatjuk egy szimmetrikus tridiagonális mátrix összes sajátértékét.

Legyen

$$p(x) = a^{(n)}x^n + a^{(n-1)}x^{n-1} + \cdots + a^{(0)} \tag{8.19}$$

egy valós polinom és a továbbiakban tegyük fel, hogy

$$a^{(n)} = 1.$$

Tegyük fel továbbá, hogy a polinomnak $n$ valós gyöke van, $\xi^{(1)}, \xi^{(2)}, \ldots, \xi^{(n)}$, tároljuk el a gyököket egy $(\xi^{(i)})$ vektorba, a többszörös gyökök a multiplicitásaiknak megfelelően. Tegyük fel, hogy minden gyökhöz ismert egy tartalamzó intervallum

$$\xi^{(j)} \in [x^{(0,j)}] = [\underline{x}^{(0,j)}, \overline{x}^{(0,j)}], \quad 1 \leq j \leq n.$$

Először tegyük fel, hogy ezek a tartalmazó intervallumok páronként diszjunktak, vagyis

$$[x^{(0,j)}] \cap [x^{(0,k)}] = \emptyset \quad 1 \leq j < k \leq n. \tag{8.20}$$

A $p(x)$ polinom

$$p(x) = \prod_{j=1}^{n}(x - \xi^{(j)})$$

alakban, vagy

$$p(x) = (x - \xi^{(i)}) \prod_{j=1, j\neq i}^{n} (x - \xi^{(j)})$$

alakban írható, ahonnét

$$\xi^{(i)} = \frac{x - p(x)}{\prod_{j=1, j\neq i}^{n}(x - \xi^{(j)})}$$

következik. Ha $x = x^{(0,i)} \in [x^{(0,i)}]$-t választjuk, akkor

$$0 \notin \prod_{j=1, j\neq i}^{n} (x^{(0,i)} - [x^{(0,j)}])$$

összefüggést kapjuk, és (1.9) felhasználásával következik

$$\xi^{(i)} \in [x^{(1,i)}] = \left\{ \frac{x^{(0,i)} - p(x^{(0,i)})}{\prod_{j=1, j\neq i}^{n}(x^{(0,i)} - [x^{(0,j)}])} \right\} \cap [x^{(0,i)}].$$

A jobb oldalon álló intervallum-kifejezés szintén egy tartalmazó intervallum $[x^{(1,i)}]$, amelyre

$$\xi^{(i)} \in [x^{(1,i)}] \subseteq [x^{(0,i)}]$$

szintén teljesül. Ez a reláció ad lehetőséget az alábbi iterációra:

$$[x^{(k+1,i)}] = \left\{ \frac{x^{(k,i)} - p(x^{(k,i)})}{\prod_{j=1, j\neq i}^{n}(x^{(k,i)} - [x^{(k,j)}])} \right\} \cap [x^{(k,i)}], \tag{8.21}$$

ahol

$$x^{(k,i)} \in [x^{(k,i)}], \quad 1 \leq i \leq n, \quad k \geq 0.$$

A nevezőben szereplő intervallum kifejezés helyett a továbbiakban röviden írjunk

$$[q^{(k,i)}] = \prod_{j=1, j\neq i}^{n} (x^{(k,i)} - [x^{(k,j)}]).$$

A (8.21)-ben adott iterációs rendszer egy ún. *total step* eljárás a polinom $\xi^{(i)},\ 1 \leq i \leq n$ gyökeinek szimultán befoglalására.

Ha mindig a legfrissebben számolt tartalmazó intervallum értékeit használjuk $[q^{(k,i)}]$ felírásakor, akkor

$$[r^{(k,i)}] = \prod_{j=1}^{i-1}(x^{(k,i)} - [x^{(k+1,j)}]) \prod_{j=i+1}^{n} (x^{(k,i)} - [x^{(k,j)}])$$

az egylépéses iterációval összefüggő eredményre vezet. $p(x^{(k+1,i)})$ és $[r^{(k,i)}]$ előjelétől függően az $[x^{(k+1,i)}]$ tartalmazó intervallumok az $[y^{(k+1,i)}]$ intervallumokra húzódnak. Az előjelfüggvény intervallumokra legyen az alábbi módon értelmezett

$$\operatorname{sign}([x]) = \begin{cases} 1 & \text{ha } \underline{x} > 0 \\ -1 & \text{ha } \underline{x} < 0 \\ 0 & \text{egyébként} \end{cases} \tag{8.22}$$

Az $[y^{(k+1,i)}]$ intervallumhalmaz, mely tartalmazza a $\xi^{(i)}$ gyököket legyen definiálva az alábbi módon

$$[y^{(k+1,i)}] = \begin{cases} [\underline{x}^{(k+1,i)}, x^{(k+1,i)}] & \text{ha } \operatorname{sign}([r^{(k,i)}])\operatorname{sign}(p(x^{(k+1,i)})) > 0 \\ [x^{(k+1,i)}, \overline{x}^{(k+1,i)}] & \text{ha } \operatorname{sign}([r^{(k,i)}])\operatorname{sign}(p(x^{(k+1,i)})) < 0 \\ [x^{(k+1,i)}] & \text{egyébként.} \end{cases}$$

Jegyezzük meg, hogy

$$\operatorname{sign}([r^{(0,i)}]) = \operatorname{sign}([r^{(1,i)}]) = \cdots, \quad 1 \le i \le n,$$

mindig igaz, azaz az egyes intervallumok előjele nem változik. Az új tartalmazó intervallumokat felhasználva újraszámolhatjuk a nevezőben található kifejezést:

$$[s^{(k+1,i)}] = \prod_{j=1}^{i-1}(x^{(k+1,i)} - [y^{(k+2,j)}]) \cdot \prod_{j=i+1}^{n}(x^{(k+1,i)} - [y^{(k+1,j)}]).$$

Ezt alkalmazva az alábbi módosított egylépéses eljáráshoz jutunk:

$$\begin{cases} [y^{(0,i)}] = [x^{(0,i)}], x^{(0,i)} \in [x^{(0,i)}], \\ [x^{(k+1,i)}] = \{x^{(k,i)} - p(x^{(k,i)})/[s^{(k,i)}]\} \cap [x^{(k,i)}], \\ \text{ahol} \\ [s^{(k,i)}] = \prod_{j=1}^{i-1}(x^{(k,i)} - [y^{(k+1,j)}]) \cdot \prod_{j=i+1}^{n}(x^{(k,i)} - [y^{(k,j)}]), \\ [y^{(k+1,i)}] = \begin{cases} [\underline{x}^{(k+1,i)}, x^{(k+1,i)}] & \text{ha } \operatorname{sign}([r^{(k,i)}])\operatorname{sign}(p(x^{(k+1,i)})) > 0 \\ [x^{(k+1,i)}, \overline{x}^{(k+1,i)}] & \text{ha } \operatorname{sign}([r^{(k,i)}])\operatorname{sign}(p(x^{(k+1,i)})) < 0 \\ [x^{(k+1,i)}] & \text{egyébként} \end{cases} \\ 1 \le i \le n, \quad k \ge 0. \end{cases} \tag{8.23}$$

Meggondolható, hogy mind a (8.21) mind pedig a (8.23) eljárás a polinomok gyökeinek szimultán meghatározására szolgáló ismert eljárások intervallumos megfelelője. Az eljárások intervallumos változatának előnye, hogy nem csak egy tartalmazó intervallumot ad, hanem az említett feltételek mellett mindig konvergens. Ezt mutatjuk be a következő tételben.

**8.6. Tétel.** *Legyen adott a (8.19) polinom $n$ darab egyszeres valós gyökkel, melyek legyenek $\xi^{(i)}$, $1 \leq i \leq n$. Továbbá legyenek $[x^{(0,i)}] \ni \xi^{(i)}$, $1 \leq i \leq n$ tartalmazó intervallumok, melyekre (8.20) teljesül. Ekkor a (8.21)-ben (illetve (8.23)-ban) megadott $\{[x^{(k,i)}]\}_{k=0}^{\infty}$ iterációs sorozatra teljesül*

$$\xi^{(i)} \in [x^{(k,i)}], \quad k \geq 0$$

*és*

$$[x^{(0,i)}] \supset [x^{(1,i)}] \supset [x^{(2,i)}] \supset \cdots \quad \textit{ahol} \lim_{k\to\infty} [x^{(k,i)}] = \xi^{(i)},$$

*vagy az eljárás véges lépésben lecseng és a $[\xi^{(i)}, x^{(i)}]$ intervallumra vezet.*

A 8.6. tétel állítása a 8.1. szakasz megfelelő tételével (8.1. tétel) megegyező módon kapható.

Behelyettesítve

$$x^{(k,i)} = \frac{1}{2}(\underline{x}^{(k,i)} + \overline{x}^{(k,i)})$$

a megfelelő eljárásokba és követve a (8.21) és (8.23) konstrukciót, azonnal adódik, hogy a gyököket tartalmazó intervallumok szélessége legalább feleződik minden iterációs lépésben.

A 8.6. tétel részben igaz marada akkor is, ha a polinomnak vannak többszörös gyökei is. Ha összegyűjtjük ezeket a többszörös gyököket:

$$\xi^{(m)}, \xi^{(m+1)}, \ldots, \xi^{(n)},$$

akkor mind a (8.21), mind pedig a (8.23) eljárást meg kell változtatnunk, úgy, hogy a számításokat csak az $1 \leq i \leq m$ indexű tartalmazó intervallumokra hajtjuk végre. A 8.6. tétel állításai igazak azokra az egyszeres gyököket tartalmazó intervallumokra, amelyeken az egyes iterációs lépések számításait végezzük. A többi intervallum változatlan marad.

