# Alkalmazott matematikai lapok

1984/1-2

A MAGYAR TUDOMÁNYOS AKADÉMIA
MATEMATIKAI ÉS FIZIKAI TUDOMÁNYOK
OSZTÁLYÁNAK KÖZLEMÉNYEI

10.

KÖTET

AKADÉMIAI KIADÓ, BUDAPEST

# ALKALMAZOTT MATEMATIKAI LAPOK

## A MAGYAR TUDOMÁNYOS AKADÉMIA MATEMATIKAI ÉS FIZIKAI TUDOMÁNYOK OSZTÁLYÁNAK KÖZLEMÉNYEI





X. kötet 1—2. szám

Szerkesztőség: 1502 Budapest XI., Kende u. 13—17.
Kiadóhivatal: 1055 Budapest V., Alkotmány u. 21.


Az Alkalmazott Matematikai Lapok változó terjedelmű füzetekben jelenik meg, és olyan eredeti tudományos cikkeket publikál, amelyek a gyakorlatban, vagy más tudományokban közvetlenül felhasználható új matematikai eredményt tartalmaznak, illetve már ismert, de színvonalas matematikai apparátus újszerű és jelentős alkalmazását mutatják be. A folyóirat közöl cikk formájában megírt, új tudományos eredménynek számító programokat, és olyan, külföldi folyóiratban már publikált dolgozatokat, amelyek magyar nyelven történő megjelentetése elősegítheti az elért eredmények minél előbbi, széles körű hazai felhasználását.

A folyóirat feladata a Magyar Tudományos Akadémia III. (Matematikai és Fizikai) Osztályának munkájára vonatkozó közlemények, könyvismertetések stb. publikálása is.

A kéziratok a főszerkesztőhöz, vagy a szerkesztő bizottság bármely tagjához beküldhetők. A főszerkesztő címe:

Prékopa András, főszerkesztő
1502 Budapest, Kende u. 13—17.

Közlésre el nem fogadott kéziratokat a szerkesztőség lehetőleg visszajuttat a szerzőhöz, de a beküldött kéziratok megőrzéséért vagy továbbításáért felelősséget nem vállal.

Az Alkalmazott Matematikai Lapok előfizetési ára kötetenként 100 forint. Belföldi megrendelések az Akadémiai Kiadó, 1055 Budapest V., Alkotmány u. 21. címen (pénzforgalmi jelzőszám 215—11 488), külföldi megrendelések a Kultúra Külkereskedelmi Vállalat, H-1389 Budapest, Pf. 149. címen (pénzforgalmi jelzőszám 218—10 990) lehetségesek.

A Magyar Tudományos Akadémia III. (Matematikai és Fizikai) Osztálya a következő idegen nyelvű folyóiratokat adja ki:

1. Acta Mathematica Hungaricae,
2. Acta Physica Hungaricae,
3. Studia Scientiarum Mathematicarum Hungarica.

*Alkalmazott Matematikai Lapok* **10** (1984) 1—13

# EGY NAGY ADATRENDSZER KARBANTARTÁSÁNAK VIZSGÁLATA

BENCZÚR ANDRÁS ÉS STAHL JÁNOS

Budapest

A dolgozat egyrészt egy adatkezelési problémák elemzésére alkalmas keretet ismertet, másrészt pedig ennek egy konkrét alkalmazásával foglalkozik.

Adatkezelő rendszerek működése leképezések sorozataként tekinthető. A leképezések lényegében programok, amiket költség és meghibásodás nélküli végrehajtás valószínűsége szerint vizsgálhatunk. A konkrét esetben azzal foglalkozunk, hogy a kimentési műveletek gyakorisága, illetve az adatrendszer egymástól nagyjából függetlenül kezelhető kisebb részekre bontása miképpen befolyásolja a várható költségeket. A megfelelő leképezések és paramétereik meghatározása után a meghibásodást követő helyreállítási folyamat zavartalan végrehajtását feltételezve kiszámítható a gyakoriság és a részek számának optimális értéke. Ezt követően pedig megmutatjuk, hogy az előbbi eredmények akkor sem változnak lényegesen, ha nem tekintünk el a helyreállítás alatti meghibásodás lehetőségétől sem.

## 1. Bevezetés

Egy adatrendszer állapotai címek egy sorozatával és azok tartalmával írhatók le. Az adatrendszer különféle változtatásai az állapotokat állapotokba átvivő leképezéseknek tekinthetők. A leképezéseket, amelyek mindegyike nagyjából programokat takar, költség és a meghibásodás nélküli végrehajtás valószínűsége szempontjából vizsgáljuk.

Egy leképezés költsége két részből áll. Egy fix részből, amely a leképezés realizálásánál használt adathordozók biztosításának, mozgatásának felel meg, és egy olyan részből, amely a megváltoztatott tartalmú címek számának nagyságától függ. Ez a kapcsolat az általunk vizsgált esetben olyan, hogy a költség ezen része arányosnak vehető a megváltoztatott tartalmú címek számával, illetve a leképezés végrehajtásához szükséges idővel. Nagy adatrendszer esetén leképezések adminisztrálása is része a számítógépes rendszernek. Ez nagyjából az összes figyelembe vett leképezések számának logaritmusával arányos mennyiséggel növeli az egyes leképezések költségét. (Ugyanakkor nagyon sok leképezés esetén, magának az adminisztrációnak a megoldása is nagyszámú címet köthet le.) Ugyancsak a végrehajtáshoz szükséges idő függvényének tekintjük a meghibásodás nélküli végrehajtás valószínűségét is.

Egy valóságos adatkezelési feladatnál általában többféleképpen is megadhatók a feladat megoldásához szükséges leképezések vagy eljárások. Az ismertetendő feladatnál mi egy viszonylag szűk osztályra szorítkozunk, és az ide tartozókat elemezzük a fenti szempontokból, bár az osztály megválasztásánál is már érvényesültek ezek a szempontok.

A 2. pontban azt az adatkezelési feladatot ismertetjük röviden, amelynél a megoldás újratervezése vizsgálataink kiinduló pontja volt. A 3. pontban a leképezéseknek a feladat megoldására kiválasztott osztályát ismertetjük. A 4. és 5. pontokban az egyes leképezések, illetve az adatkezelő rendszer várható költségének meghatározásával foglalkozunk.

## 2. Egy konkrét feladat rövid ismertetése

Lényegében az *Állami Népességnyilvántartás* adatrendszerének folyamatos karbantartását vizsgáltuk. A teljes állomány egy rögzített születési év szerinti bontásban 43 db 100 M byte-os lemezt kötött le, minden lemez egy-egy személyi-szám sorrendes indexszekvenciális file. Utóbbi csak az akkori helyzet adottsága, a bontás viszont rögzített volt.

A karbantartás új rekordoknak a bontásból adódó helyre történő behelyezését és az adatrendszer bizonyos rekordjainak a változásjelentő rekordok tartalmával történő módosítását jelenti. Utóbbi változtatás előidézheti egy rekordnak egy másik lemezre történő áthelyezését is.

A karbantartásba beleértendő az új rekordok, illetve a régiekre vonatkozó változásjelentő rekordok ellenőrzése és javítása is. A következők szempontjából tekinthetjük úgy, hogy ez az általunk vizsgált számítógépes rendszeren kívül történik. Ugyanígy figyelmen kívül hagyhatók az ún. szótárváltoztatásokhoz kapcsolódó karbantartások. (A szótár(ak) a rekordok bizonyos mezőinek, pl. lakcím, utónév ellenőrzésére és kódolására szolgál(nak).

A számítógépes rendszernek azt a lehetőséget is le kell kezelnie, hogy egy korábban érkezett (és az ellenőrzés során jónak talált) változásjelentő rekord tartalmát nem lehetett az adatrendszeren végigvezetni. Az ilyen rekordok ún. várakozó file(ok)-ba kerülnek és az adatrendszer módosítása mindig a jónak bizonyult újonnan érkezett és a várakozó file(ok)-ban tartalmazott változásjelentő rekordokkal történik.

A változásjelentő tételek egy része lekérdezés, és az ezekre adott válaszok jelentik a számítógépes rendszer nagyszámú output-jának egyik részét. Az output következő része az új rekordok sikeres vagy sikertelen behelyezéséről, illetve a változásjelentő rekordokkal végrehajtott módosításokról tájékoztat. (Sikertelen behelyezés esetén maga a rekord is elhagyja a rendszert.) A számítógépes rendszer outputjai végül még a pillanatnyilag hibásnak bizonyult rekordokról és a várakozó file-okban levő változásjelentő rekordokról készülő táblázatok.

Az adatrendszer a rögzített bontásnak megfelelően tetszőlegesen szakaszolható, és az így adódó részek karbantartása már egymástól függetlenül végezhető. Mivel az output első két részét az adatrendszerétől eltérő és sokféle bontásban kell szolgáltatni, ez a karbantartás során nyert karbantartási sorrendű elsődleges output további feldolgozását teszi szükségessé. Ugyanakkor ez az elsődleges output olyan, hogy segítségével az adatrendszer korábbi állapota is visszaállítható.

Havonta kb. 15 000 új rekord és 200 000 változásjelentő rekord fogadásával kell foglalkozni. Minden végrehajtott behelyezésről és módosításról átlagban végül két outputrekord készül. (A továbbiakban nem teszünk különbséget változásjelentő és beléptetendő rekord között.) Az új számítógépes karbantartó rendszerrel szemben két főbb elvárás volt: megbízhatóan és könnyen üzemeltethetőnek kell lennie. Ezeket elsősorban az akkori rendszerben használt nagyszámú mágnesszalagnak mágneslemezekkel történő helyettesítésével és a megoldás során használt leképezések adminisztrálása segítségével lehet biztosítani.

Azt fogjuk vizsgálni, hogy az adatrendszer részekre bontása és a kimentési gyakoriság miképpen hat a karbantartás várható költségére. Bár maga a karbantartási ciklus hossza is lehetne döntési változó, esetünkben a számítógépes rendszert körülvevő információs rendszer ezt nagyjából két hét — egy hónapban határozta meg. Mindenesetre, a későbbi elemzési módszer használható abban az esetben is, ha a kar-

bantartási ciklus hossza (is) meghatározandó változó. Ugyancsak használható a módszerünk abban az esetben is, ha a vázolt feladatot a következő részben tárgyaltaktól eltérő eljárásokkal oldjuk meg.

### 3. A feladat megoldásakor alkalmazandó leképezések

Legyen $S$ egy adatrendszer lehetséges állapotának halmaza, $\mathscr{F}$ az egyes állapotokat egymásba átvivő leképezések egy osztálya.

Egy $s \in S$ állapotot címekből és a címek tartalmából álló párok egy sorozata ír le. Az $s$ állapot megadásakor felsorolt $A_s$ címek halmaza az állapot címtere. Esetünkben úgy tekintjük, hogy $A_s = A$ minden $s$-re.

Minden $f \in \mathscr{F}$ leképezés esetén $A$-t felbonthatjuk három, a leképezéstől függő $A_1$, $A_2$ és $A_3$ részre, amelyekre $A_1 \cup A_2 \cup A_3 = A$ és $(A_1 \cup A_2) \cap A_3 = \varnothing$. $A_1$ a leképezés meghatározó címtere, $A_2$ $A$ azon része, melynek tartalmát $f$ megváltoztathatja, $A_3$ pedig $A$-nak $f$-től független része. Más szavakkal; az $f$ leképezés végrehajtásához csak az $A_1$ tartalmának végigolvasása szükséges; nevezhetnénk $A_1$-et olvasási címtérnek is. Az $f$ leképezés nem változtatja meg a teljes $s$ állapotot, csak $A_2$ címterének tartalmát; nevezhetnénk $A_2$-t írási címtérnek is. Az $A_3$ címtér tartalma teljesen közömbös az $f$ leképezésre nézve, tehát a leképezést végrehajtó eljárás se nem olvas $A_3$-ról, se nem ír rá. Formálisabban, $s_B$-vel jelölve egy $s$ állapot $B$-beli címeknek megfelelő részét, $s_{A_3} = f(s)_{A_3}$ és $s_{A-A_3}$ nem függ $s_{A_3}$-tól, $s_{A-A_2} = f(s)_{A-A_2}$ és $f(s)_{A_2}$ csak $s_{A_1}$-től függ. (Tehát $s_{A_1}$ és $s_{A_3}$ meghatározza $f(s)$-t.) A továbbiakban nem okoz félreértést, ha $f(A)$-t írva abba a címtér tartalmát is beleértjük, illetve az $f(A_1) = A_2$ jelölést fogjuk használni. Mindezekről bővebben vö. [1]-t.

A következőkben felsoroljuk azokat a címtereket és leképezéseket, melyekre az előző pontban vázolt feladat leírásához illetve megoldásához szükségünk lesz.

Az INP címtér tartalmazza az ellenőrzött változásjelentő tételeket, DB az adatrendszert, OUT pedig az adatrendszer karbantartásakor adódott (és még további feldolgozást igénylő) outputrekordokat. A további feldolgozás eredménye az OUTP címtérre kerül. Jelölje továbbá WINP a várakozó változásjelentő tételeket, INPS pedig az összerendezett input és várakozó rekordokat tartalmazó címteret.

Ahhoz, hogy a rendszer karbantartását megbízható módon lehessen végrehajtani, szükséges a rendszer egyes állapotainak kimentése. Az INPSC, DBC, WINPC és OUTC címterek tartalmazzák az INPS, DB, WINP és OUT címterek másolatait.

Az adatrendszer karbantartása az alábbi leképezésekből tehető össze:

$f_1(\mathrm{INP} \cup \mathrm{WINP}) = \mathrm{INPS}$ — az új és várakozó változásjelentő tételek előkészítése a karbantartáshoz

$f_2(\mathrm{INPS} \cup \mathrm{DB}) = \mathrm{DB} \cup \mathrm{OUT} \cup \mathrm{WINP}$ — a szűkebb értelemben vett karbantartás

$f_3(\mathrm{OUT}) = \mathrm{OUTP}$ — a karbantartáskor adódott outputrekordok további feldolgozása

$f_{41}(\mathrm{INPS}) = \mathrm{INPSC}$
$f_{42}(\mathrm{DB}) = \mathrm{DBC}$
$f_{43}(\mathrm{WINP}) = \mathrm{WINPC}$
$f_{44}(\mathrm{OUT}) = \mathrm{OUTC}$
} — kimentő eljárások

$f_{51}(\mathrm{INPSC}) = \mathrm{INPS}$
$f_{52}(\mathrm{DBC}) = \mathrm{DB}$
$f_{53}(\mathrm{WINPC}) = \mathrm{WINP}$
$f_{54}(\mathrm{OUTC}) = \mathrm{OUT}$
} — visszamentő eljárások

$n$-nel jelölve az adatrendszer két kimentése közötti karbantartási ciklusok számát

$$f = f_{42} + n(f_1 + f_{41} + f_2 + f_{43} + f_{44} + f_3)$$

az adatrendszer két kimentése között végrehajtandó leképezés sorozat, ha nem fordul elő meghibásodás. (Az „összeadások" sorrendje nagyjából a leképezések végrehajtási sorrendjének felel meg. Az sem okoz félreértést, illetve nem jelent problémát, hogy minden ciklusban pl. különböző INPSC címterekre van szükség INPS-kimentésénél.)

Ha az adatrendszer $j$-edik ciklusbeli karbantartásakor meghibásodás történt, akkor a működő rendszerben a

$$\mathrm{HF}' = f_{52} + (j-1)(f_{51} + f_2)$$

leképezéssel történt az adatrendszer $j$-edik karbantartási ciklus előtti állapotának helyreállítására. Hasonlóan építhetők fel az $f$ leképezésekből egyéb szükséges helyreállító leképezések is. Az adatrendszer helyreállítása másképpen is történhet. Amint azt már említettük, a karbantartás során adódó elsődleges output tartalmaz a karbantartás előtti állapot visszaállításához szükséges minden információt. Jelölje $f_6$ azt a leképezést, amivel ezt a visszaállítást végrehajtjuk, azaz

$$f_6(\mathrm{DB} \cup \mathrm{OUT}) = \mathrm{DB}.$$

Ennek segítségével a

$$\mathrm{HF}^{(j)} = f_{52} + (j-1)(f_{54} + f_6)$$

leképezéssel állítható helyre az adatrendszer $j$-edik karbantartási ciklus előtti állapota. A következő rész alapján ez a fajta helyreállítás ráfordítások szempontjából is előnyösebb.

Ha az adatrendszert — egyébként a rögzített bontásnak megfelelő — $K$ részre osztjuk, a végleges bontású és formájú outputot elkészítő $f_3$ kivételével valamennyi eddig bevezetett leképezés ezekre egymástól függetlenül végezhető. (Az elsődleges outputot biztosító $f_{44}$ is!) Pl. az $i$-edik rész karbantartó eljárására bevezetve az $f_2,.i$ jelölést $f_2 = \sum_{i=1}^{K} f_2,.i$

Minthogy változásjelentő tételek átvezetése kiválthatja egyes rekordok egyik részből másik részbe történő áthelyezését, pontosabb lenne leírásunk, ha erre is bevezettünk volna egy külön címteret stb. Ez is olyan, ami valójában összeköti a különböző részekre vonatkozó eljárásokat. Ez a címtér azonban olyan kicsi, hogy biztosításának stb. figyelmen kívül hagyása nem jelent lényeges elhanyagolást. (Ugyanakkor az eddigiekkel már összhangban van az, hogy nem foglalkozunk a karbantartás inputjának új részét tartalmazó INP tartalmának biztosításával, minthogy az — elsődleges — input ellenőrzésével együtt ezt sem tekintjük a vizsgált rendszer részének.)

Figyelembe kell venni még egy, az eddigi eljárások adminisztrálásához szükséges címteret. Ennek nagysága a bevezetett eljárások számával arányos. Ebbe természetesen belejátszik az a $K$ szám ki, ahány részre osztottuk az adatrendszert.

## 4. A karbantartás várható költségének vizsgálata

A továbbiakban nem foglalkozunk az ellenőrzött — új — input és várakozó változásjelentő tételeket egyberendező $f_1$ és a karbantartás során keletkezett outputtételek további feldolgozását végző $f_3$ eljárással sem. Ez megengedhető, mivel egyrészt ezek súlya kisebb, mint a szűkebb értelemben vett karbantartásé (nem érintik közvetlenül az adatrendszert, az eljárások időigénye is kisebb), másrészt pedig nem is könnyen illeszthetők az általunk eddig kialakított, illetve a továbbiakban kialakítandó keretekbe. Ugyanis ezek alapvetően mágnesszalagos eljárások lévén (a megfelelő címtereket mágnesszalagok tartalmazzák) a hibafolyamat paramétere(i) más(ok), valamint nem teljesül az a már említett feltevésünk, hogy ezen eljárásokra a költség legfőbb tényezőjét jelentő időigény az érintett címterek nagyságával arányos. Rendezésről ($f_1$), illetve főleg átrendezésről ($f_3$) lévén szó, az időigény érintett címterek nagyságának négyzetével vehető arányosnak. $r$-rel jelölve az inputrekordok, $R$-rel jelölve az adatrendszer rekordjainak számát, az adatrendszert $K$ részre osztva, az $i$-edik rész esetén a teljes karbantartást megadó $f$ leképezésben szereplő további leképezések költségei a következők:

$$c(f_{42}, .i) \approx c(f_{52}, .i) = \alpha_{42}\frac{R}{K} + \delta$$

$$c(f_{41}, .i) = \alpha_{41}\frac{r}{K} + \delta$$

$$c(f_2, .i) = \frac{\alpha_{41}}{2}\frac{r}{K} + \frac{\alpha_{43}}{2}\frac{r}{K} + \alpha_2\frac{R}{K} + \frac{\alpha_{44}}{2}\frac{r}{K} + \delta$$

$$c(f_{43}, .i) = \alpha_{43}\frac{r}{K} + \delta$$

$$c(f_{44}, .i) \approx c(f_{54}, .i) = \alpha_{44}\frac{r}{K} + \delta$$

A fix részbe ($\delta$) beleértettük az adminisztrációs költségeket is, ami ily módon már $K$-tól függő. Elsősorban a jelöléseket egyszerűsítendő vettük ezt a költségrészt minden leképezésnél azonosnak: a későbbiekben látszik majd, hogy a költségkifejezésekben valójában csak egy további $K$-val növekvő tag létezése az érdekes. Az INPS, WINP és OUT címterek kimentésénél olvasás és kiírás is van, míg az $f_2$ szűkebb értelemben karbantartó leképezésnél ezeken a címtereken csak az egyik művelet kerül végrehajtásra. (Valójában pontosabb azt mondani, hogy a költségek a címtér érintett lapjainak a függvényei. Az $f_2, .i$ leképezés költségében szereplő $\alpha_2\frac{R}{K}$ tag a $DB_i$ címtér teljes végigolvasása miatt szerepel, tehát nem a megváltoztatandó rekordok $\frac{r}{K}$ mennyiségével arányos. Ezt a nagy blokkolási faktor okozza, mert az $R/r$ hányados közel azonos a blokkolási faktorral, s így gyakorlatilag minden blokkban van módosítandó rekord. Maga a blokkolási faktor is lehetne döntési változó, az ezt behatároló feltételek egy része azonban mindenképpen kívül esik az itteni vizsgálatokon. Nagyjából a következő megállapítás érzékelteti a helyzetet: a karbantartás

szempontjából — legalábbis egy bizonyos mértékig — a blokkméret csökkentése a kedvező, az adatrendszert érintő más tevékenységek esetén pedig a blokkméret növelése. A megállapítás első része a következők alapján látható is.)

Az adatrendszer $i$-edik részére vonatkozó tényleges karbantartási költség tehát

$$c(f_2, .i) = \frac{\alpha_{41}+\alpha_{43}+\alpha_{44}}{2}\frac{r}{K}+\alpha_2\frac{R}{K}+\delta\sim\alpha_4\frac{r}{K}+\alpha_2\frac{R}{K} = \tilde{a}_4\frac{\overline{T}_F}{K}+\tilde{\alpha}_2\frac{\overline{T}_F}{K},$$

ahol $\overline{T}_F$ a teljes adatrendszer szűkebb értelemben vett karbantartásához szükséges idő.

Csak a karbantartás közben jelentkező meghibásodásokkal foglalkozva, ha a meghibásodás egy $j$-edik ciklusbeli $t$ időpontban történt, az adatrendszer $i$-edik része $j$-edik ciklus előtti állapotának az OUT címtér tartalmát felhasználó HF, $.i$ leképezéssel történő helyreállításának költsége

$$C(\mathrm{HF}^{(j)}\ .i,\ t) = \alpha_{42}\frac{R}{K}+(j-1)\left(\alpha_{44}\frac{r}{K}+\frac{\alpha_{44}}{2}\frac{r}{K}+\alpha_{\mathrm{HF}}\frac{R}{K}\right)$$

(A „fix" költségeket ismét elhagytuk.) Ugyanúgy, mint az előbb, ez is írható

$$C(\mathrm{HF}^{(j)}\ .i,\ t) = \alpha_{42}\frac{R}{K}+(\tilde{\alpha}_{44}+\tilde{\alpha}_{\mathrm{HF}})(j-1)\frac{\overline{T}_F}{K}$$

alakban. (Hasonló kifejezés adódik a karbantartás megismétlésének megfelelő HF′ leképezés esetén is.)

A továbbiakban folytonos időre térünk át. Úgy képzeljük, hogy az adatrendszer két kimentése között az adatrendszer valamennyi részére egymástól függetlenül és egymással párhuzamosan $T/K$ időn keresztül végezzük a szűkebb értelemben vett karbantartást. (A kimentés is részenként történik.) Valamennyi rész ilyen karbantartása közben történhetnek meghibásodások. Az ilyenkor szükséges helyreállításoknak, valamint az ezeket lehetővé tevő kimentéseknek csak költségkihatásuk van. A teljes $K\cdot T/K=T$ idő alatt jelentkező költségeket $C(T)$-vel jelöljük. Minthogy valamennyi leképezésben (így a helyreállításokban, valamint az ezekhez szükséges kimentéseknél is) a fő költségtényező az idő, mindezt úgy is képzelhetjük, hogy $T$ idő helyett $C(T)$ időre van szükség a $T$ idő alatt egymást követő karbantartási ciklusok végrehajtásához.

Ha a $T/K$ idő során az $i$-edik rész karbantartásakor a $j$-edik ciklusban a $t_{jl}$ időpontokban lépett fel meghibásodás, akkor

$$c(T) = \sum_{i=1}^{K}\big(c(f_{42}, .i)+c(f_{41}, .i)+c(f_2, .i)+c(f_{43}, .i)+c(f_{44}, .i)+$$

$$+\sum_{j=1}^{n}\sum_{l}(HF^{(j)}\ .i,\ t_{jl})) \approx$$

$$\approx \tilde{\alpha}_{42}T_F+(2\tilde{\alpha}_4+\tilde{\alpha}_2)T+\sum_{i=1}^{K}\sum_{j=1}^{n}\sum_{l}\left(\tilde{\alpha}_{42}\frac{\overline{T}_F}{K}+\tilde{\tilde{\alpha}}_{HF}t_{jl}\right),$$

ahol $\approx$ arra utal, hogy a fix költségeket ismét elhagytuk. A $t_{jl}$ időpontban történt meghibásodás költségének megfelelő tagban azért szerepel $t_{jl}$, mert nemcsak a $j$-edik

ciklust megelőző állapotot kell helyreállítanunk, hanem a $j$-edik ciklusban a $t_j$ időpontig végrehajtott karbantartások elvesztése is növeli a költséget.

Ha feltesszük, hogy a meghibásodási folyamat egy $\lambda$ paraméterű *Poisson folyamat*, akkor $c(T)$ várható értékére ([1], [2])

$$E(c(T)) = \tilde{\alpha}_{42}\overline{T}_F + \tilde{\tilde{\alpha}}_4 T + \tilde{\alpha}_{42}\frac{\lambda \overline{T}_F T}{K} + \frac{\tilde{\tilde{\alpha}}_{HF}}{2}\frac{\lambda_T^2}{K}$$

Legyen a továbbiakban $\tilde{\tilde{\alpha}}_4 = 1$. A $T$-hez tartozó relatív költség azt mutatja, hogy a két egymást követő kimentés között a biztosítás miatt szükséges többletidők milyen mértékben növelik a szorosabb értelemben vett karbantartáshoz szükséges időt. Az $E(c(T))/T$ hányadost minimalizáló $T$ értékre

$$T_{\text{opt}} = K\left(\frac{2\tilde{\alpha}_{42}\overline{T}_F}{\tilde{\tilde{\alpha}}_{HF}\lambda K}\right)^{1/2}$$

adódik.

A karbantartási ciklusok száma ezen időtartam alatt

$$n_{\text{opt}} = \frac{T_{\text{opt}}}{\overline{T}_F} = \left(\frac{2\tilde{\alpha}_{42}K}{\tilde{\tilde{\alpha}}_{KF}\lambda \overline{T}_F}\right)^{1/2}.$$

Végül a relatív költség minimumára

$$\bar{c}(T_{\text{opt}}) = \frac{E(c(T_{\text{opt}}))}{T_{\text{opt}}} = 1 + \left(\frac{2\tilde{\alpha}_{42}\tilde{\tilde{\alpha}}_{HF}\lambda \overline{T}_F}{K}\right)^{1/2} + \frac{\tilde{\alpha}_{42}\lambda \overline{T}_F}{K}$$

adódik. (A fix költségek elhanyagolásakor két, a többiekhez képest kicsiny tagot hagytunk el $c(T)$-ből és így $E(c(T))$-ből, és $E(c(T_{\text{opt}}))$-ból is. Az első $T$-től független és $K$-val növekvő, a $c(\text{HF}^{(j)}{}_{,i}, t_j)$-ekből adódó második is $K$-val növekvő, bár $T$-től nem független. Ezen tagok miatt $\bar{c}(T_{\text{opt}})$ nem monoton csökkenő függvénye $K$-nak, ugyanakkor a szóba jövő $K$ értékek mellett a csökkenés még fennáll. Mindenesetre, a szóban forgó tagok „pontos" figyelembevétele csak bonyolította volna az adódó kifejezéseket.) $\tilde{\tilde{\alpha}} = 1$-et véve tulajdonképpen az $f_2$-beli karbantartás költségét tekintettük egységnyinek, mivel $r \approx 2 \cdot 10^5$ és $R \approx 10$ miatt $\tilde{\alpha}_4\tilde{\alpha}_2$-höz képest elhanyagolható. Az OUTP címtér tartalma alapján történő helyreállítást a karbantartásnál kétszer gyorsabbnak tekintve $\tilde{\tilde{\alpha}}_{HF} = 0.5$.

Ha a kimentést a karbantartásnál négyszer lassabbnak tekintjük, akkor $\tilde{\alpha}_{42} = 4$. Továbbá, egy karbantartási ciklus várható időszükséglete 320 perc és legyen $\lambda = 1/320$. (Átlagosan egy meghibásodás van egy karbantartási ciklusban.) Ilyen értékek mellett meghatároztuk az optimális kimentés időközének, az optimális ciklusszámnak és a megfelelő relatív költségnek az értékét a rendszer részekre bontásának néhány értéke mellett.

1. TÁBLÁZAT

| $K$ | $T_{\text{opt}}$ | $n_{\text{opt}}$ | $\bar{c}(T_{\text{opt}})$ |
|---|---|---|---|
| 40 | 8 095 | 24 | 1,42 |
| 160 | 16 190 | 50 | 1,18 |
| 320 | 22 897 | 71 | 1,12 |

Egy másik táblázat különböző nem optimális kimentés időközök és adatrendszer egymástól független részei számának különböző értékei esetén adódó relatív költségeket tartalmaz.

2. TÁBLÁZAT

| $K$ | $\bar{c}(T_{opt}/2)$ | $\bar{c}(T_{opt}/4)$ |
|---|---|---|
| 40 | 1,625 (=1+0,4+0,1+0,125) | 1,962 (=1+0,8+0,1+0,062) |
| 160 | 1,287 (=1+0,2+0,0257+0,062) | 1,456 (=1+0,4+0,025+0,031) |
| 320 | 1,196 (=1+0,14+0,012+0,041) | 1,317 (=1+0,283+0,012+0,022) |

(A második tag a kimentés, a harmadik a visszamentés, a negyedik tag pedig a helyreállítás relatív többletköltsége.)

A működő rendszer némileg eltért a tervezettől. A leglényegesebb eltérés, hogy az outputkészítés nem vált el teljesen a karbantartástól, így $f_2$ viszonylag hosszú. Nagyrészt ennek, és az ettől részben független nagyobb mértékű mágnesszalag használat következtében gyakrabban volt meghibásodás is. Meghibásodás esetén a helyreállítás feldolgozások megismétlésével történt.

A következő táblázat összeállításakor az alábbi paramétereket használtuk:

$\tilde{\tilde{\alpha}}_4=1$

$\tilde{\tilde{\alpha}}_{HF}=J$

$\tilde{\alpha}_{42}=2$ (a kimentés a jelenlegi lassú karbantartáshoz képest csak fele olyan lassú)

$T_F=600$

$K=40$

3. TÁBLÁZAT

| $\lambda$ | $T_{opt}$ | $n_{opt}$ | $\bar{c}(T_{opt})$ | $\bar{c}(600)$ |
|---|---|---|---|---|
| 1/15 | 1200 | 2 | 5,000 | 5,500 |
| 1/30 | 1697 | 3 | 3,414 | 3,654 |
| 1/150 | 3795 | 6 | 1,832 | 3,250 |
| 1/300 | 5366 | 9 | 1,547 | 3,125 |
| 1/600 | 7590 | 12 | 1,366 | 3,062 |

Korábban minden karbantartási ciklus után történt kimentés. Ez mindenképpen feleslegesnek tűnik, ugyanis a negyedik és ötödik oszlopbeli számok különbségei csak irreálisan magas meghibásodási gyakoriság esetén hagyhatók figyelmen kívül.

## 5. Egy alternatív modell

Az előzőek során feltettük, hogy egy meghibásodást követő helyreállítás során nem történik újabb meghibásodás. Ennek egyik oka technikai volt: az [1] és [2]-beli eredmények alkalmazása kívánta ezen feltétel bevezetését. Ugyanakkor könnyen felírható egy valamivel több számolást igénylő olyan modell, amely az ilyen meghibáso-

dás lehetőségét is figyelembe veszi. Valójában a két modell eredményének összevetése alapján dönthető el, hogy a szóban forgó feltétel az előzőekben megszorítást jelentett-e. A következő vizsgálatok egyébként tetszőleges költségfüggvény esetén elvégezhetők.

Az adatrendszer egy részének karbantartását vizsgáljuk az adatrendszer két kimentése között. Az adatrendszer egyes részeinek karbantartása egymástól függetlenül végezhető, illetve egy ciklusban az adatrendszer karbantartása valamennyi részének karbantartását jelenti.

Feltesszük, hogy egy karbantartási ciklushoz tartozó bármely leképezés meghibásodási valószínűsége csak a leképezés megkezdése óta eltelt időtől függ, amit minden $j$-edik ciklusbeli leképezés esetén ugyanaz a $P_j(t)$ függvény ír le: $1-P_j(t)$ annak a valószínűsége, hogy a $t$ idő alatt nem következett be meghibásodás. A $j$-edik karbantartási ciklus várható költségét meghatározandó azt további részciklusokra bontjuk. Egy részciklus egy sikertelen (meghibásodott) feldolgozást követő sikertelen helyreállítások és egy sikeres helyreállítás sorozata. A $j$-edik ciklus részciklusok olyan sorozata lesz, amit egy sikeres feldolgozás zár le. (Szélső esetben a $j$-edik ciklus egyetlen sikeres feldolgozás, illetve egy részciklus nem szükségképpen tartalmaz sikertelen helyreállítást.)

Legyen tehát egy karbantartó feldolgozás időszükséglete — hibamentes feldolgozást feltételezve — $T_F$, $t(\leqq T_F)$ időpontig a feldolgozás költségét a $c_F(t)$ függvény adja meg. Sikertelen feldolgozás többféle lehet attól függően, hogy a feldolgozás mely részében (korábban vagy később) következett be hiba. Legyen ezek száma $L$.

Úgy képzeljük, hogy a $T_F$ időszükségletű $F$-karbantartás $L$ számú, rendre $T_{F_1}$, $T_{F_2}$...$T_{F_L}$ időszükségletű $F_1$, $F_2$, ... $F_L$ leképzés egymásutáni végrehajtásával adódik $(T_F=T_{F_1}+T_{F_2}+\ldots+T_{F_L})$. A meghibásodást mindig valamelyik $T_{F_k}$ $(1\leqq k\leqq L)$ intervallum végén észleljük. Legyen továbbá $t_k=\sum_{i\leqq k} T_{F_i}$, akkor a bevezetett leképezéseknél $p_j(F_1)$, $p_j(F_2)$...$p_j(F_L)$ meghibásodási valószínűségekként a $p_j(F_k)=P_j(t_k)-P_j(t_{k-1})$, a megfelelő $c(F_1)$, $c(F_2)$.....$c(F_L)$ költségekként pedig a $c(F_k)=c_F(t_k)$ értékekkel számolunk. A sikeres feldolgozás valószínűsége tehát $p_j(F)=1-P_j(T_F)$, költsége pedig $c_F(T_F)$. Egy $j$-edik ciklusbeli sikeres helyreállítás időszükséglete $T_H^{(j)}=T_{RS}+(j-1)T_{HF}$, ahol $T_{RS}$ a visszaállítás, $T_{HF}$ pedig egy (megelőző) ciklus karbantartásainak helyreállításához szükséges idő. $t(\leqq T_H^j)$ időpontig a helyreállítás költségét a $c_H^{(j)}(t)$ függvény írja le. Ugyancsak többféle sikertelen helyreállítást veszünk figyelembe. Ezek számát $L_j$-vel jelölve legyenek a megfelelő bekövetkezési valószínűségek, illetve költségek $p_j(H_1^{(j)})...p_j(H_{L_j}^{(j)})$, illetve $(c(H_1^{(j)})...$ $...c(H_{L_j}^{(j)}))$. A $j$-edik ciklusbeli sikeres helyreállítás valószínűsége legyen $p_j(H^{(j)})=$ $=1-P_j(T_H^{(j)})$, költsége pedig $c(H^{(j)})=c_H^{(j)}(T_H^{(j)})$.