A (8.21) iteráció általánosítható, oly módon, hogy a 8.6. tétel $[x^{(0,i)}]$, $1 \leq i \leq n$ intervallumokra vonatkozó (8.20) kikötését egy gyengébb feltételre cseréljük. Eközben alaposan kihasználjuk, hogy $x^{(k,i)} \in [x^{(k,i)}]$ tetszőleges, és nem valamely konkrét szabály szerint választjuk, például mindig az intervallum középpontját. Egy ilyen általánosítással foglalkozik Alefeld és Herzberger.

Most részletesebben végigondoljuk a

$$\{d([x^{(k,i)}])\}_{k=0}^{\infty}, \quad 1 \leq i \leq n$$

szélességsorozat tulajdonságait. Ezért, felhasználva az (1.19), (1.20) és (1.24) összefüggéseket a (8.21) során az alábbi becslés tehető

$$\begin{aligned} d([x^{(k+1,i)}]) &\leq d(\{x^{(k,i)} - p(x^{(k,i)})/[q^{(k,i)}]\}) = \\ &= d(p(x^{(k,i)})/[q^{(k,i)}]) = |p(x^{(k,i)})| d(1/[q^{(k,i)}]). \end{aligned}$$

Mivel

$$\begin{aligned} |p(x^{(k,i)})| &= |p(x^{(k,i)}) - p(\xi^{(i)})| = |(x^{(k,i)} - \xi^{(i)}) p'(\tilde{\eta}^{(k,i)})| \leq \\ &\leq d([x^{(k,i)}]) |p'(\tilde{\eta}^{(k,i)})| \leq d([x^{(k,i)}]) |p'([x^{(0,i)}])|, \end{aligned}$$

következik

$$d([x^{(k+1,i)}]) \leq d([x^{(k,i)}]) |p'([x^{(0,i)}])| d(1/[q^{(k,i)}]).$$

Felhasználva az 1.3. szakasz 1.24. tételt igaz a következő becslés:

$$d(1/[q^{(k,i)}]) \leq \gamma^{(k,i)} d([q^{(k,i)}]),$$

és mivel

$$[q^{(k,i)}] \subseteq \prod_{j=1, j \neq i}^{n} ([x^{(0,i)}] - [x^{(0,j)}]),$$

a következő összefüggést kapjuk

$$d\left(\frac{1}{[q^{(k,i)}]}\right) \leq \gamma^{(i)} d([q^{(k,i)}]) = \gamma^{(i)} d\left(\prod_{j=1, j \neq i}^{n} (x^{(k,i)} - [x^{(k,j)}])\right)$$

ahol a $\gamma^{(i)}$ konstans csak $[x^{(0,j)}]$, $1 \le j \le n$ intervallumtól függ. Ekkor a következőkhöz jutunk

$$d\left(\frac{1}{[q^{(k,i)}]}\right) \le \gamma^{(i)} \sum_{j=1, j\neq i}^{n} \eta^{(i,j)} d([x^{(k,j)}])$$

egy alkalmas $\eta^{(i,j)}$ konstanssal, amely csak $[x^{(0,j)}]$, $1 \le j \le n$ intervallumtól függ, mivel $[x^{(k,j)}] \subseteq [x^{(0,j)}]$. A fentieket összegyűjtve az alábbi egyenlőtlenséget kapjuk

$$d([x^{(k+1,i)}]) \le |p'([x^{(0,i)}])|\gamma^{(i)} d([x^{(k,i)}]) \sum_{j=1, j\neq i}^{n} \eta^{(i,j)} d([x^{(k,j)}]), \quad 1 \le i \le n. \tag{8.24}$$

Ugyanez a meggondolás vihető végig (8.23)-ra is, ahol az egyetlen kiegészítés amit szem előtt kell tartani, hogy $[y^{(k,i)}] \subseteq [x^{(k,i)}]$. Ez az alábbi összefüggéshez vezet:

$$\begin{aligned} d([x^{(k+1,i)}]) \;\le\; & |p'([x^{(0,i)}])|\gamma^{(i)} d([x^{(k,i)}]) \Big(\textstyle\sum_{j=1}^{i-1} \eta^{(i,j)} d([x^{(k+1,j)}]) + \\ + \;& \textstyle\sum_{j=i+1}^{n} \eta^{(i,j)} d([x^{(k,j)}])\Big), \quad 1 \le i \le n. \end{aligned} \tag{8.25}$$

A következő tétel a (8.21) és (8.23) iterációk konvergencia rendjével kapcsolatos állításokat igazol.

**8.7. Tétel.** *A feltételek és megjegyzések ugyanazok, mint a 8.6. tétel esetében voltak. A (8.21)-ben definiált iteráció legalább másodrendben, a (8.23)-ban leírt iteráció pedig legalább* $1 + \sigma^{(n)}$*-rend rendben konvergál, ahol* $\sigma^{(n)} > 1$ *a*

$$\tilde{q}^{(n)}(y) = y^n - y - 1.$$

*polinom egyetlen pozitív gyöke.*

**Bizonyítás:** Az első állítás igazolása: (8.24) állításból azonnal kap-

ható:

$$\begin{aligned} d([x^{(k+1,i)}]) &\leq |p'([x^{(0,i)}])|\gamma^{(i)} \left( \sum_{j=1,j\neq i}^{n} \eta^{(i,j)} \right) (d^{(k)})^2 \\ &\leq \max_{1\leq i\leq n} \left\{ |p'([x^{(0,i)}])|\gamma^{(i)} \left( \sum_{j=1,j\neq i}^{n} \eta^{(i,j)} \right) \right\} (d^{(k)})^2 \\ &\leq \gamma (d^{(k)})^2, \quad 1 \leq i \leq n, \end{aligned}$$

ahol

$$d^{(k)} = \max_{1\leq i\leq n} \{ d([x^{(k,i)}]) \}.$$

Amiből következik, hogy

$$d^{(k+1)} = \max_{1\leq i\leq n} \{ d([x^{(k+1,i)}]) \} \leq \gamma (d^{(k)})^2,$$

és pont ezt állítottuk.

A második állítás igazolása sem igényel nagyobb erőfeszítést, mint az előző állításé. Legyen

$$\gamma = \max_{1\leq i,j\leq n} \{ \eta^{(i,j)} |p'([x^{(0,i)}])|\gamma^{(i)} \}.$$

Ekkor visszaírva (8.25)-be:

$$d([x^{(k+1,i)}]) \leq \gamma d([x^{(k,i)}]) \left( \sum_{j=1}^{i-1} d([x^{(k+1,j)}]) + \sum_{j=i+1}^{n} d([x^{(k,j)}]) \right).$$

Felhasználva a

$$d([x^{(k,i)}]) = \frac{1}{(n-1)\gamma} h^{(k,i)}, \quad q \leq i \leq n, \quad \hat{\varepsilon} = \frac{1}{n-1},$$

helyettesítést, az alábbi formában írható

$$h^{(k+1,i)} \leq \hat{\varepsilon} h^{(k,i)} \left( \sum_{j=1}^{i-1} h^{(k+1,j)} + \sum_{j=i+1}^{n} h^{(k,j)} \right).$$

Az általánosság megszorítása nélkül feltehető, hogy

$$h^{(0,i)} \leq h \leq 1, \quad 1 \leq i \leq n.$$

Ekkor

$$h^{(k+1,i)} \leq h^{u^{(k+1,i)}}, \quad 1 \leq i \leq n, \ k \geq 0.$$

Az $\dot{u}^{(k+1)}$ egészkoordinátás vektor az alábbi szabály alapján számolható

$$\dot{u}^{(k+1)} = \dot{\mathbf{A}}\dot{u}^{(k)},$$

az $\dot{u}^{(0)} = (1, 1, \ldots, 1)^T$ kiindulási vektor segítségével. (A valós kordinátájú vektorok, úgynevezett *pont-vektorok* jelölésére a $\dot{a}, \dot{b} \ldots$ jelölést használjuk, hogy meg tudjuk őket különböztetni az intervallum-vektoroktól. Hasonló jelölést alkalmazunk a ,,pont mátrixokra" is.

Az $\dot{\mathbf{A}}$ mátrix az alábbi alakú

$$\dot{\mathbf{A}} = \begin{pmatrix} 1 & 1 & & & & \\ & 1 & 1 & & & \\ & & 1 & 1 & & \\ & & & \ddots & \ddots & \\ & & & & 1 & 1 \\ 1 & 1 & 0 & \cdots & 0 & 1 \end{pmatrix}.$$

A reláció teljes indukcióval igazolható, amitől itt eltekintünk. Az $\dot{\mathbf{A}}$ mátrix nem-negatív és az irányított gráfja tisztán, erősen összefüggő. Ebből következik, hogy az $\dot{\mathbf{A}}$ felbonthatatlan. A Perron-Frobenius tételből következik, hogy az $\dot{\mathbf{A}}$ mátrix egy $\lambda^{(1)}$ sajátértéke megegyezik a $\rho(\dot{\mathbf{A}})$ spektrálsugárral. Az $\dot{\mathbf{A}}$ primitív, lásd [17]. Az $\dot{\mathbf{A}}$ mátrix többi sajátértéke kielégíti az alábbi összefüggést

$$\lambda^{(1)} = \rho(\dot{\mathbf{A}}) > |\lambda^{(2)}| \geq \cdots \geq |\lambda^{(n)}|.$$

Mivel $\dot{\mathbf{A}}$ primitív, ezért egy $k^{(0)}$ természetes számra

$$\dot{\mathbf{A}}^{(k)} = (a_{ij}^{(k)}) > \dot{\mathbf{O}}, \quad k \geq k^{(0)}.$$

Ahogy Gröbner megmutatta, az ilyen mátrixokra, melyek ez utóbbi két tulajdonsággal rendelkeznek, igaz

$$\lim_{k\to\infty}(a_{ij}^{(k+1)}/a_{ij}^{(k)}) = \lambda^{(1)}.$$

Egy adott $\varepsilon > 0$ esetén vagy

$$a_{ij}^{(k+1)}/a_{ij}^{(k)} \geq \rho(\underset{\cdot}{\mathbf{A}}) - \varepsilon, \quad k \geq k(\varepsilon) \geq k^{(0)}$$

igaz, vagy

$$a_{ij}^{(k+1)} \geq \alpha(\rho(\underset{\cdot}{\mathbf{A}}) - \varepsilon), \quad 1 \leq i, j \leq n$$

igaz, ahol

$$\alpha = \min_{1\leq i,j\leq n} a_{ij}^{(k)} > 0.$$

Ebből következik, hogy

$$a_{ij}^{(k+2)} \geq a_{ij}^{(k+1)}(\rho(\underset{\cdot}{\mathbf{A}}) - \varepsilon) \geq \alpha(\rho(\underset{\cdot}{\mathbf{A}}) - \varepsilon)^2$$

vagy általánosan

$$a_{ij}^{(k+r)} \geq \alpha(\rho(\underset{\cdot}{\mathbf{A}}) - \varepsilon)^r, \quad 1 \leq i,\ j \leq n,\ r \geq 0.$$

Ha ezt felhasználjuk az $\underset{\cdot}{u}$ vektor kiszámítási szabályában, akkor

$$\underset{\cdot}{u}^{(k+r)} = \underset{\cdot}{\mathbf{A}}^{k+r}\underset{\cdot}{u}^{(0)} = \left(\sum_{j=1}^{n} a_{ij}^{(k+r)}\right) \geq (n\alpha(\rho(\underset{\cdot}{\mathbf{A}}) - \varepsilon)^r)\underset{\cdot}{e}$$

kapjuk, ahol $\underset{\cdot}{e} = (1, 1, \ldots, 1)^T$. És így azt kapjuk, hogy

$$h^{(k+r,i)} \leq h^{u^{(k+r,i)}} \leq h^{n\alpha(\rho(\underset{\cdot}{\mathbf{A}})-\varepsilon)^r},$$

$$1 \leq i \leq n, \quad r \geq 0, k \geq k(\varepsilon) \geq k^{(0)}.$$

Másképp kifejezve ez azt jelenti, hogy

$$d([x^{(k+r,i)}]) \leq (\hat{\varepsilon}/\gamma)h^{n\alpha(\rho(\underset{\cdot}{\mathbf{A}})-\varepsilon)^r}.$$

Legyen most

$$d^{(k)} = \max_{1\le i\le n}\{d([x^{(k,i)}])\}.$$

Ekkor azt kapjuk, hogy

$$d^{(k+r)} \le (\hat{\varepsilon}/\gamma)h^{n\alpha(\rho(\dot{\mathbf{A}})-\varepsilon)^r}.$$

Tehát megállapíthatjuk, hogy az $R$ tényező kielégíti az alábbiakat

$$\begin{aligned} R_{\rho(\dot{\mathbf{A}})-\varepsilon}\{d^{(k)}\} &= \limsup_{r\to\infty}(d^{(k+r)})^{[1/(\rho(\dot{\mathbf{A}})-\varepsilon)^r]} \\ &\le \limsup_{r\to\infty}\left(\frac{\hat{\varepsilon}}{\gamma}h^{n\alpha(\rho(\dot{\mathbf{A}})-\varepsilon)^r}\right)^{[1/(\rho(\dot{\mathbf{A}})-\varepsilon)^r]} \\ &= h^{\alpha n} < 1. \end{aligned}$$

Ebből következik, hogy a konvergencia rend legalább $\rho(\dot{\mathbf{A}}) - \varepsilon$ bármely $\varepsilon > 0$ esetén és innen, hogy nem kisebb $\rho(\dot{\mathbf{A}})$-nál.

Vizsgáljuk most az $\dot{\mathbf{A}}$ mátrix $q^{(n)}(\lambda)$ karakterisztikus polinomját

$$q^{(n)}(\lambda) = (\lambda-1)^n - (\lambda-1) - 1.$$

$\tau = \lambda - 1$ helyettesítés mellett ez

$$\tilde{q}^{(n)}(\tau) = \tau^n - \tau - 1,$$

alakban írható.

A $\tilde{q}^{(n)}(\tau)$ polinomnak a Descartes-szabály értelmében pontosan egy $\sigma^{(n)}$ pozitív gyöke van, amelyre

$$1 < \sigma^{(n)} < 2$$

mivel

$$\tilde{q}^{(n)}(1) = -1, \quad \text{és} \quad \tilde{q}^{(n)}(2) = 2^n - 3 \ge 1 > 0$$

igaz, ha $n \ge 2$. Az $\dot{\mathbf{A}}$ mátrix spektrálsugara tehát kielégíti a

$$\rho(\dot{\mathbf{A}}) = 1 + \sigma^{(n)} > 2$$

összefüggést, amiből az állítás második része következik. □

A (8.23) iteráció egy alkalmazását szeretnénk az alábbiakban bemutatni. Adott egy valós, szimmetrikus $n \times n$-es $\mathbf{A}' = (a_{ij})$ mátrix. Az $\mathbf{A}'$ mátrix sajátértékeinek nevezzük azokat a $\lambda$ számokat, melyekre

$$\mathbf{A}'x = \lambda x, \quad \text{ahol} \quad x \neq o$$

igaz. Ezek meghatározásához véges sok ortogonális hasonlósági transzformációt hajtunk végre

$$\tilde{\mathbf{A}} = \mathbf{U}^T \mathbf{A} \mathbf{U},$$

melyekkel az általános teli mátrixot az alábbi alakú $\mathbf{A}$ mátrixra transzformáljuk

$$\mathbf{A} = \begin{pmatrix} a^{(1)} & b^{(1)} & & & \\ b^{(1)} & a^{(2)} & b^{(2)} & & \\ & \ddots & \ddots & \ddots & \\ & & & b^{(n-1)} & a^{(n)} \end{pmatrix}.$$

Az $\mathbf{A}$ mátrix sajátértékei (és így az $\mathbf{A}'$ mátrixé is) az $\mathbf{A}$ mátrix

$$p(\lambda) = \det(\lambda \mathbf{I} - \mathbf{A})$$

karakterisztikus polinomjának gyökeiként számolhatóak. $p(\lambda)$ értéke az alábbi rekurzióval határozható meg

$$\begin{cases} f^{(0)}(\lambda) &= 1, \quad f^{(0)}(\lambda) = \lambda - a^{(1)}, \\ f^{(k)}(\lambda) &= (\lambda - a^{(k)}) f^{(k-1)}(\lambda) - (b^{(k-1)})^2 f^{(k-2)}(\lambda), \quad 2 \leq k \leq n, \\ p(\lambda) &= f^{(n)}(\lambda). \end{cases} \tag{8.26}$$

Ha az $\mathbf{A}$ mátrix sajátértékei mind egyszeresek és ismertek páronként diszjunkt tartalmazó intervallumok, például a Gersgorin tétel alapján, akkor alkalmazható a (8.23) eljárás. A következő példa ezt demonstrálja.

**Példa:** ($\alpha$) Tekintsük az alábbi mátrixot

$$\dot{\mathbf{A}} = \begin{pmatrix} 15 & 1 & & & & & & & \\ 1 & 10 & 1 & & & & & & \\ & 1 & 7 & 1 & & & & & \\ & & 1 & 4 & 1 & & & & \\ & & & 1 & 0 & 1 & & & \\ & & & & 1 & -4 & 1 & & \\ & & & & & 1 & -7 & 1 & \\ & & & & & & 1 & -10 & 1 \\ & & & & & & & 1 & -15 \end{pmatrix}.$$

A Gersgorin tételt alkalmazva, az $\dot{\mathbf{A}}$ mátrix sajátértékeire az alábbi tartalmazó intervallumokat nyerjük:

$$\begin{aligned}
[x^{(0,1)}] &= [+13.99999999995, +16.00000000005], \\
[x^{(0,2)}] &= [+7.999999999974, +12.00000000005], \\
[x^{(0,3)}] &= [+4.999999999981, +9.000000000015], \\
[x^{(0,4)}] &= [+1.999999999992, +6.000000000022], \\
[x^{(0,5)}] &= [-2.000000000008, +2.000000000008], \\
[x^{(0,6)}] &= [-6.000000000022, -1.999999999992], \\
[x^{(0,7)}] &= [-9.000000000015, -4.999999999981], \\
[x^{(0,8)}] &= [-12.00000000005, -7.000000000074], \\
[x^{(0,9)}] &= [-17.00000000005, -12.99999999995].
\end{aligned}$$

Ezekkel a kiindulási intervallumokkal a (8.23) iterációt alkalmazva a

következő eredményeket kapjuk:

$$\begin{aligned}
[x^{(5,1)}] &= [+\underline{15.19709300}868, && +\underline{15.19709300}872],\\
[x^{(4,2)}] &= [+\underline{10.13174515}464, && +\underline{10.13174515}471],\\
[x^{(4,3)}] &= [+\underline{7.001927580}904, && +\underline{7.001927580}971],\\
[x^{(4,4)}] &= [+\underline{3.920346203}678, && +\underline{3.920346203}715],\\
[x^{(5,5)}] &= [-0,1096791595101\cdot 10^{-10}, && +0,1096791595101\cdot 10^{-10}],\\
[x^{(4,6)}] &= [-\underline{3.920346203}719, && -\underline{3.920346203}674],\\
[x^{(4,7)}] &= [-\underline{7.001927580}969, && -\underline{7.001927580}895],\\
[x^{(3,8)}] &= [-\underline{10.13174515}473, && -\underline{10.13174515}463],\\
[x^{(3,9)}] &= [-\underline{15.19709300}876, && -\underline{15.19709300}866],
\end{aligned}$$

Ezek az intervallumok nem javíthatóak a program további alkalmazásával semmiképpen. Az alsó és fölső határokban megegyező jegyeket aláhúzással jelöltük.