Legyenek a különböző részciklusok bekövetkezésének valószínűségei $p_1, p_2 \ldots$, költségei $c_1, c_2 \ldots$. Ha az $s$-edik részciklus az $F_l$ sikertelen feldolgozással kezdődik, amit $N$ sikertelen helyreállítás követ, melyek között $l_1$ számú $H_1^{(j)}$, $l_2$ számú $H_2^{(j)}$ ... helyreállítási kísérlet van, akkor

$$p_s = p_j(F_l)\frac{N!}{l_1!\ldots l_{L_j}!}\, p_j(H_1^{(j)})^{l_1}\ldots p_j(H_{L_j}^{(j)})^{l_{L_j}} p_j(H^{(j)})$$

és

$$c_s = c(F_l)+l_1 c(H_1^{(j)})+\ldots+l_{L_j} c(H_{L_j}^{(j)})+c(H^{(j)})$$

A $j$-edik ciklus várható költsége

$$(5.1)\qquad \sum_{\substack{l_1+\ldots+l_s=M \\ S,\,M\to\infty}} \frac{M!}{l_1!\ldots l_s!}\, p_1^{l_1}\ldots p_s^{l_s} p_j(F)\left(l_1c_1+\ldots+l_sc_s+c(F)\right) =$$

$$= \frac{\left(\sum_s p_sc_s\right)p_j(F)}{\left(1-\sum_s p_s\right)^2} + \frac{c(F)p_j(F)}{1-\sum_s p_s}$$

Az itt szereplő összegekre — az előzőhöz hasonlóan —

$$(5.2)\qquad \sum_s p_s = \left(\sum_l p_j(F_l)\right)\frac{p_j(H_l^{(j)})}{1-\sum_l p_j(H^{(j)})}$$

$$(5.3)\qquad \sum_s p_sc_s = \sum_l \left[\frac{\left(\sum_{l'} p_j(H_{l'}^{(j)})c(H_{l'}^{(j)})\right)p_j(H^{(j)})\,p_j(F_l)}{1-\sum_{l'} p_j(H_{l'}^{(j)})^2} + \right.$$

$$\left. + \frac{\left(c(F_l)+c(H^{(j)})\right)p_j(H^{(j)})\,p_j(F_l)}{1-\sum_{l'} p(H_{l'}^{(j)})}\right]$$

adódik. Minthogy

$$\sum_{l'} p_j(H_{l'}^{(j)}) = 1-p_j(H^{(j)}) = P_j(T_H^{(j)}),$$

$$\sum_l p_j(F_l) = 1-p_j(F) = P_j(T_F),$$

$$\sum_{l'} p_j(H_{l'}^{(j)})c(H_{l'}^{(j)}) = \int_0^{T_H^{(j)}} c_H^{(j)}(t)\,dP_j(t),$$

$$\sum_l p_j(F_l)c(F_l) = \int_0^{T_F} c_F(t)\,dP_j(t),$$

ezeket az (5.2) és (5.3) összefüggésekbe behelyettesítve, a $j$-edik ciklus várható költségére a

$$(5.4)\qquad \frac{\sum_s p_sc_s}{1-\sum_s p_s} + c(F) = \frac{1}{p_j(F)}\left[\frac{\int_0^{T_H^{(j)}} c_H^{(j)}(t)\,dP_j(t)}{p_j(H^{(j)})}\left(1-p_j(F)\right) + \right.$$

$$\left. + \int_0^{T_H} c_F(t)\,dP_j(t) + c(H^{(j)})\left(1-p_j(F)\right)\right] + c(F) =$$

$$= \frac{1-p_i(F)}{p_j(F)\,p_j(H^{(j)})}\int_0^{T_H^{(j)}} c_H^{(j)}(t)\,dP_j(t) + \frac{1-p_j(F)}{p_j(F)}\,c(H^{(j)}) +$$

$$+ \frac{1}{p_j(F)}\int_0^{T_H} c_F(t)\,dP_j(t) + c(F)$$

kifejezést kapjuk.

Ha $\overline{T}_{F'}$, $\overline{T}_{RS}$ és $\overline{T}_{HF}$ jelöli a teljes adatrendszerre vonatkozó megfelelő időket és az adatrendszert $K$ — egyenlő — részre osztottuk, akkor

$$T_F = \overline{T}_{F/K}$$

$$T_{RS} = \overline{T}_{RS/K}$$

$$T_{HF} = \overline{T}_{HF/K}$$

Legyen továbbá $c(S)$ a teljes adatrendszer kimentésének költsége és $n$ a két kimentés közötti karbantartó ciklusok száma. Akkor, a két kimentés közötti teljes költség várható értéke, osztva a teljes hasznos idővel:

$$\bar{c}(n) = \frac{1}{nK\overline{T}_{F/K}} \left\{ c(S) + K \left[ \sum_{j=1}^{n} \frac{P_j(\overline{T}_{F/K})}{1-P_j(\overline{T}_{F/K})} \times \right.\right.$$

$$\times \left( \frac{1}{1-P_j(\overline{T}_H^{(j)}/K)} \int_0^{T_H^{(j)}/K} c_H^{(j)}(t)\, dP_j(t) + c_H(\overline{T}_H^{(j)}/K) \right) +$$

$$\left.\left. + \frac{1}{1-P_j(\overline{T}_{F/K})} \int_0^{T_{F/K}} c_F(t)\, dP_j(t) + c_F(\overline{T}_{F/K}) \right] \right\}$$

A fenti $\bar{c}(n)$ mutató megadja, hogy a meghibásodásokat követő helyreállítások és az ezekhez szükséges kimentések miatt milyen mértékben nő a karbantartáshoz szükséges idő.

A következőkben az eddigieknek megfelelően feltételezzük, hogy

$$P_j(t) = 1 - e^{-\lambda_j t}$$

$$c_F(t) = t + \delta_F$$

és

$$c_H^{(j)}(t) = \begin{cases} t + \delta_{RS}, & \text{ha} \quad 0 < t \leqq T_{RS} \\ t - T_{RS} - lT_{HF} + \delta_{HF} + c_{RS} + lc_{HF}, & \end{cases}$$

$$\text{ha} \quad T_{RS} + lT_{HF} < t \leqq T_{RS} + (l+1)T_{HF} \quad \text{és} \quad 0 \leqq l \leqq j-2,$$

ahol

$$c_{RS} = T_{RS} + \delta_{RS}$$

és

$$c_{HF} = T_{HF} + \delta_{HF}$$

az egy ciklusbeli visszamentés, illetve az egy ciklusbeli karbantartás helyreállításának költsége és a $\delta$-k a különböző eljárások fix költségeit jelölik.

Ebben az esetben (5.2), (5.3) és (5.4) összefüggésekben szereplő integrálok kiszámíthatók, nevezetesen

$$\frac{1}{1-P_j(\overline{T}_{F/K})} \int_0^{T_{F/K}} c_F(t)\, dP_j(t) + c_F(\overline{T}_{F/K}) = \left(\delta_F + \frac{1}{\lambda_j}\right) e^{\lambda_j \overline{T}_{F/K}} - \frac{1}{\lambda_j}$$

és

$$\frac{P_j(\overline{T}_{F/K})}{1-P_j(\overline{T}_{F/K})}\left(\frac{1}{1-P_j(\overline{T}_H^{(j)}/_K)}\int_0^{T_H^{(j)}/_K} c_H^{(j)}(t)\,dP_j(t)+c_H(\overline{T}_H^{(j)}/_K)\right)=$$

$$=(e^{\lambda_j T_{F/K}}-1)\left[\left(\delta_{RS}+\delta_{HF}e^{-\lambda_j T_{RS/K}}\frac{1-e^{-\lambda_j(j-1)T_{HF/K}}}{1-e^{-\lambda_j T_{HF/K}}}+\frac{1}{\lambda_j}\right)\times\right.$$

$$\left.\times e^{\lambda_j\frac{T_{RS}+(j-1)T_{HF}}{K}}-\frac{1}{\lambda_j}\right]=(e^{\lambda_j T_{F/K}}-1)\times$$

$$\times\left[\left(\delta_{RS}+\frac{1}{\lambda_j}\right)e^{\lambda_j\frac{T_{RS}+(j-1)T_{HF}}{K}}+\delta_{HF}\left(\sum_{l=1}^{j-1}e^{\lambda_j l T_{HF/K}}\right)-\frac{1}{\lambda_j}\right],$$

azaz

$$\bar{c}(n)=\frac{c(S)+\sum_{j=1}^{n}K\left\{\delta_F e^{\lambda_j T_{F/K}}+(e^{\lambda_j T_{F/K}}-1)\left[\left(\delta_{RS}+\right.\right.\right.}{n\overline{T}_F}+$$

$$\left.\left.\left.+\frac{1}{\lambda_j}\right)e^{\lambda_j\frac{T_{RS}+(j-1)T_{HF}}{K}}+\delta_{HF}\left(\sum_{l=1}^{j-1}e^{\lambda_j l T_{HF/K}}\right)\right]\right\}}{n\overline{T}_F}.$$

A $\delta$-kat fixnek véve ez a kifejezés $K$-nak csökkenő függvénye. A bevezetés szerint az eljárások költségének a végrehajtásához szükséges időtől függő részén túli és annál lényegesen kisebb fix részeit kifejező $\delta$-k tartalmaznak egy $K$-ban növekvő tagot és így ezt a megállapítást csak $K$ „szóbajövő" értékei mellett tekintjük igaznak, ami ugyanaz az eredmény, mint korábban. Megtartva az eddig bevezetett költségfüggvényeket, $\lambda_j=\lambda$-t feltételezve és a $\delta$-kat elhagyva $\bar{c}(n)$ számlálójára

$$c(S)+\frac{K}{\lambda}(e^{\lambda T_{F/K}}-1)e^{\lambda T_{RS/K}}\left(\sum_{j=1}^{n}e^{\lambda(j-1)T_{HF/K}}\right)=$$

$$=c(S)+\frac{K}{\lambda}e^{\lambda T_{RS/K}}\frac{e^{\lambda T_{F/K}}-1}{e^{\lambda T_{HF/K}}-1}(e^{n\lambda T_{HF/K}}-1)= \tag{5.5}$$

$$=c(S)+\frac{K}{\lambda}\left(1+\frac{\lambda\overline{T}_{RS}}{K}\right)\frac{\overline{T}_F}{\overline{T}_{HF}}\left[\frac{1}{2}\left(\frac{n\lambda\overline{T}_{HF}}{K}\right)^2+\left(\frac{n\lambda\overline{T}_{HF}}{K}\right)\right]+\sigma\left(\left(\frac{\lambda}{K}\right)^2\right)+\sigma\left(\left(\frac{n\lambda}{K}\right)^3\right)$$

adódik. Mint a 4. pontban legyen

$$c(S)\approx\tilde{\alpha}_{42}\overline{T}_F$$

$$\overline{T}_{RS}(=c(S))=\tilde{\alpha}_{42}\overline{T}_F$$

$$\overline{T}_{HF}=\frac{\tilde{\tilde{\alpha}}_{HF}}{\tilde{\tilde{\alpha}}_4}\overline{T}_F$$

és $\tilde{\tilde{\alpha}}_4 = 1$. Akkor (5.5) helyett

$$\tilde{\alpha}_{42}\overline{T}_F + \left(\frac{K}{\lambda} + \tilde{\alpha}_{42}\overline{T}_F\right)\frac{1}{\tilde{\tilde{\alpha}}_{HF}}\left[\frac{1}{2}\left(\frac{n\tilde{\tilde{\alpha}}_{HF}\lambda\overline{T}_F}{K}\right)^2 + \frac{n\tilde{\tilde{\alpha}}_{HF}\lambda\overline{T}_F}{K}\right]$$

írható, és így

$$\bar{c}(n) = \frac{\tilde{\alpha}_{42}}{n} + \frac{n}{2}\left[\frac{\tilde{\tilde{\alpha}}_{HF}\lambda\overline{T}_F}{K} + \frac{\tilde{\alpha}_{42}\tilde{\tilde{\alpha}}_{HF}\lambda^2\overline{T}_F^2}{K^2}\right] + 1 + \frac{\tilde{\alpha}_{42}\lambda\overline{T}_F}{K}$$

Könnyen ellenőrizhető, hogy — a paraméterek megadott értékei, illetve a változókra szóba jövő értékek mellett — az alkalmazott közelítések jogosak.

Innen

$$n_{\text{opt}} = \left(\frac{2\tilde{\alpha}_{42}K}{\tilde{\tilde{\alpha}}_{HF}\lambda\overline{T}_F + \frac{\tilde{\alpha}_{42}\tilde{\tilde{\alpha}}_{HF}\lambda^2\overline{T}_F^2}{K}}\right)^{1/2}$$

ami — érthetően — kisebb, mint a megfelelő 4. pontbeli érték. Az $n_{\text{opt}}$-hoz tartozó relatív költség

$$\bar{c}(n_{\text{opt}}) = 1 + \left(\frac{2\tilde{\alpha}_{42}\tilde{\tilde{\alpha}}_{HF}\lambda\overline{T}_F}{K} + \frac{2\tilde{\alpha}_{42}\tilde{\tilde{\alpha}}_{HF}\lambda^2\overline{T}_F^2}{K}\right)^{1/2} + \frac{\tilde{\alpha}_{42}\lambda\overline{T}_F}{K},$$

ami viszont — ugyancsak érthetően — nagyobb, mint 4. pontbeli megfelelője. A különbség az $\tilde{\tilde{\alpha}}_{HF}=0{,}5$, $\alpha_{42}=4$, $\overline{T}_F=320$ és $\lambda=\frac{1}{320}$ értékek mellett (amik a tervezett valóságos rendszert várhatóan jellemzik) lényegében nulla. Az adott esetben tehát jogosnak tekinthető a helyreállítás alatti meghibásodás figyelmen kívül hagyása. Ugyanakkor a működő rendszerre vonatkozó értékek ($\tilde{\tilde{\alpha}}_{HF}=1$, $\tilde{\alpha}_{42}=2$, $\overline{T}_F=600$ és $K=40$) mellett még $\lambda=\frac{1}{600}$-nál is számottevően nagyobb a jelen modellből nyert optimális relatív költség érték. De még ez a nagyobb $\bar{c}(n_{\text{opt}})$ is nagyon messze van a gyakorlatnak megfelelő $\bar{c}(1)$-től.

## IRODALOM

[1] BENCZÚR, A., „Adatkezelő rendszerek biztonsági problémái", Kandidátusi értekezés, 1977.

[2] CHANDY, K., M., BROWNE, J. C., DISSLY, C. W., and UHRIG, W. R., „Analytic models for roll back and recovery in data base systems", *IEEE Trans. on Software Eng.* **1** (1975) 100—110.

(Beérkezett: 1982. március 17.)
(Átdolgozva beérkezett: 1983. július 6.)

BENCZÚR ANDRÁS
ÁLLAMI NÉPESSÉGNYILVÁNTARTÓ HIVATAL
1094 BUDAPEST, MÁRTON U. 15.

STAHL JÁNOS
SZÁMÍTÁSTECHNIKA-ALKALMAZÁSI VÁLLALAT
1536 BUDAPEST, CSALOGÁNY U. 30—32.

## ON UPDATING A LARGE-SCALE DATASYSTEM

A. BENCZÚR and J. STAHL

The paper deals with the expected updating cost of a real-life system. The effect of changing the save frequency and splitting the whole system into independent parts is considered in two cases. In the first case we assume that no failures occur in the recovery process. In the second case we drop this assumption. We compare the obtained results to the corresponding values of the real system.

*Alkalmazott Matematikai Lapok* **10** (1984) 15—34

# ÉRDEKLŐDÉS-IRÁNYÍTOTT TÖBBVÁLTOZÓS DETERMINÁCIÓS EGYÜTTHATÓ

JÓZSA SÁNDOR

Keszthely

A cikk első négy fejezetében a többszörös korrelációs együttható négyzetének általánosításaként egy nyomtípusú predikciómérő mutatót vezetünk be, melyben meghatározó szerepet kap egy — a predikátumhoz rendelt — ún. érdeklődés-mátrix. Áttekintjük a mutató legfontosabb tulajdonságait, szólunk az irodalomból ismert speciális eseteiről.

A cikk második részében megmutatjuk, hogyan kapcsolódik a bevezetett mutató a főkomponensanalízis, a kanonikus korrelációanalízis, a faktoranalízis és a diszkriminanciaanalízis alapfeladatához. Eközben interpretációs kérdéseket is érintünk.

## 1. Bevezetés

Két valószínűségi változó lineáris kapcsolatának szorosságát méri a korrelációs együttható ($r$). Ennek egyenes általánosítása a többszörös korrelációs együttható ($R$), amikor az egyik változó vektor. Az utóbbi időben a többszörös korrelációs együttható több általánosításával találkozhattunk két valószínűségi vektorváltozóra [9], [14], [19]. Ezek többnyire a (particionált) korrelációs mátrixból épülnek fel, egyrészük determinánssal operál, másrészük mátrixnyommal. A determináns képzésű mutatók a *Wilks—Fresch-féle általánosított variancia* előnyös tulajdonságain alapszanak, reguláris lineáris transzformációkkal szemben invariánsak, szimmetrikusak a két vektorváltozóban. Jelentős statisztikai alkalmazásuk a két vektorváltozó korrelálatlanságának ellenőrzése *(Wilks-kritérium)*, a kapcsolat szorosságának mérésére azonban nem látszanak alkalmasnak. A mátrixnyom képzésű mutatókról áttekintést ad LENGYEL TAMÁS [14], ezek a RAO által javasolt mátrixnyommal mérik az összvarianciát. (Lásd még KRZYSKO, Biom. J., 1982.)

ANDERSON [2—245] rámutat arra, hogy két változó-csoport között a kapcsolatot leírni egyetlen mutatóval nem lehet. A próbálkozások célja nem az, hogy a strukturális kapcsolatot kifejezzék, csupán annak globális zártságát kívánják számszerűsíteni. Dolgozatom e kérdéshez kapcsolódva egy olyan általános nyomtípusú, irányított kapcsolatszorosság-mérő mutatót ismertet, amely szemléletmódbeli pontosítást tesz lehetővé. A szóban forgó mérőszám tartalmaz egy paraméter-mátrixot, így nem egyértelmű, de éppen ebből adódóan alkalmas arra, hogy néhány többváltozós módszer interpretációjában segítsen. Úgy gondolom, különösen aktuális ez napjainkban, amikor a számítóközpontok program-csomagjai sajátos helyzetet teremtettek: a felhasználóra csak az eredmények interpretálásának feladata hárul, az alkalmazott módszert mélységeiben nem szükséges ismernie.

*Jelölések, terminológia.* A tárgyalásban végig lineáris algebrai eszközöket alkalmazok jórészt a [17] munkára támaszkodva. A gyakoribb jelölések: $\mathbf{A}(m\times n)=$

$=m\times n$-es mátrix, $\mathbf{A}'=\mathbf{A}$ transzponáltja, $\mathbf{A}^{-1}=\mathbf{A}$ inverze, $|\mathbf{A}|=\mathbf{A}$ determinánsa, tr $\mathbf{A}=\mathbf{A}$ nyoma, diag $\mathbf{A}=$az $\mathbf{A}$ mátrix diagonális elemeivel képzett diagonális mátrix, ugyanezt jelenti diag $(a_{11}, \ldots, a_{nn})$; rang $\mathbf{A}=\mathbf{A}$ rangja, a $q$-adrendű egységmátrix $\mathbf{E}_q$. Adatmátrix helyett valószínűségi változóról (röviden: v. változó), illetve valószínűségi vektorváltozóról (röviden: v. vektor) beszélek. A két v. vektort, amelyek kapcsolatát vizsgáljuk, $\mathbf{X}$, illetve $\mathbf{Y}$ jelöli, komponenseik $x_1, x_2, \ldots, x_p$, illetve $y_1, y_2, \ldots, y_q$. A szokásos elnevezéssel $\mathbf{X}$ a predikátor-, $\mathbf{Y}$ a predikátum v. vektor.

## 2. Érdeklődés-irányított többváltozós determinációs együttható

$R^2$-nek, az ún. determinációs együtthatónak az általánosításához $R^2$ varianciamagyarázó interpretációjából indulunk ki. Legyen $\mathbf{X}$ v. vektor $\boldsymbol{\Sigma}_{\mathbf{XX}}$ nemszinguláris kovarianciamátrixszal és $y$ v. változó, var $y=\sigma^2$ varianciával. $\mathbf{X}$ és $y$ kovarianciavektorát jelölje $\boldsymbol{\sigma}_{\mathbf{X}y}=\boldsymbol{\sigma}'_{y\mathbf{X}}\cdot y$-hoz a legkisebb négyzetek értelmében legközelebb álló $a+\mathbf{b}'\mathbf{X}$ regressziós függvény négyzetelt „távolsága" $y$-tól:

$$\sigma^2_{y\cdot\mathbf{X}} = \operatorname{var}(y-a-\mathbf{b}'\mathbf{X}) = \sigma^2-\boldsymbol{\sigma}_{y\mathbf{X}}\boldsymbol{\Sigma}^{-1}_{\mathbf{XX}}\boldsymbol{\sigma}_{\mathbf{X}y} = (1-R^2)\sigma^2.$$

Ez az eltérés $y$ varianciájának az $\mathbf{X}$ által nem magyarázott része, így a magyarázott rész $\sigma^2-\sigma^2_{y\cdot\mathbf{X}}$, amit $\sigma^2_{y:\mathbf{X}}$-szel jelölve

$$\sigma^2_{y:\mathbf{X}} = \boldsymbol{\sigma}_{y\mathbf{X}}\boldsymbol{\Sigma}^{-1}_{\mathbf{XX}}\boldsymbol{\sigma}_{\mathbf{X}y} = R^2\sigma^2. \tag{2.1}$$

Az utóbbi összefüggés mutatja $R^2$ varianciamagyarázó jelentését. A $\sigma^2_{y:\mathbf{X}}$ mennyiséget $y$ $\mathbf{X}$-re vonatkoztatott redundáns varianciájának nevezik, ez az $\mathbf{X}$ által predikált variancia.

Legyen most $\mathbf{Y}$ v. vektor $y_1, y_2, \ldots, y_q$, páronként korrelálatlan komponensekkel, $\sigma_1^2, \sigma_2^2, \ldots, \sigma_p^2$ varianciákkal. $\mathbf{X}$ és $y_k$ determinációs együtthatóját jelölje $R_k^2$. Írjuk fel (2.1)-et $\mathbf{Y}$ valamennyi komponensére:

$$\begin{cases} \sigma^2_{y_1:\mathbf{X}} = \boldsymbol{\sigma}_{y_1\mathbf{X}}\boldsymbol{\Sigma}^{-1}_{\mathbf{XX}}\boldsymbol{\sigma}_{\mathbf{X}y_1} = R_1^2\sigma_1^2 \\ \sigma^2_{y_2:\mathbf{X}} = \boldsymbol{\sigma}_{y_2\mathbf{X}}\boldsymbol{\Sigma}^{-1}_{\mathbf{XX}}\boldsymbol{\sigma}_{\mathbf{X}y_2} = R_2^2\sigma_2^2 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \sigma^2_{y_q:\mathbf{X}} = \boldsymbol{\sigma}_{y_q\mathbf{X}}\boldsymbol{\Sigma}^{-1}_{\mathbf{XX}}\boldsymbol{\sigma}_{\mathbf{X}y_q} = R_q^2\sigma_q^2. \end{cases} \tag{2.2}$$

Az $R_1^2, R_2^2, \ldots, R_q^2$ determinációkat egyetlen „közös" $\bar{R}^2$-be szeretnénk tömöríteni. Valamilyen átlagot keresünk. Feltéve, hogy $\mathbf{Y}$ komponenseinek fizikai dimenziója azonos, az $R_k^2\sigma_k^2$ mennyiségek összeadhatók. Az a közepes $\bar{R}^2$, melyet az $R_k^2$ determinációk helyére írva az összeg nem változik, az $R_k^2$-ek súlyozott átlaga:

$$\bar{R}^2 = \sum_{k=1}^{q}\sigma_k^2 R_k^2 \Big/ \sum_{k=1}^{q}\sigma_k^2 \tag{2.3}$$

Nem megengedett viszont az összegzés, ha $\mathbf{Y}$ eltérő dimenziójú komponensekből áll. A dimenziók elhagyása sem segít, mert $\bar{R}^2$ mértékegység-függő: ha egy-egy komponensnél módosítjuk a mértékegységet, $\bar{R}^2$ értéke megváltozik. Elemezzük ezt a tulajdonságát.

$\bar{R}^2$-ben az egyes komponensek a varianciáikkal arányos súllyal szerepelnek. Ezt úgy is fogalmazhatjuk, hogy az egyes komponensek saját mértékegységükben kifejezett egységnyi megváltozásai azonos mértékben járulnak hozzá $\bar{R}^2$-hez. A gyakorlatban viszont ritkán fordul elő, hogy a vizsgált változók egységnyi megváltozásainak azonos jelentősége lenne. Ez a szemléletmód útmutatást ad arra, hogy miként lehet kiküszöbölni az eltérő fizikai dimenziók zavaró hatását és a mértékegység-függőséget.

Tételezzük föl, hogy az egyes komponensek $\Delta_1, \Delta_2, \ldots, \Delta_q$ megváltozásaira irányuló „érdeklődésünk" azonos mérvű, röviden e módosulások érdeklődés-ekvivalensek. Ekkor az egységnyi megváltozásokra eső érdeklődések rendre $1/\Delta_1^2$, $1/\Delta_2^2, \ldots, 1/\Delta_q^2$, ha varianciákra vonatkoztatjuk. A fentebb képzett $\bar{R}^2$-et ezekkel a „súlyokkal" korrigálva nyerjük az

$$\bar{R}^2_{\Delta^{-2}} = \sum_{k=1}^{q}\left(\frac{\sigma_k}{\Delta_k}\right)^2 R_k^2 \Big/ \sum_{k=1}^{q}\left(\frac{\sigma_k}{\Delta_k}\right)^2 = \sum_{k=1}^{q}\tilde{\sigma}_k^2 R_k^2 \Big/ \sum_{k=1}^{q}\tilde{\sigma}_k^2 \tag{2.4}$$

átlagos determinációs együtthatót, melyben

$$\tilde{\sigma}_k^2 = (\sigma_k/\Delta_k)^2 = \operatorname{var}(y_k/\Delta_k).$$

Ebben a $\sigma_k/\Delta_k$ hányadosból a fizikai dimenzió kiesik, ugyanakkor az előírt érdeklődés-arányoknak megfelelő információ-arányokat biztosít. $\bar{R}^2_{\Delta^{-2}}$ formuláját (2.2) felhasználásával tömörebben is felírhatjuk. A számláló:

$$\sum_{k=1}^{q}\left(\frac{\sigma_k}{\Delta_k}\right)^2 R_k^2 = \sum_{k=1}^{q}\frac{1}{\Delta_k^2}\boldsymbol{\sigma}_{y_k\mathbf{X}}\boldsymbol{\Sigma}_{\mathbf{XX}}^{-1}\boldsymbol{\sigma}_{\mathbf{X}y_k} = \operatorname{tr}\boldsymbol{\Delta}^{-2}\boldsymbol{\Sigma}_{\mathbf{YX}}\boldsymbol{\Sigma}_{\mathbf{XX}}^{-1}\boldsymbol{\Sigma}_{\mathbf{XY}},$$

ahol $\boldsymbol{\Delta}=\operatorname{diag}(\Delta_1, \Delta_2, \ldots, \Delta_q)$ és $\boldsymbol{\Sigma}_{\mathbf{XY}}=\boldsymbol{\Sigma}'_{\mathbf{YX}}=(\boldsymbol{\sigma}_{\mathbf{X}y_1}, \boldsymbol{\sigma}_{\mathbf{X}y_2}, \ldots, \boldsymbol{\sigma}_{\mathbf{X}y_p})$, a kovarianciamátrix $\mathbf{X}$ és $\mathbf{Y}$ között. A nevező:

$$\sum_{k=1}^{q}(\sigma_k/\Delta_k)^2 = \operatorname{tr}\boldsymbol{\Delta}^{-2}\boldsymbol{\Sigma}_{\mathbf{YY}},$$

itt $\boldsymbol{\Sigma}_{\mathbf{YY}}=\operatorname{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_q^2)$, egyben az $\mathbf{Y}$ v. vektor kovarianciamátrixa. Innen (2.4) tömörebb alakja:

$$R^2_{\Delta^{-2}} = \frac{\operatorname{tr}\boldsymbol{\Delta}^{-2}\boldsymbol{\Sigma}_{\mathbf{YX}}\boldsymbol{\Sigma}_{\mathbf{XX}}^{-1}\boldsymbol{\Sigma}_{\mathbf{XY}}}{\operatorname{tr}\boldsymbol{\Delta}^{-2}\boldsymbol{\Sigma}_{\mathbf{YY}}}. \tag{2.5}$$

Az utóbbi formula alapján definiáljuk az érdeklődés-irányított többváltozós determinációs együtthatót (l. még [9]). Eltekintünk attól a megszorítástól, hogy $\boldsymbol{\Delta}^{-2}$ és $\boldsymbol{\Sigma}_{\mathbf{YY}}$ diagonálisak legyenek. Tekintsük tehát a $p+q$ komponensű $\begin{pmatrix}\mathbf{X}\\ \mathbf{Y}\end{pmatrix}$ v. vektor particionált kovarianciamátrixát:

$$\begin{pmatrix}\boldsymbol{\Sigma}_{\mathbf{XX}} & \boldsymbol{\Sigma}_{\mathbf{XY}}\\ \boldsymbol{\Sigma}_{\mathbf{YX}} & \boldsymbol{\Sigma}_{\mathbf{YY}}\end{pmatrix} \quad (p+q\times p+q),$$

$\boldsymbol{\Sigma}_{\mathbf{XX}}$ és $\boldsymbol{\Sigma}_{\mathbf{YY}}$ legyenek nemszingulárisak. (2.1) analógiájára vezessük be az $\mathbf{Y}$ prediká-

tumnak az $\mathbf{X}$ predikátorra vonatkozó redundanciamátrixát:

$$\boldsymbol{\Sigma}_{\mathbf{Y}:\mathbf{X}} = \boldsymbol{\Sigma}_{\mathbf{YX}} \boldsymbol{\Sigma}_{\mathbf{XX}}^{-1} \boldsymbol{\Sigma}_{\mathbf{XY}}, \tag{2.6}$$

amely — mint ismeretes — $\mathbf{Y}$ $\mathbf{X}$-re vonatkozó regressziójának kovarianciamátrixa.

2.1. DEFINÍCIÓ. Legyen $\mathbf{G}(q \times q)$ tetszőleges szimetrikus pozitív szemidefinit mátrix. Az $\mathbf{X}$-ben $\mathbf{Y}$-ról a $\mathbf{G}$ „érdeklődés" mellett tartalmazott „információ"

$$R_{\mathbf{G}}^2(\mathbf{Y}:\mathbf{X}) = \frac{\operatorname{tr} \mathbf{G}\boldsymbol{\Sigma}_{\mathbf{Y}:\mathbf{X}}}{\operatorname{tr} \mathbf{G}\boldsymbol{\Sigma}_{\mathbf{YY}}} \tag{2.7}$$

formulával értelmezett mérőszámát $\mathbf{G}$-irányított (többváltozós) determinációs együtthatónak (röviden: $\mathbf{G}$-irányított determináció) nevezzük.

Megjegyezzük, hogy $\mathbf{G}$ (a továbbiakban: érdeklődés-mátrix) most nem $\mathbf{Y}$ komponenseire, hanem a komponensek együttesére, mégpontosabban a $\boldsymbol{\Sigma}_{\mathbf{YY}}$ kovarianciamátrixra irányuló érdeklődést fejez ki. Ez önmagában mutatja $R_{\mathbf{G}}^2(\mathbf{Y}:\mathbf{X})$ aszimmetriáját.

A további kifejtés egyszerűsítése végett feltettük, hogy $\boldsymbol{\Sigma}_{\mathbf{XX}}$ és $\boldsymbol{\Sigma}_{\mathbf{YY}}$ nemszinguláris, ettől a megszorítástól azonban el lehet tekinteni. A 4. pontban erről lesz szó.

$R_{\mathbf{G}}^2(\mathbf{Y}:\mathbf{X})$ részletesebb elemzése előtt lássuk néhány speciális esetét. A definíció alapján könnyen ellenőrizhető, hogy $q=1$ esetén $\mathbf{G}=g$ tetszőleges konstanssal $R_g^2(y:\mathbf{X})=R^2$. $\mathbf{G}=\mathbf{E}_q$ mellett kapjuk az $R_k^2$ determinációs együtthatók $\sigma_k^2=\operatorname{var}(y_k)$ súlyokkal képzett átlagát (formálisan (2.3)-at), amely egység-ekvivalens érdeklődésnek felel meg. A fentebb értelmezett $\mathbf{G}=\boldsymbol{\Delta}^{-2}$ mátrixszal, (2.4)-hez jutunk ($\Delta$-ekvivalens érdeklődés). A $\mathbf{G}=(\operatorname{diag} \boldsymbol{\Sigma}_{\mathbf{YY}})^{-1}=\operatorname{diag}\,(\sigma_1^{-2}, \sigma_2^{-2}, \ldots, \sigma_q^{-2})$ választással (2.7) az $\frac{1}{q}\sum_{k=1}^{q} R_k^2$ átlaghoz vezet, amely $\Delta_k=\sigma_k$-nak felel meg, azaz $\sigma_1, \sigma_2, \ldots, \sigma_q$ érdeklődés-ekvivalensek (szórás-ekvivalens érdeklődés). Kevés meggondolással belátható, hogy a STEWART és LOVE [18] által bevezetett

$$R_*^2 = \frac{1}{q} \sum_k \varrho_k^2 \mathbf{v}_k' \mathbf{R}_{\mathbf{YY}}^2 \mathbf{v}_k$$

mérőszám is megegyezik $\frac{1}{q}\sum_k R_k^2$-tel, ahol $\mathbf{R}_{\mathbf{YY}}$ az $\mathbf{Y}$ v. vektor korrelációmátrixa, $\varrho_k$ az $\mathbf{X}$ és $\mathbf{Y}$ között értelmezett $k$-adik kanonikus korreláció és $\mathbf{v}_k$ a $\varrho_k$-hoz tartozó, $\mathbf{Y}$-oldalon képzett kanonikus komponens együttható-vektora.

Végül, könnyű megmutatni, hogy a $\mathbf{G}=\boldsymbol{\Sigma}_{\mathbf{YY}}^{-1}$ választással („információ-mátrix" [7])

$$R_{\boldsymbol{\Sigma}_{\mathbf{YY}}^{-1}}^2(\mathbf{Y}:\mathbf{X}) = \frac{1}{q} \sum_{k=1}^{s} \varrho_k^2, \tag{2.8}$$

az összegzés felső határa $s=\min(p, q)$. Ezzel a mutatóval több helyütt találkozhatunk (lásd pl. [14]), a fenti terminológiával „kanonikus, szórás-ekvivalens érdeklődés" mellett fejezi ki a determináció mértékét.

### 3. $R^2_G(\mathbf{Y}:\mathbf{X})$ fontosabb tulajdonságai

Ebben a fejezetben a G-irányított determináció invariancia-tulajdonságait, dekompozíció-kérdéseit vizsgáljuk és egy — az előző pont bevezetésében adottal megegyező alapelvű — interpretációját tárgyaljuk.

A kifejtésben néhol külön említés nélkül felhasználjuk azt, hogy $\mathbf{\Sigma}_{XX}$ és $\mathbf{\Sigma}_{YY}$ pozitív definit szimmetrikus, $\mathbf{\Sigma}_{Y:X}$ és $\mathbf{G}$ pozitív szemidefinit szimmetrikus, $\mathbf{G\Sigma}_{YY}$ és $\mathbf{G\Sigma}_{Y:X}$ pozitív szemidefinit mátrixok. Így a szereplő mátrixok sajátértékei nemnegatívok. Ebből azonnal adódik, hogy $R^2_G(\mathbf{Y}:\mathbf{X})\geqq 0$.

3.1. TÉTEL. $R^2_G(\mathbf{Y}:\mathbf{X})$ invariáns $\mathbf{X}$ reguláris lineáris transzformációval szemben, azaz $R^2_G(\mathbf{Y}:\mathbf{L}'\mathbf{X})=R^2_G(\mathbf{Y}:\mathbf{X})$, ha $\mathbf{L}(p\times p)$ nemszinguláris.

*Bizonyítás.* A tétel abból következik, hogy $\tilde{\mathbf{X}}=\mathbf{L}'\mathbf{X}$-re

$$\mathbf{\Sigma}_{Y:\tilde{X}} = \mathbf{\Sigma}_{Y\tilde{X}}\mathbf{\Sigma}^{-1}_{\tilde{X}\tilde{X}}\mathbf{\Sigma}_{\tilde{X}Y} = \mathbf{\Sigma}_{YX}\mathbf{L}(\mathbf{L}'\mathbf{\Sigma}_{XX}\mathbf{L})^{-1}\mathbf{L}'\mathbf{\Sigma}_{XY} = \mathbf{\Sigma}_{YX}\mathbf{\Sigma}^{-1}_{XX}\mathbf{\Sigma}_{XY} = \mathbf{\Sigma}_{Y:X}.$$

KÖVETKEZMÉNY: $R^2_G(\mathbf{Y}:\mathbf{X})$-ben $\mathbf{\Sigma}_{XX}$ minden esetben helyettesíthető az $\mathbf{R}_{XX}$ korrelációmátrixszal.

3.2. TÉTEL (dekompozíció $\mathbf{X}$ szerint). Ha $\mathbf{X}$ két korrelálatlan részvektorra bontható, azaz $\mathbf{X}'=(\mathbf{X}_1', \mathbf{X}_2')$ és $\mathbf{\Sigma}_{X_1X_2}=\mathbf{0}$, akkor tetszőleges $\mathbf{G}$-vel

$$R^2_G(\mathbf{Y}:\mathbf{X}) = R^2_G(\mathbf{Y}:\mathbf{X}_1)+R^2_G(\mathbf{Y}:\mathbf{X}_2).$$

*Bizonyítás.* Csak azt kell igazolni, hogy $\mathbf{\Sigma}_{Y:X}=\mathbf{\Sigma}_{Y:X_1}+\mathbf{\Sigma}_{Y:X_2}$.
Ez viszont abból következik, hogy

$$\mathbf{\Sigma}_{YX} = (\mathbf{\Sigma}_{YX_1}|\mathbf{\Sigma}_{YX_2}) \quad \text{és} \quad \mathbf{\Sigma}^{-1}_{XX} = \begin{pmatrix} \mathbf{\Sigma}^{-1}_{X_1X_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}^{-1}_{X_2X_2} \end{pmatrix}.$$

KÖVETKEZMÉNY. Válasszuk az $\mathbf{L}(p\times p)$ nemszinguláris mátrixot úgy, hogy $\mathbf{L}'\mathbf{\Sigma}_{XX}\mathbf{L}$ diagonális legyen. A 3.1 tétel szerint $\mathbf{X}$ helyettesíthető az $\mathbf{X}^*=(\mathbf{x}_1^*, \mathbf{x}_2^*, \ldots$ $\ldots \mathbf{x}_p^*)'=\mathbf{L}'\mathbf{X}$ v. vektorral és mivel az $x_i^*$ komponensek páronként korrelátlanok, fennáll:

$$R^2_G(\mathbf{Y}:\mathbf{X}) = R^2_G(\mathbf{Y}:x_1^*)+R^2_G(\mathbf{Y}:x_2^*)+\ldots+R^2_G(\mathbf{Y}:x_p^*). \tag{3.1}$$

3.3. TÉTEL. Bármely $\mathbf{G}$-re $R^2_G(\mathbf{Y}:\mathbf{X})=0$ akkor és csak akkor, ha $\mathbf{\Sigma}_{YX}=\mathbf{0}$.

A tétel annak a természetes elvárásnak a teljesülését biztosítja, hogy $R^2_G(\mathbf{Y}:\mathbf{X})$ határozottan pozitív legyen, ha van legalább egy korrelált $(x_i, y_k)$ pár. Megjegyezzük, hogy a determináns-képzésű mérőszámok nem felelnek meg ennek az elvárásnak, ha $q>p$.

*Bizonyítás.* Ha $\mathbf{\Sigma}_{YX}=\mathbf{0}$, akkor $\mathbf{\Sigma}_{Y:X}=\mathbf{0}$ is áll, így tetszőleges $\mathbf{G}$-vel $\operatorname{tr}\mathbf{G\Sigma}_{Y:X}=0$. Megfordítva, ha $R^2_G(\mathbf{Y}:\mathbf{X})=0$, helyettesítsük $\mathbf{X}$-et egy $\mathbf{L}'\mathbf{\Sigma}_{XX}\mathbf{L}=\mathbf{E}_p$-nek megfelelő $\mathbf{X}^*=\mathbf{L}'\mathbf{X}$ reguláris transzformáltjával (ilyen $\mathbf{L}$ létezik, mert $\mathbf{\Sigma}_{XX}$ szimmetrikus, pozitív definit mátrix). $\mathbf{G}=\mathbf{E}_q$-val ekkor

$$0 = \operatorname{tr}\mathbf{E}_q\mathbf{\Sigma}_{Y:X} = \operatorname{tr}\mathbf{\Sigma}_{YX^*}\mathbf{\Sigma}'_{YX^*},$$

amiből nyilvánvalóan $\mathbf{\Sigma}_{YX^*}=0$. Tekintve, hogy $\mathbf{\Sigma}_{YX^*}=\mathbf{\Sigma}_{YX}\mathbf{L}$, felhasználva a

rang $(\mathbf{AB}) \leqq$ rang $\mathbf{A}$ relációt, fennáll:

$$\text{rang } \boldsymbol{\Sigma}_{\mathbf{YX}} = \text{rang } \boldsymbol{\Sigma}_{\mathbf{YX}^*}\mathbf{L}^{-1} \leqq \text{rang } \boldsymbol{\Sigma}_{\mathbf{YX}^*} = 0.$$

Innen $\boldsymbol{\Sigma}_{\mathbf{YX}} = \mathbf{0}$, mint állítottuk.

Áttérve a prediktumra vonatkozó vizsgálatokra, nyilvánvaló, hogy $R^2_{\mathbf{G}}(\mathbf{Y}:\mathbf{X})$ nem invariáns $\mathbf{Y}$ lineáris transzformációval szemben. Ezek a leképezések a $\mathbf{G}$ mátrixhoz csatolhatók, vagyis az érdeklődést módosítják. A 3.2 tétellel analóg $R^2_{\mathbf{G}}(\mathbf{Y}:\mathbf{X})$ alábbi tulajdonsága.

3.4. TÉTEL. Ha $\mathbf{G} = \begin{pmatrix} \mathbf{G}_1 \mathbf{O} \\ \mathbf{O}\ \mathbf{G}_2 \end{pmatrix}$, ahol $\mathbf{G}_1(q_1 \times q_1)$ és $\mathbf{G}_2(q_2 \times q_2)$, $(q_1 + q_2 = q)$ parciális érdeklődés mátrixok, akkor a predikátum $\mathbf{Y}' = (\mathbf{Y}_1', \mathbf{Y}_2')$ $q_1 + q_2$ komponensű bontásával

$$R^2_{\mathbf{G}}(\mathbf{Y}:\mathbf{X}) = \frac{C_1^2 R^2_{\mathbf{G}_1}(\mathbf{Y}_1:\mathbf{X}) + C_2^2 R^2_{\mathbf{G}_2}(\mathbf{Y}_2:\mathbf{X})}{C_1^2 + C_2^2},$$

ahol

$$C_i^2 = \text{tr } \mathbf{G}_i \boldsymbol{\Sigma}_{\mathbf{YY}} \quad (i = 1, 2).$$

A bizonyítás a 2.1 definíció alapján egyszerű.

$R^2_{\mathbf{G}}(\mathbf{Y}:\mathbf{X})$-et most előállítjuk (közönséges) determinációs együtthatók súlyozott átlagaként.

3.5. TÉTEL. Legyen $\boldsymbol{\Gamma}\boldsymbol{\Gamma}'$ a $\mathbf{G}$ érdeklődés-mátrix tetszőleges faktorizációja, ahol $r = \text{rang } \mathbf{G}$-vel $\boldsymbol{\Gamma}$ $q \times r$-es. Képezzük az $\tilde{\mathbf{Y}} = (\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_r)' = \boldsymbol{\Gamma}'\mathbf{Y}$ v. vektort és legyen $\tilde{R}_k^2$ a determinációs együttható $\mathbf{X}$ és $\tilde{y}_k$ között, továbbá $\tilde{\sigma}_k^2 = \text{var } \tilde{y}_k$. Ekkor

$$R^2_{\mathbf{G}}(\mathbf{Y}:\mathbf{X}) = R^2_{\mathbf{E}_r}(\tilde{\mathbf{Y}}:\mathbf{X}) = \frac{\sum_{k=1}^{r} \tilde{\sigma}_k^2 \tilde{R}_k^2}{\sum_{k=1}^{r} \tilde{\sigma}_k^2}.$$

*Bizonyítás.* A $\mathbf{G}$-re tett feltevések biztosítják a $\mathbf{G} = \boldsymbol{\Gamma}\boldsymbol{\Gamma}'$ alakú faktorizáció(k) létezését.

A 2.1. definícióban a számláló és a nevező az alábbi módon alakíthatók:

$$\text{tr } \mathbf{G}\boldsymbol{\Sigma}_{\mathbf{Y}:\mathbf{X}} = \text{tr } \boldsymbol{\Gamma}\boldsymbol{\Gamma}'\boldsymbol{\Sigma}_{\mathbf{Y}:\mathbf{X}} = \text{tr } \boldsymbol{\Gamma}'\boldsymbol{\Sigma}_{\mathbf{Y}:\mathbf{X}}\boldsymbol{\Gamma} = \text{tr } \boldsymbol{\Sigma}_{\tilde{\mathbf{Y}}:\mathbf{X}} = \sum_{k=1}^{r} \tilde{\sigma}_k^2 \tilde{R}_k^2,$$

és

$$\text{tr } \mathbf{G}\boldsymbol{\Sigma}_{\mathbf{YY}} = \text{tr } \boldsymbol{\Gamma}\boldsymbol{\Gamma}'\boldsymbol{\Sigma}_{\mathbf{YY}} = \text{tr } \boldsymbol{\Gamma}'\boldsymbol{\Sigma}_{\mathbf{YY}}\boldsymbol{\Gamma} = \text{tr } \boldsymbol{\Sigma}_{\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}} = \sum_{k=1}^{r} \tilde{\sigma}_k^2.$$

Felhasználtuk, hogy $\text{tr } \mathbf{AB} = \text{tr } \mathbf{BA}$. Ezt az összefüggést a továbbiakban gyakran alkalmazzuk.

KÖVETKEZMÉNY. A tétel közvetlen folyamánya, hogy

$$0 \leqq R^2_{\mathbf{G}}(\mathbf{Y}:\mathbf{X}) \leqq 1.$$

Láttuk, hogy az alsó határt $\boldsymbol{\Sigma}_{\mathbf{XY}} = \mathbf{0}$ esetén kapjuk, egyszerű algebrai meggondolás

mutatja, hogy ha **Y** **X**-nek valamely lineáris képe, akkor a **G**-re irányított detaminációbármely **G**-re 1.

Az előző tételben bevezetett $\tilde{\mathbf{Y}}$ v. vektor komponensei általában nem korrelálatlanok. Ha a 2. pontban vázolt interpretációhoz kívánunk jutni, a $\mathbf{\Sigma}_{\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}}$ kovarianciamátrixot diagonálissá kell transzformálnunk. Ezáltal a **G** érdeklődési mátrix valóságos „iránya" is feltárható. Az alábbi alapvető tétel erre vonatkozik.

3.6. TÉTEL (dekompozíció **Y** szerint). Található olyan, páronként korrelálatlan komponensekből álló $\mathbf{Y}^*=(y_1^*, y_2^*, \dots y_r^*)'$ v. vektor, mellyel

$$R_{\mathbf{G}}^2(\mathbf{Y}:\mathbf{X}) = R_{\mathbf{E},r}^2(\mathbf{Y}^*:\mathbf{X}) = \sum_{k=1}^{r} \sigma_k^{*2} R_k^{*2} \Big/ \sum_{k=1}^{r} \sigma_k^{*2}$$

(az előző tételben adott jelölések értelemszerű módosítása mellett).

Tételünk szerint a **G** érdeklődésmátrix ténylegesen az $y_k^*$ „idegen" komponensek felé egység-ekvivalens érdeklődést fejez ki. A formula (2.3) általános megfelelője.

*Bizonyítás.* **G** pozitív szemidefinit és $\mathbf{\Sigma}_{\mathbf{YY}}$ (ezzel együtt $\mathbf{\Sigma}_{\mathbf{YY}}^{-1}$) pozitív definit, szimmetrikus mátrixok, ezért létezik olyan nemszinguláris $\mathbf{F}(q\times q)$ és $\mathbf{D}^2(q\times q)$ nemnegatív elemekből képzett diagnális mátrix (l. [2], [17]), melyekkel

$$\mathbf{F}'\mathbf{G}\mathbf{F} = \mathbf{D}^2 \quad \text{és} \quad \mathbf{F}'\mathbf{\Sigma}_{\mathbf{YY}}^{-1}\mathbf{F} = \mathbf{E}_q. \tag{3.2}$$

Legyen **D** a $\mathbf{D}^2$ mátrix diagnális négyzetgyöke, továbbá

$$\mathbf{\Gamma} = \mathbf{F}'^{-1}\mathbf{D} \quad \text{és} \quad \tilde{\mathbf{Y}} = \mathbf{\Gamma}'\mathbf{Y}. \tag{3.3}$$

(3.2)-ből $\mathbf{\Sigma}_{\mathbf{YY}}^{-1}=\mathbf{F}'^{-1}\mathbf{F}^{-1}=(\mathbf{F}\mathbf{F}')^{-1}$, innen $\mathbf{\Sigma}_{\mathbf{YY}}=\mathbf{F}\mathbf{F}'$.

Felhasználva $\mathbf{\Gamma}$ és $\tilde{\mathbf{Y}}$ értelmezését:

$$\mathbf{\Sigma}_{\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}} = \mathbf{\Gamma}'\mathbf{\Sigma}_{\mathbf{YY}}\mathbf{\Gamma} = \mathbf{\Gamma}'\mathbf{F}\mathbf{F}'\mathbf{\Gamma} = \mathbf{D}'\mathbf{D} = \mathbf{D}^2 = \operatorname{diag}(\delta_1^2, \delta_2^2, \dots, \delta_q^2),$$

ebből $\tilde{\sigma}_k^2=\operatorname{var}\tilde{y}_k=\delta_k^2$ $(1\leqq k\leqq q)$ és $\operatorname{cov}(\tilde{y}_j, \tilde{y}_k)=0$ $j\neq k$-ra. (3.2)-ből (3.3) felhasználásával

$$\mathbf{G} = \mathbf{F}'^{-1}\mathbf{D}^2\mathbf{F}^{-1} = \mathbf{\Gamma}\mathbf{\Gamma}',$$

melyből (3.3) ismételt alkalmazásával kapjuk, hogy

$$\operatorname{tr}\mathbf{G}\mathbf{\Sigma}_{\mathbf{Y}:\mathbf{X}} = \operatorname{tr}\mathbf{\Gamma}'\mathbf{\Sigma}_{\mathbf{Y}:\mathbf{X}}\mathbf{\Gamma} = \operatorname{tr}\mathbf{\Sigma}_{\tilde{\mathbf{Y}}:\mathbf{X}} = \sum_{k=1}^{q} \tilde{\sigma}_k^2 \tilde{R}_k^2, \tag{3.4}$$

és

$$\operatorname{tr}\mathbf{G}\mathbf{\Sigma}_{\mathbf{YY}} = \operatorname{tr}\mathbf{\Gamma}'\mathbf{\Sigma}_{\mathbf{YY}}\mathbf{\Gamma} = \operatorname{tr}\mathbf{\Sigma}_{\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}} = \sum_{k=1}^{q} \tilde{\sigma}_k^2. \tag{3.5}$$

Feltehetjük, hogy $\delta_1^2\geqq\delta_2^2\geqq\dots\geqq\delta_q^2$, ekkor $\operatorname{rang}\mathbf{D}^2=\operatorname{rang}\mathbf{G}=r$ miatt $\tilde{\sigma}_k^2=\delta_k^2=0$ $r<k\leqq q$-ra. A megfelelő tagok (3.4) és (3.5) utolsó összegeiből elhagyhatók. Ezáltal az $\mathbf{Y}^*=(y_1^*, y_2^*, \dots, y_r^*)'=(\tilde{y}_1, \tilde{y}_2, \dots, \tilde{y}_r)'$ redukált v. vektornak megfelelő összegeket kapjuk. A $\sigma_k^*=\tilde{\sigma}_k$, $R_k^*=\tilde{R}_k$ $(k=1, 2, \dots, r)$ átjelölésekkel a tétel bizonyítása teljes.

Végül előállítjuk $R_{\mathbf{G}}^2(\mathbf{Y}:\mathbf{X})$-et a kanonikus korrelációkból is. A formula analóg (2.8)-cal.

3.7. TÉTEL. A G-irányított determináció és az X és Y között értelmezett $\varrho_k$ kanonikus korrelációk kapcsolata:

$$R^2_{\mathbf{G}}(\mathbf{Y}:\mathbf{X}) = \sum_{k=1}^{s} c_k^2 \varrho_k^2 \Big/ \sum_{k=1}^{q} c_k^2, \quad \text{ahol} \quad s = \min(p, q).$$

A $c_k^2$ súlyokat a bizonyítás közben adjuk meg. Az összegzés valamennyi kanonikus korrelációt tartalmazza.

*Bizonyítás.* Ismert [2], hogy a $\boldsymbol{\Sigma}_{\mathbf{Y}:\mathbf{X}}\mathbf{v}_k = \lambda_k \boldsymbol{\Sigma}_{\mathbf{YY}}\mathbf{v}_k$ feladat $\lambda_k$ megoldása $1 \leqq k \leqq s$-re $\varrho_k^2$-tel azonos, ahol $\varrho_k$ az X és Y között értelmezett $k$-adik kanonikus korreláció, ha a $\lambda_k$ sajátértékek nemnövekvő sorrendben következnek. A további $k$ indexekre $\lambda_k = 0$. A $\mathbf{v}_k$ sajátvektorok úgy választhatók, hogy $\mathbf{v}_j' \boldsymbol{\Sigma}_{\mathbf{YY}} \mathbf{v}_k = \delta_{jk}$ (a *Kronecker-féle delta*) teljesüljön. Mindezt a

$$\boldsymbol{\Sigma}_{\mathbf{Y}:\mathbf{X}} = \boldsymbol{\Sigma}_{\mathbf{YY}} \mathbf{V} \boldsymbol{\Lambda} \mathbf{V}^{-1} \quad \text{és} \quad \mathbf{V}' \boldsymbol{\Sigma}_{\mathbf{YY}} \mathbf{V} = \mathbf{E}_q \tag{3.6}$$

formában írhatjuk, ahol $\mathbf{V} = (\mathbf{v}_1, \mathbf{v}_2 \ldots, \mathbf{v}_q)$ és $\boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_q)$.

(3.6)-ból $\boldsymbol{\Sigma}_{\mathbf{Y}:\mathbf{X}} = \mathbf{V}'^{-1} \boldsymbol{\Lambda} \mathbf{V}^{-1}$ és $\boldsymbol{\Sigma}_{\mathbf{YY}} = \mathbf{V}'^{-1} \mathbf{V}^{-1}$. Legyen $\mathbf{G} = \boldsymbol{\Gamma}\boldsymbol{\Gamma}'$, ahol $\boldsymbol{\Gamma}(q \times r)$, $\operatorname{rang} \boldsymbol{\Gamma} = \operatorname{rang} \mathbf{G} = r$, és vezessük be az $\mathbf{L}' = \mathbf{V}^{-1} \boldsymbol{\Gamma}$ jelölést. Ekkor írhatjuk:

$$\operatorname{tr} \mathbf{G} \boldsymbol{\Sigma}_{\mathbf{Y}:\mathbf{X}} = \operatorname{tr} \boldsymbol{\Gamma}' \mathbf{V}'^{-1} \boldsymbol{\Lambda} \mathbf{V}^{-1} \boldsymbol{\Gamma} = \operatorname{tr} \mathbf{L} \boldsymbol{\Lambda} \mathbf{L}' = \operatorname{tr} \mathbf{L}' \mathbf{L} \boldsymbol{\Lambda} = \sum_{k=1}^{s} c_k^2 \varrho_k^2$$

és

$$\operatorname{tr} \mathbf{G} \boldsymbol{\Sigma}_{\mathbf{YY}} = \operatorname{tr} \boldsymbol{\Gamma}' \mathbf{V}'^{-1} \mathbf{V}^{-1} \boldsymbol{\Gamma} = \operatorname{tr} \mathbf{L} \mathbf{L}' = \operatorname{tr} \mathbf{L}' \mathbf{L} = \sum_{k=1}^{q} c_k^2,$$

ahol $c_k^2$ az $\mathbf{L}'\mathbf{L} = \mathbf{V}^{-1} \mathbf{G} \mathbf{V}'^{-1}$ mátrix $k$-adik diagonális eleme.

A két nyom hányadosaként a tételben adott formulát kapjuk.

## 4. $R^2_{\mathbf{G}}(\mathbf{Y}:\mathbf{X})$ kiterjesztése szinguláris esetre

Mindeddig az egyszerűség kedvéért föltettük, hogy $\boldsymbol{\Sigma}_{\mathbf{XX}}$ és $\boldsymbol{\Sigma}_{\mathbf{YY}}$ nemszingulárisak. Bár a gyakorlati alkalmazások során a becsült $\hat{\boldsymbol{\Sigma}}_{\mathbf{XX}}$ és $\hat{\boldsymbol{\Sigma}}_{\mathbf{YY}}$ kovarianciamátrixok ritkán szingulárisak [19], elméleti szempontból érdemes e megszorításoktól is eltekinteni.

Ha Y degenerált eloszlású, különösebb módosításra nincs szükség $R^2_{\mathbf{G}}(\mathbf{Y}:\mathbf{X})$ 2.1. definíciójában, csak azt kell előírni, hogy $\operatorname{tr} \mathbf{G} \boldsymbol{\Sigma}_{\mathbf{YY}} \neq 0$ legyen. Ha X degenerált eloszlású, a 2.1. definíció természetes kiterjesztéseként $\boldsymbol{\Sigma}_{\mathbf{XX}}$ közönséges inverzét a *Moore—Penrose-* vagy *pszeudoinverzével* ($\boldsymbol{\Sigma}^{+}_{\mathbf{XX}}$) helyettesítjük a redundancia-mátrix értelmezésében. A G-irányított determináció így kapott kiterjesztése:

$$R^2_{\mathbf{G}}(\mathbf{Y}:\mathbf{X}) = \frac{\operatorname{tr} \mathbf{G} \boldsymbol{\Sigma}_{\mathbf{Y}:\mathbf{X}}}{\operatorname{tr} \mathbf{G} \boldsymbol{\Sigma}_{\mathbf{YY}}}, \quad \text{ahol} \quad \boldsymbol{\Sigma}_{\mathbf{Y}:\mathbf{X}} = \boldsymbol{\Sigma}_{\mathbf{YX}} \boldsymbol{\Sigma}^{+}_{\mathbf{XX}} \boldsymbol{\Sigma}_{\mathbf{XY}}. \tag{4.1}$$

*Egyszerű példa:* természetes elvárás, hogy az $\mathbf{X} = \begin{pmatrix} x \\ x \end{pmatrix}$ predikátor egyenértékű legyen az $x$ predikátorral. Könnyű belátni, hogy (4.1) X-szel és $x$-szel ugyanazt az eredményt adja. E megjegyzés általánosabb megfogalmazását adjuk.

Értelmezünk egy nemdegenerált eloszlású $\mathbf{X}_0$ v. vektort, amellyel $\mathbf{X}$ helyettesíthető. Legyen $p_0 = \text{rang}\, \Sigma_{XX}$ és tekintsük $\Sigma_{XX}$ (redukált) spektrálfelbontását:

$$\Sigma_{XX} = \mathbf{AMA}', \quad \mathbf{A}'\mathbf{A} = \mathbf{E}_{p_0}, \quad \mathbf{M}(p_0 \times p_0) \text{ diagonális.} \tag{4.2}$$

($\mathbf{M}$ a $\Sigma_{XX}$ mátrix pozitív sajátértékeiből áll. $\mathbf{A}(p \times p_0)$ oszlopvektorai egy megfelelő ortonormált sajátvektor-rendszer). Definiáljuk az $\mathbf{X}_0$ v. vektort az

$$\mathbf{X} = \mathbf{AX}_0 \quad (\text{azaz } \mathbf{X}_0 = \mathbf{A}'\mathbf{X}) \tag{4.3}$$

leképezéssel. $\mathbf{X}_0$ $p_0$ komponensből áll és kovarianciamátrixa (4.2)-ből:

$$\Sigma_{X_0X_0} = \mathbf{M}, \quad \text{nemszinguláris.}$$

4.1. TÉTEL. $\mathbf{X}_0$ egyenértékű $\mathbf{X}$-szel, azaz

$$R^2_G(\mathbf{Y}:\mathbf{X}) = R^2_G(\mathbf{Y}:\mathbf{X}_0).$$

*Bizonyítás.* Megmutatjuk, hogy $\Sigma_{Y:X} = \Sigma_{Y:X_0}$. Felhasználjuk (l. [3], [13]), hogy a fenti $\mathbf{A}$ és $\mathbf{M}$ mátrixokkal

$$\mathbf{A}^+ = \mathbf{A}' \quad \text{és} \quad (\mathbf{AMA}')^+ = \mathbf{A}'^+\mathbf{M}^{-1}\mathbf{A}^+. \tag{4.4}$$

Tekintettel a (4.1—4) formulákra:

$$\Sigma_{Y:X} = \Sigma_{YX_0}\mathbf{A}'(\mathbf{AMA}')^+\mathbf{A}\Sigma_{X_0Y} = \Sigma_{YX_0}\mathbf{A}'\mathbf{A}'^+\mathbf{M}^{-1}\mathbf{A}^+\mathbf{A}\Sigma_{X_0Y} = \Sigma_{YX_0}\mathbf{M}^{-1}\Sigma_{X_0Y} = \Sigma_{Y:X_0}.$$

Hasonló eljárással küszöbölhető ki $\mathbf{Y}$ esetleges degenerált volta. Legyen $q_0 = \text{rang}\, \Sigma_{YY}$ és a (4.2)-nek megfelelő felbontás:

$$\Sigma_{YY} = \mathbf{BNB}', \quad \mathbf{B}'\mathbf{B} = \mathbf{E}_{q_0}, \quad \mathbf{N}(q_0 \times q_0) \text{ diagonális,}$$

ahol $\mathbf{N}$ $\Sigma_{YY}$ pozitív sajátértékeiből épül fel. Válasszuk az $\mathbf{Y}_0$ v. vektort és a $\mathbf{G}_0$ érdeklődés-mátrixot úgy, hogy

$$\mathbf{Y} = \mathbf{BY}_0 \quad (\mathbf{Y}_0 = \mathbf{B}'\mathbf{Y}) \quad \text{és} \quad \mathbf{G} = \mathbf{BG}_0\mathbf{B}' \quad (\mathbf{G}_0 = \mathbf{B}'\mathbf{GB}) \tag{4.5}$$

teljesüljenek. Ekkor $\Sigma_{Y_0Y_0} = \mathbf{N}$, nemszinguláris.

4.2. TÉTEL. Az $(\mathbf{Y}, \mathbf{G})$ pár helyettesíthető az $(\mathbf{Y}_0, \mathbf{G}_0)$ párral: $R^2_G(\mathbf{Y}:\mathbf{X}) = R^2_{G_0}(\mathbf{Y}_0:\mathbf{X})$.

*Bizonyítás.*

$$\text{tr}\, \mathbf{G}\Sigma_{Y:X} = \text{tr}\, \mathbf{BG}_0\mathbf{B}'\mathbf{B}\Sigma_{Y_0X}\mathbf{B}' = \text{tr}\, \mathbf{G}_0\Sigma_{Y_0:X}\mathbf{B}'\mathbf{B} = \text{tr}\, \mathbf{G}_0\Sigma_{Y_0:X}$$

és

$$\text{tr}\, \mathbf{G}\Sigma_{YY} = \text{tr}\, \mathbf{BG}_0\mathbf{B}'\mathbf{B}\Sigma_{Y_0Y_0}\mathbf{B}' = \text{tr}\, \mathbf{G}_0\Sigma_{Y_0Y_0}\mathbf{B}'\mathbf{B} = \text{tr}\, \mathbf{G}_0\Sigma_{Y_0Y_0}.$$

Ezekből a tétel következik.

Az utóbbi két tétel folyamánya, hogy a 3. pontban megfogalmazott tételek a lényeget nem érintő korrekciókkal a kiterjesztett $\mathbf{G}$-irányított determinációra is érvényben maradnak.

A továbbiakban néhány többváltozós diszciplinált vizsgálunk a $\mathbf{G}$-irányított determináció szemszögéből. $R^2_G(\mathbf{Y}:\mathbf{X})$-re megfogalmazott maximumfeladatok megoldásaként eljutunk a kanonikus komponensek, a főkomponensek, a faktoranalízis fak-

torainak G-irányított kiterjesztéséhez és a diszkriminanciaanalízishez. A kiterjesztés kapcsán interpretációs problémák merülnek fel. Egyszerűség kedvéért nemszinguláris $\Sigma_{XX}$ és $\Sigma_{YY}$ kovarianciamátrixokra szorítkozunk, továbbá kényelmi okokból **O** várható értékű **X** és **Y** v. vektorokra vonatkozik a megállapítások jórésze.

## 5. Kanonikus komponensek

Két természetes kérdés merül fel: a) adott **X** és **Y** mellett mely **G** érdeklődésre maximális a G-irányított determináció, b) adott **Y** és **G** mellett **X** alacsonyabb dimenziójú leképezettjei közül melyek predikálják legerősebben az **Y** v. vektort.

Könnyű megmutatni, hogy rang **G**=1-re $R^2_G(\mathbf{Y}:\mathbf{X})$ maximuma $\varrho_1^2$, ahol $\varrho_1$ a legnagyobb kanonikus korrelációs együttható (**X** és **Y** között), a megfelelő $\mathbf{v}_1'\mathbf{Y}$ kanonikus komponens $\mathbf{v}_1$ együtthatóvektorával $\mathbf{G}=\mathbf{v}_1\mathbf{v}_1'$ a maximumot adó érdeklődés-mátrix ($\mathbf{v}_1$ normája tetszőleges). Keresve $R^2_G(\mathbf{Y}:\mathbf{X})$ maximumát az $r$-edrangú **G** mátrixok fölött, az alábbi tételhez jutunk.

5.1. TÉTEL. Adott **X** és **Y** mellett

$$\sup_{\operatorname{rang}\mathbf{G}=r} R^2_G(\mathbf{Y}:\mathbf{X}) = \varrho_1^2.$$

*Bizonyítás.* Az $r$-edrangú **G** mátrix faktorizálható úgy, hogy $\mathbf{G}=\mathbf{\Gamma}\mathbf{\Gamma}'$ ($\mathbf{\Gamma}q\times r$-es) mellett $\mathbf{\Gamma}'\mathbf{\Sigma}_{YY}\mathbf{\Gamma}=\mathbf{D}^2$, diagonális legyen. Ehhez a 3.6. tétel bizonyításában konstruált $\mathbf{\Gamma}$ első oszlopából álló mátrixot választjuk $\mathbf{\Gamma}$-nak. $\mathbf{D}^2$ diagonális elemei **G**-től függenek, nem növekvő sorrendben $\delta_1^2, \delta_2^2, \ldots \delta_r^2 > 0$. $\mathbf{\Gamma}$ $k$-adik oszlopvektorát $\gamma_k$-val jelölve, rögzített $\mathbf{D}^2$ mellett $R^2_G(\mathbf{Y}:\mathbf{X})$ számlálójának maximumát keressük a $\gamma_k'\mathbf{\Sigma}_{YY}\gamma_k=\delta_k^2$ ($k=1, 2, \ldots, r$) feltételekkel. A $\gamma_j'\mathbf{\Sigma}_{YY}\gamma_k=0 (j\neq k)$ feltételektől eltekinthetünk. A feladat az

$$\mathcal{L} = \sum_{k=1}^{r} \gamma_k'\mathbf{\Sigma}_{Y:X}\gamma_k - \sum_{k=1}^{r} \lambda_k(\gamma_k'\mathbf{\Sigma}_{YY}\gamma_k - \delta_k^2)$$

*Lagrange-függvény* maximálása $\mathbf{\Gamma}=(\gamma_1, \gamma_2, \ldots, \gamma_r)$-ben, $\lambda_1, \lambda_2, \ldots, \lambda_r$ a feltételekhez kapcsolt multiplikátorok.

Deriválással és rendezéssel $\mathbf{\Gamma}$ oszlopvektoraira az

$$\mathbf{\Sigma}_{Y:X}\gamma_k = \lambda_k\mathbf{\Sigma}_{YY}\gamma_k \quad (1 \leqq k \leqq r) \tag{5.1}$$

kritériumokat kapjuk, amelyek a kanonikus korrelációk és komponensek alapegyenleteiként ismertek. Megoldásai: $\lambda_1, \lambda_2, \ldots, \lambda_q$ az általánosított sajátértékek, és $\gamma_1=\mathbf{v}_1, \gamma_2=\mathbf{v}_2, \ldots, \gamma_q=\mathbf{v}_q$ a megfelelő sajátvektorok. Utóbbiakra $\mathbf{v}_j'\mathbf{\Sigma}_{YY}\mathbf{v}_k=\delta_{jk}\delta_k^2$ ($\delta_{jk}$ a *Kronecker-féle delta*) is teljesül.

Tetszőlegesen kiválasztva $r(\lambda_k, \mathbf{v}_k)$ párt, (5.1) egy megoldásához jutunk. (5.1)-et balról $\gamma_k'$-vel szorozva és $k$-ra összegezve a kiválasztott párokkal $R^2_G(\mathbf{Y}:\mathbf{X}) = \Sigma\delta_k^2\lambda_k/\Sigma\delta_k^2$. Feltehetjük, hogy $\lambda_1$ a legnagyobb sajátérték, azaz $\lambda_1=\varrho_1^2$. A $\delta_k$ elemek tetszőlegesen választhatók, ezért $R^2_G(\mathbf{Y}:\mathbf{X})$ szuprémuma valóban $\varrho_1^2$.

A tétel bizonyításából kiolvasható, hogy a $\delta_1^2=\delta_2^2\ldots=\delta_r^2$ előírás mellett nyerjük maximumként az $r$ legnagyobb kanonikus korreláció négyzetének átlagát, ha emellett $q=p$, $G$-re $\mathbf{\Sigma}_{YY}^{-1}$-et kapjuk.

Az előző feladatban a $\mathbf{G}$ mátrixot a' posteriori értelmeztük. A probléma természetének azonban az érdeklődés a' priori választása felel meg. Legyen tehát $\mathbf{G}$ rögzített és keressük az $\mathbf{u}'\mathbf{X}$ lineáris kombinációt, amely az $\mathbf{Y}$ v. vektort maximálisan predikálja. Látni fogjuk, hogy ez a feladat nem az első kanonikus korrelációhoz és komponenshez vezet, hanem annak bizonyos általánosításához.

Feltehetjük, hogy $\operatorname{var}(\mathbf{u}'\mathbf{X}) = \mathbf{u}'\boldsymbol{\Sigma}_{\mathbf{XX}}\mathbf{u} = 1$. A $\lambda$ multiplikátort bevezetve az

$$\mathscr{L} = \operatorname{tr} \mathbf{G}\boldsymbol{\Sigma}_{\mathbf{Y}:\mathbf{u}'\mathbf{X}} - \lambda(\mathbf{u}'\boldsymbol{\Sigma}_{\mathbf{XX}}\mathbf{u} - 1)$$

függvény maximumát keressük. A feltétel és $\boldsymbol{\Sigma}_{\mathbf{Y}:\mathbf{X}}$ definíciója szerint $\boldsymbol{\Sigma}_{\mathbf{Y}:\mathbf{u}'\mathbf{X}} = \boldsymbol{\Sigma}_{\mathbf{YX}}\mathbf{u}\mathbf{u}'\boldsymbol{\Sigma}_{\mathbf{XY}}$, ebből kapjuk $\mathscr{L}$ alábbi alakját:

$$\mathscr{L} = \mathbf{u}'\boldsymbol{\Sigma}_{\mathbf{XY}}\mathbf{G}\boldsymbol{\Sigma}_{\mathbf{YX}}\mathbf{u} - \lambda(\mathbf{u}'\boldsymbol{\Sigma}_{\mathbf{XX}}\mathbf{u} - 1).$$

A $\mathbf{u}'$-szerinti deriválásával átrendezés után jutunk az

$$\boldsymbol{\Sigma}_{\mathbf{XY}}\mathbf{G}\boldsymbol{\Sigma}_{\mathbf{YX}}\mathbf{u} = \lambda\boldsymbol{\Sigma}_{\mathbf{XX}}\mathbf{u} \tag{5.2}$$

általánosított sajátérték feladathoz. Innen a szokásos módon következik az alábbi tétel.

5.2. TÉTEL. (5.2) legnagyobb $\lambda_{\mathbf{G}}$ sajátértékével és a hozzá tartozó bármely $\mathbf{u}_G$ sajátvektorral

$$\max_{(\mathbf{u})} R^2_{\mathbf{G}}(\mathbf{Y}:\mathbf{u}'\mathbf{X}) = R^2_{\mathbf{G}}(\mathbf{Y}:\mathbf{u}'_{\mathbf{G}}\mathbf{X}) = \lambda_{\mathbf{G}}/\operatorname{tr}\mathbf{G}\boldsymbol{\Sigma}_{\mathbf{YY}}.$$

Megfigyelhetjük, hogy $\lambda_{\mathbf{G}}$ általában nem azonos $\varrho_1^2$-tel és $\mathbf{u}'_{\mathbf{G}}\mathbf{X}$ is különbözik az $\mathbf{X}$-oldali első kanonikus komponenstől. Mindkettő a $\mathbf{G}$-érdeklődés függvénye, speciálisan a mértékegység megválasztásával szemben sem invariánsak. A $\mathbf{G} = \boldsymbol{\Sigma}_{\mathbf{YY}}^{-1}$ érdeklődés mellett (5.2)-ből láthatóan $\lambda_{\mathbf{G}} = \varrho_1^2$ és $\mathbf{u}'_{\mathbf{G}}\mathbf{X}$ valóban az első $\mathbf{X}$-oldali kanonikus komponens. Mindezt összefoglalva: (5.2) a kanonikus korrelációk és komponensek általánosított alapegyenletének tekinthető.