($\beta$) Tekintsük most a következő mátrixot

$$\underset{\cdot}{\mathbf{A}} = \begin{pmatrix} 12 & 1 & & & \\ 1 & 9 & 1 & & \\ & 1 & 6 & 1 & \\ & & 1 & 3 & 1 \\ & & & 1 & 0 \end{pmatrix}.$$

Újra használjuk a Gersgorin tételt és így az alábbi tartalmazó intervallumokat kapjuk az $\underset{\cdot}{\mathbf{A}}$ mátrix sajátértékeire:

$$\begin{aligned}
[x^{(0,1)}] &= [+10.99999999998, +13.00000000003],\\
[x^{(0,2)}] &= [+6.999999999970, +11.00000000003],\\
[x^{(0,3)}] &= [+3.999999999989, +8.000000000021],\\
[x^{(0,4)}] &= [+0.9999999999945, +5.000000000019],\\
[x^{(0,5)}] &= [-1.000000000004, -1.000000000004].
\end{aligned}$$

A következő javított intervallumok adódtak, ha a (8.23) iterációs eljárást használtuk. (Hasonlítsuk össze az eredményt a következő 8.6. tétel megjegyzéseivel):

$$
\begin{aligned}
[x^{(1,1)}] &= [+\underline{12}.11013986010, +\underline{12}.55506993010],\\
[x^{(1,2)}] &= [+\underline{9.0}06328989416, +\underline{9.0}, 48379503166],\\
[x^{(1,3)}] &= [+\underline{5.999999999}958, +\underline{6.000000000}041],\\
[x^{(1,4)}] &= [+\underline{2.9}79804773200, +\underline{2.9}87022580008],\\
[x^{(1,5)}] &= [-0.\underline{3}230758693540, -0.\underline{3}162523763767],
\end{aligned}
$$

$$
\begin{aligned}
[x^{(2,1)}] &= [+\underline{12.31}617201370, +\underline{12.31}774922532],\\
[x^{(2,2)}] &= [+\underline{9.016}110401580, +\underline{9.016}149094187],\\
[x^{(2,3)}] &= [+\underline{5.999999999}958, +\underline{6.000000000}013],\\
[x^{(2,4)}] &= [+\underline{2.983860}239266, +\underline{2.983864}788268],\\
[x^{(2,5)}] &= [-0.\underline{3168759526}293, -0.\underline{3168759526}051],
\end{aligned}
$$

$$
\begin{aligned}
[x^{(3,1)}] &= [+\underline{12.31687595}112, +\underline{12.31687595}546],\\
[x^{(3,2)}] &= [+\underline{9.016136303}134, +\underline{9.016136303}198],\\
[x^{(3,3)}] &= [x^{(2,3)}]\\
[x^{(3,4)}] &= [+\underline{2.983863696}823, +\underline{2.983863696}853],\\
[x^{(3,5)}] &= [-0.\underline{3168759526}293, -0.\underline{3168759526}051],
\end{aligned}
$$

$$
\begin{aligned}
[x^{(4,1)}] &= [+\underline{12.3168759526}58, +\underline{12.3168759526}66],\\
[x^{(4,2)}] &= [+\underline{9.016136303}134, +\underline{9.016136303}181],\\
[x^{(4,3)}] &= [x^{(3,3)}]\\
[x^{(4,4)}] &= [x^{(3,4)}]\\
[x^{(4,5)}] &= [-0.\underline{3168759526}284, -0.\underline{3168759526}051],
\end{aligned}
$$

## 8.6. Polinomok komplex zérushelyeinek szimultán meghatározása

Ebben a fejezetben egy polinom általában komplex gyökeinek szimultán meghatározására szolgáló eljárást fogunk tárgyalni Gargantini és Henrici által ismertetett módon [16].
Legyen adott egy $p(z)$ polinom

$$p(z) = a^{(n)}z^n + a^{(n-1)}z^{n-1} + \cdots + a^{(1)}z + a^{(0)}, \tag{8.27}$$

ahol $a^{(i)} \in \mathbb{C},\ 0 \leq i \leq n,\ n \geq 2$. Továbbá tegyük fel, hogy adott $n$ intervallum,

$$[w^{(0,i)}] =< z^{(0,i)}, r^{(0,i)} >\in K\mathbb{C},$$

melyekre

$$\zeta^{(i)} \in [w^{(0,i)}], \quad p(\zeta^{(i)}) = 0, \quad 1 \leq i \leq n, \tag{8.28}$$

$$[w^{(0,i)}] \cap [w^{(0,j)}] = \emptyset, \quad 1 \leq i < j \leq n, \tag{8.29}$$

Egy $[z] \in K\mathbb{C}$ a továbbiakban $[z] = \langle m([z]), r([z])\rangle$-vel is reprezentálható.

Tekintsük a következő iterációt

$$\begin{cases} z^{(k,i)} = m([w^{(k,i)}]), \\ [c^{(k,i)}] = \sum\limits_{j=1, j\neq i}^{n} \frac{1}{z^{(k,i)} - [w^{(k,j)}]}, \\ q(z^{(k,i)}) = \frac{p'(z^{(k,i)})}{p(z^{(k,i)})}, \quad \text{ha} \quad p(z^{(k,i)}) \neq 0, \\ [w^{(k+1,i)}] =< z^{(k+1,i)}, r^{(k,i)} >= -\frac{1}{q(z^{(k,i)}) - [c^{(k,i)}]}, \end{cases} \tag{8.30}$$

$$1 \leq i \leq n, \quad k \geq 0,$$

és legyen

$$r^{(k)} = \max_{1\leq i\leq n}\{r^{(k,i)}\}, \tag{8.31}$$

$$\rho^{(k)} = \min_{1\leq i<j\leq n}\{\min\{|z| \mid z \in z^{(k,i)} - [w^{(k,j)}]\}\}. \tag{8.32}$$

$i \neq j$ esetén (8.29)-ből következik, hogy

$$\min\{|z| \ |z \in z^{(0,i)} - [w^{(0,j)}]\} = |z^{(0,i)} - z^{(0,j)}| - r^{(0,j)} \geq \rho^{(0)}. \tag{8.33}$$

Továbbá legyen $\eta^{(k)}$ az alábbi módon definiálva

$$\rho^{(k)} = (n-1)\eta^{(k)}. \tag{8.34}$$

Ekkor a következő igaz a (8.30) iterációs rendszerre.

**8.8. Tétel.** *Legyen $p(z)$ egy (8.27)-ben felírt polinom, melynek gyökei $\zeta^{(i)}$, $1 \le i \le n$, és amely kielégíti a (8.28) és (8.29) feltételeket. (8.31), (8.32) és (8.34) jelöléseivel legyen*

$$6r^{(0)} \le \eta^{(0)}. \tag{8.35}$$

*(a) Ekkor a (8.30) iteráció mindig végrehajtható, továbbá*

$$\zeta^{(i)} \in [w^{(k,i)}], \quad 1 \le i \le n, \quad k \ge 0.$$

*(b) Mindig igaz az*

$$r^{(k+1)} \le \frac{1}{\rho^{(0)}(\eta^{(0)} - 4r^{(0)})}(r^{(k)})^3 \le \frac{1}{12(n-1)}r^{(k)}, \quad k \ge 0,$$

*egyenlőtlenség.*

Megjegyzés: (b)-ből következően $\lim_{k\to\infty} r^{(k)} = 0$, valamint (a) miatt szükségszerűen teljesül, hogy

$$\lim_{k\to\infty} [w^{(k,i)}] = \zeta^{(i)}, \quad 1 \le i \le n.$$

A (8.30) iteráció legalább harmadrendben konvergens.

**Bizonyítás:** (a) bizonyítása: Mivel

$$\begin{aligned} |z^{(0,i)} - \zeta^{(i)}| &\le r^{(0,i)} \le r^{(0)}, \\ |z^{(0,i)} - \zeta^{(j)}| &\ge |z^{(0,i)} - z^{(0,j)}| - |z^{(0,j)} - \zeta^{(j)}| \\ &\ge |z^{(0,i)} - z^{(0,j)}| - r^{0,j} \ge \rho^{(0)}, \end{aligned}$$

következik, hogy

$$\begin{aligned} |q(z^{(0,i)})| &= \left| \sum_{j=1}^{n} \frac{1}{z^{(0,i)} - \zeta^{(j)}} \right| \\ &\ge \left| \frac{1}{z^{(0,i)} - \zeta^{(i)}} \right| - \sum_{j=1, j\ne i}^{n} \left| \frac{1}{z^{(0,i)} - \zeta^{(j)}} \right| \\ &\ge \frac{1}{r^{(0)}} - \frac{1}{\eta^{(0)}}, \quad \text{ha } z^{(0,i)} \ne \zeta^{(i)}. \end{aligned} \tag{8.36}$$

$$|z^{(0,i)} - z^{(0,j)}| - r^{(0,j)} \geq \rho^{(0)} > 0,$$

relációból kiindulva

$$0 \notin z^{(0,i)} - [w^{(0,j)}]$$

kapjuk, éppúgy, mint

$$\frac{1}{z^{(0,i)} - [w^{(0,i)}]} \subset \left\langle 0, \frac{1}{\rho^{(0)}} \right\rangle,$$

$$\begin{aligned} [c^{(0,i)}] &= \sum_{j=1, j\neq i}^{n} \frac{1}{z^{(0,i)} - [w^{(0,i)}]} \subset \\ &\subset \sum_{j=1, j\neq i}^{n} \left\langle 0, \frac{1}{\rho^{(0)}} \right\rangle = \left\langle 0, \frac{1}{\rho^{(0)}} \right\rangle, \end{aligned}$$

$$q(z^{(0,i)}) - [c^{(0,i)}] \subset < q(z^{(0,i)}), 1/\eta^{(0)} > . \tag{8.37}$$

Mivel

$$|q(z^{(0,i)})| - 1/\eta^{(0)} \geq 1/r^{(0)} - 2/\eta(0) > 0,$$

nyilvánvalóan

$$0 \notin q(z^{(0,i)}) - [c^{(0,i)}]$$

és ezért

$$[w^{(1,i)}], \quad 1 \leq i \leq n,$$

meghatározott. Mivel

$$\frac{p'(z^{(0,i)})}{p(z^{(0,i)})} = \sum_{j=1}^{n} \frac{1}{z^{(0,i)} - \zeta^{(j)}},$$

ezért (8.28)-t és a tartalmazás monotonitását felhasználva következik, hogy

$$\begin{aligned} \zeta^{(i)} &= \frac{z^{(0,i)} - p(z^{(0,i)})}{p'(z^{(0,i)}) - p(z^{(0,i)}) \sum\limits_{j=1, j\neq i}^{n} \frac{1}{z^{(0,i)} - \zeta^{(j)}}} \in \\ &\in z^{(0,i)} - \frac{1}{q(z^{(0,i)}) - [c^{(0,i)}]} = [w^{(0,i)}], \quad 1 \leq i \leq n. \end{aligned}$$

Ezzel az (a) részt bizonyítottuk $k = 1$ esetre.