Legyen $s = \min(p, q, r)$, $p, q, r$ rendre $\boldsymbol{\Sigma}_{\mathbf{XX}}$ $\boldsymbol{\Sigma}_{\mathbf{YY}}$ és $\mathbf{G}$ rangja. (5.2) első $r$ sajátértéke és sajátvektora legyen $\lambda_{\mathbf{G}}^{(1)}, \lambda_{\mathbf{G}}^{(2)}, \ldots, \lambda_{\mathbf{G}}^{(s)}$ és $\mathbf{u}_{\mathbf{G}}^{(1)}, \mathbf{u}_{\mathbf{G}}^{(2)}, \ldots, \mathbf{u}_{\mathbf{G}}^{(s)}$, melyekre $\mathbf{u}_{\mathbf{G}}^{(i)}\boldsymbol{\Sigma}_{\mathbf{XX}}\mathbf{u}_{\mathbf{G}}^{(j)} = \delta_{ij}$, $1 \leqq i, j \leqq s$.

5.1. DEFINÍCIÓ. A $\varrho_{\mathbf{G}}^{(i)} = \sqrt{\lambda_{\mathbf{G}}^{(i)}}$ mennyiségeket $\mathbf{X}$ és $\mathbf{Y}$ $\mathbf{G}$ érdeklődésre vonatkozó kanonikus korrelációinak nevezzük, az $\mathbf{u}_{\mathbf{G}}^{(i)\prime}\mathbf{X}$ kombinációk pedig a $\mathbf{G}$ érdeklődésre vonatkozó kanonikus predikátor komponensek.

Az általánosított kanonikus komponensekkel képzett $\mathbf{G}$-irányított determinációk $\varrho_{\mathbf{G}}^{(i)2}/\operatorname{tr}\mathbf{G}\boldsymbol{\Sigma}_{\mathbf{YY}}$, $1 \leqq i \leqq s$.

Minthogy a komponensek páronként korrelálatlanok, a 3.2. tétel szerint több komponens együttes „információja" a megfelelő tagok összegével azonos. Az $s$ tag összege (5.2)-ből következően $R^2_{\mathbf{G}}(\mathbf{Y}:\mathbf{X})$. Megjegyezzük, hogy a $\operatorname{var}(\mathbf{u}_{\mathbf{G}}^{(i)\prime}\mathbf{X}) = 1$ előírásnál célszerűbb a $\operatorname{var}(\mathbf{u}_{\mathbf{G}}^{(i)\prime}\mathbf{X}) = \varrho_{\mathbf{G}}^{(i)2}$ normálás.

## 6. Főkomponensek

A főkomponens analízis viszonylag egyszerű technikája és a főkomponensek optimális tulajdonságai miatt [5], [6], [12], [15] gyakran alkalmazott elemző módszer. Itt is felmerül azonban a mértékegység-függőség problémája [1], [2], [10], [16]. Ehhez kapcsolódva a $\mathbf{G}$-irányított determináció tükrében vizsgáljuk a főkomponenseket.

Egyetlen v. vektort szeretnénk komponenseinek bizonyos kombinációival predikálni. Y-nal jelölve a v. vektort, adott G mellett a legjobb $\mathbf{b}'\mathbf{Y}$ predikátort keressük. $R^2_{\mathbf{G}}(\mathbf{Y}:\mathbf{b}'\mathbf{Y})$ nevezőjében b-nek nincs szerepe, a számláló pedig egyszerű meggondolással

$$\operatorname{tr} \mathbf{G}\boldsymbol{\Sigma}_{\mathbf{Y}:\mathbf{b}'\mathbf{Y}} = \mathbf{b}'\boldsymbol{\Sigma}_{\mathbf{YY}}\mathbf{G}\boldsymbol{\Sigma}_{\mathbf{YY}}\mathbf{b}/\mathbf{b}'\boldsymbol{\Sigma}_{\mathbf{YY}}\mathbf{b}. \tag{6.1}$$

Feltehetjük, hogy b $\boldsymbol{\Sigma}_{\mathbf{YY}}$-ra normált, ehhez a $\nu$ multiplikátort rendelve és (6.1) b fölötti maximumát keresve kapjuk a

$$\mathbf{G}\boldsymbol{\Sigma}_{\mathbf{YY}}\mathbf{b} = \nu\mathbf{b} \tag{6.2}$$

egyenletet. Ezt balról $\mathbf{b}'\boldsymbol{\Sigma}_{\mathbf{YY}}$-nal szorozva a (6.1) alatti mátrixnyomra $\nu$ adódik. Eszérint (6.1) maximuma $\mathbf{G}\boldsymbol{\Sigma}_{\mathbf{YY}}$ legnagyobb sajátértéke, jelöljük ezt $\nu_{\mathbf{G}}$-vel. b tetszőlegesen nyújtható, válasszunk $\nu_{\mathbf{G}}$-hez sajátvektort úgy, hogy $\operatorname{var}(\mathbf{b}'_{\mathbf{G}}\mathbf{Y}) = \mathbf{b}'_{\mathbf{G}}\boldsymbol{\Sigma}_{\mathbf{YY}}\mathbf{b}_{\mathbf{G}}$ pontosan $\nu_{\mathbf{G}}$ legyen. Eredményünket tételben mondjuk ki.

6.1. TÉTEL. Legyen $\nu_{\mathbf{G}}$ $\mathbf{G}\boldsymbol{\Sigma}_{\mathbf{YY}}$ legnagyobb sajátértéke és $\mathbf{b}_{\mathbf{G}}$ a hozzá tartozó sajátvektor, melyre $\mathbf{b}'_{\mathbf{G}}\boldsymbol{\Sigma}_{\mathbf{YY}}\mathbf{b}_{\mathbf{G}} = \nu_{\mathbf{G}}$. Fennáll:

$$\max_{(\mathbf{b})} R^2_{\mathbf{G}}(\mathbf{Y}:\mathbf{b}'\mathbf{Y}) = R^2_{\mathbf{G}}(\mathbf{Y}:\mathbf{b}'_{\mathbf{G}}\mathbf{Y}) = \nu_{\mathbf{G}}/\operatorname{tr}\mathbf{G}\boldsymbol{\Sigma}_{\mathbf{YY}} = \operatorname{var}(\mathbf{b}'_{\mathbf{G}}\mathbf{Y})/\operatorname{tr}\mathbf{G}\boldsymbol{\Sigma}_{\mathbf{YY}}.$$

$\mathbf{G} = \mathbf{E}_q$ mellett $\mathbf{b}'_{\mathbf{G}}\mathbf{Y}$ az Y v. vektoron értelmezett első főkomponens, különben annak általánosítása. A legtöbb számítógépes program a korrelációs mátrixot bontja főkomponensekre, ami a $\mathbf{G} = (\operatorname{diag}\boldsymbol{\Sigma}_{\mathbf{YY}})^{-1}$ érdeklődést tünteti ki. Mivel (6.2) megoldásai G-től nem függetlenek, célszerű bevezetni a főkomponensek alábbi általánosítását.

6.1. DEFINÍCIÓ. A $\mathbf{G}\boldsymbol{\Sigma}_{\mathbf{YY}}$ mátrix pozitív $\nu^{(i)}_{\mathbf{G}}$ sajátértékeihez tartozó jobb oldali $\mathbf{b}^{(i)}_{\mathbf{G}}$ sajátvektorokat válasszuk úgy, hogy $\mathbf{b}^{(i)\prime}_{\mathbf{G}}\boldsymbol{\Sigma}_{\mathbf{YY}}\mathbf{b}^{(j)}_{\mathbf{G}} = \delta_{ij}\nu^{(i)}_{\mathbf{G}}$ teljesüljön. Az $x^{(i)}_{\mathbf{G}} = \mathbf{b}^{(i)\prime}_{\mathbf{G}}\mathbf{Y}$ kombinációkat az Y v. vektor G-re vonatkoztatott főkomponenseinek nevezzük.

Az általánosított főkomponensek száma nyilvánvalóan $r = \operatorname{rang}\mathbf{G}$, páronként korrelálatlanok és

$$R^2_{\mathbf{G}}(\mathbf{Y}:x^{(i)}_{\mathbf{G}}) = \nu_i \Big/ \sum_{k=1}^{r} \nu_k, \quad (i = 1, 2, \ldots, r). \tag{6.3}$$

$i$-re összegezve 1-et kapunk (teljes determináció).

## 7. Hipotetikus faktorok

Az előző fejezetben az Y v. vektor lineáris kombinációiból indultunk ki. Most a faktoranalízis alapfeltevéséből kiindulva fogalmazzuk meg a G-irányított determinációra vonatkozó maximumfeladatot és enyhébb feltételek mellett az általánosított főkomponensekhez jutunk, szigorúbb feltételek előírása pedig a faktoranalízis faktorainak kiterjesztéséhez vezet.

Adott az Y $q$-dimenziós v. vektor és a $\mathbf{G}(q\times q)$ $r$-edrangú érdeklődés-mátrix. Feltételezünk egy $p$-dimenziós ($p \leqq r$) háttér v. vektort, melyet $\underset{\sim}{\mathbf{X}}$-mal jelölünk és amellyel Y elfogadhatóan predikálható. $R^2_{\mathbf{G}}(\mathbf{Y}:\underset{\sim}{\mathbf{X}})$ maximálásával keressük $\underset{\sim}{\mathbf{X}}$-ot.

A 3.1. tételre tekintettel feltehetjük, hogy a hipotetikus $\underset{\sim}{\mathbf{X}}$ v. vektor páronként korrelátlan, egységszórású komponensekből áll, ekkor $\boldsymbol{\Sigma}_{\underset{\sim}{X}\underset{\sim}{X}}=\boldsymbol{\Sigma}_{\underset{\sim}{X}\underset{\sim}{X}}^{-1}=\mathbf{E}_p$ és

$$\text{(7.1)} \qquad \operatorname{tr}\mathbf{G}\boldsymbol{\Sigma}_{\mathbf{Y}:\underset{\sim}{X}}=\operatorname{tr}\mathbf{G}\boldsymbol{\Sigma}_{\mathbf{Y}\underset{\sim}{X}}\boldsymbol{\Sigma}_{\underset{\sim}{X}\mathbf{Y}}=\operatorname{tr}\boldsymbol{\Sigma}_{\underset{\sim}{X}\mathbf{Y}}\mathbf{G}\boldsymbol{\Sigma}_{\mathbf{Y}\underset{\sim}{X}}=\sum_{i=1}^{p}\boldsymbol{\sigma}_{\underset{\sim}{x}_i\mathbf{Y}}\mathbf{G}\boldsymbol{\sigma}_{\mathbf{Y}\underset{\sim}{x}_i}$$

a **G**-irányított determináció számlálója. A nevezőben $\underset{\sim}{\mathbf{X}}$ nem szerepel. (7.1) $\underset{\sim}{\mathbf{X}}$ fölötti maximálásakor az $\underset{\sim}{x}_i$ komponensek korrelátlan voltát nem szükséges feltenni. A $\operatorname{var}\underset{\sim}{x}_i=1 (i=1, 2, \ldots, p)$ feltételekhez kapcsolva a $v_i$ multiplikátorokat, a feladat *Lagrange-függvénye*:

$$\text{(7.2)} \qquad \mathscr{L}=\sum_{i=1}^{p}\boldsymbol{\sigma}_{\underset{\sim}{x}_i\mathbf{Y}}\mathbf{G}\boldsymbol{\sigma}_{\mathbf{Y}\underset{\sim}{x}_i}-\sum_{i=1}^{p}v_i(\operatorname{var}\underset{\sim}{x}_i-1).$$

Felhasználva, hogy $\boldsymbol{\sigma}_{\mathbf{Y}\underset{\sim}{x}_i}$ $\underset{\sim}{x}_i$ szerinti deriváltja $\mathbf{Y}$, $\partial\mathscr{L}/\partial\underset{\sim}{x}_i=0$-ból kapjuk, hogy $\mathbf{Y}'\mathbf{G}\boldsymbol{\sigma}_{\mathbf{Y}\underset{\sim}{x}_i}=\underset{\sim}{x}_i v_i$ $(i=1, 2, \ldots, p)$, amit az $\mathbf{N}=\operatorname{diag}(v_1, v_2, \ldots, v_p)$ jelöléssel az alábbi módon írhatunk:

$$\text{(7.3)} \qquad \mathbf{Y}'\mathbf{G}\boldsymbol{\Sigma}_{\mathbf{Y}\underset{\sim}{X}}=\underset{\sim}{\mathbf{X}}'\mathbf{N}.$$

Bevezetve a $\boldsymbol{\Sigma}_{\mathbf{Y}\underset{\sim}{X}}=\mathbf{A}(q\times p)$ jelölést, határozzuk meg előbb az **A** mátrixot. Ismeretes, hogy **A** a faktorsúlyok mátrixa [5] a faktoranalízisben és a főkomponens súlyok mátrixa a főkomponens analízisben; meghatározása elsődleges feladat. Szorozzuk (7.3)-at balról **Y**-nal, illetve **X**-szel és képezzük a várható értékeket.

Mivel $\boldsymbol{\Sigma}_{\underset{\sim}{X}\underset{\sim}{X}}=\mathbf{E}_p$, nyerjük a

$$\text{(7.4)} \qquad \boldsymbol{\Sigma}_{\mathbf{YY}}\mathbf{GA}=\mathbf{AN} \quad \text{és} \quad \mathbf{A}'\mathbf{GA}=\mathbf{N}$$

összefüggéseket. Megjegyezzük, hogy (7.4)-et kielégítő $\mathbf{A}(q\times p)$ és $\mathbf{N}(p\times p)$ létezik, pl. a $\mathbf{Ga}=v\boldsymbol{\Sigma}_{\mathbf{YY}}^{-1}\mathbf{a}$ általánosított feladat $\boldsymbol{\Sigma}_{\mathbf{YY}}^{-1}$-re normált $\{\mathbf{a}_i\}$ megoldásrendszeréből kiválasztott tetszőleges $p$ sajátvektor megfelel **A** oszlopainak, **N**-nek pedig az ezekhez tartozó sajátértékek diagonális mátrixa. (7.4) alapján bizonyítjuk a következő tételt.

7.1. TÉTEL. $R_{\mathbf{G}}^2(\mathbf{Y}:\underset{\sim}{\mathbf{X}})$ maximális értékét azon a $p$-dimenziós hipotetikus $\underset{\sim}{\mathbf{X}}_{\mathbf{G}}$ v. vektoron (és $\underset{\sim}{\mathbf{X}}_{\mathbf{G}}$ reguláris leképzettjein) veszi fel, melynek komponensei a 6.1. definícióban bevezetett, a $p$ legnagyobb $v_i$ sajátértékhez tartozó általánosított főkomponensek. A maximumra: $R_{\mathbf{G}}^2(\mathbf{Y}:\underset{\sim}{\mathbf{X}}_{\mathbf{G}})=\sum_{i=1}^{p}v_{\mathbf{G}}^{(i)}/\sum_{k=1}^{r}v_{\mathbf{G}}^{(k)}$.

*Bizonyítás.* (7.1)-ből (7.4)-re tekintettel $\operatorname{tr}\mathbf{G}\boldsymbol{\Sigma}_{\mathbf{Y}:\underset{\sim}{X}}=\operatorname{tr}\mathbf{A}'\mathbf{GA}=\operatorname{tr}\mathbf{N}=\sum_{i=1}^{p}v_i$.

(7.4) szerint $v_i$ $\boldsymbol{\Sigma}_{\mathbf{YY}}\mathbf{G}$ (és egyben $\mathbf{G}\boldsymbol{\Sigma}_{\mathbf{YY}}$) sajátértéke, a maximális összeget tehát a legnagyobb $p$ sajátértékkel kapjuk, jelöljük ezeket $v_{\mathbf{G}}^{(1)}, v_{\mathbf{G}}^{(2)}, \ldots, v_{\mathbf{G}}^{(p)}$-vel. Ugyancsak (7.4) szerint **A** oszlopvektorai a $\mathbf{G}\boldsymbol{\Sigma}_{\mathbf{YY}}$ mátrix bal oldali sajátvektorai, továbbá **GA** oszlopvektorai $\mathbf{G}\boldsymbol{\Sigma}_{\mathbf{YY}}$ jobb oldali sajátvektorai $(\mathbf{G}\boldsymbol{\Sigma}_{\mathbf{YY}}(\mathbf{GA})=(\mathbf{GA})\mathbf{N})$. Legyen $\mathbf{N}_{\mathbf{G}}$ az első $p$ sajátértékből képzett diagonális mátrix, $\mathbf{A}_{\mathbf{G}}$ a megfelelő sajátvektorokból álló mátrix. Fejezzük ki (7.3)-ból $\underset{\sim}{\mathbf{X}}$-ot:

$$\text{(7.5)} \qquad \underset{\sim}{\mathbf{X}}_{\mathbf{G}}=\mathbf{N}_{\mathbf{G}}^{-1}\mathbf{A}_{\mathbf{G}}'\mathbf{GY},$$

azaz a hipotetikus változók **Y** lineáris függvényei.

Könnyen ellenőrizhető, hogy $\Sigma_{\underset{\sim}{X}_G \underset{\sim}{X}_G} = \mathbf{E}_p$. A (7.5) alatt adott $\underset{\sim}{X}_G$ a maximum feladatnak csak alapmegoldása, tetszőleges nemszinguláris $\mathbf{L}(p \times p)$-vel $\mathbf{L}'\underset{\sim}{X}_G$ egyenértékű $\underset{\sim}{X}_G$-vel. Ha $\mathbf{L}=\mathbf{T}$, ortogonális, $\underset{\sim}{X}=\mathbf{T}'\underset{\sim}{X}_G$ kovarianciamátrixa $\mathbf{E}_p$ marad, a főkomponens súlyok $\mathbf{A}$ mátrixa viszont megváltozik: $\mathbf{A}=\mathbf{A}_G\mathbf{T}$ (forgatás).

Legyen $\mathbf{L}=\mathbf{N}_G^{1/2}$ és $\mathbf{B}_G$ jelölje a $\mathbf{G}\mathbf{A}_G\mathbf{N}^{-1/2}$ mátrixot. $\mathbf{B}_G$ oszlopvektorai nyilvánvalóan a 6.1. definícióban adott $\mathbf{b}_G^{(i)}$ vektorok. Az

$$\mathbf{X}_G = \mathbf{L}'\mathbf{T}\underset{\sim}{X}_G = \mathbf{N}_G^{-1/2}\mathbf{A}_G'\mathbf{G}\mathbf{Y} = \mathbf{B}_G'\mathbf{Y}$$

v. vektor komponensei: $x_G^{(i)} = \mathbf{b}_G^{(i)\prime}\mathbf{Y}$ $(i=1, 2, \ldots, p)$ a $\mathbf{G}$-irányított főkomponensek. Ezzel a bizonyítás teljes.

Áttérünk a faktoranalízis modelljére. (7.1) szerint a kapott $\mathbf{A}_G$-vel $\operatorname{tr} \mathbf{G}\Sigma_{\mathbf{Y}:\underset{\sim}{X}} = \operatorname{tr} \mathbf{G}\mathbf{A}_G\mathbf{A}_G'$ maximális, azaz $\operatorname{tr} \mathbf{G}(\Sigma_{\mathbf{YY}} - \mathbf{A}_G\mathbf{A}_G')$ minimális. Tetszőleges $\underset{\sim}{X}=\mathbf{T}'\underset{\sim}{X}_G$ ($\mathbf{T}$ ortogonális) esetén $\mathbf{A}=\Sigma_{\mathbf{Y}\underset{\sim}{X}}=\Sigma_{\mathbf{Y}\underset{\sim}{X}_G}\mathbf{T}=\mathbf{A}_G\mathbf{T}$, amiből $\mathbf{A}\mathbf{A}'=\mathbf{A}_G\mathbf{A}_G'$.

A $\Sigma_{\mathbf{YY}} - \mathbf{A}\mathbf{A}'$ különbség az $\mathbf{Y}-\mathbf{A}\underset{\sim}{X}$ eltérésvektor kovarianci-mátrixa, amely eltérés a faktoranalízisben az egyedi faktorok $\mathbf{U}$ vektora, tehát

$$\Sigma_{\mathbf{UU}} = \Sigma_{\mathbf{YY}} - \mathbf{A}\mathbf{A}'. \tag{7.6}$$

A faktoranalízis modeljének alapvető feltevése $\Sigma_{\mathbf{UU}}$-ra vonatkozik: $\mathbf{U}$ komponensei egymással és a közös $(\underset{\sim}{X}_i)$ faktorokkal korrelálatlanok, [5], [10], [12]. Az utóbbi kitétel a fentebb képzett $\mathbf{A}$-val teljesül, hiszen $\Sigma_{\mathbf{U}\underset{\sim}{X}}=\Sigma_{\mathbf{Y}\underset{\sim}{X}}-\mathbf{A}\Sigma_{\underset{\sim}{X}\underset{\sim}{X}}=\mathbf{A}-\mathbf{A}=\mathbf{0}$.

Nem biztos viszont, hogy a (7.6) alatti különbség diagonális, még az sem biztos, hogy kovarianciamátrix (ti., hogy pozitív szemidefinit). A főkomponensanalízis és a faktoranalízis modelljének lényegi eltéréséről van szó.

Mindezek összefoglalásaként megfogalmazhatjuk a faktoranalízis $\mathbf{G}$-irányított alapmodelljét. Keressük azt az $\underset{\sim}{X}p(\leqq r)$ dimenziós hipotetikus v. vektort, amelynek kovarianciamátrixa $\mathbf{E}_p$ és amelyre

1) $R_G^2(\mathbf{Y}:\underset{\sim}{X})$ maximális, feltéve, hogy

2) az $\mathbf{A}=\Sigma_{\mathbf{Y}\underset{\sim}{X}}(q \times p)$ mátrixszal $\Sigma_{\mathbf{YY}} - \mathbf{A}\mathbf{A}'$ diagonális, nemnegatív diagonális elemekkel.

Mint láttuk, az 1) kitétel egyenértékű a $\operatorname{tr} \mathbf{G}(\Sigma_{\mathbf{YY}} - \mathbf{A}\mathbf{A}')$ nyom minimalizálásával.

A feladat nem minden $p$-re oldható meg, ha léteznek megoldások, azok az irodalomból ismert eljárások értelemszerű módosításával találhatók meg [5], [12]. Számunkra a megoldás mikéntjénél lényegesebb szempont az, hogy a $\mathbf{G}$ érdeklődésnek meghatározó szerepe van a faktoranalízisben is. A mértékegység megváltoztatása is (mint speciális $\mathbf{G}$) módosítja a faktorsúlyok $\mathbf{A}$ mátrixát, végeredményben az interpretációt.

Miután meghatároztuk a faktorsúlyok $\mathbf{A}$ mátrixát, felírható az $\mathbf{Y}$ v. vektor $\mathbf{G}$-irányított determinációja az egyes faktorokra, illetve a faktorok együttesére vonatkozóan. Felhasználva, hogy $\Sigma_{\underset{\sim}{X}\underset{\sim}{X}}=\mathbf{E}_p$ és $\Sigma_{\mathbf{Y}\underset{\sim}{x}_i}=\mathbf{a}_i$, az $\mathbf{A}$ mátrix $i$-edik oszlopvektora, a 3.1. definícióból kapjuk:

$$R_G^2(\mathbf{Y}:\underset{\sim}{x}_i) = \mathbf{a}_i'\mathbf{G}\mathbf{a}_i / \operatorname{tr} \mathbf{G}\Sigma_{\mathbf{YY}'} \quad (1 \leqq i \leqq p), \tag{7.7}$$

továbbá

$$R^2_G(\mathbf{Y}:\underset{\sim}{\mathbf{X}}) = \operatorname{tr} \mathbf{A}'\mathbf{G}\mathbf{A}/\operatorname{tr} \mathbf{G}\boldsymbol{\Sigma}_{\mathbf{YY}} = \sum_{i=1}^{p} R^2_G(\mathbf{Y}:\underset{\sim}{x}_i). \tag{7.8}$$

Megjegyezzük, hogy $\mathbf{G}$ tetszőleges $\boldsymbol{\Gamma}\boldsymbol{\Gamma}'$ faktorizációjával $\mathbf{a}'\mathbf{G}\mathbf{a}$ a $\boldsymbol{\Gamma}'\mathbf{A}$ mátrix $i$-edik oszlopában álló elemek négyzetösszege.

Fejezzük ki $\mathbf{Y}$ komponenseinek determinációs együtthatóit is $\underset{\sim}{\mathbf{X}}$-ra vonatkozóan:

$$R^2(\mathbf{y}_k:\underset{\sim}{\mathbf{X}}) = \mathbf{a}^{(k)'}\mathbf{a}^{(k)}/\operatorname{var} \mathbf{y}_k = h_k^2/\operatorname{var} \mathbf{y}_k, \quad 1 \leqq k \leqq q, \tag{7.9}$$

ahol $\mathbf{a}^{(k)'}$ az $\mathbf{A}$ mátrix $k$-adik sorvektora és $h_k^2 = \mathbf{a}^{(k)'}\mathbf{a}^{(k)}$, az $y_k$ változóhoz tartozó kommunalitás. Tisztább képet kapunk, ha az $\tilde{\mathbf{Y}} = \boldsymbol{\Gamma}'\mathbf{Y}$ v. vektor komponenseire vonatkozó kommunalitásokat állítjuk elő. $\tilde{\mathbf{Y}}\,\tilde{y}_k$ komponensére

$$\tilde{h}_k^2 = \gamma_k' \mathbf{A}\mathbf{A}'\gamma_k = (\operatorname{var} \tilde{y}_k) R^2(\underset{\sim}{y}_k:\underset{\sim}{\mathbf{X}}), \quad 1 \leqq k \leqq r, \tag{7.10}$$

ahol $\gamma_k$ a $\boldsymbol{\Gamma}(q \times r)$ mátrix $k$-adik oszlopvektora. Innen $\tilde{h}_k$ a $\boldsymbol{\Gamma}'\mathbf{A}$ mátrix $k$-adik sorában álló elemek négyzetösszege, $R^2_G(\mathbf{Y}:\underset{\sim}{\mathbf{X}})$ megfelelő előállítása pedig

$$R^2_G(\mathbf{Y}:\underset{\sim}{\mathbf{X}}) = R^2_{E_r}(\tilde{\mathbf{Y}}:\underset{\sim}{\mathbf{X}}) = \sum_{k=1}^{r} \tilde{h}_k^2 \Big/ \sum_{k=1}^{r} \operatorname{var} \tilde{y}_k, \tag{7.11}$$

a relatív $h_k^2$ kommunalitások súlyozott átlaga.

A fenti fejtegetés rámutat a $\mathbf{G}$-irányított faktoranalízis egy célszerű technikai megoldására: $\mathbf{G}$ faktorizációja után az $\tilde{\mathbf{Y}} = \boldsymbol{\Gamma}'\mathbf{Y}$ v. vektor faktorsúly-mátrixát állítjuk elő, ez $\tilde{\mathbf{A}}$, $\boldsymbol{\Gamma}'\mathbf{A}$-val azonos. $\mathbf{G}$ legjobban interpretálható faktorizációját a 3.6. tételben megfogalmazott módon kapjuk.

## 8. Lineáris diszkriminátorok

E pontban az [1], [4], [6], [10] és [11] munkákhoz kapcsolódunk. Legyenek $\mathbf{Z}$ és $\mathbf{E}$ $q$-dimenziós v. vektorok és $\mathbf{E}$ várható értéke $\mathbf{0}$. $\mathbf{Z}$ véges sok értékű, értékei egy-egy „osztályt" reprezentálnak, $\mathbf{E}$ komponensei az osztályokon belüli hibakomponensek. Feltesszük, hogy $\boldsymbol{\Sigma}_{\mathbf{ZE}} = \mathbf{0}$, továbbá bevezetjük a $\boldsymbol{\Sigma}_{\mathbf{ZZ}} = \mathbf{B}$, $\boldsymbol{\Sigma}_{\mathbf{EE}} = \mathbf{W}$ jelöléseket. A diszkriminancia analízis terminológiájával $\mathbf{B}$ az osztályok közötti *(between)*, $\mathbf{W}$ pedig az osztályokon belüli *(within)* kapcsolatok mátrixa. Feltételezzük még, hogy $\mathbf{W}$ és $\mathbf{B}+\mathbf{W}$ nemszingulárisak.

Elöljáróban emlékeztetünk a lineáris diszkriminancia analízis alapegyenletére [2], [11]. $\mathbf{B}$ és $\mathbf{W}$ statisztikai minták alapján becsülhetők és meghatározható az a $\mathbf{d}$ vektor, amelyre $\operatorname{var}(\mathbf{d}'\mathbf{Z})$ és $\operatorname{var}(\mathbf{d}'\mathbf{E})$ aránya maximális. E feladat a $\mathbf{d}'\mathbf{B}\mathbf{d}/\mathbf{d}'\mathbf{W}\mathbf{d}$ hányados maximalizálását jelenti, amely a $\mathbf{B}\mathbf{d} = \lambda\mathbf{W}\mathbf{d}$ egyenlethez vezet. Ennek legnagyobb $\lambda$ megoldása a hányados maximuma, a hozzá tartozó $\mathbf{d}$ (általánosított) sajátvektor pedig a keresett $\mathbf{d}$ vektor. A $\mathbf{d}'\mathbf{X} = \mathbf{d}'(\mathbf{Z}+\mathbf{E})$ kombinációt lineáris diszkriminátornak nevezik. $\mathbf{d}'\mathbf{X}$ alapján adott $\mathbf{X}$ mintáról bizonyos megbízhatósággal megállapítható, hogy melyik osztályhoz „tartozik".

A továbbiakban a $\mathbf{G}$-irányított determináció néhány maximum feladatát elemezzük. Az előzőektől eltérően ebben a pontban a $\mathbf{G}$-érdeklődésmátrixot rögzítjük: $\mathbf{G} = \mathbf{W}^{-1}$. Az így megfogalmazott maximum feladatok megoldásaiként a diszkrimi-

nancia analízis tárgyköréhez tartozó megállapításokra jutunk. A kifejtés során többször hivatkozunk az alábbi ismert segédtételre.

8.1. SEGÉDTÉTEL. Ha $\mathbf{Bv}=\lambda\mathbf{Wv}$, akkor $\mu=\lambda^2/(1+\lambda)$-val $\mathbf{BW}^{-1}\mathbf{Bv}=\mu(\mathbf{B}+\mathbf{W})\mathbf{v}$, azaz a két általánosított sajátérték-sajátvektor feladat sajátvektorai megegyeznek, a sajátértékek között pedig a felírt kapcsolat van.

Közvetlen következmény, hogy ha $\lambda_1 \geqq \lambda_2 \geqq \ldots \geqq \lambda_q$, akkor $\mu_1 \geqq \mu_2 \geqq \ldots \geqq \mu_q$, hiszen a $\lambda^2/(1+\lambda)$ függvény $\lambda$-ban monoton növekvő.

*Bizonyítás.* $\mathbf{Bv}=\lambda\mathbf{Wv}$-ből adódik, hogy

$$\mathbf{BW}^{-1}\mathbf{Bv} = \lambda\mathbf{Bv} = \lambda^2\mathbf{Wv}$$

és

$$\lambda\mathbf{BW}^{-1}\mathbf{Bv} = \lambda^2\mathbf{Bv}.$$

A két kapcsolat összeadásával

$$(1+\lambda)\mathbf{BW}^{-1}\mathbf{Bv} = \lambda^2(\mathbf{B}+\mathbf{W})\mathbf{v}$$

amiből $1+\lambda$-val átosztással a segédtétel állításához jutunk. Megjegyezzük, hogy a $\lambda$ és $\mu$ sajátértékek mind pozitívak.

Először tekintsük az $\mathbf{X}=\mathbf{Z}+\mathbf{E}$ predikátort szemben az $\mathbf{Y}=\mathbf{Z}$ predikátummal. Legyen $\mathbf{G}=\mathbf{W}^{-1}$ és határozzuk meg az $R^2_{\mathbf{G}}(\mathbf{Y}:\mathbf{X})$ G-irányított determinációt. Az alapmátrixok, tekintettel a $\mathbf{\Sigma}_{\mathbf{ZE}}=\mathbf{0}$ feltevésre:

$$(8.1)\quad \mathbf{\Sigma}_{\mathbf{XX}} = \mathbf{\Sigma}_{\mathbf{ZZ}}+\mathbf{\Sigma}_{\mathbf{EE}} = \mathbf{B}+\mathbf{W}, \quad \mathbf{\Sigma}_{\mathbf{YY}} = \mathbf{\Sigma}_{\mathbf{ZZ}} = \mathbf{B} \quad \text{és} \quad \mathbf{\Sigma}_{\mathbf{YX}} = \mathbf{\Sigma}_{\mathbf{ZX}} = \mathbf{\Sigma}_{\mathbf{ZZ}} = \mathbf{B}.$$

8.1. TÉTEL. $\mathbf{X}=\mathbf{Z}+\mathbf{E}$, $\mathbf{Y}=\mathbf{Z}$ és $\mathbf{G}=\mathbf{W}^{-1}$-re,

$$R^2_{\mathbf{I}} = R^2_{\mathbf{G}}(\mathbf{Y}:\mathbf{X}) = \sum_{k=1}^{q}\mu_k \Big/ \sum_{k=1}^{q}\lambda_k = \frac{\Sigma\lambda_k^2/(1+\lambda_k)}{\Sigma\lambda_k},$$

ahol mindegyik $\lambda_k$ a $|\mathbf{B}-\lambda\mathbf{W}|=0$ karakterisztikus egyenlet megoldása.

*Bizonyítás.* $R^2_{\mathbf{G}}(\mathbf{Y}:\mathbf{X})$ számlálója, felhasználva a segédtételt:

$$\operatorname{tr}\mathbf{W}^{-1}\mathbf{B}(\mathbf{B}+\mathbf{W})^{-1}\mathbf{B} = \operatorname{tr}(\mathbf{B}+\mathbf{W})^{-1}\mathbf{BW}^{-1}\mathbf{B} = \sum_{k=1}^{q}\mu_k.$$

A nevezőre hasonló módon:

$$\operatorname{tr}\mathbf{W}^{-1}\mathbf{B} = \sum_{k=1}^{q}\lambda_k.$$

Figyelembevéve $\mu_k$ és $\lambda_k$ kapcsolatát, a tételbeli $R^2_{\mathbf{I}}$ utolsó alakját nyerjük.

$R^2_{\mathbf{I}}$-et úgy foghatjuk fel, mint annak az információnak egy mérőszámát, amelyet $\mathbf{X}$ egyetlen $\hat{\mathbf{X}}$ mintája átlagosan tartalmaz $\mathbf{Y}$-ról, vagyis arról, hogy melyik osztályból származik az $\hat{\mathbf{X}}$ minta. Ha az $\hat{\mathbf{X}}$ minta valamely $\mathbf{d}'\hat{\mathbf{X}}$ egydimenziós függvénye alapján szeretnénk „diszkriminálni" az osztályok között, akkor természetszerűen a legtöbb információt tartalmazó (a legjobban predikáló) függvényt keressük.

8.2. TÉTEL. $R^2_{\mathbf{G}}(\mathbf{Y}:\mathbf{v}'\mathbf{X})$ maximumát az ismert lineáris diszkriminátoron veszi fel, a maximális érték pedig $\mu_1 \Big/ \sum_{k=1}^{q}\lambda_k$.

*Bizonyítás.* $R^2_G(\mathbf{Y}:\mathbf{v}'\mathbf{X})$ nevezője nem függ $\mathbf{v}'\mathbf{X}$-től, a számláló elemei pedig: $\mathbf{G}=\mathbf{W}^{-1}$, $\mathbf{\Sigma}_{\mathbf{Y}\mathbf{v}'\mathbf{X}}=\mathbf{B}\mathbf{v}$, $\mathbf{\Sigma}_{\mathbf{v}'\mathbf{X}\mathbf{v}'\mathbf{X}}=\mathbf{v}'(\mathbf{B}+\mathbf{W})\mathbf{v}$, innen a számláló:

$$\text{(8.2)} \qquad \operatorname{tr}\mathbf{W}^{-1}\mathbf{\Sigma}_{\mathbf{Y}:\mathbf{v}'\mathbf{X}} = \operatorname{tr}\mathbf{W}^{-1}\mathbf{B}\mathbf{v}\big(\mathbf{v}'(\mathbf{B}+\mathbf{W})\mathbf{v}\big)^{-1}\mathbf{v}'\mathbf{B} = \frac{\mathbf{v}'\mathbf{B}\mathbf{W}^{-1}\mathbf{B}\mathbf{v}}{\mathbf{v}'(\mathbf{B}+\mathbf{W})\mathbf{v}}.$$

Az utóbbi hányados $\mathbf{v}$ fölötti maximumát keresve a szokott módon a $\mathbf{B}\mathbf{W}^{-1}\mathbf{B}\mathbf{v}=$ $=\mu(\mathbf{B}+\mathbf{W})\mathbf{v}$ egyenlethez jutunk, ahol $\mu$ a hányados értéke. Ebből következik, hogy (8.2) maximuma $\mu_1$, a legnagyobb általánosított sajátérték, és $\mathbf{v}=\mathbf{v}_1$ a hozzá tartozó sajátvektor. A segédtétel szerint $\mathbf{v}_1$ egyben a $\mathbf{B}\mathbf{v}=\lambda\mathbf{W}\mathbf{v}$ egyenlet legnagyobb $(\lambda_1)$ sajátértékéhez tartozó sajátvektor, vagyis $\mathbf{v}_1=\mathbf{d}$, amellyel $\mathbf{d}'\mathbf{X}$ a lineáris diszkriminátor. $R^2_G(\mathbf{Y}:\mathbf{d}'\mathbf{X})$ nevezője $\sum_1^q \lambda_k$, ezzel a tételt bizonyítottuk.

$\mathbf{d}$ tetszőlegesen normálható. Ha $\mathbf{W}$-re normáljuk, akkor $\operatorname{var}(\mathbf{d}'\mathbf{Y})=\mathbf{d}'\mathbf{B}\mathbf{d}=\lambda_1$. A $\mathbf{B}\mathbf{v}=\lambda\mathbf{W}\mathbf{v}$ egyenlet minden megoldása egy-egy diszkriminátor. $\lambda_1$-hez a $\mathbf{d}_1'\mathbf{X}=\mathbf{d}'\mathbf{X}$, első diszkriminátor tartozik, $\lambda_2$-höz $\mathbf{d}_2'\mathbf{X}(\mathbf{d}_2=\mathbf{v}_2)$ a második diszkriminátor, és így tovább. A diszkriminátorok páronként korrelátlanok és $R^2_G(\mathbf{Y}:\mathbf{d}_i'\mathbf{X})=\mu_i/\Sigma\lambda_k$. Célszerű mindegyik $\mathbf{d}_i$ vektort $\mathbf{W}$-re normálni, így $\operatorname{var}(\mathbf{d}_i'\mathbf{X})=\lambda_i$ és a $\mathbf{G}$-irányított determinációk nevezője a diszkrimátorok összvarianciája.

Kimutatható, hogy $R^2_G(\mathbf{v}'\mathbf{Y}:\mathbf{d}'\mathbf{Y})$ $\mathbf{v}=\mathbf{d}$-vel maximális, a maximum értéke:

$$\text{(8.3)} \qquad \max_{(\mathbf{v})} R^2(\mathbf{v}'\mathbf{Y}:\mathbf{d}'\mathbf{x}) = R^2(\mathbf{d}'\mathbf{Y}:\mathbf{d}'\mathbf{x}) = \frac{\lambda_1}{1+\lambda_1} = r_1^2.$$

Ezzel analóg állítás érvényes a második, harmadik, stb. diszkriminátorokra is. (8.3)-ban az $r^2$ jelölés arra utal, hogy $\lambda_1/(1+\lambda_1)$ $\mathbf{d}'\mathbf{Y}$ és $\mathbf{d}'\mathbf{X}$ korrelációs együtthatójának négyzete.

Érdekes eredményeket kapunk, ha fordított szereposztásban elemezzük a fenti maximum feladatokat: legyen most

$$\mathbf{X} = \mathbf{Z} \quad \text{és} \quad \mathbf{Y} = \mathbf{Z}+\mathbf{E}.$$

$\mathbf{G}$-ként továbbra is $\mathbf{W}^{-1}$-et választjuk. A fordított szereposztásban

$$\text{(8.2)} \quad \mathbf{\Sigma}_{\mathbf{XX}} = \mathbf{\Sigma}_{\mathbf{ZZ}} = \mathbf{B}, \quad \mathbf{\Sigma}_{\mathbf{YY}} = \mathbf{\Sigma}_{\mathbf{ZZ}}+\mathbf{\Sigma}_{\mathbf{EE}} = \mathbf{B}+\mathbf{W} \quad \text{és} \quad \mathbf{\Sigma}_{\mathbf{YX}} = \mathbf{\Sigma}_{\mathbf{XY}} = \mathbf{\Sigma}_{\mathbf{ZZ}} = \mathbf{B}.$$

Feltéve, hogy $\mathbf{B}$ invertálható, $\mathbf{\Sigma}_{\mathbf{Y}:\mathbf{X}}=\mathbf{B}\mathbf{B}^{-1}\mathbf{B}=\mathbf{B}$. Néhány lépésben nyerjük a 8.1. tétel megfelelőjét: az $\mathbf{X}=\mathbf{Z}$, $\mathbf{Y}=\mathbf{Z}+\mathbf{E}$ és $\mathbf{G}=\mathbf{W}^{-1}$ szereposztással

$$\text{(8.4)} \qquad R^2_{II} = R^2_G(\mathbf{Y}:\mathbf{X}) = \frac{\operatorname{tr}\mathbf{W}^{-1}\mathbf{B}}{\operatorname{tr}\mathbf{W}^{-1}(\mathbf{B}+\mathbf{W})} = \frac{\sum_{i=1}^q \lambda_i}{\sum_{i=1}^q (1+\lambda_i)}.$$

$\mathbf{X}$ és $\mathbf{Y}$ értelmezésére tekintettel $R^2_{II}$ azt méri, hogy az adott osztályból vett $\hat{\mathbf{Y}}$ mintát átlagosan milyen mértékben determinálja maga az osztály. Hasonlítsuk össze $R^2_I$-et és $R^2_{II}$-et.

Az $r_k^2=\dfrac{\lambda_k}{1+\lambda_k}$ $(k=1, 2, \ldots, q)$ jelöléssel

$$R_{\mathrm{I}}^2=\frac{\sum_1^q \lambda_k r_k^2}{\sum_1^q \lambda_k} \quad \text{és} \quad R_{\mathrm{II}}^2=\frac{\sum_1^q (1+\lambda_k) r_k^2}{\sum_1^q (1+\lambda_k)},$$

mindkettő az $r_k^2$ korreláció-négyzetek súlyozott átlaga. Minthogy $r_k^2$ $\lambda_k$ monoton növő függvénye, heurisztikus meggondolás szerint $R_{\mathrm{I}}^2 \geqq R_{\mathrm{II}}^2$, hiszen $R_{\mathrm{I}}^2$-ben a nagyobb $r_k^2$ értékek nagyobb súllyal szerepelnek, mint $R_{\mathrm{II}}^2$-ben. Az egyenlőtlenség szabatos igazolása sem nehéz. Ez a reláció arra utal, hogy a minta alapján pontosabban behatárolható az ismeretlen osztály, amelyből a minta származik, mint fordítva.

Keresve azt a $\mathbf{v}'\mathbf{X}$-et, amely legjobban predikálja az $\mathbf{Y}$ v. vektort, megoldásként az $\mathbf{X}=\mathbf{Z}$ vektor $\mathbf{d}'\mathbf{X}$ függvényét kapjuk, mind fentebb. A $\mathbf{G}$-irányított determinációra pedig

$$\max_{(\mathbf{v})} R_{\mathbf{G}}^2(\mathbf{Y}:\mathbf{v}'\mathbf{X}) = R_{\mathbf{G}}^2(\mathbf{X}:\mathbf{d}'\mathbf{X}) = \frac{\lambda_1}{\sum_{k=1}^q (1+\lambda_k)} = \frac{(1+\lambda_1) r_1^2}{\sum_{k=1}^q (1+\lambda_k)}. \tag{8.5}$$

A maximumfeladat további lokális megoldásai $\mathbf{d}_2'\mathbf{X}$, $\mathbf{d}_3'\mathbf{X}$ stb., a megfelelő $\mathbf{G}$-irányított determinációk: $\lambda_i / \sum_{k=1}^q (1+\lambda_k)$, $1 \leqq i \leqq q$.

A lokális megoldások páronként korrelátlanok.

A $\max_{(\mathbf{u})} R_{\mathbf{G}}^2(\mathbf{u}'\mathbf{Y}:\mathbf{d}'\mathbf{X})$ feladat megoldása $\mathbf{u}=\mathbf{d}$, azaz $\mathbf{u}'\mathbf{Y}=\mathbf{d}'\mathbf{Y}$, az első diszkriminátor, a maximum természetesen $r_1^2=\lambda_1/(1+\lambda_1)$, megegyezően (8.3)-mal. Ebben az értelemben a diszkriminátor, $\mathbf{d}'(\mathbf{Z}+\mathbf{E})$ és a „diszkriminátum", $\mathbf{d}'\mathbf{Z}$ szimmetrikusan viselkednek.

## 9. Záró megjegyzések

1. A $\mathbf{G}$-irányított determinációs együttható gyakorlati alkalmazása alapvetően az érdeklődésmátrix adequat megválasztását feltételezi. Az alkalmazásnak talán ez a legnagyobb körültekintést igénylő szakasza, amelynek vizsgálatára ebben a dolgozatban nem vállalkoztunk. A probléma elsősorban szakmai (nem matematikai) természetű. A legegyszerűbb esetről a 2. pontban szóltunk, amikoris az $\mathbf{Y}$ v. vektor egyes komponenseire szakmailag egyenértékű $\Delta_1, \Delta_2, \ldots, \Delta_q$ megváltozások előzetes ismeretét feltételeztük. Ilyenkor a $\mathbf{G}$ érdeklődésmátrix diagonális, $\Delta_k^{-2}$ $(1 \leqq k \leqq q)$ elemekkel, a célszerű technika pedig a $\boldsymbol{\Gamma}=\mathrm{diag}\,(\Delta_1^{-1}, \Delta_2^{-1}, \ldots, \Delta_q^{-1})$ mátrixszal képzett $\tilde{\mathbf{Y}}=\boldsymbol{\Gamma}\mathbf{Y}$ v. vektor vizsgálata $\mathbf{E}_q$ érdeklődésmátrix mellett. Bizonyos esetekben előzetes szakmai tapasztalatok alapján rögzíthetők a $\Delta_k$ érdeklődésekvivalens értékek. Egy mezőgazdasági példával élve, tegyük fel, hogy $\mathbf{Y}$ komponensei valamely növény bizonyos beltartalmi paramétereit jelentik. Számos vizsgálat alapján megállapíthatók azok az eltérés-küszöbértékek, amelyeken belül az egyes paraméterek még a termés minőségének lényeges változása nélkül mozoghatnak. Ezek a küszöbértékek beltartalmi paraméterenként eltérőek és választhatók $\Delta_k$-ként. Az általánosabb esetre vonatkozóan — amikoris $\mathbf{G}$ nem diagonális — a diszkriminancia analízisnél alkalmazott

érdeklődésmátrix ($\mathbf{W}^{-1}$) tartalma tekinthető a $\mathbf{G}$ mátrix megválasztásának irányelvéül. Egyidejűleg tekintetbe vehetők méréstechnikai nehézségek, költségek, illetve fontossági sorrend is.

2. A dolgozatban csak a legfontosabbnak ítélt tételeket fogalmaztuk meg. Különösen a 4. pontban tárgyalt szinguláris eset teljesebb vizsgálata lehetséges TUSNÁDY G. (20) dolgozatára támaszkodva. Az érdeklődésirányított determinációs együttható további alkalmazási lehetőségét illetően csak egy témakörre utalok: kvalitatív változók „skálázása" is visszavezethető a $R^2_{\mathbf{G}}$-re megfogalmazott maximumfeladatra. E kérdéssel egy későbbi cikkben kívánok foglalkozni.

## IRODALOM


[1] ANDERBERG, M. R., *Cluster Analysis for Applications* (Academic Press, New York, London, 1973).
[2] ANDERSON, T. W., *An Introduction to Multivariate Statistical Analysis* (Wiley, New York, 1958).
[3] EGERVÁRY, J., „Az inverz mátrix általánosítása", *A Matematikai Kutató Intézet Közleményei* **3** (1956) 315—324.
[4] GUPTA, R. D. and GUPTA, R. P., "Adequacy of discriminant functions and their Applications", in *Multivariate Statistical Inference Ed. D. G. Kabe and R. P. Gupta* (Elsevier Publishing Company, Amsterdam—New York, 1973), 109—128.
[5] HARMAN, H. H., *Modern Factor Analysis* (The Univ. of Chicago Press, Chicago and London, 1967).
[6] HARRIS, R. H., *A Primer of Multivariate Statistics* (Academic Press, New York—London, 1975).
[7] JAMES, A. T., "The Variance Information Manifold and the functions on it", in *Multivariate Analysis—III. Ed. P. R. Krishnaiah* (Academic Press, New York—London, 1973) 157—170.
[8] JÓZSA, S., "A method for seeking the most informative characters", *Acta Agronomica Ac. Sci. Hung.* **21** (1972) 335—344.
[9] JÓZSA, S., "A multivariate coefficient of determination", in: *Proceedings of the 3rd Hung. Biometric Conference*, Bp., (1981) 291—294.
[10] KENDALL, M. G., *A Course in Multivariate Analysis* (Hafner Publ. Co., New York, 1968).
[11] KRZYSKO, M., "Discriminant variables", *Biometrical Journal* **21** (1979) 227—242.
[12] LAWLEY, D. N. and MAXWELL, A. E., *Factor Analysis as a Statistical Method* (Buttterworths, London, 1971).
[13] LEE, A., „Mátrixok általánosított inverzeiről", *SZIGMA* **6** (1973) 127—144.
[14] LENGYEL, T., „A kanonikus korrelációanalízis és néhány kapcsolódó probléma", *Alk. Mat. Lapok* **5** (1979) 385—393.
[15] OKAMOTO, M., "Optimality of Principal Components", in: *Multivariate Analysis—II. Ed. P. R. Krishnaiah* (Academic Press, New York—London, 1969) 673—685.
[16] RAO, C. R., "The use and interpretation of PCA in applied research", *Sankhya Ser. A* **26** (1964) 329—358.
[17] RÓZSA, P., *Lineáris algebra és alkalmazásai* (Műszaki Könyvkiadó, Bp., 1974).
[18] STEWART, D. K. and LOVE, W. A., "A general correlation index", *Psychological Bulletin* **70** (1968) 160—163.
[19] SVÁB, J., *Többváltozós módszerek a biometriában* (Mezőgazdasági Kiadó, Bp., 1979).
[20] TUSNÁDY, G., „Mátrixok szinguláris felbontása", *Alkalmazott Matematikai Lapok* **5** (1979) 375—384.


(Beérkezett: 1983. március 15.)

JÓZSA SÁNDOR
AGRÁRTUDOMÁNYI EGYETEM
8361 KESZTHELY, DEÁK F. U. 16.

# A BIMULTIVARIATE INTEREST-ORIENTATED COEFFICIENT OF DETERMINATION

S. JÓZSA

Some of the multivariate techniques are not invariant under linear transformation of the set of variables. As a special case, even alterations in the scale have an influence on the numerical results. This fact leads to confusion in the interpretation — as it has been pointed out by many authors.

In order to get rid of this confusion we introduce a bimultivariate "interest-orientated" correlation measurement. It gives a measure of information contained in the predictor vector variable (**X**) on the criterion vector variable (**Y**) along an arbitrary interest defined on **Y** and given in form of matrix (**G**). **X** and **Y** may be degenerate as well.

While discussing multivariate techniques on the basis of the interest-orientated measurement, some of them will be modified and some aspects of interpretation cleared according to the actual interest. How to choose the matrix of interest — it is an essential practical question hardly belonging to the field of statistical methodology.

*Alkalmazott Matematikai Lapok* **10** (1984) 35—60

# ÚJ ALGORITMUS A TÖBBDIMENZIÓS GAMMA ELOSZLÁS EMPIRIKUS ADATOKHOZ TÖRTÉNŐ ILLESZTÉSÉRE

SZÁNTAI TAMÁS

Budapest

A dolgozatban szükséges feltételeket adunk meg az empirikus kovarianciamátrix elemeire ahhoz, hogy létezzen az adatokhoz pontosan illeszkedő többdimenziós gamma eloszlás. Ezen szükséges feltételek segítségével egy, az eddig ismerteknél lényegesen hatékonyabb, heurisztikus algoritmust adunk a pontosan illeszkedő többdimenziós gamma eloszlás meghatározására.

A kidolgozott algoritmus hatékonyságát számítógépes futási eredmények és gépidő adatok közlésével szemléltetjük.

Külön fejezetben, illetve a dolgozat függelékében foglalkozunk a szükséges feltételek elégséges volta igazolásának a problémájával. A bizonyítást eddig csak a 2, 3 és 4 dimenziós gamma eloszlás esetére sikerült elvégezni.

## 1. Bevezetés

A [3] dolgozatban a szerzők három különböző módszert javasoltak a bevezetett többdimenziós gamma eloszlás empirikus adatokhoz történő illesztésére. Ezek az eltérések abszolút értékei összegének a minimalizálásán; az eltérések négyzetösszegének a minimalizálásán; illetve a maximális abszolút eltérés minimalizálásán alapulnak. Mindegyik módszer közös vonása az, hogy az illesztés végrehajtásához igen nagyméretű lineáris programozási feladatot kell megoldani. Ugyanez igaz akkor is, ha az illeszkedés feltételrendszerének egy megengedett megoldását a szimplex algoritmus első fázisának az alkalmazásával keressük meg.

Ebben a dolgozatban szükséges feltételeket adunk meg az empirikus kovarianciamátrix elemeire ahhoz, hogy létezzen az adatokhoz pontosan illeszkedő többdimenziós gamma eloszlás. Ezen szükséges feltételek segítségével egy, az eddig ismerteknél lényegesen hatékonyabb, heurisztikus algoritmust tudunk adni a pontosan illeszkedő többdimenziós gamma eloszlás meghatározására. Az algoritmus alapgondolata az, hogy úgy építjük fel lépésről lépésre a pontosan illeszkedő többdimenziós gamma eloszlást, hogy a fennmaradó rész kovarianciamátrixa mindig eleget tegyen a rá vonatkozó szükséges feltételeknek, vagy legalábbis azok közül a leglényegesebbeknek. Minthogy az egyes lépések során mindig csak az általunk eddig ismert szükséges feltételek — sőt azok közül sem feltétlen mindegyik — teljesülését követeljük meg a fennmaradó rész kovarianciamátrixára vonatkozóan, semmi sem garantálja, hogy az algoritmus nem akad el úgy, hogy a szükséges feltételek teljesülnek ugyan, de mégsem lehet a fennmaradó részt pontosan előállítani. Ezért az így felépített illesztő algoritmust mindaddig heurisztikusnak kell tekintenünk, amíg be nem bizonyítjuk, hogy az illeszthetőség általunk ismert szükséges feltételei elégségesek is. Ezt jelenleg csak a 2, 3 és 4 dimenziós esetre tudjuk bizonyítani, magasabb dimenziószámokra csak sejtjük, hogy igaz a feltételek elégségessége. Külön fejezetben foglalkozunk az elégségesség igazolásának a problémájával.

Megjegyezzük, hogy az ismertetésre kerülő heurisztikus illesztő algoritmus — a szükséges feltételek igen nagy száma miatt — akkor tud csak igazán hatékony lenni, ha az ismert szükséges feltételeknek csak egy szűkített halmazára támaszkodva építjük fel. A számítási tapasztalatok azt mutatják, hogy az így felépített illesztő algoritmus a feladatok nagy százalékára igen gyorsan és eredményesen adja meg a keresett többdimenziós gamma eloszlás előállítását. Ez a számítási tapasztalat azt mutatja, hogy a szükséges feltételek elégségességének a bizonyítása inkább csak elméleti, semmint gyakorlati jelentőségű.

Köszönettel tartozom KÉRI GERZSONnak, aki észrevételeivel jelentősen bővítette az általam előzőleg ismert szükséges feltételek számát.

## 2. A többdimenziós gamma eloszlás illeszthetőségének szükséges feltételei

A [3] dolgozatban a szerzők bevezették az

$$\mathbf{A}\boldsymbol{\eta}$$

valószínűségi vektorváltozót, mely többdimenziós gamma eloszlású, ha az $\mathbf{A}$ mátrix 0, 1 elemekből áll és az $\boldsymbol{\eta}$ valószínűségi vektorváltozó $\eta_1, \ldots, \eta_n$ komponensei függetlenek és standard gamma eloszlásúak. Az így bevezetett többdimenziós gamma eloszlás empirikus adatokhoz történő illesztésének a feladata abban áll, hogy olyan — előbb említett tulajdonságú — $\mathbf{A}$ mátrixot és $\boldsymbol{\eta}$ valószínűségi vektorváltozót kell keresnünk, hogy $\mathbf{A}\boldsymbol{\eta}$ komponenseinek az eloszlása megegyezzen a többdimenziós statisztikai sokaság komponenseinek a feltételezett gamma eloszlásával, továbbá, hogy $\mathbf{A}\boldsymbol{\eta}$ kovariancia mátrixa legyen egyenlő $\mathbf{C}$-vel, az adatok empirikus kovariancia mátrixával. Minthogy nyilvánvalóan akkor van a legnagyobb esélyünk a fent leírt követelmények teljesítésére, ha az $\mathbf{A}$ mátrix a lehető legtöbb oszlopból áll, azért az $\mathbf{A}$ mátrix oszlopainak a számát $2^n-1$-ben rögzíthetjük, hiszen az $n$-méretű, 0, 1 komponensű különböző vektorok száma $2^n$, ám a zéró vektort figyelmen kívül lehet hagyni.

Az így bevezetett többdimenziós gamma eloszlás illesztésének a feladata tehát abból áll, hogy az $\boldsymbol{\eta}$ valószínűségi vektorváltozó $E(\boldsymbol{\eta})=\boldsymbol{\vartheta}$ paramétervektorát úgy kell megválasztani, hogy a

$$\begin{aligned} \tilde{\mathbf{A}}\boldsymbol{\vartheta} &= \mathbf{c} \\ \boldsymbol{\vartheta} &\geqq \mathbf{0} \end{aligned} \tag{2.1}$$

feltételek teljesüljenek, ahol $\mathbf{c}$ a $\mathbf{C}$ kovariancia mátrix $c_{11}, \ldots, c_{nn}, c_{12}, \ldots, c_{1n}, c_{23}, \ldots, c_{2n}, \ldots, c_{n-1,n}$ elemeiből, mint egymást ebben a sorrendben követő komponensekből alkotott $\frac{1}{2}n(n+1)$ méretű vektor, $\tilde{\mathbf{A}}$ pedig egy $\frac{1}{2}n(n+1)\times(2^n-1)$ méretű mátrix. Az utóbbit úgy származtatjuk, hogy az első $n$ sorába leírjuk a 0 és 1 elemekből képezhető $n$-méretű vektorokat (eltekintve a 0 vektortól), majd az egyes sorok elemenkénti szorzatai révén alkotunk új sorokat; előbb az első sort szorozzuk a másodikkal, ..., $n$-edikkel, majd a második sort szorozzuk a harmadikkal, ..., $n$-edikkel, ..., végül az $(n-1)$-edik sort szorozzuk az $n$-edikkel.

A (2.1) feltételrendszer a következő alakban is felírható:

$$\sum_{l=1}^{p} \mathbf{a}_l \mathbf{a}_l' \vartheta_l = \mathbf{C}$$

$$\vartheta_l \geqq 0, \quad l = 1, \ldots, p, \tag{2.2}$$

ahol $p = 2^n - 1$ és az $\mathbf{a}_l \in R^n$, $l = 1, \ldots, p$ vektorok az összes olyan $R^n$-beli nem nulla vektort jelentik, amely koordinátái mind nullával, vagy eggyel egyenlők.

Megjegyezzük, hogy a $\mathbf{C}$ mátrix és az $\mathbf{a}_l \mathbf{a}_l'$, $l = 1, \ldots, p$ diádok szimmetrikussága miatt a (2.2) feltételrendszerben szereplő mátrix egyenlőség alsó és felső háromszög része ugyanazokat a feltételeket szolgáltatja. Elég ezért azok közül csak például a felső háromszög részhez tartozó $\frac{1}{2} n(n+1)$ darab feltételt tekinteni, melyek így valóban teljesen azonosak lesznek a (2.1) feltételrendszer egyenlőségeivel.

Az $n = 3$ esetben például az $\mathbf{a}_1, \ldots, \mathbf{a}_7$ vektorok rendre a következők:

$$\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}.$$

Az ezeknek megfelő diádok:

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 1 \end{pmatrix},$$

$$\begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}.$$

A (2.2) feltételrendszer pedig (csak a felső háromszög részhez tartozóan):

$$\begin{aligned}
\vartheta_1 \quad + \vartheta_4 + \vartheta_5 \quad + \vartheta_7 &= c_{11} \\
\vartheta_2 \quad + \vartheta_4 \quad + \vartheta_6 + \vartheta_7 &= c_{22} \\
\vartheta_3 \quad + \vartheta_5 + \vartheta_6 + \vartheta_7 &= c_{33} \\
\vartheta_4 \quad + \vartheta_7 &= c_{12} \\
\vartheta_5 \quad + \vartheta_7 &= c_{13} \\
\vartheta_6 + \vartheta_7 &= c_{23} \\
\vartheta_1 \geqq 0, \vartheta_2 \geqq 0, \vartheta_3 \geqq 0, \vartheta_4 \geqq 0, \vartheta_5 \geqq 0, \vartheta_6 \geqq 0, \vartheta_7 \geqq 0.
\end{aligned}$$

Ahhoz, hogy létezzenek a (2.2) feltételrendszernek eleget tevő $\vartheta_1, \ldots, \vartheta_p$ paraméterek, a $\mathbf{C}$ empirikus kovariancia mátrixnak a következő tétel feltételeit kell kielégíteni:

2.1. TÉTEL. Ha egy $\mathbf{C}$ empirikus kovariancia mátrixra létezik a (2.2) feltételrendszernek eleget tevő $\vartheta_1, \ldots, \vartheta_p$ paraméter halmaz, akkor $\mathbf{C}$ elemeire teljesülni kell a

$$\sum_{i\in I_1} c_{ii} - \sum_{i\in I_1, k\in I_2} c_{ik} + \sum_{\substack{i\in I_1, k\in I_1 \\ i<k}} c_{ik} + \sum_{\substack{i\in I_2, k\in I_2 \\ i<k}} c_{ik} \geqq 0 \tag{2.3}$$

feltételeknek, ahol $I_1, I_2 \subset I = \{1, 2, \ldots, n\}$ és $I_1 \cap I_2 = \emptyset$.

*Bizonyítás.* Legyen $\mathbf{a}_i' = (a_{1i}, \ldots, a_{ni})$, és írjuk a (2.2) feltételrendszert a következő (elemenkénti) alakba:

$$\begin{gathered} \sum_{l=1}^{p} a_{il} a_{kl} \vartheta_l = c_{ik}, \quad i = 1, \ldots, n; \quad k = 1, \ldots, n \\ \vartheta_l \geqq 0, \quad (l = 1, \ldots, p. \end{gathered} \tag{2.4}$$

Helyettesítsük be a $c_{ik}$ mátrixelemek (2.4) alatti előállítását a (2.3) egyenlőtlenségekbe. Rendezés után a következő egyenlőtlenségeket kapjuk:

$$\begin{aligned} 0 \leqq & \sum_{i\in I_1} c_{ii} - \sum_{i\in I_1, k\in I_2} c_{ik} + \sum_{\substack{i\in I_1, k\in I_1 \\ i<k}} c_{ik} + \sum_{\substack{i\in I_2, i\in I_2 \\ i<k}} c_{ik} = \\ & = \sum_{l=1}^{p} \Big( \sum_{i\in I_1} a_{il}^2 - \sum_{i\in I_1, k\in I_2} a_{il} a_{kl} + \sum_{\substack{i\in I_1, k\in I_1 \\ i<k}} a_{il} a_{kl} + \sum_{\substack{i\in I_2, k\in I_2 \\ i<k}} a_{il} a_{kl} \Big) \vartheta_l, \end{aligned} \tag{2.5}$$

ahol tudjuk, hogy $\vartheta_l \geqq 0$, $l = 1, \ldots, p$. A tétel állításának igazolásához elég azt megmutatni, hogy a (2.5) egyenlőtlenségekben $\vartheta_l$ együtthatója minden $l$-re nemnegatív. Jelölje

$$\begin{aligned} I_1(l) &= \{i:\ i \in I_1 \quad \text{és} \quad a_{il} = 1\} \\ I_2(l) &= \{i:\ i \in I_2 \quad \text{és} \quad a_{il} = 1\}, \end{aligned} \qquad l = 1, \ldots, p,$$

valamint legyen $n_1(l)$ az $I_1(l)$ és $n_2(l)$ az $I_2(l)$ halmaz elemeinek a száma. Ekkor $\vartheta_l$ együtthatójának az értéke a (2.5) egyenlőtlenségekben:

$$\begin{aligned} & n_1(l) - n_1(l) n_2(l) + \binom{n_1(l)}{2} + \binom{n_2(l)}{2} = \\ & = \frac{1}{2} [2n_1(l) - 2n_1(l) n_2(l) + n_1(l)(n_1(l) - 1) + n_2(l)(n_2(l) - 1)] = \\ & = \frac{1}{2} [n_1(l) - 2n_1(l) n_2(l) + n_1^2(l) + n_2^2(l) - n_2(l)] = \\ & = \frac{1}{2} [(n_1(l) - n_2(l))(n_1(l) - n_2(l) + 1)]. \end{aligned}$$

Ez az érték pedig biztosan nemnegatív, mert $n_1(l) - n_2(l)$ csak egész értéket vehet fel.

A 2.1. tételben kimondott szükséges feltételek szemléltetésére tekintsük ismét az $n = 3$ esetet. Az $I_1$ és $I_2$ indexhalmazok elemszámaira az $n_1$ és $n_2$ jelölést használva a (2.3) feltételek a következők lesznek:

a) $n_1=0,\ n_2=0$ esetén — nem kapunk feltételt,
$n_1=0,\ n_2=1$ esetén — nem kapunk feltételt,
$n_1=0,\ n_2=2$ esetén — $c_{12}\geqq 0,\ c_{13}\geqq 0,\ c_{23}\geqq 0,$
$n_1=0,\ n_2=3$ esetén — $c_{12}+c_{13}+c_{23}\geqq 0,$

b) $n_1=1,\ n_2=0$ esetén — $c_{11}\geqq 0,\ c_{22}\geqq 0,\ c_{33}\geqq 0,$
$n_1=1,\ n_2=1$ esetén — $c_{11}-c_{12}\geqq 0,\ c_{11}-c_{13}\geqq 0,\ c_{22}-c_{12}\geqq 0,$
$c_{22}-c_{23}\geqq 0,\ c_{33}-c_{13}\geqq 0,\ c_{33}-c_{23}\geqq 0,$

$n_1=1,\ n_2=2$ esetén — $c_{11}-c_{12}-c_{13}+c_{23}\geqq 0,$
$c_{22}-c_{12}-c_{23}+c_{13}\geqq 0,$
$c_{33}-c_{13}-c_{23}+c_{12}\geqq 0,$

c) $n_1=2,\ n_2=0$ esetén — $c_{11}+c_{22}+c_{12}\geqq 0,$
$c_{11}+c_{33}+c_{13}\geqq 0,$
$c_{22}+c_{33}+c_{23}\geqq 0,$

$n_1=2,\ n_2=1$ esetén — $c_{11}+c_{22}-c_{13}-c_{23}+c_{12}\geqq 0,$
$c_{11}+c_{33}-c_{12}-c_{23}+c_{13}\geqq 0,$
$c_{22}+c_{33}-c_{12}-c_{13}+c_{23}\geqq 0,$

d) $n_1=3,\ n_2=0$ esetén — $c_{11}+c_{22}+c_{33}+c_{12}+c_{13}+c_{23}\geqq 0.$

A példaként tekintett $n=3$ esetből leolvasható, hogy a 2.1. tételben megadott szükséges feltételek között sok olyan feltétel is található, mely a többinek következménye. (Pl. a $c_{12}+c_{13}+c_{23}\geqq 0$ feltétel következik a $c_{12}\geqq 0,\ c_{13}\geqq 0,\ c_{23}\geqq 0$ feltételekből, stb....).

A következmény feltételeket egyszerűen ki lehet zárni az $I_1$ és $I_2$ halmazok elemszámára tett korlátozó feltételekkel. Ez a következő szűkebb, de az előzőkkel ekvivalens szükséges feltételrendszert eredményezi:

2.1′. Tétel. Ha egy $\mathbf{C}$ empirikus kovariancia mátrixra létezik a (2.2) feltételrendszernek eleget tevő $\vartheta_1, \ldots, \vartheta_p$ paraméter halmaz, akkor annak az elemeire teljesülni kell a

$$\sum_{i\in I_1} c_{ii} - \sum_{i\in I_1, k\in I_2} c_{ik} + \sum_{\substack{i\in I_1, k\in I_1 \\ i<k}} c_{ik} + \sum_{\substack{i\in I_2, k\in I_2 \\ i<k}} c_{ik} \geqq 0 \tag{2.6}$$

feltételeknek, ahol $I_1, I_2 \subset I=\{1, \ldots, n\}$, $I_1 \cap I_2 = \emptyset$ és az $I_1$ és $I_2$ indexhalmazok elemszámaira ($n_1$-re és $n_2$-re) a következő esetek egyikének kell fennállnia:

$$\text{(i)}\quad n_1 = 0,\quad n_2 = 2$$

$$\text{(ii)}\quad n_1 = 1,\quad n_2 \geqq 1 \tag{2.7}$$

$$\text{(iii)}\quad n_1 \geqq 2,\quad n_2 \geqq 2$$

és természetesen mindegyik esetben $n_1+n_2\leqq n$.

A következő tételben összeszámoljuk, hogy az $I_1$ és $I_2$ halmazok elemszámára tett (2.7) korlátozó feltételek mellett hány elemű a szükséges feltételek (2.6) alatt leírt rendszere.

2.2. TÉTEL. Jelölje $S_n$ a (2.6) típusú szükséges feltételek (2.7) korlátozás melletti számát, akkor

$$S_n = 3^n - 2^{n+1} - n2^{n-1} + \frac{n(3n-1)}{2} + 1. \tag{2.8}$$

*Bizonyítás.* A (2.7) megkötések egyes eseteihez tartozó feltételek száma a következő:

— az (i) esethez tartozó feltételek száma: $\binom{n}{2}$,

— az (ii) esethez tartozó feltételek száma: $n\left[\binom{n-1}{1} + \ldots + \binom{n-1}{n-1}\right]$,

— az (iii) esethez tartozó feltételek száma: $\sum_{n_1=2}^{n-2} \sum_{n_2=2}^{n-n_1} \binom{n}{n_1}\binom{n-n_1}{n_2}$.

Ezeket összeadva $S_n$-re azt kapjuk, hogy

$$S_n = \binom{n}{2} + n\left[\binom{n-1}{1} + \ldots + \binom{n-1}{n-1}\right] + \sum_{n_1=2}^{n-2} \sum_{n_1=2}^{n-n_1} \binom{n}{n_1}\binom{n-n_1}{n_2} =$$

$$= \frac{n(n-1)}{2} + n(2^{n-1}-1) + \sum_{n_1=2}^{n-2} \binom{n}{n_1}(2^{n-n_1}-1-n+n_1) =$$

$$= \frac{n(n-1)}{2} + n(2^{n-1}-1) + \sum_{n_1=2}^{n-2} \binom{n}{n_1}2^{n-n_1} - (n+1)\sum_{n_1=2}^{n-2}\binom{n}{n_1} + \sum_{n_1=2}^{n-2} n_1\binom{n}{n_1} =$$

$$= \frac{n(n-1)}{2} + n(2^{n-1}-1) + \sum_{n_1=2}^{n-2} \binom{n}{n_1}2^{n-n_1} - (n+1)\sum_{n_1=2}^{n-2}\binom{n}{n_1} + n\sum_{n_1=2}^{n-2}\binom{n-1}{n_1-1} =$$

$$= \frac{n(n-1)}{2} + n(2^{n-1}-1) + 3^n - 2^n - n2^{n-1} - 2n - 1 - (n+1)(2^n - 1 - n - n - 1) +$$

$$+ n(2^{n-1} - 1 - n + 1 - 1) =$$

$$= \frac{n(n-1)}{2} + n(2^{n-1}-1) + 3^n - n(2^{n-1}+2) - 2^n - 1 - (n+1)(2^n - 2(n+1)) +$$

$$+ n(2^{n-1} - n - 1) =$$

$$= 3^n - 2^{n+1} - n2^{n-1} + \frac{n^2 - n - 2n - 4n - 2 + 4n^2 + 8n + 4 - 2n^2 - 2n}{2} =$$

$$= 3^n - 2^{n+1} - n2^{n-1} + \frac{3n^2 - n + 2}{2} =$$

$$= 3^n - 2^{n+1} - n2^{n-1} + \frac{n(3n-1)}{2} + 1.$$

*Megjegyzés.* Bár az $S_n$-re adott (2.8) képlet levezetése csak $n=4$ esetén helyes, a képlet érvényes $n=2, 3$ esetén is. $S_n$ értéke az első néhány $n$-re:

$$S_2 = 3, \quad S_3 = 12, \quad S_4 = 40, \quad S_5 = 135, \quad S_6 = 461, \quad S_7 = 1554,$$

$$S_8 = 5118, \quad S_9 = 16\,473, \quad S_{10} = 52\,027, \ldots$$

### 3. Algoritmus a többdimenziós gamma eloszlás illesztésére

Mint azt a bevezetésben említettük, az algoritmus alapgondolata az, hogy úgy építjük fel lépésről lépésre a pontosan illeszkedő többdimenziós gamma eloszlást, hogy a fennmaradó rész kovariancia mátrixa mindig eleget tegyen a rá vonatkozó illeszthetőségi feltételeknek.