(b) bizonyítása: Kiindulva az

$$|z^{(0,i)} - z^{(0,j)}|^2 - (r^{(0,j)})^2 \geq (\rho^{(0)} + r^{(0,j)})^2 - (r^{(0,j)})^2 \geq (\rho^{(0)})^2,$$

egyenlőtlenségből kapjuk, hogy

$$r\left(\frac{1}{z^{(0,i)} - [w^{(0,j)}]}\right) = \frac{r^{(0,j)}}{|z^{(0,i)} - z^{(0,j)}|^2 - (r^{(0,j)})^2} \leq \frac{r^{(0)}}{(\rho^{(0)})^2}$$

és ezért

$$r([c^{(0,i)}]) \leq \frac{n-1}{\rho^{(0)}} \cdot \frac{r^{(0)}}{\rho^{(0)}} = \frac{r^{(0)}}{\eta^{(0)}\rho^{(0)}}.$$

Felhasználva ezt az egyenlőtlenséget úgy mint (8.37) most

$$r(q(z^{(0,i)}) - [c^{(0,i)}]) = r([c^{(0,i)}]),$$

$$\begin{aligned} |m(q(z^{(0,i)}) - [c^{(0,i)}])| &\geq 1/r^{(0)} - 2/\eta^{(0)} + r(q(z^{(0,i)}) - [c^{(0,i)}]) = \\ &= 1/r^{(0)} - 2/\eta^{(0)} + r([c^{(0,i)}]) \end{aligned}$$

kapjuk, ezért az

$$\begin{aligned} r([w^{(0,i)}]) &= r\left(\frac{1}{q(z^{(0,i)}) - [c^{(0,i)}]}\right) = \\ &= \frac{r(q(z^{(0,i)}) - [c^{(0,i)}])}{|m(q(z^{(0,i)}) - [c^{(0,i)}])|^2 - (r(q(z^{(0,i)}) - [c^{(0,i)}]))^2} \leq \\ &\leq \frac{(r^{(0)})^3}{\rho^{(0)}(\eta^{(0)} - 4r^{(0)})}, \end{aligned}$$

egyenlőtlenségből kapjuk, hogy

$$r^{(1)} \leq \frac{(r^{(0)})^3}{\rho^{(0)}(\eta^{(0)} - 4r^{(0)})}. \tag{8.38}$$

Felhasználva (8.35)-t kapjuk az alábbi egyenlőtlenséget a fenti becslésből

$$r^{(1)} \leq \frac{1}{12(n-1)} r^{(0)}.$$

Legyen

$$\delta^{(0)} = \max_{1\le i\le n}\{|z^{(0,i)} - z^{(1,i)}|\}.$$

Ekkor (8.32) felhasználásával kapjuk

$$\rho^{(1)} \ge \rho(0) - \delta^{(0)} - 2r^{(1)}. \tag{8.39}$$

$\delta^{(0)}$ becsléséhez felhasználjuk (8.36), (8.37) és az alábbi relációkat

$$z^{(1,i)} - z^{(0,i)} \in \frac{1}{q(z^{(0,i)}) - [c^{(0,i)}]},$$

hogy a következőt nyerjük

$$\begin{aligned} |z^{(1,i)} - z^{(0,i)}| \le \left|\frac{1}{< q(z^{(0,i)}), 1/\eta^{(0)} >}\right| &= \frac{1}{|q(z^{(0,i)})| - 1/\eta^{(0)}} \\ &\le \frac{r^{(0)\eta^{(0)}}}{\eta^{(0)} - 2r^{(0)}}, \end{aligned}$$

amely végül az alábbi becslést adja

$$\delta^{(0)} \le \frac{r^{(0)\eta^{(0)}}}{\eta^{(0)} - 2r^{(0)}}. \tag{8.40}$$

A (8.35) egyenlőtlenségből kiindulva és felhasználva (8.38), (8.39) és (8.40) egyenlőtlenségeket következik az alábbi

$$\begin{aligned} \eta^{(1)} - 6r^{(1)} &= \rho^{(1)}/(n-1) - 6r^{(1)} \\ &\ge \eta^{(0)} - r^{(0)}\left(\tfrac{\eta^{(0)}}{\eta^{(0)}-2r^{(0)}} + \tfrac{8(r^{(0)})^2}{\rho^{(0)}(\eta^{(0)}-4r^{(0)})}\right) \\ &\ge \eta^{(0)} - 3r^{(0)} \ge 0; \end{aligned} \tag{8.41}$$

ami alapján

$$\eta^{(1)} \ge 6r^{(1)}.$$

Ezt felhasználva, a fentiekhez hasonló módon megmutatható, hogy

$$r^{(2)} \le \frac{1}{\rho(1)(\eta^{(1)} - 4r^{(1)})}(r^{(1)})^3 \le \frac{1}{12(n-1)}r^{(1)}.$$

A (8.39)-ből kiindulva a (8.41)-hez hasonló módon következik

$$\eta^{(1)} - 4r^{(1)} \geq \eta^{(0)} - r^{(0)} \left( \frac{\eta^{(0)}}{\eta^{(0)} - 2r^{(0)}} + \frac{6(r^{(0)})^2}{\rho^{(0)}(\eta^{(0)} - 4r^{(0)})} \right), \qquad (8.42)$$

úgy mint

$$\eta^{(1)} \geq \eta^{(0)} - r^{(0)} \left( \frac{\eta^{(0)}}{\eta^{(0)} - 2r^{(0)}} + \frac{2(r^{(0)})^2}{\rho^{(0)}(\eta^{(0)} - 4r^{(0)})} \right) \geq 0. \qquad (8.43)$$

Felhasználva mindkét fenti egyenlőtlenséget, (8.35)-ből kiindulva kapjuk

$$\begin{aligned} \eta^{(1)}(\eta^{(1)} - 4r^{(1)}) &\geq (\eta^{(0)})^2 - \eta^{(0)} r^{(0)} \left( \frac{2\eta^{(0)}}{\eta^{(0)} - 2r^{(0)}} + \frac{8(r^{(0)})^2}{\rho^{(0)}(\eta^{(0)} - 4r^{(0)})} \right) \\ &\geq \eta^{(0)}(\eta^{(0)} - 4r^{(0)}) \end{aligned}$$

és ezért

$$r^{(2)} \leq \frac{1}{\rho^{(0)}(\eta^{(0)} - 4r^{(0)})} (r^{(})^3.$$

A tételt a megmaradt esetekre teljes indukcióval lehet bizonyítani. □

Most (8.30) iteráció egy alkalmazását fogjuk bemutatni. Ehhez egy alsó Hessenberg mátrix sajátértékeinek kiszámításának problémáját fogjuk vizsgálni, felhasználva a tartalmazó intervallumok egy sorozatát. Az iterációhoz szükségesek a karakterisztikus polinom és a deriváltjának helyettesítési értékei. Konkrét példaként tekintsük az alábbi mátrixot

$$\mathbf{H} = \begin{pmatrix} 12 + 16i & 1 & 0 & 0 \\ 0 & 9 + 12i & 1 & 0 \\ 0 & 0 & 6 + 8i & 1 \\ 1 & 0 & 0 & 3 + 4i \end{pmatrix},$$

ahol $i = \sqrt{-1}$. A Gersgorin-tétel értelmében a

$$\begin{aligned} [w^{(0,1)}] &= \langle 12 + 16i, 1 \rangle, \quad [w^{(0,2)}] = \langle 9 + 12i, 1 \rangle, \\ [w^{(0,3)}] &= \langle 6 + 8i, 1 \rangle, \quad [w^{(0,4)}] = \langle 3 + 4i, 1 \rangle \end{aligned}$$

körlapok pontosan egy-egy sajátértékét tartalmazzák a $\mathbf{H}$ mátrixnak. A (8.30) eljárás segítségével a következő $[w^{(k,i)}]$ javított tartalmazó halmazokat kapjuk a $\mathbf{H}$ mátrix sajátértékeire, ahol

$$[w^{(k,i)}] =< m([w^{(k,i)}]), r([w^{(k,i)}]) >,$$

reprezentáció az alábbi jelöléseket használva

$$m([w^{(k,i))}]) = \Re(m([w^{(k,i)}])) + i\Im(m([w^{(k,i)}])).$$

A számítások eredményeit a 8.6. táblázat tartalmazza.

| k | i | Re | Im | r |
|---|---|---|---|---|
| 1 | 1 | +11.99875131516 | +15.99953080496 | $0.1001255 \cdot 10^{-6}$ |
| | 2 | +9.003742419628 | +12.00140833328 | $0.1494005 \cdot 10^{-5}$ |
| | 3 | +5.996257580383 | +7.998591666711 | $0.1493969 \cdot 10^{-5}$ |
| | 4 | +3.001248654837 | +4.000469195035 | $0.1000782 \cdot 10^{-6}$ |
| 2 | 1 | +11.99875136181 | +15.99953080159 | $0.1019500 \cdot 10^{-9}$ |
| | 2 | +9.003742437190 | +12.00140832752 | $0.8760740 \cdot 10^{-10}$ |
| | 3 | +5.996257562811 | +7.998591672458 | $0.3665239 \cdot 10^{-10}$ |
| | 4 | +3.001248638204 | +4.000469198423 | $0.2555951 \cdot 10^{-10}$ |
| 3 | 1 | +11.99875136181 | +15.99953080159 | $0.1019496 \cdot 10^{-9}$ |
| | 2 | +9.003742437190 | +12.00140832752 | $0.8760740 \cdot 10^{-10}$ |
| | 3 | +5.996257562811 | +7.998591672458 | $0.3665353 \cdot 10^{-10}$ |
| | 4 | +3.001248638204 | +4.000469198423 | $0.2556093 \cdot 10^{-10}$ |

8.6. táblázat.