Vizsgáljuk meg ezért, hogy ha az

$$\mathbf{A}\boldsymbol{\eta} = \sum_{l=1}^{p} \mathbf{a}_l \eta_l$$

előállításban szereplő $\eta_l$ valószínűségi változók $\vartheta_l$ paraméterét valamely $l=j$-re $\vartheta_j^* > 0$ értékűnek rögzítjük, akkor milyen lesz a tovább fennmaradó illesztési feladat. A többdimenziós gamma eloszlás illesztéséhez eredetileg megoldandó (2.2) feltételrendszer a következőre redukálódik:

$$\sum_{l=1}^{p} \mathbf{a}_l \mathbf{a}_l' \vartheta_l = \mathbf{C} - \mathbf{a}_j \mathbf{a}_j' \vartheta_j^* \tag{3.1}$$

$$\vartheta_l \geqq 0, \quad l = 1, \ldots, p.$$

Megjegyezzük, hogy a (3.1) feltételrendszer a $\vartheta_j$ paramétert továbbra is tartalmazhatja azzal a megkötéssel, hogy az illesztés befejezése után az azonos indexű pozitív $\vartheta_l$ paramétereket és a nekik megfelelő $\eta_l$ valószínűségi változókat összevonjuk. (Később látni fogjuk, hogy erre sohasem fog sor kerülni, ugyanis az illesztés algoritmusa garantálja azt, hogy egy pozitív szinten rögzített $\vartheta_l$ paraméter többé nem lesz rögzíthető nullától különböző szinten.) Ezáltal el tudjuk érni azt, hogy bármely

$\vartheta_l$, $l=1, \ldots, p$ paraméter értékének pozitív szinten történő rögzítése után egy, az eredetivel ekvivalens illesztési feladat marad vissza, melyre csupán a többdimenziós gamma eloszlás kovariancia mátrixa lesz a (3.1) feltételrendszer jobb oldalának megfelelő alakú.

Nyilvánvaló ezért, hogy bármely $\vartheta_l$, $l=1, \ldots, p$ paraméter értékét csak olyan pozitív szinten szabad rögzíteni, hogy a fennmaradó illesztési feladat kovariancia mátrixa az illeszthetőség szükséges feltételeinek eleget tegyen. Minthogy a fennmaradó illesztési feladat kovariancia mátrixának az elemei (3.1) szerint

$$c_{ik}-a_{ij}a_{kj}\vartheta_j^*, \quad i=1, \ldots, n; \quad k=1, \ldots, n,$$

azért az illeszthetőség (2.6) alatti szükséges feltételei a következők lesznek:

$$\sum_{i\in I_1}(c_{ii}-a_{ij}^2\vartheta_j^*)-\sum_{i\in I_1,k\in I_2}(c_{ik}-a_{ij}a_{kj}\vartheta_j^*)+ \tag{3.2}$$
$$+\sum_{\substack{i\in I_1,k\in I_1\\ i<k}}(c_{ik}-a_{ij}a_{kj}\vartheta_j^*)+\sum_{\substack{i\in I_2,k\in I_2\\ i<k}}(c_{ik}-a_{ij}a_{kj}\vartheta_j^*)\geqq 0,$$

ahol $I_1, I_2 \subset I=\{1, \ldots, n\}$, $I_1 \cap I_2=\emptyset$ és az $I_1$ és $I_2$ indexhalmazok $n_1$ és $n_2$ elemszámaira teljesülnek a (2.7) feltételek.

Írjuk a (3.2) feltételeket a következő alakba:

$$\Big(\sum_{i\in I_1}a_{ij}^2-\sum_{i\in I_1,k\in I_2}a_{ij}a_{kj}+\sum_{\substack{i\in I_1,k\in I_1\\ i<k}}a_{ij}a_{kj}+\sum_{\substack{i\in I_2,k\in I_2\\ i<k}}a_{ij}a_{kj}\Big)\vartheta_j^*\leqq \tag{3.3}$$
$$\leqq \sum_{i\in I_1}c_{ii}-\sum_{i\in I_1,k\in I_2}c_{ik}+\sum_{\substack{i\in I_1,k\in I_1\\ i<k}}c_{ik}+\sum_{\substack{i\in I_2,k\in I_2\\ i<k}}c_{ik},$$

ahol az $I_1$ és $I_2$ indexhalmazokra az előbb felsorolt feltételeknek kell továbbra is teljesülni.

Ahhoz, hogy az illesztés menete során az egyes $\vartheta_j$ paramétereket olyan $\vartheta_j^*>0$ értéken rögzíthessük, hogy a (3.3) feltételek teljesüljenek, tekintsük a 2.1. tétel bizonyításakor felírt (2.5) egyenletrendszert:

$$\sum_{l=1}\Big(\sum_{i\in I_1}a_{il}^2-\sum_{i\in I_1,k\in I_2}a_{il}a_{kl}+\sum_{\substack{i\in I_1,k\in I_1\\ i<k}}a_{il}a_{kl}+\sum_{\substack{i\in I_2,k\in I_2\\ i<k}}a_{il}a_{kl}\Big)\vartheta_l= \tag{3.4}$$
$$=\sum_{i\in I_1}c_{ii}-\sum_{i\in I_1,k\in I_2}c_{ik}+\sum_{\substack{i\in I_1,k\in I_1\\ i<k}}c_{ik}+\sum_{\substack{i\in I_2,k\in I_2\\ i<k}}c_{ik}.$$

A fent mondottak értelmében bármely $\vartheta_j$ paraméter értékének pozitív szinten történő rögzítésekor a következőképpen kell eljárni. Meg kell keresni az összes olyan, a (2.7) feltételeknek eleget tevő elemszámú $I_1$ és $I_2$ indexhalmazt, amelyre a (3.4) egyenletrendszerben $\vartheta_j$ együtthatója nem nulla. Képezni kell a megfelelő jobb oldali értékek és a nem nulla együtthatók hányadosait és ki kell választani ezen hányadosok legkisebbikét. Ezen a szinten rögzíthetjük a $\vartheta_j$ paraméter értékét, mely szintet jelölje $\vartheta_j^*$. Ha ezután a (3.4) egyenletrendszerben a jobb oldalt úgy módosítjuk, hogy levonjuk belőle a $\vartheta_j^*$ megfelelő együtthatóval felszorzott értékét, akkor a (3.2) és (3.3)

egyenlőtlenségek algebrai azonossága biztosítja, hogy a keletkező egyenletrendszer éppen a $\vartheta_j$ paraméter $\vartheta_j^*$ szinten történt rögzítésével keletkező új illesztési feladatnak megfelelő (3.4) egyenletrendszert fogja szolgáltatni. Ezért a (3.4) egyenletrendszer jobboldalának a módosítása után áttérhetünk további $\vartheta_j$ paraméterek hasonló elv szerinti rögzítésére. A paraméterértékek rögzítését addig kell folytatni, amíg lehetőség kínálkozik pozitív szinten történő rögzítés végrehajtására. Ezzel az illesztési algoritmust lényegében megadtuk.

A módszer gyakorlati alkalmazásához célszerű a (3.4) egyenletrendszert olyan alakban is felírni, hogy benne figyelembe vesszük a $\vartheta_l$ együtthatójára a 2.1. tétel bizonyításában levezetett átalakítást:

$$\sum_{l=1}^{p} \frac{1}{2}[(n_1(l)-n_2(l))(n_1(l)-n_2(l)+1)]\vartheta_l = \tag{3.4'}$$
$$= \sum_{i\in I_1} c_{ii} - \sum_{i\in I_1, k\in I_2} c_{ik} + \sum_{\substack{i\in I_1, k\in I_1 \\ i<k}} c_{ik} + \sum_{\substack{i\in I_2, k\in I_2 \\ i<k}} c_{ik},$$

ahol $n_1(l)$ és $n_2(l)$ jelentése ugyanaz, mint a 2. szakaszban volt. Az egyenletrendszernek ez az alakja lehetőséget ad arra, hogy $\vartheta_l$ együtthatóit lényegesen kevesebb számolással határozzuk meg, és ami talán még ennél is fontosabb, az együtthatók oszloponkénti előállítására is mód nyílik. Ez utóbbi azért nagyon fontos, mert a (3.4) egyenletrendszer sorainak a száma a 2.2. tétel értelmében $3^n - 2^{n+1} - n2^{n-1} + \frac{n(3n-1)}{2} + n + 1$, az oszlopok száma pedig $2^n - 1$, ami miatt nagyobb $n$ értékek mellett az együtthatómátrix explicit tárolása gyakorlatilag lehetetlenné válik. Mindemellett azt is tudjuk, hogy csak elenyészően kevés $\vartheta_l$ paraméter fog pozitív szinten szerepelni a végső előállításban, vagyis az együtthatómátrix minden oszlopára nem is lesz ténylegesen szükségünk az illesztés végrehajtása során. Ugyanakkor nem haszontalan az sem, ha a (3.4) egyenletrendszer alkalmas sorának az együtthatóit is elő tudjuk önmagukban állítani, ha ugyanis a (3.4) egyenletrendszer jobb oldalán egy érték nullára redukálódik, akkor az egyenletrendszernek abban a sorában pozitív együtthatóval szereplő $\vartheta_l$ paraméterekre már nem kell vizsgálni a pozitív szinten történő rögzítés lehetőségét.

A fenti meggondolások alapján érthető, hogy bár az illesztés algoritmusát pontosan megfogalmaztuk, nagyon sok múlik a konkrét számítógépes megvalósításon. Az algoritmus minden részletét jól átgondolt, hatékony programmal kell megvalósítani ahhoz, hogy nagyobb $n$ értékekre is elfogadható számítógép idő felhasználásával lehessen az illesztést végrehajtani.

A módszer illusztrálására tekintsük a következő 4 dimenziós illesztési feladatot. Legyen az előállítandó 4 dimenziós gamma eloszlás kovariancia mátrixa

$$\mathbf{C} = \begin{pmatrix} 4{,}39 & 1{,}30 & 0{,}79 & 0{,}66 \\ 1{,}30 & 2{,}00 & 1{,}40 & 0{,}37 \\ 0{,}79 & 1{,}40 & 1{,}76 & 0{,}48 \\ 0{,}66 & 0{,}37 & 0{,}48 & 0{,}90 \end{pmatrix}.$$

A jobb áttekinthetőség kedvéért írjuk fel explicite a (3.4) egyenletrendszert:

$$
\begin{array}{cccccccccccccccc}
\vartheta_1 & & & + \vartheta_5 & + \vartheta_6 & + \vartheta_7 & & & + \vartheta_{11} & + \vartheta_{12} & + \vartheta_{13} & & + \vartheta_{15} & = c_{11} \\
& \vartheta_2 & & + \vartheta_5 & & & + \vartheta_8 & + \vartheta_9 & + \vartheta_{11} & + \vartheta_{12} & & + \vartheta_{14} & + \vartheta_{15} & = c_{22} \\
& & \vartheta_3 & & + \vartheta_6 & & + \vartheta_8 & + \vartheta_{10} & + \vartheta_{11} & & + \vartheta_{13} & + \vartheta_{14} & + \vartheta_{15} & = c_{33} \\
& & & \vartheta_4 & & + \vartheta_7 & & + \vartheta_9 + \vartheta_{10} & & + \vartheta_{12} & + \vartheta_{13} & + \vartheta_{14} & + \vartheta_{15} & = c_{44} \\
& & & & \vartheta_5 & & & & + \vartheta_{11} & + \vartheta_{12} & & & + \vartheta_{15} & = c_{12} \\
& & & & & \vartheta_6 & & & + \vartheta_{11} & & + \vartheta_{13} & & + \vartheta_{15} & = c_{13} \\
& & & & & & \vartheta_7 & & & + \vartheta_{12} & + \vartheta_{13} & & + \vartheta_{15} & = c_{14} \\
& & & & & & & \vartheta_8 & + \vartheta_{11} & & & + \vartheta_{14} & + \vartheta_{15} & = c_{23} \\
& & & & & & & \vartheta_9 & & + \vartheta_{12} & & + \vartheta_{14} & + \vartheta_{15} & = c_{24} \\
& & & & & & & & \vartheta_{10} & & + \vartheta_{13} & + \vartheta_{14} & + \vartheta_{15} & = c_{34} \\
\vartheta_1 & & & & + \vartheta_6 & + \vartheta_7 & & & & & + \vartheta_{13} & & & = c_{11} - c_{12} \\
\vartheta_1 & & & + \vartheta_5 & & + \vartheta_7 & & & & + \vartheta_{12} & & & & = c_{11} - c_{13} \\
\vartheta_1 & & & + \vartheta_5 & + \vartheta_6 & & & & + \vartheta_{11} & & & & & = c_{11} - c_{14} \\
& \vartheta_2 & & & & & + \vartheta_8 & + \vartheta_9 & & & & + \vartheta_{14} & & = c_{22} - c_{12} \\
& \vartheta_2 & & + \vartheta_5 & & & & + \vartheta_9 & & + \vartheta_{12} & & & & = c_{22} - c_{23} \\
& \vartheta_2 & & + \vartheta_5 & & & + \vartheta_8 & & + \vartheta_{11} & & & & & = c_{22} - c_{24} \\
& & \vartheta_3 & & & & + \vartheta_8 & & + \vartheta_{10} & & & + \vartheta_{14} & & = c_{33} - c_{13} \\
& & \vartheta_3 & & + \vartheta_6 & & & & + \vartheta_{10} & & + \vartheta_{13} & & & = c_{33} - c_{23} \\
& & \vartheta_3 & & + \vartheta_6 & & + \vartheta_8 & & + \vartheta_{11} & & & & & = c_{33} - c_{34} \\
& & & \vartheta_4 & & & & + \vartheta_9 & + \vartheta_{10} & & & + \vartheta_{14} & & = c_{44} - c_{14} \\
& & & \vartheta_4 & & + \vartheta_7 & & & + \vartheta_{10} & & + \vartheta_{13} & & & = c_{44} - c_{24} \\
& & & \vartheta_4 & & + \vartheta_7 & & + \vartheta_9 & & + \vartheta_{12} & & & & = c_{44} - c_{34}
\end{array}
$$

$$
\begin{aligned}
&\vartheta_1 + \vartheta_7 + \vartheta_8 + \vartheta_{14} = c_{11} - c_{12} - c_{13} + c_{23}\\
&\vartheta_1 + \vartheta_6 + \vartheta_9 + \vartheta_{14} = c_{11} - c_{12} - c_{14} + c_{24}\\
&\vartheta_1 + \vartheta_5 + \vartheta_{10} + \vartheta_{14} = c_{11} - c_{13} - c_{14} + c_{34}\\
&\vartheta_2 + \vartheta_6 + \vartheta_9 + \vartheta_{13} = c_{22} - c_{12} - c_{23} + c_{13}\\
&\vartheta_2 + \vartheta_7 + \vartheta_8 + \vartheta_{13} = c_{22} - c_{12} - c_{24} + c_{14}\\
&\vartheta_2 + \vartheta_5 + \vartheta_{10} + \vartheta_{13} = c_{22} - c_{23} - c_{24} + c_{34}\\
&\vartheta_3 + \vartheta_5 + \vartheta_{10} + \vartheta_{12} = c_{33} - c_{13} - c_{23} + c_{12}\\
&\vartheta_3 + \vartheta_7 + \vartheta_8 + \vartheta_{12} = c_{33} - c_{13} - c_{34} + c_{14}\\
&\vartheta_3 + \vartheta_6 + \vartheta_9 + \vartheta_{12} = c_{33} - c_{23} - c_{34} + c_{24}\\
&\vartheta_4 + \vartheta_5 + \vartheta_{10} + \vartheta_{11} = c_{44} - c_{14} - c_{24} + c_{12}\\
&\vartheta_4 + \vartheta_6 + \vartheta_9 + \vartheta_{11} = c_{44} - c_{14} - c_{34} + c_{13}\\
&\vartheta_4 + \vartheta_7 + \vartheta_8 + \vartheta_{11} = c_{44} - c_{24} - c_{34} + c_{23}\\
&\vartheta_1 + \vartheta_8 + \vartheta_9 + \vartheta_{10} + 3\vartheta_{14} + \vartheta_{15} = c_{11} - c_{12} - c_{13} - c_{14} + c_{23} + c_{24} + c_{34}\\
&\vartheta_2 + \vartheta_6 + \vartheta_7 + \vartheta_{10} + 3\vartheta_{13} + \vartheta_{15} = c_{22} - c_{12} - c_{23} - c_{24} + c_{13} + c_{14} + c_{34}\\
&\vartheta_3 + \vartheta_5 + \vartheta_7 + \vartheta_9 + 3\vartheta_{12} + \vartheta_{15} = c_{33} - c_{13} - c_{23} - c_{34} + c_{12} + c_{14} + c_{24}\\
&\vartheta_4 + \vartheta_5 + \vartheta_6 + \vartheta_8 + 3\vartheta_{11} + \vartheta_{15} = c_{44} - c_{14} - c_{24} - c_{34} + c_{12} + c_{13} + c_{23}\\
&\vartheta_1 + \vartheta_2 + 3\vartheta_5 + \vartheta_{10} + \vartheta_{11} + \vartheta_{12} = c_{11} + c_{22} - c_{13} - c_{14} - c_{23} - c_{24} + c_{12} + c_{34}\\
&\vartheta_1 + \vartheta_3 + 3\vartheta_6 + \vartheta_9 + \vartheta_{11} + \vartheta_{13} = c_{11} + c_{33} - c_{12} - c_{14} - c_{23} - c_{34} + c_{13} + c_{24}\\
&\vartheta_1 + \vartheta_4 + 3\vartheta_7 + \vartheta_8 + \vartheta_{12} + \vartheta_{13} = c_{11} + c_{44} - c_{12} - c_{13} - c_{24} - c_{34} + c_{14} + c_{23}\\
&\vartheta_2 + \vartheta_3 + \vartheta_7 + 3\vartheta_8 + \vartheta_{11} + \vartheta_{14} = c_{22} + c_{33} - c_{12} - c_{24} - c_{13} - c_{34} + c_{23} + c_{14}\\
&\vartheta_2 + \vartheta_4 + \vartheta_6 + 3\vartheta_9 + \vartheta_{12} + \vartheta_{14} = c_{22} + c_{44} - c_{12} - c_{23} - c_{14} - c_{34} + c_{24} + c_{13}\\
&\vartheta_3 + \vartheta_4 + \vartheta_5 + 3\vartheta_{10} + \vartheta_{13} + \vartheta_{14} = c_{33} + c_{44} - c_{13} - c_{23} - c_{14} - c_{24} + c_{34} + c_{12}
\end{aligned}
$$

Az illesztés menetét a jobb oldali értékek változásainak a táblázata segítségével mutatjuk be.

A táblázat oszlopai alatt az éppen aktuális paraméter érték rögzítéseket adtuk meg, az oszlopokban a dőlt számok azt mutatják, hogy mely feltételekben szerepel nem nulla együtthatóval a pozitív szinten rögzíteni kívánt paraméter.

Vegyük észre, hogy az ötödik oszlopig eljutva már öt paraméter értékét lehetett nulla szinten rögzíteni ($\vartheta_6=\vartheta_9=\vartheta_{14}=\vartheta_{13}=\vartheta_{12}=0$), és minthogy $n=4$ esetén a (2.2) lineáris egyenletrendszer együttható mátrixának a rangja 10, a további paraméter értékek a következő lineáris egyenletrendszer megoldásával is egyértelműen meghatározhatók lettek volna:

$$\begin{aligned}
\vartheta_1 \quad &+\vartheta_5+\vartheta_7 \quad &&+\vartheta_{11}+\vartheta_{15} = 4{,}39\\
\vartheta_2 \quad &+\vartheta_5 \quad &+\vartheta_8 \quad &+\vartheta_{11}+\vartheta_{15} = 2{,}00\\
\vartheta_3 \quad &\quad &+\vartheta_8+\vartheta_{10} \quad &+\vartheta_{11}+\vartheta_{15} = 1{,}76\\
\vartheta_4 \quad &+\vartheta_7 \quad &+\vartheta_{10} \quad &+\vartheta_{15} = 0{,}90\\
\vartheta_5 \quad &\quad &\quad &+\vartheta_{11}+\vartheta_{15} = 1{,}30\\
\quad &\quad &\quad &\vartheta_{11}+\vartheta_{15} = 0{,}79\\
\quad &\vartheta_7 \quad &\quad &+\vartheta_{15} = 0{,}66\\
\quad &\quad &\vartheta_8 \quad &+\vartheta_{11}+\vartheta_{15} = 1{,}40\\
\quad &\quad &\quad &\vartheta_{15} = 0{,}37\\
\quad &\quad &\vartheta_{10} \quad &+\vartheta_{15} = 0{,}48
\end{aligned}$$

A fenti egyenletrendszer megoldásakor figyelembe vehetjük azt is, hogy az ötödik oszlopig jutva már végrehajtottuk a $\vartheta_1=2{,}80$, $\vartheta_2=0{,}09$, $\vartheta_3=0{,}25$ pozitív szinten történő értékrögzítéseket is. Nyilvánvaló, hogy az illesztő algoritmus számítógépes megvalósítása során célszerű a nulla rögzítéseket számolni, és amikor a további paraméterek meghatározása egyértelművé válik, azokat a megfelelő lineáris egyenletrendszer megoldásával kell kiszámolni.

Az illesztő algoritmus számítógépes megvalósítása során az első problémát a (2.2) feltételrendszerben bevezetett $\mathbf{a}_l$, $l=1, \ldots, p$ vektorok előállítása jelenti. Erre a következő egyszerű képletet használtuk:

$$a_{kl} = \begin{cases} 0, & \text{ha } (l-1)/2^{n-k} \text{ páratlan} \\ 1, & \text{ha } (l-1)/2^{n-k} \text{ páros,} \end{cases}, \quad k = 1, \ldots, n.$$

Az $\mathbf{a}_l$ vektorok ily módon történő előállításának az az előnye, hogy az $l$ index ismeretében egyetlen $\mathbf{a}_l$ vektort önmagában is gyorsan elő tudunk állítani. Célszerűnek mutatkozott az is, hogy a 2 hatványokat előre meghatározva, egy tömbben tároljuk. Megjegyezzük, hogy az $\mathbf{a}_l$ vektorok fenti módon történő előállításával azok sorrendje nem lesz az eddig megszokott, ez azonban csak a $\vartheta$ paraméterek indexezésében fog változást okozni.

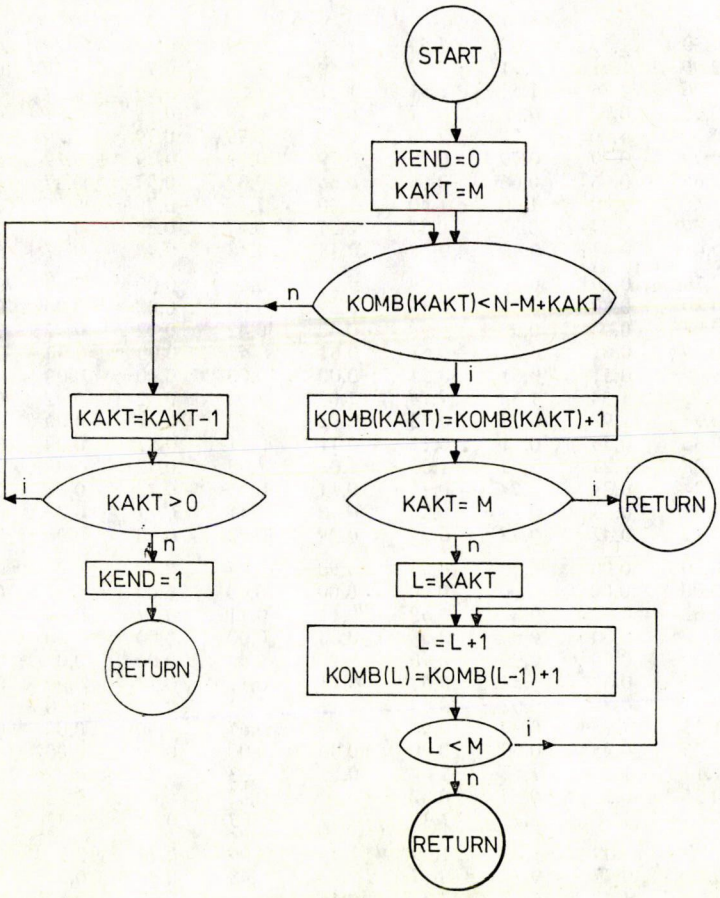
A számítógépes megvalósítás második alapvető problémája azon $I_1$ és $I_2$ indexhalmazok előállítása, amelyek elemszámai eleget tesznek a (2.7) feltételeknek. Ennek alapját az az általános célú szubrutin képezi, amely az $1, 2, \ldots, n$ számok $k$-adosztályú kombinációit állítja elő úgy, hogy egy kezdeti kombinációból kiindulva, minden egyes hívásakor egy új kombinációt állít elő, illetve ha már nincs több, még elő nem állított kombináció, azt jelzi. A szubrutin blokkdiagramja az 1. ábrán látható, a kezdeti kombinációt az $1, 2, \ldots, k$ számok kell, hogy alkossák. A változók jelentése a következő:

2.1 TÁBLÁZAT

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| *4,39* | 1,59 | 1,59 | 1,59 | *1,59* | *1,08* | 0,79 | 0,79 | *0,79* | *0,37** | 0,00 |
| 2,00 | *2,00* | 1,91 | 1,91 | *1,91* | 1,40 | *1,40* | 0,79 | *0,79* | *0,37* | 0,00 |
| 1,76 | 1,76 | *1,76* | 1,51 | 1,51 | 1,51 | *1,51* | *0,90* | *0,79* | *0,37* | 0,00 |
| 0,90 | 0,90 | 0,90 | *0,90* | 0,77 | *0,77* | 0,48 | *0,48* | 0,37 | *0,37* | 0,00 |
| 1,30 | 1,30 | 1,30 | 1,30 | *1,30* | 0,79 | 0,79 | 0,79 | *0,79* | *0,37* | 0,00 |
| 0,79 | 0,79 | 0,79 | 0,79 | 0,79 | 0,79 | 0,79 | 0,79 | *0,79* | *0,37* | 0,00 |
| 0,66 | 0,66 | 0,66 | 0,66 | 0,66 | *0,66* | 0,37 | 0,37 | 0,37 | *0,37* | 0,00 |
| 1,40 | 1,40 | 1,40 | 1,40 | 1,40 | 1,40 | *1,40* | 0,79 | *0,79* | *0,37* | 0,00 |
| 0,37 | 0,37 | 0,37 | 0,37 | 0,37 | 0,37 | 0,37 | 0,37 | 0,37 | *9,37* | 0,00 |
| 0,48 | 0,48 | 0,48 | 0,48 | 0,48 | 0,48 | 0,48 | *0,48* | 0,37 | *0,37* | 0,00 |
| *3,09* | 0,29 | 0,29 | 0,29 | 0,29 | *0,29** | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 |
| *3,60* | 0,80 | 0,80 | 0,80 | *0,80* | *0,29* | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 |
| *3,73* | 0,93 | 0,93 | 0,93 | *0,93* | 0,42 | 0,42 | 0,42 | *0,42** | 0,00 | 0,00 |
| 0,70 | *0,70* | 0,61 | 0,61 | 0,61 | 0,61 | *0,61** | 0,00 | 0,00 | 0,00 | 0,00 |
| 0,60 | *0,60* | 0,51 | 0,51 | *0,51** | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 |
| 1,63 | *1,63* | 1,54 | 1,54 | *1,54* | 1,03 | *1,03* | 0,42 | *0,42* | 0,00 | 0,00 |
| 0,97 | 0,97 | *0,97* | 0,72 | 0,72 | 0,72 | *0,72* | *0,11** | 0,00 | 0,00 | 0,00 |
| 0,36 | 0,36 | *0,36* | 0,11 | 0,11 | 0,11 | 0,11 | *0,11* | 0,00 | 0,00 | 0,00 |
| 1,28 | 1,28 | *1,28* | 1,03 | 1,03 | 1,03 | *1,03* | 0,42 | *0,42* | 0,00 | 0,00 |
| 0,24 | 0,24 | 0,24 | *0,24* | 0,11 | 0,11 | 0,11 | *0,11* | 0,00 | 0,00 | 0,00 |
| 0,53 | 0,53 | 0,53 | *0,53* | 0,40 | *0,40* | 0,11 | *0,11* | 0,00 | 0,00 | 0,00 |
| 0,42 | 0,42 | 0,42 | *0,42* | 0,29 | *0,29* | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 |
| *3,70* | 0,90 | 0,90 | 0,90 | 0,90 | *0,90* | *0,61* | 0,00 | 0,00 | 0,00 | 0,00 |
| *2,80** | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 |
| *3,42* | 0,62 | 0,62 | 0,62 | *0,62* | 0,11 | 0,11 | *0,11* | 0,00 | 0,00 | 0,00 |
| 0,09 | *0,09** | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 |
| 0,99 | *0,99* | 0,90 | 0,90 | 0,90 | *0,90* | *0,61* | 0,00 | 0,00 | 0,00 | 0,00 |
| 0,71 | *0,71* | 0,62 | 0,62 | *0,62* | 0,11 | 0,11 | *0,11* | 0,00 | 0,00 | 0,00 |
| 0,87 | 0,87 | *0,87* | 0,62 | *0,62* | 0,11 | 0,11 | *0,11* | 0,00 | 0,00 | 0,00 |
| 1,15 | 1,15 | *1,15* | 0,90 | 0,90 | *0,90* | *0,61* | 0,00 | 0,00 | 0,00 | 0,00 |
| 0,25 | 0,25 | *0,25** | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 |
| 1,17 | 1,17 | 1,17 | *1,17* | *1,04* | 0,53 | 0,53 | *0,53* | *0,42* | 0,00 | 0,00 |
| 0,55 | 0,55 | 0,55 | *0,55* | 0,42 | 0,42 | 0,42 | 0,42 | *0,42* | 0,00 | 0,00 |
| 1,45 | 1,45 | 1,45 | *1,45* | 1,32 | *1,32* | *1,03* | 0,42 | *0,42* | 0,00 | 0,00 |
| *3,89* | 1,09 | 1,09 | 1,09 | 1,09 | 1,09 | *1,09* | *0,48* | 0,37 | *0,37* | 0,00 |
| 0,86 | *0,86* | 0,77 | 0,77 | 0,77 | *0,77* | 0,48 | *0,48* | 0,37 | *0,37* | 0,00 |
| 1,42 | 1,42 | *1,42* | 1,17 | *1,17* | *0,66* | 0,37 | 0,37 | 0,37 | *0,37* | 0,00 |
| 2,88 | 2,88 | 2,88 | *2,88* | *2,75* | 2,24 | *2,24* | 1,63 | *1,63* | *0,37* | 0,00 |
| *4,95* | *2,15* | 2,06 | 2,06 | *2,06* | 0,53 | 0,53 | *0,53* | *0,42* | 0,00 | 0,00 |
| *3,47* | 0,67 | *0,67* | 0,42 | 0,42 | 0,42 | 0,42 | 0,42 | *0,42* | 0,00 | 0,00 |
| *4,41* | 1,61 | 1,61 | *1,61* | 1,48 | *1,48* | *0,61* | 0,00 | 0,00 | 0,00 | 0,00 |
| 2,88 | *2,88* | *2,79* | 2,54 | 2,54 | *2,54* | *2,25* | 0,42 | *0,42* | 0,00 | 0,00 |
| 0,22 | *0,22* | 0,13 | *0,13** | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 |
| 1,22 | 1,22 | *1,22* | *0,97* | *0,84* | 0,33 | 0,33 | *0,33* | 0,00 | 0,00 | 0,00 |

$\vartheta_1=2{,}80$ $\vartheta_6=0$ $\vartheta_{13}=0$ $\vartheta_{12}=0$ $\vartheta_5=0{,}51$ $\vartheta_7=0{,}29$ $\vartheta_8=0{,}61$ $\vartheta_{10}=0{,}11$ $\vartheta_{11}=0{,}42$ $\vartheta_{15}=0{,}37$
$\vartheta_9=0$ $\vartheta_3=0{,}25$ $\vartheta_4=0{,}13$
$\vartheta_{14}=0$
$\vartheta_2=0{,}09$

N — az összes elem száma,
M — a kiválasztandó elemek száma,
KOMB — M-elemű tömb az aktuálisan kiválasztott kombináció tárolására,
KEND — jelzi, hogy az összes kombinációt előállítottuk-e már (KEND=0, ha nem és KEND=1, ha igen),
KAK,L — belső segédváltozók.




*1. ábra*

Az illesztő algoritmus egy iterációs lépésének a végrehajtásakor lényegében a következő információkat tároljuk:

(i) mely $\vartheta_j$ paraméter értékét rögzítettük, milyen pozitív szinten,
(ii) mely $\vartheta_j$ paraméterekről került ki a pozitív szintű rögzítés hatására, hogy nulla értékűeknek kell lenniük,
(iii) hogyan módosult a további illesztésre váró kovariancia mátrix.

Az első típusú információ tárolásához elegendő egy $\frac{1}{2}n(n+1)$ méretű egész és egy ugyanekkora méretű valós tömb, a kovariancia mátrix tárolása is megoldható ugyanekkora méretű valós tömbben. A legtöbb gondot a nulla rögzítések tárolása okozza, mivel erre elegendő ugyan paraméterenként egy 1 byte-os logikai változó, azonban a $\vartheta$ paraméterek száma $2^n-1$, ami miatt 15-nél nagyobb $n$ értékekre általában háttértároló alkalmazása válik szükségessé.

A program jelenlegi változata két menetben vizsgálja át a (3.4′) egyenletrendszert. Első menetben meghatározza az éppen vizsgált $\vartheta$ paraméter lehetséges pozitív rögzítési szintjét, a második menetben először felderíti azokat a feltételi egyenleteket, amelyek a vizsgált $\vartheta$ paraméter értékének a rögzítése után nulla jobb oldali értékkel fognak rendelkezni, és megjelöli az ezen egyenletekben szereplő $\vartheta$ paramétereket, mint nulla értékűvé válókat, majd végrehajtja a kovariancia mátrix módosítását.

A fent leírt algoritmus — amint azt a bevezetésben is jeleztük — a szükséges feltételek 2.2. tételben bizonyított nagy száma miatt nagyobb $n$ értékekre irreálisan sok számítási időt igényel. Ezért elkészítettük az algoritmus egy olyan változatának a számítógépes implementálását is, amely az illeszthetőség 2.1′ tételben megadott szükséges feltételei közül a (2.7) (iii) alattiakat nem veszi figyelembe. Ezáltal az algoritmus heurisztikus jellege erősödik, azonban a számítógépes program hatékonyabban valósítható meg, és ezáltal a számítási idők még 15—20 dimenziós esetekre is reális korlátok alatt maradnak. Ugyanakkor az eddigi számítási tapasztalatok — mind a gyakorlatban előfordult, mind véletlenszerűen generált feladatokra — azt mutatják, hogy az így korlátozott algoritmus is megbízhatóan működik. Elkészítettünk továbbá egy, a szimplex módszer első fázisát megvalósító algoritmust is, melynél a nehézséget az jelentette, hogy az együtthatómátrix explicit tárolása magasabb dimenziókra nem oldható meg. Ezért az algoritmus egy olyan változatát kellett kidolgozni, amely az együtthatómátrix egyes oszlopait csak akkor állítja elő, amikor azokra éppen szükség van. Ez az oszlopgenerálási technika a matematikai programozásban jól ismert, ezért a részletek ismertetésétől eltekintünk.

Mindhárom programot egy TPA 1140-es kisszámítógépen teszteltük, ezért a 3.1. táblázatban közölt időértékeknek nem az abszolút értéke, hanem az egymáshoz való viszonyuk érdekes. Ezek alapján a gyakorlati alkalmazások céljára azt javasoljuk, hogy először meg kell kísérelni a redukált heurisztikus algoritmust alkalmazni az illesztési feladat megoldására, ha ez sikertelen, akkor ellenőrizni kell az összes ismert szükséges feltétel teljesülését, és ha azok teljesülnek, akkor érdemes a szimplex módszer első fázisát alkalmazni. Ha pedig a szükséges feltételek nem teljesülnek, akkor a [3] dolgozatban javasolt időigényesebb módszerek egyikét kell alkalmazni, melyek előnye éppen az, hogy ilyen esetekben is valamilyen értelemben a „valódihoz legközelebbi" eloszlást találják meg.

3.1 TÁBLÁZAT

A három illesztő algoritmus számítógépes futási időeredményei másodpercben

| Dimenzió | Heurisztikus (redukált | Szimplex módszer első fázisa | Heurisztikus (teljes) |
|---|---|---|---|
| 3 | 0,1406 | 0,7188 | 0,4590 |
| 4 | 0,1797 | 1,8203 | 2,4395 |
| 5 | 0,2207 | 7,0000 | 20,5000 |
| 6 | 0,3984 | 21,8008 | 166,7402 |
| 7 | 0,6217 | 59,3045 | — |
| 8 | 1,0908 | 173,1427 | — |

## 4. Az illeszthetőség szükséges feltételei elégségességére vonatkozó eredmények

Amint azt a bevezetésben jeleztük, az előző szakaszban leírt algoritmust mindaddig heurisztikusnak kell tekinteni, amíg be nem bizonyítjuk, hogy az illeszthetőség szükséges feltételei elégségesek is. Minthogy azonban az illesztés előző szakaszban leírt algoritmusa csak akkor működik igazán hatékonyan, ha nem vesszük figyelembe az összes szükséges feltételt, az elégségesség igazolásának a problémája csak elméleti jelentőséggel bír.

$n=2$ *esetén* triviálisan igaz, hogy az illeszthetőség szükséges feltételei ($c_{12}\geqq 0$, $c_{11}-c_{12}\geqq 0$, $c_{22}-c_{12}\geqq 0$) elégségesek is, hiszen ekkor $\vartheta_1=c_{11}-c_{12}$, $\vartheta_2=c_{22}-c_{12}$ és $\vartheta_3=c_{12}$ paramétereket választva mindig megoldható az illesztés feladata.

$n=3$ *esetén* még mindig elemi módon bizonyítható az illeszthetőségi feltételek elégségessége. Ekkor ugyanis az illesztés végrehajtásához az alábbi lineáris egyenletrendszernek kell nemnegatív komponensekből álló megoldást keresni:

$$\begin{aligned}
\vartheta_1 \quad &+\vartheta_4+\vartheta_5 \quad &&+\vartheta_7 = c_{11}\\
\vartheta_2 \quad &+\vartheta_4 \quad &&+\vartheta_6+\vartheta_7 = c_{22}\\
\vartheta_3 \quad &+\vartheta_5+\vartheta_6+\vartheta_7 = c_{33}\\
\vartheta_4 \quad &+\vartheta_7 = c_{12}\\
\vartheta_5 \quad &+\vartheta_7 = c_{13}\\
\vartheta_6+\vartheta_7 &= c_{23}
\end{aligned}$$

Fejezzük ki a fenti egyenletrendszerből a $\vartheta_1, \ldots, \vartheta_6$ változókat:

$$\begin{aligned}
\vartheta_6 &= c_{23}-\vartheta_7\\
\vartheta_5 &= c_{13}-\vartheta_7\\
\vartheta_4 &= c_{12}-\vartheta_7\\
\vartheta_3 &= c_{33}-c_{13}-c_{23}+\vartheta_7\\
\vartheta_2 &= c_{22}-c_{12}-c_{23}+\vartheta_7\\
\vartheta_1 &= c_{11}-c_{12}-c_{13}+\vartheta_7
\end{aligned}$$

Azt kell csak belátni ezután, hogy $\vartheta_7$-nek mindig lehet olyan nemnegatív értéket találni, hogy $\vartheta_1, \ldots, \vartheta_6$ értékei is nemnegatívak legyenek. Ennek az a feltétele, hogy

$$\min\{c_{12}, c_{13}, c_{23}\} \geqq \max\{-c_{33}+c_{13}+c_{23}, -c_{22}+c_{12}+c_{23}, -c_{11}+c_{12}+c_{13}\} \tag{4.1}$$

teljesüljön. Ez a feltétel pedig következik az illeszthetőség szükséges feltételeiből, melyek $n=3$ esetén a következők:

$$\begin{aligned}
c_{11}-c_{12} &\geqq 0\\
c_{11}-c_{13} &\geqq 0\\
c_{22}-c_{12} &\geqq 0\\
c_{22}-c_{23} &\geqq 0
\end{aligned}$$

$$c_{33}-c_{13} \geqq 0$$

$$c_{33}-c_{23} \geqq 0$$

$$c_{11}-c_{12}-c_{13}+c_{23} \geqq 0$$

$$c_{22}-c_{12}-c_{23}+c_{13} \geqq 0$$

$$c_{33}-c_{13}-c_{23}+c_{12} \geqq 0$$

Meg lehet ugyanis mutatni, hogy a (4.1) feltétel bármely bal oldali mennyisége nagyobb vagy egyenlő, mint bármely jobb oldali mennyisége. Például $c_{12}$ esetén ez a következőképpen történhet:

$$c_{12} \geqq -c_{33}+c_{13}+c_{23}, \quad \text{mivel} \quad c_{33}-c_{13}-c_{23}+c_{12} \geqq 0,$$

$$c_{12} \geqq -c_{22}+c_{12}+c_{23}, \quad \text{mivel} \quad c_{22} \geqq c_{23},$$

$$c_{12} \geqq -c_{11}+c_{12}+c_{13}, \quad \text{mivel} \quad c_{11} \geqq c_{13}.$$

$n \geqq 4$ *esetén* a feladat bonyolultabb. Ezekben az esetekben célszerűnek látszik a problémát átfogalmazni. Az átfogalmazás módja önmagában is érdekes, és a két probléma azonossága emlékeztet a kombinatorikus optimalizálás témakörének azon eredményeire, amelyek diszkrét ponthalmaz köré írható legszűkebb konvex poliéder meghatározására vonatkoznak. A probléma átfogalmazásához néhány definíció kimondására és egyszerű állítások igazolására van szükségünk.

4.1. DEFINÍCIÓ. Legyen $\mathscr{C}_n$ azoknak az $n \times n$-es szimmetrikus, nemnegatív elemű $\mathbf{C}=(c_{ij})_{i,j=1}^{n}$ mátrixoknak a halmaza, amelyekre a (2.2) feltételrendszernek létezik megengedett megoldása.

4.1. ÁLLÍTÁS. A $\mathscr{C}_n$ mátrix halmaz az $n^2$-dimenziós euklideszi térben egy konvex kúpot alkot.

*Bizonyítás.* A $\mathscr{C}_n$ halmaz definíció szerint az $\mathbf{a}_l \mathbf{a}_l'$, $l=1, \ldots, p$ diádok konvex kúp burkával egyenlő.

4.2. DEFINÍCIÓ. Legyen $\mathscr{D}_n$ a $\mathscr{C}_n$ konvex kúp pozitív poláris halmaza, azaz

$$\mathscr{D}_n = \mathscr{C}_n^* = \{\mathbf{D}: \sum_{i=1}^{n}\sum_{k=1}^{n} c_{ik}d_{ik} \geqq 0, \quad \text{minden} \quad \mathbf{C} \in \mathscr{C}_n \quad \text{esetén}\}.$$

4.2. ÁLLÍTÁS. $\mathbf{C} \in \mathscr{C}_n$ akkor és csak akkor, ha

$$\sum_{i=1}^{n}\sum_{k=1}^{n} c_{ik}d_{ik} \geqq 0, \quad \text{minden} \quad \mathbf{D} \in \mathscr{D}_n \quad \text{esetén.} \tag{4.2}$$

*Bizonyítás.* A *Farkas-tétel* értelmében $\mathscr{D}_n$ poláros halmaza egyenlő $\mathscr{C}_n$ konvex kúp burkával, vagyis magával a $\mathscr{C}_n$ halmazzal, hiszen a 4.1. állítás értelmében ő maga is konvex kúp. Az állításunk pedig éppen ennek a ténynek a megfogalmazása.

Mivel a $\mathscr{D}_n$ mátrix halmaz definíciójánál fogva az $n^2$-dimenziós euklideszi térben egy konvex kúpot alkot, azért a 4.2. állítás tovább élesíthető.

4.2'. ÁLLÍTÁS. $\mathbf{C}\in\mathscr{C}_n$ akkor és csak akkor, ha

$$\sum_{i=1}^{n}\sum_{k=1}^{n} c_{ik}d_{ik} \geqq 0,$$

minden olyan $\mathbf{D}=(d_{ik})_{i,k=1}^{n}$ mátrix esetén, amely a $\mathscr{D}_n$ konvex kúp extremális iránya.

Foglalkozzunk ezért a továbbiakban a $\mathscr{D}_n$ konvex kúp extremális irányainak a meghatározásával.

4.3. ÁLLÍTÁS. $\mathbf{D}\in\mathscr{D}_n$ akkor és csak akkor, ha

$$\sum_{i=1}^{n}\sum_{k=1}^{n} d_{ik}a_{il}a_{kl} \geqq 0, \quad l = 1, \ldots, p,$$

ahol az $\mathbf{a}_l'=(a_{1l}, \ldots, a_{nl})$, $l=1, \ldots, p$ vektorok a (2.2) összefüggésben bevezetettekkel azonosak.

*Bizonyítás.* Ha $\mathbf{D}\in\mathscr{D}_n$, vagyis $\sum_{i=1}^{n}\sum_{k=1}^{n} c_{ik}d_{ik}\geqq 0$ minden $\mathbf{C}\in\mathscr{C}_n$-re, akkor mint-hogy $\mathbf{a}_l\mathbf{a}_l'\in\mathscr{C}_n$, $l=1, \ldots, p$, ezekre felírva a $\sum_{i=1}^{n}\sum_{k=1}^{n} c_{ik}d_{ik}\geqq 0$ feltételt, éppen az állításban szereplő feltételekre jutunk.

Fordítva, ha $\sum_{i=1}^{n}\sum_{k=1}^{n} d_{ik}a_{il}a_{kl}\geqq 0$, $l=1, \ldots, p$ és $\mathbf{C}$ a $\mathscr{C}_n$ halmaz egy tetszőleges eleme, azaz $c_{ik}=\sum_{l=1}^{p} a_{il}a_{kl}\vartheta_l$, ahol $\vartheta_l\geqq 0$, $l=1, \ldots, p$, akkor

$$\sum_{i=1}^{n}\sum_{k=1}^{n} c_{ik}d_{ik} = \sum_{i=1}^{n}\sum_{k=1}^{n}\Big(\sum_{l=1}^{p} a_{il}a_{kl}\vartheta_l\Big)d_{ik} = \sum_{l=1}^{p}\Big(\sum_{i=1}^{n}\sum_{k=1}^{n} d_{ik}a_{il}a_{kl}\Big)\vartheta_l \geqq 0,$$

ami éppen azt jelenti, hogy $\mathbf{D}\in\mathscr{D}_n$.

Ha figyelembe vesszük az $\mathbf{a}_l$, $l=1, \ldots, p$ vektorok definícióját, akkor a 4.3. állítás a következő alakra hozható:

4.3'. ÁLLÍTÁS. $\mathbf{D}\in\mathscr{D}_n$ akkor és csak akkor, ha

$$\sum_{j=1}^{s}\sum_{k=1}^{s} d_{i_ji_k} \geqq 0, \quad \text{tetszőleges} \quad \{i_1, \ldots, i_s\}\subset\{1, \ldots, n\} \quad \text{esetén.} \tag{4.3}$$

Ugyanez szavakkal fogalmazva azt jelenti, hogy $\mathbf{D}\in\mathscr{D}_n$ akkor és csak akkor, ha az elemeinek az összege és minden szimmetrikus részmátrixa elemeinek az összege nemnegatív.

Ha még azt is figyelembe vesszük, hogy a $\mathbf{D}\in\mathscr{D}_n$ mátrixok szimmetrikusak (ezt a $\mathbf{C}$ mátrixok szimmetrikussága miatt tehetjük meg), akkor a $\mathscr{D}_n$ konvex kúpot leíró (4.3) feltételrendszer a következő alakot ölti:

$$\begin{gathered}\sum_{j=1}^{s} d_{i_ji_j}+2\sum_{j=1}^{s}\sum_{k=j+1}^{s} d_{i_ji_k} \geqq 0, \quad \{i_1, \ldots, i_s\}\subset\{1, \ldots, n\}, \quad s = 2, \ldots, n,\\ d_{ii} \geqq 0, \quad i = 1, \ldots, n.\end{gathered} \tag{4.4}$$

A (4.4) feltételrendszer által leírt konvex kúp extremális irányai keresésének a problémáját visszavezetjük egy alkalmasan definiált konvex poliéder extremális pontjai keresésének a problémájára. Erre a következő állítás ad lehetőséget.

4.4. ÁLLÍTÁS. A (4.4) feltételrendszer által definiált konvex kúp extremális irányait a

$$\sum_{j=1}^{s} d_{i_j i_j} + 2 \sum_{j=1}^{s} \sum_{k=j+1}^{s} d_{i_j i_k} \geqq 0, \quad \{i_1, \ldots, i_s\} \subset \{1, \ldots, n\}, \quad s = 2, \ldots, n,$$

$$\sum_{i=1}^{n} d_{ii} + \sum_{i=1}^{n} \sum_{k=i+1}^{n} d_{ik} = 1, \quad d_{ii} = 0, \quad i \geqq 1, \ldots, n \tag{4.5}$$

konvex poliéder extremális pontjaiba mutató helyvektorok szolgáltatják.

*Bizonyítás.* Csak azt kell belátni, hogy a (4.4) konvex kúp bármely extremális iránya által definiált, origó kezdőpontú félegyenesnek van közös pontja a

$$\sum_{i=1}^{n} d_{ii} + \sum_{i=1}^{n} \sum_{k=i+1}^{n} d_{ik} = 1 \tag{4.6}$$

hipersíkkal. A $d_{ii} \geqq 0$, $i = 1, \ldots, n$ és az $s = 2$ esetnek megfelelő $d_{ii} + d_{jj} + 2d_{ij} \geqq 0$, $\{i, j\} \subset \{1, \ldots, n\}$ feltételekből látható, hogy a (4.4) feltételek közül legalább egynek szigorú egyenlőtlenséggel kell teljesülni ahhoz, hogy ne legyen minden $d_{ij}$, $i, j = 1, \ldots, n$, $i \leqq j$ komponens nulla értékű. Minthogy bármely extremális iránynak különböznie kell a nullvektortól, és ki kell elégítenie a (4.4) feltételeket, azért egy tetszőleges extremális irány $d_{ij}^*$ komponenseire a (4.4) feltételeket felírva, és összeadva azt kapjuk, hogy

$$2^{n-1} \sum_{i=1}^{n} d_{ii}^* + 2^{n-1} \sum_{i=1}^{n} \sum_{k=i+1}^{n} d_{ik}^* > 0,$$

vagyis alkalmas $\lambda > 0$ szorzóval mindig elérhető, hogy az extremális irány $\lambda$-szorosa a (4.6) hipersík pontja legyen.

*Példa.* $n = 4$ esetén a (4.5) feltételrendszer a következő:

$$\begin{array}{rrrrrrrrrrl}
d_{11} + d_{22} & & & + 2d_{12} & & & & & & & \geqq 0 \\
d_{11} & + d_{33} & & & + 2d_{13} & & & & & & \geqq 0 \\
d_{11} & & + d_{44} & & & + 2d_{14} & & & & & \geqq 0 \\
 & d_{22} + d_{33} & & & & & + 2d_{23} & & & & \geqq 0 \\
 & d_{22} & + d_{44} & & & & & + 2d_{24} & & & \geqq 0 \\
 & & d_{33} + d_{44} & & & & & & + 2d_{34} & & \geqq 0 \\
d_{11} + d_{22} + d_{33} & & & + 2d_{12} + 2d_{13} & & & + 2d_{23} & & & & \geqq 0 \\
d_{11} + d_{22} & & + d_{44} & + 2d_{12} & & + 2d_{14} & & + 2d_{24} & & & \geqq 0 \\
d_{11} & + d_{33} + d_{44} & & & + 2d_{13} + 2d_{14} & & & & + 2d_{34} & & \geqq 0 \\
 & d_{22} + d_{33} + d_{44} & & & & & + 2d_{23} + 2d_{24} + 2d_{34} & & & & \geqq 0 \\
d_{11} + d_{22} + d_{33} + d_{44} & & & + 2d_{12} + 2d_{13} + 2d_{14} & & & + 2d_{23} + 2d_{24} + 2d_{34} & & & & \geqq 0 \\
d_{11} + d_{22} + d_{33} + d_{44} & & & + d_{12} & + d_{13} & & + d_{14} & + d_{23} & + d_{24} & + d_{34} & = 1
\end{array}$$

$$d_{11} \geqq 0, \quad d_{22} \geqq 0, \quad d_{33} \geqq 0, \quad d_{44} \geqq 0$$

Végül a 4.5. állításban egy sejtést fogalmazunk meg a (4.5) feltételrendszer által definiált konvex poliéder extremális pontjaira vonatkozóan. Az eddigi állítások alapján nyilvánvaló, hogy a 4.5. állítás bizonyítása a többdimenziós gamma eloszlás illeszthetőségére a 2. szakaszban kimondott szükséges feltételek elégségességének a bizonyítását is eredményezné. A 4.5. állítást csak $n \leqq 4$ esetén tudjuk bizonyítani. Az $n=4$ eset bizonyítását a dolgozat függelékében közöljük.

4.5. ÁLLÍTÁS. A (4.5) feltételrendszer által definiált konvex poliéder összes extremális pontját a következő $\mathbf{D}=(d_{ij})_{i,j=1}^{n}$ mátrixok felső háromszög részei szolgáltatják:

$$\mathbf{D} = \left[\begin{array}{cccc:cccc:c}
a & b & \dots & b & -b & -b & \dots & -b & -b \\
b & a & \dots & b & -b & -b & \dots & -b & -b \\
\vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots \\
b & b & \dots & a & -b & -b & \dots & -b & -b \\
b & b & \dots & b & -b & -b & \dots & -b & -b \\
\hdashline
-b & -b & \dots & -b & 0 & b & \dots & b & b \\
-b & -b & \dots & -b & b & 0 & \dots & b & b \\
\vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots \\
-b & -b & \dots & -b & b & b & \dots & 0 & b \\
-b & -b & \dots & -b & b & b & \dots & b & 0 \\
\hdashline
\multicolumn{4}{c:}{0} & \multicolumn{4}{c:}{0} & 0
\end{array}\right]
\begin{array}{l}
\left.\vphantom{\begin{array}{c}a\\a\\a\\a\\a\end{array}}\right\}I_1 \\
\left.\vphantom{\begin{array}{c}a\\a\\a\\a\\a\end{array}}\right\}I_2 \\
\left.\vphantom{\begin{array}{c}a\\a\\a\\a\\a\end{array}}\right\}I_3
\end{array}$$

$$\underbrace{\qquad\qquad}_{I_1} \qquad \underbrace{\qquad\qquad}_{I_2} \qquad \underbrace{\qquad\qquad}_{I_3}$$

ahol $I_1, I_2, I_3$ az $I=\{1, \dots, n\}$ indexhalmaz tetszőleges részhalmazai úgy, hogy $I_1 \cup I_2 \cup I_3 = I$, $I_1 \cap I_2 = \emptyset$, $I_1 \cap I_3 = \emptyset$, $I_2 \cap I_3 = \emptyset$, és ha $n_1$ és $n_2$ jelöli az $I_1$ és $I_2$ halmazok elemszámát, akkor azokra a (2.7) feltételeknek kell teljesülni. Az $a$ és $b$ számok minden esetben a következők:

$$a = \frac{4}{(n_1-n_2)^2+3n_1-n_2}, \quad b = \frac{a}{2}.$$

*Megjegyzés.* A 4.5. állítás igazolásának egy elvileg lehetséges módja az, hogy ismert algoritmusok (lásd [1[) alkalmazásával számítógéppel keressük meg a (4.5) feltételrendszer által definiált konvex poliéder összes extremális pontját. Ez a módszer eddig szintén csak az $n=2, 3$ és 4 esetekben vezetett eredményre, $n=5$ esetén az összes extremális pont megkeresése (a nyilvánvaló szimmetriák kihasználása mellett is) az általunk használható leggyorsabb számítógépeken is irreálisan sok CPU időt igényel.

## IRODALOM

[1] MANAS, M. and NEDOMA, J., "Finding all vertices of a convex polyhedron", *Numerische Mathematik* **12** (1968) 226—229.
[2] PRÉKOPA, A., *Lineáris programozás I.* (Bolyai János Matematikai Társulat, Budapest, 1968).
[3] PRÉKOPA, A. és SZÁNTAI, T., „Egy új, többdimenziós gamma eloszlás és annak illesztése empirikus adatokhoz", *Alkalmazott Matematikai Lapok* **1** (1975) 299—318.

(Beérkezett: 1983. június 14.)

SZÁNTAI TAMÁS
BME GÉPÉSZMÉRNÖKI KAR MATEMATIKA TANSZÉK
1521 BUDAPEST, STOCZEK U. H ÉP. IV. E. 43.

## AN EFFICIENT ALGORITHM FOR FITTING MULTIVARIATE GAMMA DISTRIBUTION TO EMPIRICAL DATA

T. SZÁNTAI

In this paper necessary conditions are proved for the existence of multivariate gamma distribution to a given empirical covariance matrix. Using these conditions we give an efficient heuristic algorithm for fitting the multivariate gamma distribution.

In a separate section there are published some ideas for proving the sufficiency of our conditions. In the appendix of the paper we give a complete proof for the case of four dimensional gamma distribution.

## FÜGGELÉK

### *A (4.5) konvex poliéder összes extremális pontjának meghatározása $n=4$ esetén*

A 4.5. állítás $n=4$ esetén azt mondja ki, hogy a (4.5) feltételrendszernek a következő **D** mátrixok alkotják az extremális pontjait (a **D** mátrixoknak csak a felső háromszög részeit soroljuk fel $d_{11}, d_{22}, d_{33}, d_{44}, d_{12}, d_{13}, d_{14}, d_{23}, d_{24}, d_{34}$ sorrendben):

$n_1 = 0, \quad n_2 = 2$ *eset:*

$$\begin{matrix} 0&0&0&0&1&0&0&0&0&0 \\ 0&0&0&0&0&1&0&0&0&0 \\ 0&0&0&0&0&0&1&0&0&0 \\ 0&0&0&0&0&0&0&1&0&0 \\ 0&0&0&0&0&0&0&0&1&0 \\ 0&0&0&0&0&0&0&0&0&1 \end{matrix}$$

$n_1 = 1, \quad n_2 = 1$ *eset:*

$$\begin{matrix} 2&0&0&0&-1&0&0&0&0&0 \\ 2&0&0&0&0&-1&0&0&0&0 \\ 2&0&0&0&0&0&-1&0&0&0 \\ 0&2&0&0&-1&0&0&0&0&0 \\ 0&2&0&0&0&0&0&-1&0&0 \\ 0&2&0&0&0&0&0&0&-1&0 \end{matrix} \qquad \begin{matrix} 0&0&2&0&0&-1&0&0&0&0 \\ 0&0&2&0&0&0&0&-1&0&0 \\ 0&0&2&0&0&0&0&0&0&-1 \\ 0&0&0&2&0&0&-1&0&0&0 \\ 0&0&0&2&0&0&0&0&-1&0 \\ 0&0&0&2&0&0&0&0&0&-1 \end{matrix}$$

$n_1 = 1, \quad n_2 = 2$ *eset:*

$$\begin{array}{cccccccc} 2&0&0&0&-1&-1&0&1&0&0 \\ 2&0&0&0&-1&0&-1&0&1&0 \\ 2&0&0&0&0&-1&-1&0&0&1 \\ 0&2&0&0&-1&1&0&-1&0&0 \\ 0&2&0&0&-1&0&1&0&-1&0 \\ 0&2&0&0&0&0&0&-1&-1&1 \end{array} \qquad \begin{array}{cccccccccc} 0&0&2&0&1&-1&0&-1&0&0 \\ 0&0&2&0&0&-1&1&0&0&-1 \\ 0&0&2&0&0&0&0&-1&1&-1 \\ 0&0&0&2&1&0&-1&0&-1&0 \\ 0&0&0&2&0&1&-1&0&0&-1 \\ 0&0&0&2&0&0&0&1&-1&-1 \end{array}$$

$n_1 = 1, \quad n_2 = 3$ *eset:*

$$\begin{array}{cccccccccc} 1&0&0&0&-\frac{1}{2}&-\frac{1}{2}&-\frac{1}{2}&\frac{1}{2}&\frac{1}{2}&\frac{1}{2} \\ 0&1&0&0&-\frac{1}{2}&\frac{1}{2}&\frac{1}{2}&-\frac{1}{2}&-\frac{1}{2}&\frac{1}{2} \end{array} \qquad \begin{array}{cccccccccc} 0&0&1&0&\frac{1}{2}&-\frac{1}{2}&\frac{1}{2}&-\frac{1}{2}&\frac{1}{2}&-\frac{1}{2} \\ 0&0&0&1&\frac{1}{2}&\frac{1}{2}&-\frac{1}{2}&\frac{1}{2}&-\frac{1}{2}&-\frac{1}{2} \end{array}$$

$n_1 = 2, \quad n_2 = 2$ *eset:*

$$\begin{array}{cccccccccc} 1&1&0&0&\frac{1}{2}&-\frac{1}{2}&-\frac{1}{2}&-\frac{1}{2}&-\frac{1}{2}&\frac{1}{2} \\ 1&0&1&0&-\frac{1}{2}&\frac{1}{2}&-\frac{1}{2}&-\frac{1}{2}&\frac{1}{2}&-\frac{1}{2} \\ 1&0&0&1&-\frac{1}{2}&-\frac{1}{2}&\frac{1}{2}&\frac{1}{2}&-\frac{1}{2}&-\frac{1}{2} \end{array} \qquad \begin{array}{cccccccccc} 0&1&1&0&-\frac{1}{2}&-\frac{1}{2}&\frac{1}{2}&\frac{1}{2}&-\frac{1}{2}&-\frac{1}{2} \\ 0&1&0&1&-\frac{1}{2}&\frac{1}{2}&-\frac{1}{2}&-\frac{1}{2}&\frac{1}{2}&-\frac{1}{2} \\ 0&0&1&1&\frac{1}{2}&-\frac{1}{2}&-\frac{1}{2}&-\frac{1}{2}&-\frac{1}{2}&\frac{1}{2} \end{array}$$

Az állítás bizonyításához egészítsük ki a (4.5) feltételrendszert $u_{12}, u_{13}, \ldots, u_{1234}$ nemnegatív segédváltozókkal, és keressük a keletkező lineáris egyenletrendszer megengedett bázismegoldásait, melyeknek csak a $d$ komponensei lesznek számunkra érdekesek. A segédváltozókkal kiegészített feltételrendszer a következő alakot ölti:

$$\begin{array}{rcl}
d_{11}+d_{22}+2d_{12}-u_{12} & = & 0 \\
d_{11}+d_{33}+2d_{13}-u_{13} & = & 0 \\
d_{11}+d_{44}+2d_{14}-u_{14} & = & 0 \\
d_{22}+d_{33}+2d_{23}-u_{23} & = & 0 \\
d_{22}+d_{44}+2d_{24}-u_{24} & = & 0 \\
d_{33}+d_{44}+2d_{34}-u_{34} & = & 0 \\
d_{11}+d_{22}+d_{33}+2d_{12}+2d_{13}+2d_{23}-u_{123} & = & 0 \\
d_{11}+d_{22}+d_{44}+2d_{12}+2d_{14}+2d_{24}-u_{124} & = & 0 \\
d_{11}+d_{33}+d_{44}+2d_{13}+2d_{14}+2d_{34}-u_{134} & = & 0 \\
d_{22}+d_{33}+d_{44}+2d_{23}+2d_{24}+2d_{34}-u_{234} & = & 0 \\
d_{11}+d_{22}+d_{33}+d_{44}+2d_{12}+2d_{13}+2d_{14}+2d_{23}+2d_{24}+2d_{34}-u_{1234} & = & 0 \\
d_{11}+d_{22}+d_{33}+d_{44}+d_{12}+d_{13}+d_{14}+d_{23}+d_{24}+d_{34} & = & 1
\end{array}$$

$$d_{11} \geqq 0, \quad d_{22} \geqq 0, \quad d_{33} \geqq 0, \quad d_{44} \geqq 0$$

$$u_{12} \geqq 0, \quad u_{13} \geqq 0, \quad u_{14} \geqq 0, \quad u_{23} \geqq 0, \quad u_{24} \geqq 0, \quad u_{34} \geqq 0, \quad u_{123} \geqq 0, \quad u_{124} \geqq 0,$$

$$u_{134} \geqq 0, \quad u_{234} \geqq 0, \quad u_{1234} \geqq 0.$$

Az állítás bizonyítását néhány nyilvánvaló megjegyzéssel kezdjük:

*1. Megjegyzés.* Az 1, 2, 3, 4 indexekre vonatkozó teljes szimmetria miatt elég csak például azokat az extremális pontokat összeszámolni, amelyekben minden $d_{ii}$ változó nulla értékű, amelyekben pontosan egy $d_{ii}$ változó (pl. $d_{11}$) nem nulla értékű, ..., végül, amelyekben $d_{11}, d_{22}, d_{33}, d_{44}$ egyike sem nulla értékű. (Ezeket a felsorolásban vastagon szedett számokkal jelöltük.)

*2. Megjegyzés.* Ha az állításban szereplő feltételrendszerből töröljük az összes olyan $d_{ij}$ változót, amelyben $i$ vagy $j$ (akár mind a kettő) $k$-val egyenlő ($k=1, 2, 3$ vagy 4), akkor a feltételrendszer az $n=3$ esetnek megfelelő feltételrendszerre redukálódik úgy, hogy annak az utolsótól eltekintve minden feltételét megkettőzve szolgáltatja. Például $k=4$ esetén, ha külön választjuk azokat a feltételeket, amelyekben az $u$ változó indexében szerepel a 4-es és amelyekben nem, akkor külön-külön megkapjuk az $n=3$ esetre vonatkozó feltételrendszert, eltekintve az utolsó feltételtől, amely továbbra is csak egyszer és $u$ változó nélkül szerepel.

E megjegyzés értelmében, ha az állítás bizonyítása során ki tudjuk mutatni, hogy minden olyan $d_{ij}$ változónak nulla szinten kell lennie, amelyben a $k$ index szerepel, akkor felhasználhatjuk az $n=3$ esetre vonatkozó állítást, amely bizonyítása triviális.

*3. Megjegyzés.* Ismert tétel szerint (lásd [2]) a feltételrendszerünk által leírt konvex poliéder összes csúcsát azok a bázismegoldások adják, amelyekben bázisváltozóként a $d_{12}, d_{13}, d_{14}, d_{23}, d_{24}, d_{34}$ változók mindegyike szerepel, minthogy ezek nemnegativitással nem korlátozott változók.

*4. Megjegyzés.* A vizsgálandó lineáris egyenletrendszerünk nyilvánvalóan teljes rangú, hiszen az $u$-val jelölt változók és bármely $d$-vel jelölt változó oszlopvektorát tekintve, egy 12 elemű lineárisan független vektorrendszert nyerünk.

Ezek után az állítás bizonyítását négy lépésben fogjuk elvégezni.

*1. Lépés.* Tekintsük azokat a bázismegoldásokat, amelyekben *a* $d_{11}, d_{22}, d_{33}, d_{44}$ *változók mindegyike nulla szinten van.* Ekkor az első hat feltétel szerint a $d_{ij}$, $i \neq j$ változók mindegyike csak nemnegatív szinten szerepelhet a bázismegoldásban. Ha ezek közül pontosan egy szerepel pozitív szinten, akkor a 2. megjegyzés értelmében csak az $n=3$ esetre levezetett csúcsok adódhatnak, ezek pedig az állításban mind fel is lettek sorolva. Ha pedig a $d_{ij}$, $i \neq j$ változók közül egynél több is pozitív szinten szerepelne, akkor az nem lehetne bázismegoldás, hiszen minden olyan $u$ változónak is a bázisban kellene lenni, amelyik ezen $d_{ij}$ változók előfordulási sorában szerepel, ezek azonban akkor lineárisan összefüggő oszlopvektorokhoz tartoznának, mert két ilyen $d_{ij}$ változó oszlopvektorának a különbsége triviálisan kifejezhető lenne a bázisbeli $u$ változókhoz tartozó oszlopvektorok lineáris kombinációjaként.

*2. Lépés.* Tekintsük azokat a bázismegoldásokat, amelyekben *pontosan egy* $d_{ii}$ *változó, legyen ez* $d_{11}$, *szerepel pozitív szinten.* Ekkor a 4—5—6. feltételek szerint a $d_{23}, d_{24}, d_{34}$ változók csak nemnegatív szinten szerepelhetnek a bázismegoldásban. Másrészt a $d_{12}, d_{13}, d_{14}$ változók közül legalább egynek negatív szinten kell a bázismegoldásban szerepelni, ha ugyanis egyik sem szerepelne negatív szinten, akkor a $d_{11}$ változó minden előfordulási helyéhez tartozó $u$ változónak is pozitívnak, azaz bázisváltozónak kellene lenni, minthogy viszont a $d_{12}, d_{13}, d_{14}$ változók csak $d_{11}$ előfor-

dulási helyén jelenhetnek meg, azért $d_{11}$ és a $d_{12}$, $d_{13}$, $d_{14}$ változók bármelyike oszlopvektorának a különbsége előállítható lenne bázisbeli $u$ változók oszlopvektorainak a lineáris kombinációjaként, vagyis a megoldás nem lenne bázismegoldás. A továbbiakban hajtsunk végre esetszétbontást aszerint, hogy hány változó szerepel a $d_{12}$, $d_{13}$, $d_{14}$ változók közül negatív szinten a bázismegoldásban.

a) *Ha a $d_{12}$, $d_{13}$, $d_{14}$ változók közül pontosan egy szerepel negatív szinten a bázismegoldásban,* akkor a változók teljes szimmetriája miatt elég pl. a $d_{12}<0$ esetet vizsgálni. Ekkor minden olyan $u$ változónak pozitív szinten kell szerepelnie a bázismegoldásban, amely $d_{11}$ előfordulási sorában van úgy, hogy $d_{12}$ nem fordul elő abban a sorban (ezek most $u_{13}$, $u_{14}$ és $u_{134}$). Minthogy ezen $u$ változók oszlopvektorainak lineáris kombinációjaként triviálisan előállítható a $d_{11}$ és $d_{12}$ változók oszlopvektorának a különbsége (eltekintve az utolsó elemtől), azért bármilyen további $u$ bázisváltozók mellett is a bázismegoldást szolgáltató egyenletrendszer *Cramer-szabály* szerinti megoldására gondolva könnyen látható, hogy a $d_{13}$, $d_{23}$, $d_{24}$, $d_{34}$ változók mindegyike nulla értékűnek kell, hogy adódjon (természetesen fel kell tételezni, hogy a további $u$ változókkal bázist kapunk). Ismét a 2. megjegyzés értelmében csak az $n=3$ esetre levezetett csúcsok adódhatnak, ezek pedig az állításban mind fel lettek sorolva.

b) *Ha a $d_{12}$, $d_{13}$, $d_{14}$ változók közül pontosan kettő szerepel negatív szinten a bázismegoldásban,* akkor ismét a változók teljes szimmetriája miatt elég pl. a $d_{12}<0$, $d_{13}<0$ esetet vizsgálni. Ekkor minden olyan $u$ változónak pozitív szinten kell lenni a bázismegoldásban, amely $d_{11}$ előfordulási sorában van úgy, hogy $d_{12}$ és $d_{13}$ nem fordul elő abban a sorban (ez most egyedül az $u_{14}$ változó). Másrészt az első két feltétel szerint $d_{12} \geqq -\frac{1}{2} d_{11}$ és $d_{13} \geqq -\frac{1}{2} d_{11}$, és nyilván $d_{12}=d_{13}$ a változók teljes szimmetriája miatt. Ha azonban $d_{12}=d_{13} > -\frac{1}{2} d_{11}$ lenne, akkor pozitív szinten kellene lenni minden olyan $u$ változónak, amely olyan sorban van, hogy benne a $d_{12}$ és a $d_{13}$ változók közül csak az egyik fordul elő. Ekkor viszont a megfelelő oszlopvektorok nyilván lineárisan összefüggők lesznek, hiszen a pozitív szinten levő $u$ változók oszlopvektoraival $d_{12}$ és $d_{13}$ oszlopvektora is olyanra redukálható, hogy benne csak az egyszerre nem nulla (és egymással egyenlő) komponensek maradnak nullától különbözők. Így tehát a megoldás nem lenne bázismegoldás, vagyis kell, hogy $d_{12}=d_{13}=-\frac{1}{2} d_{11}$ legyen. Ekkor viszont minden olyan sorban, ahol $d_{12}$ és $d_{13}$ egyszerre fordul elő, kell hogy egy pozitív szinten levő $d_{ij}$, $i \neq j$ változó is legyen. Ez csak a $d_{23}$ változó lehet. Ez viszont maga után vonja, hogy az $u_{23}$ és $u_{234}$ változóknak is pozitív szinten kell lenni, vagyis bázisváltozónak kell lenni. Ez utóbbi $u$ változók oszlopvektorával azonban $d_{23}$ oszlopvektora éppen olyanná redukálható, hogy benne csak azokon a helyeken legyen nem nulla elem, ahol $d_{12}$ és $d_{13}$ oszlopában közösen van nem nulla elem, így ezen redukált oszlopvektorral, $d_{12}$ és $d_{13}$ oszlopvektorával, valamint $u_{14}$ oszlopvektorával $d_{11}$ oszlopvektora előállítható, eltekintve az utolsó elemtől. Ezért ismét bármilyen további (de bázist definiáló) $u$ bázisváltozók mellett is a bázismegoldást szolgáltató egyenletrendszer *Cramer-szabály* szerinti megoldására gondolva könnyen látható, hogy a $d_{14}$, $d_{24}$, $d_{34}$ változók mindegyike nulla értékűnek kell hogy adódjon. Most is a 2. megjegyzés értelmében tehát csak az $n=3$ esetre levezetett csúcsok adódhatnak, ezek pedig az állításban fel is lettek sorolva.

c) *Ha a* $d_{12}$, $d_{13}$, $d_{14}$ *változók mindegyike negatív szinten szerepel a bázismegoldásban,* akkor az első három feltétel szerint $d_{12} \geqq -\frac{1}{2} d_{11}$, $d_{13} \geqq -\frac{1}{2} d_{11}$, $d_{14} \geqq$ $\geqq -\frac{1}{2} d_{11}$ és a változók teljes szimmetriája miatt $d_{12} = d_{13} = d_{14}$. Ha azonban $-\frac{d_{11}}{2} < d_{12} = d_{13} = d_{14} < -\frac{d_{11}}{4}$ lenne, akkor az $u_{12}$, $u_{13}$, $u_{14}$ változók pozitív szinten kellene, hogy legyenek, de ugyanakkor $d_{23}$, $d_{24}$, $d_{34}$ is pozitív szinten kell hogy legyen, ami az $u_{23}$, $u_{24}$ és $u_{34}$ változók pozitívitását is maga után vonja. Ez azonban a $d_{11}$ és a $d_{ij}$, $i \neq j$ változók mellett hat további $u$ bázisváltozót jelentene, ami összesen 13 bázisváltozót adna, ez pedig lehetetlen, hiszen a feltételek száma csak 12. Ha $d_{12} =$ $= d_{13} = d_{14} = -\frac{d_{11}}{4}$ lenne, akkor az $u_{12}$, $u_{13}$ és $u_{14}$ változók továbbra is pozitív szinten kellene, hogy legyenek, emellett még a $d_{23}$, $d_{24}$ és $d_{34}$ változók egyikének is pozitívnak kellene lenni, de közülük bármelyik lenne is pozitív, az további három $u$ változó pozitivitását vonná maga után (pl. $d_{23} > 0$ esetén az $u_{23}$, $u_{123}$ és $u_{234}$ változókét), ami ismét 13 bázisváltozót jelentene, ez pedig lehetetlen. Ha pedig $d_{12} = d_{13} = d_{14} > -\frac{d_{11}}{4}$ lenne, akkor az $u_{12}$, $u_{13}$, $u_{14}$, $u_{123}$, $u_{124}$, $u_{134}$ változóknak kellene pozitívnak lenni, ez pedig most is 13 bázisváltozót jelentene, ami lehetetlen. Marad tehát az az eset, hogy $d_{12} = d_{13} = d_{14} = -\frac{d_{11}}{2}$. Ekkor viszont a $d_{23}$, $d_{24}$, $d_{34}$ változók mindegyike pozitív szinten kell hogy szerepeljen a bázismegoldában. Ez pedig maga után vonja azt, hogy az $u_{23}$, $u_{24}$, $u_{34}$, $u_{234}$, $u_{1234}$ változóknak is mind pozitív szinten kell lenni, és így egyértelműen az erre az esetre felsorolt $1\,0\,0\,0 - \frac{1}{2} - \frac{1}{2} - \frac{1}{2}\,\frac{1}{2}\,\frac{1}{2}\,\frac{1}{2}$ csúcsot nyerjük.