# 9. fejezet

# Globális optimalizáció

A fejezet célja betekintést nyújtani a többváltozós, feltétel nélküli nemlineáris optimalizálás problémájába. A feladat a következő: adott egy $f : \mathbb{R}^n \to \mathbb{R}$, nem feltétlenül lineáris függvény és egy $S \subset \mathcal{D}_f$ részhalmaz, amely felett a minimalizálást végezzük, azaz keressük az

$$f^* = \min_{x \in S} f(x),$$

illetve

$$X^* = \{x^* \in S \mid f(x^*) = f^*\}$$

értékeket, vagyis a minimum értékét és azokat az $S$-beli pontokat amelyekben ez a minimum felvétetik.

A többváltozós optimalizáció klasszikus numerikus módszerei általában közelítő megoldásokból indulnak ki és ezeket iteratívan finomítják, vagyis lényegében a célfüggvényt véges sok pontban mintavételezve próbálnak globális optimumot meghatározni. Azonban nincs biztosíték arra, hogy ezen kipróbált pontokon kívül ne lennének kiugróan alacsony értékei az optimalizálandó függvénynek.

Hansen globális optimalizációs algoritmusának ebben a fejezetben bemutatásra kerülő változata az intervallum aritmetika felhasználásával a célfüggvényt, illetve annak első és második parciális deriváltjait véges sok pont felett értékeli ki, és a végül eredményül kapott értékek *automatikusan ellenőrzött* optimum befoglaló intervallumok lesznek, azaz a kapott intervallumok garantáltan tartalmazzák a globális minimalizáló helyeket.

# 9.1. Elméleti háttér

A továbbiakban legyen $f : \mathbb{R}^n \to \mathbb{R}$ kétszer folytonosan deriválható függvény. Jelölje $\underline{f_y}$ az $f$-nek $\mathbf{y}$-on vett intervallumkiértékelésének alsó határát és legyen $\mathbf{x} \in \mathbb{IR}^n$ a minimumkeresés intervalluma. Feladatunk az összes olyan $x^* \in \text{int}(\mathbf{x})$ pont megkeresése, amelyre

$$f(x^*) = \min_{x \in \mathbf{x}} f(x),$$

azaz $x^*$ stacionárius pontja $f$-nek.

Hansen algoritmusa egy listában tárolja azon intervallumokat, amelyek tartalmazhatják a globális minimumhelyeket. Ezt a listát aztán minden iterációs lépésben tovább próbálja finomítani, egyrészt a minimumot garantáltan nem tartalmazó intervallumok eltávolításával, illetve az így megmaradtak felosztásával vagy minimumot nem tartalmazó részeik elhagyásával.

Az algoritmus hatékonysága elsősorban abban rejlik, hogy az optimumot nem tartalmazó intervallumok vagy részintervallumok eldobásának következtében gyorsan és nagy mértékben csökkenti az optimumot tartalmazó intervallumjelöltek számát.

Az intervallumfelosztás és eldobás négy teszt segítségévével valósul meg:

- középponti teszt
- monotonitási teszt
- konkavitási teszt
- intervallumos Newton Jacobi lépés

Az algoritmus iterációs része akkor áll le, ha a listában lévő intervallumok szélessége egy előre meghatározott hibaküszöb alá esik. Ezután egy verifikációs lépés során megállapítjuk, hogy a megmaradó intervallumok közül melyek azok, amelyekben létezik és egyértelmű a minimumhely.

Először azonban tárgyaljuk az itt alkalmazott Newton Jacobi lépés elméletét és az intervallum aritmetika egy számunkra szükséges kiterjesztését.

## 9.2. Newton Jacobi lépés

Legyen $f : \mathbb{R}^n \to \mathbb{R}^n$ folytonosan differenciálható vektor értékű függvény, jelölje $J_f(x)$ az $f$ Jacobi-mátrixát az $x \in \mathbf{x}$ pontban. Keressük az $f$ zérushelyeit.

A centrális alakot a Taylor-sorfejtéssel alkalmazva felírhatjuk, hogy

$$f(m(\mathbf{x})) - f(x^*) = J_f(\xi) \cdot (m(\mathbf{x}) - x^*),$$

valamely $\xi \in \mathbf{x}$-re.

Mivel zérushelyeket keresünk ezért tegyük fel, hogy $f(x^*) = 0$. Ezt felhasználva a fentiből

$$f(m(\mathbf{x})) = J_f(\xi) \cdot (m(\mathbf{x}) - x^*)$$

adódik. Tegyük fel, hogy mind $J_f(\xi)$ illetve minden részmátrixa reguláris. Ekkor a keresett $x^*$ zérushelyre $(J_f(\xi))^{-1}$-el való szorzás és átrendezés után azt kapjuk, hogy

$$\begin{aligned} x^* &= m(\mathbf{x}) - (J_f(\xi))^{-1} \cdot f(m(\mathbf{x})) \in \\ &\in m(\mathbf{x}) - (J_f(\mathbf{x}))^{-1} \cdot f(m(\mathbf{x})) =: N(\mathbf{x}). \end{aligned}$$

Nyilván az $f$ függvény minden $\mathbf{x}$-beli zérushelye egyúttal $N(\mathbf{x})$-ben is benne van.

Most relaxáljunk a regularitási feltételen! A feladatunk megoldani $x^*$-ra az

$$f(m(\mathbf{x})) = J_f(\xi) \cdot (m(\mathbf{x}) - x^*)$$

feladatot. Prekondicionáljuk ezt egy $R \in \mathbb{R}^{n \times n}$ valós mátrixszal, azaz ehelyett oldjuk meg a következőt:

$$R \cdot f(m(\mathbf{x})) = R \cdot J_f(\xi) \cdot (m(\mathbf{x}) - x^*).$$

A prekondicionálásra használt $R$ mátrixra általában az $R := (m(J_f(\mathbf{x})))^{-1}$ választás esik.

Bevezetve az $\mathbf{A} := R \cdot J_f(\mathbf{x})$, $c := m(\mathbf{x})$ illetve a $b := R \cdot f(m(\mathbf{x}))$ jelöléseket a feladat a következő befoglalás meghatározása:

$$\mathbf{A}(c - x^*) = b.$$

Ennek megoldására a Jacobi módszer egy intervallumos változatát használjuk.

A feladat azon $S$ halmaz elemeinek befoglalása, amelyre

$$S := \{x \mid A \cdot (c - x) = b, A \in \mathbf{A}\}.$$

Kiírva a mátrixszorzást a következő egyenletrendszert kapjuk:

$$\sum_{j=1}^{n} A_{ij}(c_j - x_j) = b_i, \quad i \in 1, ..., n.$$

Feltéve, hogy minden $i$-re $A_{ii} \neq 0$ az $x_i$-t kiszámolva kapjuk, hogy

$$x_i = c_i - \frac{\left(b_i + \sum_{j=1, j\neq i}^{n} A_{ij} \cdot (x_j - c_j)\right)}{A_{ii}} \in$$
$$\in c_i - \frac{\left(b_i + \sum_{j=1, j\neq i}^{n} \mathbf{A}_{ij} \cdot ([x_j] - c_j)\right)}{\mathbf{A}_{ii}}$$

Tehát az $\mathbf{x}$ intervallumból kiindulva egy Newton Jacobi lépés $N_J(\mathbf{x})$ eredményére

$$\mathbf{z} := \mathbf{x}$$
$$\mathbf{z}_i := \left(c_i - \frac{b_i + \sum_{j=1, j\neq i}^{n} \mathbf{A}_{ij} \cdot (\mathbf{z}_j - c_j)}{\mathbf{A}_{ii}}\right) \cap \mathbf{z}_i, \ i = 1, ..., n$$
$$N_J(\mathbf{y}) := \begin{cases} \mathbf{z}, \text{ ha } \mathbf{z}_i \neq \emptyset, \ i \in \{1, .., n\} \\ \emptyset, \text{ különben} \end{cases}$$

Ekkor nyilván $S \subset \mathbf{z}$. A lépés pontosságát növeli, hogy a már módosított $\mathbf{z}_i$ komponensekkel végezzük a további számításokat a $\mathbf{z}$ intervallumvektor meghatározásakor. A következő tétel néhány fontos eredményt mutat $N_J(\mathbf{x})$-ről:

**9.1. Tétel.** *Legyen $f : D \subset \mathbb{R}^n \to \mathbb{R}^n$ folytonosan differenciálható függvény, $\mathbf{x} \in \mathbb{IR}^n$, $\mathbf{x} \subset D$. Ekkor a fenti módon számított $N_J(\mathbf{x})$-re a következő három állítás teljesül:*

1. $\forall x^* \in \mathbf{x} : f(x^*) = 0 \;\Rightarrow\; x^* \in N_J(\mathbf{x})$, *azaz* $N_J(\mathbf{x})$ *az* $f$ *minden* $\mathbf{x}$*-beli zérushelyét tartalmazza*

2. *ha* $N_J(\mathbf{x}) = \emptyset$, *akkor* $f$*-nek nincs zérushelye* $\mathbf{x}$*-ben*

3. *ha* $N_J(\mathbf{x}) \subset \mathbf{x}$, *akkor* $\exists! x^* \in \mathbf{x}$, *amelyre* $f(x^*) = 0$.

## 9.3. Kiterjesztett intervallum aritmetika

Az alap intervallum aritmetikai műveletek bevezetése során kikötöttük, hogy intervallumok egymással történő osztásakor nem értelmezzük azt az esetet, amikor az osztó intervallum tartalmazza a 0-át. Most ezt a megkötést szüntetjük meg. Bővítsük ki a valós számokat a $+\infty$ és $-\infty$ elemekkel, a kibővített valós intervallumok halmazát pedig definiáljuk a következőképpen:

$$\overline{\mathbb{IR}} := \mathbb{IR} \cup \{[-\infty, r] \mid r \in \mathbb{R}\} \cup \{[l, +\infty] \mid l \in \mathbb{R}\} \cup \{[-\infty, +\infty]\}.$$

Ekkor az osztás új szabálya $0 \in [y]$ esetben:

$$\frac{[x]}{[y]} := \begin{cases} [-\infty, +\infty], & \text{ha } \underline{x} < 0 \text{ vagy } [x] = 0 \text{ vagy } [y] = 0 \\ [\overline{x}/\underline{y}, +\infty], & \text{ha } \overline{x} \leq 0 \text{ és } \underline{y} < \overline{y} = 0 \\ [-\infty, \overline{x}/\overline{y}] \cup [\overline{x}/\underline{y}, +\infty], & \text{ha } \overline{x} \leq 0 \text{ és } \underline{y} < 0 < \overline{y} \\ [-\infty, \overline{x}/\overline{y}], & \text{ha } \overline{x} \leq 0 \text{ és } 0 = \underline{y} < \overline{y} \\ [-\infty, \underline{x}/\underline{y}], & \text{ha } 0 \leq \underline{x} \text{ és } \underline{y} < \overline{y} = 0 \\ [-\infty, \underline{x}/\underline{y}] \cup [\underline{x}/\overline{y}, +\infty], & \text{ha } 0 \leq \underline{x} \text{ és } \underline{y} < 0 < \overline{y} \\ [\underline{x}/\overline{y}, +\infty], & \text{ha } 0 \leq \underline{x} \text{ és } 0 = \underline{y} < \overline{y} \end{cases}$$

A következő példa azt szemlélteti miképp kaphatjuk meg a következő szabályokat.

Legyen $[x] = [4, 5]$, $[y] = [-1, 2]$. Keressük $S := \{\frac{x}{y} \mid x \in [x], y \in [y]\}$ halmazt. Felhasználva a $[y] = [y_1] \cup [y_2] = [-1, 0] \cup [0, 2]$ felbontást $S$-re a következőt kapjuk:

$$S = \{\frac{x}{y} \mid x \in [x], y \in [y_1]\} \cup \{\frac{x}{y} \mid x \in [x], y \in [y_2]\} =$$
$$= [-\infty, -4] \cup [2, +\infty]$$

Különösen hasznos ez, ha ezeket a végtelen intervallumokat el tudjuk metszeni valamilyen véges intervallummal (mint például a Newton Jacobi módszerben a $\mathbf{z}_i$-vel).

## 9.4. Az algoritmus

### 9.4.1. Az algoritmus váza

Az algoritmus egy $L$ listában tárolja a globális optimumhely-jelölteket befoglaló intervallumokat. Kezdetben ez a lista a kiindulási $\mathbf{x}^0 := \mathbf{x}$ intervallumból áll.

Ezután a fő iteráció következik. Amíg az $L$ lista ki nem ürül, vagy minden $\mathbf{y} \in L$-re nem teljesül az, hogy egy adott tűréshatár alá nem esik az átmérőjük, a következő pontokban ismertetett négy teszt (középponti, monotonitási, konkavitási és Newton Jacobi lépés) végrehajtása következik iteratívan.

Ha az iteráció úgy ér véget, hogy $L = \emptyset$, akkor nem találtunk a kiindulási intervallumban minimumhelyét az $f$ függvénynek.

Ha az iteráció úgy ér véget, hogy $|L| > 1$, akkor egy verifikációs lépés következik, amely minden $\mathbf{y} \in L$ intervallumot megvizsgál.

A fentiek összefoglalásaként megadjuk az algoritmus rövid, programszerű leírását.

```
L := {[x]}
while ( MindenÁtmérőTűrésenBelül != igaz ÉS L != {} )
        Középpont_Teszt
        Monotonitás_Teszt
        Konkavitás_Teszt
        Newton_Jacobi
endwhile
if ( L != {})
```

Verifikáció
endif

### 9.4.2. Középponti teszt

Az algoritmus működése során számon tart és fokozatosan finomít egy felső becslést az $f^*$ globális optimum értékre. Jelölje ezt $\tilde{f}$.

Ezen felső becslést felhasználva az $L$ listából kidobható minden olyan $\mathbf{y}$ intervallum, amelyre teljesül, hogy

$$\underline{f_y} > \tilde{f},$$

hiszen ekkor

$$\underline{f_y} > \tilde{f} \geq f^*,$$

vagyis $\mathbf{y}$ nem tartalmazhat globális minimumhelyet.

A középponti teszt ennek az $\tilde{f}$ felső becslésnek kezdeti értékadását illetve finomítását hivatott szolgálni.

Kezdetben legyen $\tilde{f} = +\infty$. Válasszuk ki az $L$ listában tárolt intervallumok közül azt, amely felett a minimalizálandó célfüggvény intervallumkiértékelésének alsó korlátja a legkisebb, azaz legyen $\mathbf{y}$ olyan, hogy minden $\mathbf{z} \in L$-re

$$\underline{f_y} \leq \underline{f_z}.$$

Legyen $c = m(\mathbf{y})$, azaz az $\mathbf{y}$ intervallum középpontja és legyen $\tilde{f} := \min\{f(c), \tilde{f}\}$.

Amennyiben csökkent $\tilde{f}$ értéke eldobhatjuk a lista összes olyan $\mathbf{z}$ intervallumát, amelyre $\underline{f_z} > \tilde{f}$.

Ezen túl, amikor részintervallumokra bontunk egy listabeli $\mathbf{y}$-t szintén felhasználjuk a most kapott felsőbecslést a minimumértékre, nevezetesen a kapott $\mathbf{y}_i$ részintervallumok közül csak azokat tesszük a listába, amelyekre teljesül, hogy $\underline{f_{y_i}} \leq \tilde{f}$.

A középponti teszt ugyanúgy helyes marad, ha az intervallum középpontja helyett egy tetszőleges belső pontját vesszük.

### 9.4.3. Monotonitási teszt

A monotonitási teszt célja annak megállapítása, hogy egy $\mathbf{y}$ intervallumon a célfüggvény szigorúan monoton-e. Amennyiben az, akkor az $\mathbf{y}$ nem tartalmazhat stacionárius pontot, ami szükséges feltétele a szélsőértékhelynek, így ebben az esetben $\mathbf{y}$ kidobható az $L$ listából.

A monotonitás eldöntését a gradiens kiértékelésével végezzük. Legyen $\mathbf{g} := \nabla f(\mathbf{y})$. Ha létezik $i \in \{1, 2, ..., n\}$, hogy

$$0 \notin \mathbf{g}_i$$

akkor $f$ szigorúan monoton az $\mathbf{y}$ felett, vagyis $\mathbf{y}$ elhagyható.

Érdemes megjegyezni, hogy elég egyetlen koordinátát találni, amely mentén a fenti reláció teljesül, így általában az $n$-hez képest kevés számú intervallumkiértékelés után is dönthet a vizsgált intervallum eldobásról a monotonitási teszt.

### 9.4.4. Konkavitási teszt

Ezzel a teszttel szintén az a célunk, hogy kiszűrjük azokat az intervallumokat amelyek nem tartalmazhatnak globális minimumot, ezúttal annak az eldöntésével, hogy $f$ konkáv-e. Ehhez azt próbáljuk belátni, hogy $f$ nem konvex az $\mathbf{y}$ intervallum fölött.

Legyen $\mathbf{H} := \nabla^2 f(\mathbf{y})$, azaz legyen $\mathbf{H}$ az $f$ Hesse-mátrixának intervallum befoglalása. Amennyiben ez pozitív definit, akkor $f$ konvex. A pozitív definitség egyik szükséges feltétele, hogy a főátlóbeli elemek nullánál nagyobbak legyenek. Tehát ha létezik olyan $i \in \{1, 2, ..., n\}$, hogy

$$\overline{H}_{ii} < 0,$$

akkor $H_{ii} < 0$ minden $y \in \mathbf{y}$-ra, $H = \nabla^2 f(y)$, azaz $f$ nem lehet konvex $\mathbf{y}$-on, tehát nem tartalmazhat minimumhelyet sem, így $\mathbf{y}$ elhagyható.

### 9.4.5. Intervallumos Newton Jacobi lépés

Az algoritmus ezen lépésében az előbb bemutatott intervallumos Newton Jacobi lépés segítségével keressük egy függvény - a célfüggvényünk

gradiensének - zérushelyeit, azaz azokat az intervallumokat amelyek befoglalják az összes $y \in \mathbf{y}$ pontot, amelyre

$$\nabla f(y) = 0$$

fennáll. Ezek a helyek stacionárius pontjai lesznek a függvénynek, vagyis teljesül rájuk az optimum létezésének egy szükséges feltétele.

A lépés végrehajtásához legyen

$$\mathbf{A} := R \cdot \nabla^2 f(\mathbf{y}),$$

illetve

$$b := R \cdot \nabla f(m(\mathbf{y})),$$

ahol $R \approx (m(\nabla^2 f(\mathbf{y})))^{-1}$.

Itt $m(\nabla^2 f(\mathbf{y}))$ mátrix középponti mátrix, azaz a kifejezésben megjelenő intervallumváltozókat a középpontjaikkal helyettesítjük.

Ekkor az $\mathbf{y}$ intervallum finomításából a $N_J^{'}(\mathbf{y})$ eredményintervallum halmaz kiszámítása a következőképpen történik:

$$\begin{aligned}
\mathbf{z} &:= \mathbf{y} \\
\mathbf{z}_i &:= \left( c_i - \frac{b_i + \sum_{j=1, j\neq i}^{n} \mathbf{A}_{ij} \cdot (\mathbf{z}_j - c_j)}{\mathbf{A}_{ii}} \right) \cap \mathbf{z}_i, \; i = 1, ..., n \\
N_J^{'}(\mathbf{y}) &:= \begin{cases} \mathbf{z}, \text{ ha } \mathbf{z}_i \neq \emptyset, \; i \in \{1, .., n\} \\ \emptyset, \text{ különben} \end{cases}
\end{aligned}$$

Az algoritmushoz kiterjesztett intervallum aritmetika szükséges, ahol a 0-t tartalmazó intervallumokkal történő osztás is értelmezve van. Ekkor az adott komponens kiszámításának eredménye nem feltétlenül egy intervallum lesz, hanem lehet kettő is.

Amikor a most bemutatott Newton-szerű intervallumos módszerünk egy lépését alkalmazzuk három dolog történhet:

Ha $N_J^{'}(\mathbf{y}) = \emptyset$, akkor tudjuk a vizsgált $\mathbf{y}$ intervallumról, hogy nem tartalmaz stacionárius pontot, így kikerül a listából.

Ha $|N_J^{'}(\mathbf{y})| > 1$, akkor a lépés eleji $\mathbf{y}$ intervallum több részintervallumra esik szét. Ezeket ráhelyezzük az $L$ listára, amennyiben teljesül rájuk, hogy a célfüggvényünk intervallumkiértékelésének alsó

határa legfeljebb akkora, mint a globális minimum aktuális iterációs lépésben érvényben lévő felső becslése (ld. középponti teszt).

Ha $|N_J'(\mathbf{y})| = 1$ akkor ugyan a Newton lépés sem eldobni, sem szétszedni nem tudta az intervallumot, átmérője azonban jelentősen csökkenhetett, ezzel is növelve a többi teszt hatékonyságát.

### 9.4.6. Verifikáció

Ha $L \neq \emptyset$, akkor ebben a lépésben minden $\mathbf{y} \in L$ intervallumot megvizsgálunk a lokális minimumhely létezése és egyértelműsége szempontjából.

Amennyiben

$$N_{GS}'(\mathbf{y}) \subset \mathbf{y} \tag{9.1}$$

teljesül, akkor létezik egy egyértelmű stacionárius pont az $\mathbf{y}$ intervallumban. Ez szükséges feltétele az optimumnak.

A lokális minimumhely létezéséhez a $\nabla^2 f(\mathbf{y})$ pozitív definitségét kell belátni.

Ha a $B := I - \|A\|^{-1} \cdot A$ mátrix minden sajátértékének abszolútértéke kisebb mint 1, azaz a B spektrálsugarára igaz, hogy $\rho(B) < 1$, akkor $A$ pozitív definit. Ez utóbbira ad egy jól ellenőrizhető feltételt a következő tétel:

**9.2. Tétel.** *Legyen* $\mathbf{H} \in \mathbb{IR}^{n\times n}$, $\mathbf{S} := I - \frac{1}{\kappa}\mathbf{H}$, *ahol* $\kappa$ *olyan, hogy* $\|H\|_\infty \leq \kappa \in \mathbb{R}$. *Ha teljesül egy* $\mathbf{z} \in \mathbb{IR}^n$ *intervallum-vektorra, hogy*

$$\mathbf{S} \cdot \mathbf{z} \subset \mathbf{z}, \tag{9.2}$$

*akkor* $\rho(B) < 1$ *minden* $B \in \mathbf{S}$*-re és minden szimmetrikus* $A \in \mathbf{H}$ *mátrix pozitív definit.*

A bizonyítás a [4] cikkben található.

A (9.2) feltétel ellenőrzésére először kiszámítjuk a $\mathbf{H} = \nabla^2 f(\mathbf{y})$, $\kappa \geq \|H\|_\infty$ és $\mathbf{S} = I - \frac{1}{\kappa}\mathbf{H}$ értékeket, majd kiindulva a $\mathbf{z}^{(0)}$ intervallumvektorból, amelynek minden intervallumkomponensére $\mathbf{z}_i^{(0)} = [-1, 1]$

a következő iterációt végezzük:

$$\mathbf{z}^{(k+1)} := \mathbf{S} \cdot \mathbf{z}^{(k)},$$

amíg nem teljesül, hogy $\mathbf{z}^{(k+1)} \subset \mathbf{z}^{(k)}$. Ha ez egy bizonyos számú iterációs lépés után sem lesz igaz, akkor úgy vesszük, hogy a 9.2 feltétel nem teljesül.

A globális minimumhely egyértelműségének eldöntésére nincs lehetőség általános esetben. Ezért elégszünk meg annyival az algoritmusunk végén a verifikációs fázisban, hogy csak a lokális minimumhelyek egyértelműségét vizsgáljuk. Érdemes felhívni a figyelmet arra, hogy attól, hogy az egyértelműség teszt nem sikerül nem kell eldobni a vizsgált intervallumot, hiszen előfordulhat, hogy kontinuum sok globális minimumhelye van célfüggvényünknek és ezeket tartalmazza az aktuális intervallum.

A fentiek egyúttal azt is jelentik, hogy az algoritmus lefutása után az $L$ listán olyan intervallumok vannak, amelyek globális minimumhely-jelölt, lokálisan egyértelmű minimumhelyeket foglalnak be. Ha a végső listán csak egyetlen intervallum szerepel, ami egy egyértelmű lokális minimumhelyet foglal be, akkor az egyúttal a kiindulási $\mathbf{x}$ egyértelmű globális minimumhelye is.

## 9.5. Az algoritmus alkalmazhatósága

Az algoritmus ismertetése elején feltettük, hogy $f$ kétszer folytonosan differenciálható, azonban könnyíthetünk ezen a feltételen.

Ha nem alkalmazzuk a Newton-lépést, akkor egyszer folytonosan differenciálható függvényekre is futtathatjuk az algoritmusunkat, azonban ebben az esetben a verifikációs lépés sem használható.

Az algoritmus továbbá módosítható úgy is, hogy nem differenciálható függvényekre is alkalmazható legyen, ekkor lényegében csak felosztásokat és középponti teszteket végez már.

Tovább javítható az algoritmus középponti tesztjének hatékonysága, ha pontosítjuk a globális minimumérték felső becslését, például különböző lokális keresőeljárások segítségével.

Fontos megjegyezni, hogy az algoritmust módosítani kell, ha nem csak a kiindulási $\mathbf{x}$ intervallum belső pontjaiban keressük a minimum-

helyeket, hiszen például a határokon a globális minimumhelynek nem kell stacionáriusnak lennie.

# Irodalomjegyzék


[1] G. Alefeld and J. Herzberger, Introduction to Interval Computations, Academic Press, New York, 1983.

[2] R. Hammer M. Hocks U. Kulisch D. Ratz, Numerical Toolbox for Verified Computing, Springer-Verlag, 1993.

[3] U. Kulisch and H.J. Stetter (eds.), Scientific Computation with Automatic Result Verification, Springer-Verlag Wien New York, 1988.

[4] Ratz D., Automatische Ergebnisverifikation bei globalen Optimierungsproblemen, Dissertation, Karlsruhe, 1992

[5] J. Rohn, Solvability of Systems of Linear Interval Equations, *SIAM J. MATRIX ANAL. APPL.*, Vol. 25, No. 1, pp. 237-245, 2003.

[6] E. R. Hansen, Bounding the Solution of Interval Linear Equations, *SIAM J. NUMER. ANAL.*, Vol. 29, No. 5, pp. 1493-1503, October 1992.

[7] J. Rohn, Cheap and tight bounds: The recent result by E. Hansen can be made more efficient, *Interval Comput.*, 4 (1993), pp. 13-21.

[8] J. Rohn, An algorithm for solving the absolute value equation, *Electronic Journal of Linear Algebra*, 18 (2009), pp. 589-599.

[9] J. Rohn, An algorithm for solving the absolute value equation: An improvement, Technical Report 1063, Institute of Computer Science, Academy of Sciences of the Czech Republic, Prague, January 2010.

[10] J. Rohn, A general method for enclosing solutions of interval lin- ear equations, Technical Report 1067, Institute of Computer Science, Academy of Sciences of the Czech Republic, Prague, March 2010.

[11] J. Rohn, An Algorithm for Computing the Hull of the Solution Set of Interval Linear Equations, Technical report 1074, Institute of Computer Science, Academy of Sciences of the Czech Republic, Prague, April 2010.

[12] W. Oettli and W. Prager, Compatibility of approximate solution of linear equations with given error bounds for coefficients and right-hand sides, *Numer. Math.*, 6 (1964), pp. 405-409.

[13] J. Rohn, An existence theorem for systems of linear equations, *Linear Multilinear Algebra*, 29 (1991), pp. 141-144.

[14] S. Poljak and J. Rohn, Checking robust nonsingularity is NP-hard, *Math. Control Signals Systems*, 6 (1993), pp. 1-9.

[15] J. Rohn, Systems of linear interval equations, *Lin. Alg. Appls.* 126 (1989), 39-78

[16] I. Gargantini and P. Henrici, Circular arithmetic and the determination of polynomial zeros, *Numer. Math.*, 18 (1972), pp. 305-320.

[17] R. S. Varga, Matrix Iterative Analysis, Prentice-Hall, Englewood Cliffs, New Jersy, 1962.