*3. Lépés.* Tekintsük most azokat a bázismegoldásokat, amelyekben *pontosan két* $d_{ii}$ *változó, legyenek ezek* $d_{11}$ *és* $d_{22}$, *szerepel pozitív szinten.* Ekkor a 2, 3, 4 és 5. feltételek szerint, valamint a változók szimmetriája miatt $d_{13} = d_{14} = d_{23} = d_{24} \geqq -\frac{d_{11}}{2} =$ $= -\frac{d_{22}}{2}$. Ha $-\frac{d_{11}}{2} < d_{13} = d_{14} = d_{23} = d_{24} < -\frac{d_{11}}{4}$ lenne, akkor $u_{13}$, $u_{14}$, $u_{23}$ és $u_{24}$ pozitív szinten lenne, de a 9.10. feltételek miatt $d_{34}$-nek is pozitív szinten kellene lenni, és így $u_{34}$-nek is, ami a 8 darab $d$ változó mellett 5 darab $u$ változó bázisváltozó voltát jelentené, ez pedig lehetetlen. Ha viszont $d_{13} = d_{14} = d_{23} = d_{24} = -\frac{d_{11}}{4}$ lenne, akkor $u_{13}$, $u_{14}$, $u_{23}$, $u_{24}$ továbbra is pozitív szinten lenne és a 7, 8 feltétel szerint vagy $u_{123} > 0$ lenne, vagy $d_{12} < 0$ kellene, hogy legyen. Az első eset nem lehetséges, mert ismét öt lenne a pozitív szinten levő $u$ változók száma, ha pedig $d_{12} < 0$ lenne, akkor a 11. feltétel szerint $d_{34}$ pozitív kellene, hogy legyen, ekkor viszont $u_{34}$ lenne az ötödik olyan $u$ változó, amelynek pozitív szinten kellene lenni. Ha pedig $d_{13} = d_{14} = d_{23} = d_{24} > -\frac{d_{11}}{4}$ lenne, akkor $u_{13}$, $u_{14}$, $u_{23}$ és $u_{24}$ mellett az $u_{134}$ változónak is mindig pozitívnak kellene lennie, ami ismét lehetetlen. Marad tehát az az eset, hogy $d_{13} = d_{14} = d_{23} = d_{24} =$

$= -\frac{d_{11}}{2} = -\frac{d_{22}}{2}$, mikoris a 9, 10. feltételekből következik, hogy a $d_{34}$ változónak pozitívnak kell lenni. A 8, 9. feltételekből látható, hogy $d_{12}$ negatív nem lehet. Ha $d_{12}$ nulla értékű lenne, akkor $d_{34}=d_{11}$ kellene, hogy legyen és így $u_{12}$, $u_{34}$, $u_{134}$ és $u_{234}$ pozitív lenne, viszont könnyen ellenőrizhető, hogy ezekkel az $u$ változókkal nem kapunk bázist. Tehát kell hogy $d_{12}$ pozitív legyen, ekkor viszont az $u_{12}$, $u_{34}$, $u_{123}$ és $u_{124}$ változóknak is pozitívnak kell lenni, és így egyértelműen az erre az esetre felsorolt $1\ 1\ 0\ 0\ \frac{1}{2}\ -\frac{1}{2}\ -\frac{1}{2}\ -\frac{1}{2}\ -\frac{1}{2}\ \frac{1}{2}$ csúcsot nyerjük.

*4. Lépés.* Ha *kettőnél több $d_{ii}$ változó van pozitív szinten a bázisban,* akkor még ha egyetlen $d_{ij}$ változó negativitásával egy teljes $d_{ii}$ pozitivitását meg tudnánk szüntetni, akkor is túl sok $u$ változónak kellene pozitív szinten lenni az egyenlőségek fenntartásához. Ezért egyetlen bázismegoldás sem adódhat ilyen feltételek mellett.

*Alkalmazott Matematikai Lapok* **10** (1984) 61—63

# EGY MEGJEGYZÉS PROGRAMOK MAGNYELVEIRŐL

ÉSIK ZOLTÁN

Szeged

Kimutatásra került [1]-ben, hogy a számlálásos ciklusokból felépülő URM programok magnyelvei környezetfüggő nyelvek, vagyis az $\mathscr{L}_1$ nyelvosztályba esnek. E dolgozat célja annak igazolása, hogy ez az állítás tetszőleges URM programra is igaz.

Az URM program (továbbiakban röviden program) fogalma bevezetésre került [1]-ben. Itt egy, a vizsgálatainkat nem érintő specializálást hajtunk végre a program fogalmán. E célból rögzítsük a változóknak és címkéknek egy-egy megszámlálhatóan végtelen $X=\{x_1, x_2, \ldots\}$, illetve $C=\{c_1, c_2, \ldots\}$ halmazát, $X\cap C=\varnothing$. Tetszőleges $n\geqq 0$ egész szám esetén jelöljük $X_n$-nel az első $n$ változó, $C_n$-nel pedig az első $n$ címke halmazát. Programon egy $P=(X_k, X_l, C_n, \alpha)$ rendszert fogunk érteni, ahol $X_k$ a $P$ bemenő változóinak a halmaza; $X_l$ az $\alpha$ utasításrészben előforduló változók halmazának és $X_k$-nak az egyesítése; $\alpha$, mint már említettük, a $P$ utasításainak sorozata. Megköveteljük azt is, hogy az $\alpha$-ban előforduló utasítások sorrendben a $c_1, \ldots, c_n$, $n>0$ címkékkel legyenek címkézve. A programok jelölésére itt bevezetett formalizmustól eltekintve a jelölésmód [1]-et követi.

Legyen $P$ az előzőekben definiált program. A $P$-hez tartozó $\overline{V}_P$ nyelv egy, a $C_n\cup X_l\cup\{1\}$ halmaz feletti nyelv lesz, feltesszük, hogy $1\notin C\cup X$. $\overline{V}_P$ az összes olyan $x_1 1^{u_1}\ldots x_l 1^{u_l} w x_1 1^{v_1}\ldots x_l 1^{v_l}$ alakú jelsorozatból áll, amelyben $u_i$ és $v_i$ $(i=1, \ldots, l)$ nemnegatív egészek, $w\in C_n^*$, továbbá, ha az $x_i$ változóknak rendre az $u_i$ kezdőértékeket adjuk, úgy $P$ futása közben a vezérlés $|w|$ lépésen keresztül a $w$ címkesorozatnak megfelelő utasításokon keresztül halad át, és ezen utasítások végrehajtása után a változók értékét rendre a $v_i$ egész számok adják. Ha még azt is kikötjük, hogy $u_{k+1}=\ldots$ $\ldots=u_l=0$ teljesüljön, és hogy a $w$-hez tartozó utasítások végrehajtása után a program futása befejeződjön, úgy a $V_P$ nyelvhez jutunk. Világos, hogy $V_P\subseteqq\overline{V}_P$.

A $P$ programhoz hozzárendelünk egy $C_P(\subseteqq C_n^*)$ nyelvet is. A $C_P$-t alkotó jelsorozatok úgy állnak elő a $V_P$ elemeiből, hogy kitöröljük belőlük az $X_l\cup\{1\}$-beli jelek előfordulásait.

1. ÁLLÍTÁS. Legyen $P=(X_k, X_l, C_n, \alpha)$ tetszőleges program, $x_1 1^{u_1}\ldots x_l 1^{u_l} w x_1 1^{v_1}\ldots$ $\ldots x_l 1^{v_l}\in V_P$. Legyen továbbá $u_i'=\min\{u_i, |w|\}$ $(i=1, \ldots, l)$. Akkor léteznek olyan $v_1', \ldots, v_l'$ nemnegatív egészek, amelyekre $x_1 1^{u_1'}\ldots x_l 1^{u_l'} w x_1 1^{v_1'}\ldots x_l 1^{v_l'}\in V_P$ teljesül.

*Bizonyítás.* Jelöljük $w'$-vel azon $C_n$ feletti véges vagy végtelen sorozatot, amelyhez tartozó utasítássorozaton megy át a vezérlés $P$-ben a változók $x_1=u_1', \ldots, x_l=u_l'$ kezdeti értéke mellett. Meg kell mutatnunk, hogy $w=w'$. Indirekt módon tegyük fel, hogy $w\neq w'$. Jelölje $w_0$ a $w$ és $w'$ leghosszabb közös kezdőszeletét, és legyenek $z_i$,

illetve $z_i'$ $(i=1, \ldots, l)$ azon egyértelműen meghatározott egész számok, amelyekre

$$x_1 1^{u_1} \ldots x_l 1^{u_l} w_0 x_1 1^{z_1} \ldots x_l 1^{z_l} \in \overline{V}_P \quad \text{és} \quad x_1 1^{u_1'} \ldots x_l 1^{u_l'} w_0 x_1 1^{z_1'} \ldots x_l 1^{z_l'} \in \overline{V}_P.$$

Biztos, hogy $w_0 \neq \lambda$. Legyen $c$ a $w_0$-ban előforduló utolsó címke. Világos, hogy a $c$ címkéjű utasítás $BRx_i, c'$ alakú valamely $x_i \in X_l$ változóra és $c'$ címkére, továbbá vagy $z_i > 0$ és $z_i' = 0$ vagy fordítva, $z_i = 0$ és $z_i' > 0$.

Bontsuk fel a $w_0$ jelsorozatot $w_0 = w_1 d_1 \ldots w_r d_r w_{r+1} c$ $(r > 0)$ alakra úgy, hogy $d_j$: $RW y_j, y_{j-1} \in \alpha$ teljesül minden $j \in \{1, \ldots, r\}$ indexre, — vagyis a $d_j$ címkéjű utasítás $RW y_j, y_{j-1}$ alakú $\alpha$-ban —, ahol $y_0, \ldots, y_r$ változók, továbbá $y_r = x_i$ és $w_{j+1}$ egyetlen $j \in \{0, \ldots, r\}$ indexre sem tartalmaz $RW y_j, y$ alakú utasítást. $w_0$-nak ez a felbontása mindig létezik, és egyértelmű. Ha $y_0 \notin \{x_j | u_j \neq u_j'\}$ akkor $z_i = z_i'$, amiből viszont az következik, $w$-nek és $w'$-nek van $w_0$-nál hosszabb közös kezdőszelete is, vagy $w = w_0 = w'$. Tehát $y_0 \in \{x_j | u_j \neq u_j'\}$, mondjuk $y_0 = x_{j_0}$. Ez viszont $z_i \geqq u_{j_0} - |w_1 \ldots w_{r+1}|$, $z_i' \geqq u_{j_0}' - |w_1 \ldots w_{r+1}|$, továbbá $u_{j_0}' = |w| < u_{j_0}$ folytán azt eredményezi, hogy $z_i, z_i' > 0$. Ez ismét ellentmondás.

2. ÁLLÍTÁS. Tetszőleges $P = (X_k, X_l, C_n, \alpha)$ programra és $x_1 1^{u_1} \ldots x_l 1^{u_l} w x_1 1^{v_1} \ldots$ $\ldots x_l 1^{v_l} \in \overline{V}_P$ jelsorozatra érvényes az $|x_1 1^{v_1} \ldots x_l 1^{v_l}| \leqq l |x_1 1^{u_1} \ldots x_l 1^{u_l} w|$ egyenlőtlenség.

*Bizonyítás.* Világos, hogy $\max\{v_1, \ldots, v_l\} \leqq |w| + \max\{u_1, \ldots, u_l\}$. Ezért

$$|x_1 1^{v_1} \ldots x_l 1^{v_l}| \leqq l(1 + \max\{v_1, \ldots, v_l\}) \leqq$$

$$\leqq l(1 + \max\{u_1, \ldots, u_l\} + |w|) \leqq l(|x_1 1^{u_1} \ldots x_l 1^{u_l}| + |w|) = l|x_1 1^{u_1} \ldots x_l 1^{u_l} w|.$$

Az előző két állítás segítségével bebizonyítható a következő tétel. Ebben a tételben [1]-gyel összhangban $L_P$ a $P$-hez tartozó magnyelvet jelöli tetszőleges $P$ program esetén.

TÉTEL. A $V_P$, $C_P$ és $L_P$ nyelvek mindegyike könyezetfüggő.

*Bizonyítás.* Legyen $P$ az $(X_k, X_l, C_n, \alpha)$ program. A $V_P \in \mathscr{L}_1$ tartalmazás bizonyítása végett tervezzünk egy $l$ munkaszalaggal rendelkező *Turing-gépet*, amely tetszőleges $(C_n \cup X_l \cup \{1\})^*$-beli bemenő elsorozat esetén

a) eldönti, hogy ez $x_1 1^{u_1} \ldots x_k 1^{u_k} x_{k+1} \ldots x_l w x_1 1^{v_1} \ldots x_l 1^{v_l}$ alakú-e, ahol $u_1, \ldots, u_k$, $v_1, \ldots, v_l$ nemnegatív egészek,

b) átmásolja a változók kezdőértékeit munkaszalagjaira,

c) munkaszalagjain a változók pillanatnyi értékeit tárolva szimulálja a $w$ elemeivel címkézett utasítássorozat végrehajtásának hatását, figyelemmel kísérve azt is, hogy ténylegesen ezen utasításokon keresztül halad-e a vezérlés $P$-ben a változók adott kezdőértéke mellett,

d) eldönti, hogy a c)-ben említett utasítások végrehajtása után a program végrehajtása befejeződött-e,

e) végül eldönti, hogy a változók végértéke a munkaszalagokon rendre a $v_1, \ldots, v_l$ számokkal egyenlő vagy nem.

Ha az a) c) d) és e) valamelyikében nemleges válasz adódik, akkor az adott bemenő jelsorozat nem eleme a $V_P$ nyelvnek, különben igen.

A 2. állítás szerint egy ilyen *Turing-gép* úgy is megtervezhető, hogy működése közben a munkaszalagokon összesen is legfeljebb a bemenő jelsorozat hosszának $l$-szere-

sét használja fel segéd-memóriaként. Ezért $V_P \in \mathcal{L}_1$. (Sőt, $V_P$ determinisztikus környezetfüggő nyelv).

A $C_P \in \mathcal{L}_1$ tartalmazás igazolása végett idézzük vissza, hogy $C_P$ a $V_P$ homomorf képe azon $\varphi$: $(C_n \cup X_l \cup \{1\})^* \to C_n^*$ homomorfizmus mellett, amely a $\varphi(c)=c$ $(c \in C_n)$, $\varphi(x)=\lambda$ $(x \in X_l \cup \{1\})$ egyenlőségekkel van definiálva. Az 1. és 2. állításokból következik, hogy van olyan $K(\geqq 0)$ konstans, hogy minden $w \neq \lambda$, $w \in C_P$-hez létezik olyan $w' \in V_P$, hogy $\varphi(w')=w$ és $|w'| \leqq K|w|$, vagyis $\varphi$ gyengén $K$-lineáris $V_P$-re nézve. Belátható azonban az, hogy környezetfüggő nyelv gyengén $K$-lineáris homomorf képe környezetfüggő. (Ezen állítás egy némiképp gyengébb alakjára nézve l. [2]). Ezért $V_P \in \mathcal{L}_1$-ből $C_P \in \mathcal{L}_1$ következik.

Az $L_P \in \mathcal{L}_1$ tartalmazás hasonlóan bizonyítható. Világos ugyanis, hogy $L_P$ vagy $L_P - \{\lambda\}$ gyengén $2n-1$-lineáris homomorf képe a $C_P$ nyelvnek.

## IRODALOM

[1] DÖRNYEI, Á., „Számlálásos ciklusokból felépíthető programok magnyelveinek szerkezetéről", *Alk. Mat. Lapok* 7 (1981) 287—309.
[2] SALOMAA, A., *Formal Languages* (Academic Press, New York and London, 1973).

(Beérkezett: 1983. április 22.)

(Átdolgozva beérkezett: 1983. augusztus 30.)

ÉSIK ZOLTÁN
JATE BOLYAI INTÉZET
6720 SZEGED, ARADI VÉRTANÚK TERE 1.

## A REMARK ON THE KERNEL LANGUAGES OF PROGRAMS

Z. ÉSIK

It was shown in [1] that the kernel languages of certain URM programs are context sensitive languages. In this paper we note that this result holds for arbitrary URM programs, as well.

*Alkalmazott Matematikai Lapok* **10** (1984) 65—71

# PARAMÉTERES OPTIMALIZÁLÁSI FELADATOK EGY OSZTÁLYÁNAK MEGOLDÁSA

GERGÓ LAJOS

Budapest

Cikkünkben paraméteres optimalizálási feladatok olyan típusával foglalkozunk, amelyben a feltételrendszer jobb oldalán tetszőleges nemlineáris függvények állnak, az együtthatómátrix bizonyos elemei speciális lineáris függvények és a feltételek teljesülését egy-egy intervallumon követeljük meg. Megmutatjuk, hogy a feladat egy geometriai problémából származtatható. Bizonyítjuk, hogy a feladatnak létezik megoldása, majd a geometriai sajátosságokat kihasználva belátjuk, hogy jól közelíthető egy *LP*-feladattal. Részproblémaként véges sok pont konkáv burkának meghatározására adunk egy algoritmust, és megadjuk az algoritmus műveletigényének felső korlátját.

## 1. Bevezetés

Az említett geometriai probléma kapcsolatban áll bizonyos geodéziai mérésekkel, amikor egy terepen adottak a megfigyelőállomások. Ezek közül néhányból át kell tudnunk látni adott másikakba, ezért a terepviszonyoknak megfelelően tornyokat építünk, hogy ezt megvalósíthassuk. Célunk, hogy az építkezési összmagasság minimális legyen.

## 2. A feladat megfogalmazása, a megoldás létezése

Tekintsük a következő feladatot:

$$(P)\begin{cases} l_1(t^{12})z_1 - l_2(t^{12})z_2 \geqq f_{12}(t^{12}) \quad t^{12}\in[a_1, a_2] \\ l_1(t^{13})z_1 - l_3(t^{13})z_3 \geqq f_{13}(t^{13}) \quad t^{13}\in[a_1, a_3] \\ \vdots \\ l_i(t^{ij})z_i - l_j(t^{ij})z_j \geqq f_{ij}(t^{ij}) \quad t^{ij}\in[a_i, a_j] \quad 1 \leqq i < j \leqq N \\ \vdots \\ l_{N-1}(t^{N-1,N})z_{N-1} - l_N(t^{N-1,N})z_N \geqq f_{N-1,N}(t^{N-1,N}) \quad t^{N-1,N}\in[a_{N-1}, a_N] \\ \min \sum_{i=1}^{N} z_i \end{cases}$$

ahol

$l_i(t^{ij}) = t^{ij} - a_i \quad l_j(t^{ij}) = t^{ij} - a_j \quad z_i \in \mathbf{R}$

$[a_i, a_j]$ az $a_i$, $a_j$ számok által kifeszített zárt intervallum

$f_{ij} \in H := \{f \in LC \cap K \wedge f(a_f) = f(b_f) = 0\}$

Itt $LC$ a *Lipschitz-folytonos függvények* halmaza, $K$ a korlátosaké, $[a_f, b_f]$ az $f$ értelmezési tartománya. (Tehát függvényenként lehet más és más.)

Megmutatjuk, hogy ez a feladat mindig egy geometriai problémából származtatható. Bizonyítjuk a megoldás létezését, majd megkonstruálunk egy a megoldás közelítésére szolgáló numerikus algoritmust.

Mi csak ezzel a teljes feladattal foglalkozunk — tehát amikor minden $1 \leqq i < j \leqq N$ indexpár előfordul —, de megjegyezzük, hogy nem kell az összes feltételnek szerepelnie. Ekkor lehetnek olyan indexek, amelyek nem szerepelnek egyetlen feltételben sem, így a feladat rendje csökkenthető, vagy pedig kettő vagy több kisebb rendű feladatra esik szét.

Megmutatjuk, hogy $(P)$ a következő geometriai problémából származtatható. (Erre $(G)$-vel hivatkozunk a későbbiekben.)

Legyen adva derékszögű koordinátarendszerben egy $\Sigma \subset \mathbf{R}^3$ felület, amelyet egy $F \in LC \cap K$ kétváltozós függvény ír le. Tűzzünk ki ezen $\mathbf{p}_1, \ldots, \mathbf{p}_N \in \Sigma$ egymástól különböző pontokat, jelöljük továbbá $\mathbf{p}_i'$-vel azon pontokat, amelyeket úgy kapunk, hogy $\mathbf{p}_i$-ket felemeljük a $z$-tengellyel párhuzamosan $\Sigma$ fölé $z_i$ értékekkel. Jelöljük minden $1 \leqq i < j \leqq N$ párra $e_{ij}$-vel a $\mathbf{p}_i'$, $\mathbf{p}_j'$ pontokat összekötő szakaszt, $f_{ij}$-sal pedig $F$ leszűkítését az $e_{ij}$ szakasz $x-y$ síkra való merőleges vetületére.

Feladatunk minimalizálni a $z_i$ szakaszok összegét azon feltétel mellett, hogy minden szakasz a $\Sigma$ fölött maradjon.

Ahhoz, hogy $(P)$ a $(G)$-ből származtatható, elég belátni, hogy az $\tilde{f}_{ij} = \dfrac{f_{ij}}{|a_i - a_j|}$ függvények egy $F \in LC \cap K$ kétváltozós függvény valamilyen irányú metszetei.

Ha ez igaz, akkor felírva a $(G)$-beli feltételeket az

$$e_{ij}(t^{ij}) \geqq \tilde{f}_{ij}(t^{ij}) \quad t^{ij} \in [a_i, a_j]$$

rendszert kapjuk, ahol

$$e_{ij}(t^{ij}) = \frac{z_i - z_j}{a_i - a_j}(t^{ij} - a_j) + z_j.$$

Előzőt átrendezve kapjuk a $(P)$ feltételrendszerét

$$z_i(t^{ij} - a_j) - (t^{ij} - a_i) z_j \geqq |a_i - a_j| \tilde{f}_{ij}(t^{ij}) = f_{ij}(t^{ij}).$$

Bebizonyítjuk a következő tételt.

2.1. TÉTEL. Tetszőleges $f_1, \ldots, f_N \in C(\mathbf{R}) \cap K$ esetén létezik $F \in C(\mathbf{R} \times \mathbf{R}) \cap K$ olyan, hogy az $f_i$ függvények az $F$ függvény leszűkítései valamilyen irányú egyenes mentén.

*Bizonyítás.* Egy egyszerű konstrukcióval. Tekintsük az $y = b_1$, $y = b_2$, ..., $y = b_N$ $x$-tengellyel párhuzamos egyeneseket. Helyezzük ezekre az $f_1, \ldots, f_N$ függvényeket és terjesszük ki őket $\mathbf{R}$-re úgy, hogy az értelmezési tartományukon kívül legyenek nullák. Így folytonosak maradnak. Majd sávonként kössük össze a két adott függvényt egyenessel, ezt futtassuk végig a grafikonjukon. Az így kapott kétváltozós függvény jó lesz.

Pontosabban a következő függvényt kapjuk:

$$F(x,y)=\begin{cases}\text{folytonosan folytatjuk}\\ \text{0-ba tartóan} & \mathbf{R}\times(-\infty, b_1]\text{-ben}\\ \dfrac{f_2(x)-f_1(x)}{b_2-b_1}(y-b_1)+f_1(x) & \mathbf{R}\times[b_1, b_2]\text{-ben}\\ \dfrac{f_3(x)-f_2(x)}{b_3-b_2}(y-b_2)+f_2(x) & \mathbf{R}\times[b_2, b_3]\text{-ban}\\ \vdots\\ \dfrac{f_N(x)-f_{N-1}(x)}{b_N-b_{N-1}}(y-b_{N-1})+f_{N-1}(x) & \mathbf{R}\times[b_{N-1}, b_N]\text{-ben}\\ \text{folytonosan folytatjuk}\\ \text{0-ba tartóan} & \mathbf{R}\times[b_N, +\infty)\text{-ben}\end{cases}$$

Ezzel a tételt bizonyítottuk. Itt csak az $F$ létezése a fontos számunkra, a leszűkítések iránya nem érdekes. A két feladat ekvivalenciája fontos lesz a továbbiakban, hiszen használhatjuk a $(G)$ sajátosságait mind a megoldás létezésének bizonyításában, mind a numerikus módszer konstruálásakor.

Ezek után könnyen belátható a megoldás létezéséről szóló tétel.

2.2. TÉTEL. A $(G)$ problémának létezik megoldása.

*Bizonyítás.* Legyen $\mathbf{p}=(\mathbf{p}_1, \ldots, \mathbf{p}_N)\in\mathbf{R}^{3N}$, $Q\subset\mathbf{R}^{3N}$

$$Q=\{\mathbf{p}'\in\mathbf{R}^{3N}|\forall\leqq i<j\leqq N\ e_{ij}\geqq\bar{f}_{ij}\},$$

ahol $\mathbf{p}'=(\mathbf{p}'_1, \ldots, \mathbf{p}'_N)$, és a $(G)$ megfogalmazásában szereplő jelöléseket használtuk. Vegyük $\mathbf{R}^{3N}$-ben a szokásos metrikát $d(\mathbf{p},\mathbf{q})=\sum_{i=1}^{N} d_1(\mathbf{p}_i,\mathbf{q}_i)$, ahol $d_1$ az $\mathbf{R}$-beni euklideszi távolság. Mit jelent ekkor a $\mathbf{p}$ pontnak a $Q$ halmaztól való távolsága?

$$d(\mathbf{p},Q)=\inf_{\mathbf{p}'\in Q}\sum_{i=1}^{N} d_1(\mathbf{p}_i,\mathbf{p}'_i)=\inf_{\mathbf{p}'\in Q}\sum_{i=1}^{N} z_i.$$

Tehát a $(G)$ megoldását az approximációelmélet alapproblémája megoldásaként kaphatjuk, vagyis találnunk kell egy olyan $\mathbf{p}''\in Q$ pontot, amelyre felvetődik a $Q$ halmaz és az $f$ pont közötti távolság. Mivel $Q$ zárt halmaz, ilyen pont létezik.

## 3. $(P)$ megoldása

Mivel $(P)$ a $(G)$-ből származtatható, egy egyszerű megoldás kínálkozik. Felosztjuk az adott intervallumokat alkalmasan választott osztópontokkal és csak ezekben a pontokban követeljük meg a feltételek teljesülését. Így lineáris programozási feladathoz jutunk, melynek az az előnye, hogy nem lesz túl nagy méretű. Az alkalmasan választott pontokat a következő módon adjuk meg. Tekintsük valamelyik feltétel jobb oldalán álló függvényt. (A többi feltétel esetén ugyanígy járunk el.) Osszuk fel az adott $[a, b]$ intervallumot — a fenti $f$ függvény értelmezési tartományát —

$h$ egyenletes lépésközzel, ahol $L\frac{h}{2}<\varepsilon$. Itt $\varepsilon$ egy előre megadott tetszőleges kicsiny pozitív szám, $L$ a szóban forgó függvény *Lipschitz konstansa*. Tekintsük a függvény grafikonjának a fenti osztópontokkal kijelölt

$$G=\left\{((a+k\cdot h), f(a+kh)) \quad k=0, 1, \ldots, \frac{b-a}{h}\right\}$$

síkbeli véges részhalmazát. Vegyük ennek a véges halmaznak a konkáv burkát (a konkáv burok bármely két szomszédos pontját egyenessel összekötve, a ponthalmaz az egyenes ugyanazon oldalán marad), amely egy valódi vagy nem valódi részhalmaza lesz $G$-nek. Ha a konkáv buroknak $GK=\{(x_1, f(x_1)), \ldots (x_p, f(x_p))\}$ adódott, akkor $x_1, x_2, \ldots, x_p$ lesznek az intervallum pontjai, amelyekben megköveteljük a szóban forgó feltétel teljesülését. Ha így járunk el, akkor abból, hogy az $x_1, \ldots, x_p$ pontokban teljesül az egyenlőtlenség az is következik, hogy az egész intervallumon teljesül legfeljebb $\varepsilon$ eltéréssel. Ha viszont $\Sigma$ helyett a $\Sigma+\varepsilon$ felületből indulunk ki — itt $\Sigma+\varepsilon$ azt jelenti, hogy a felületet megemeltük $\varepsilon$ értékkel — akkor erre felírva az egyenlőtlenséget és elvégezve az osztópontok fenti kijelölését, ha az adott pontokban teljesül az egyenlőtlenség, akkor az egész intervallumban is teljesül.

Így a végleges $LP$ feladat

$$l(t_i^{12})z-l_2(t_i^{12})z_2 \geqq f_{12}(t_i^{12})+\varepsilon'_{12}, \quad i=1, \ldots, m_{12}$$

$$\vdots$$

$$l_{N-1}(t_i^{N-1,N})z_{N-1}-l_N(t_i^{N-1,N})z_N \geqq f_{N-1,N}(t_i^{N-1,N})+\varepsilon'_{N-1,N} \quad i=1, \ldots, m_{N-1,N}$$

$$\min \sum_{i=1}^{N} z_i,$$

ahol $\varepsilon'_{ij}=|a_i-a_j|\varepsilon_{ij}$

$t_i^{lk}$, $i=1, \ldots, m_{lk}$ az $[a_l, a_k]$

intervallum előbbi módon konstruált pontjai.

### 4. Adott függvény esetén a $t_i$ pontok meghatározására szolgáló algoritmus

Az algoritmus ponthármasokat vizsgál. Az intervallum bal végpontjáról indulva veszi az első 3 pontot és azt vizsgálja, hogy konkáv-e az elhelyezkedésük. Ha konkáv, akkor továbblép jobbra egy következő ponthármasra, ha nem, akkor ismételt visszalépés során törli a középső pontot egészen addig, amíg egy konkáv ponthármast nem talál vagy túl nem lépte az intervallum bal végpontját. Az algoritmus akkor fejeződik be, ha túlléptük az intervallum jobb végpontját.

Az algoritmus pontos megfogalmazásához vezessük be a következő jelöléseket.

Adott $x_1, x_2, x_3$ pont esetén $p_i=f(x_i)$ $i=1, 2, 3$

$e_{13}$ a $p_1, p_2$ pontokon átmenő egyenes

ciklus amíg $l$ olyan ciklus, amely nem hajtja végre a ciklusmagot, ha $l$ hamis,

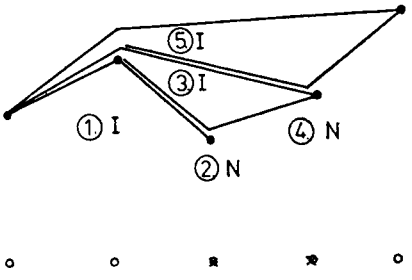
| | |
|---|---|
| *új* $x_1, x_2, x_3$ | vedd a következő jobbra fekvő ponthármast, azaz: első ráhivatkozáskor<br>$x_1=a,\ x_2=x_1+h,\ x_3=x_2+h$<br>különben<br>$x_1=x_2,\ x_2=x_3,\ x_3=x_3+h$ |
| *törlés* $x$ | töröld az $x$-et |
| *vissza* $x_1, x_2$ | $x_1, x_2$-vel lépj egyet balra, vagyis<br>$x_2=x_1$<br>$x_1=x_1-h$<br>$x_3$ változatlan |

## 5. Az algoritmus pontos megfogalmazása

1. új $x_1, x_2, x_3$
2. *ha* $(x_3>b)$ *akkor menj* 5.
3. *ha* $e_{13}\leqq p_2$ *akkor menj* 1.
   *különben törlés* $x_2$
   *vissza* $x_1, x_2$
   *ciklus amíg* $(x_1\geqq a \wedge e_{13}>p_2)$
   *törlés* $x_2$
   *vissza* $x_1, x_2$
   (ciklus) *vége*
   (ha) *vége*
4. *menj* 1.
5. *vége*

a megmaradó pontok adják a keresett $t_i$ pontokat.

Egy példa: 5 pont esetén




*1. ábra*

Bekarikázva jelöltük, hogy az algoritmus hányadik lépéséről van szó, melléírtuk a vizsgálat konkavitásra vonatkozó eredményét: I=igen, N=nem

## 6. Az algoritmus műveletigénye

Legyen $s_1, s_2, \ldots, s_n$ $n \geqq 2$ egy függvény grafikonjának a 3. pont szerint megadott pontjai.

Jelöljük ezek halmazát $P_n$-nel, azaz $P_n := \{s_1, \ldots, s_n\}$. co $(P_n)$ jelölje a $P_n$ konkáv burkát, $\mu(\text{co}\,(P_n))$ pedig a co $(P_n)$ meghatározásához szükséges műveletek számát. Itt egy olyan vizsgálatot nevezünk műveletnek, amellyel eldöntjük, hogy egy ponthármas konkáv módon helyezkedik el vagy sem. Ekkor érvényes a következő tétel.

6.1. TÉTEL. $\forall P_n$ esetén, ahol $n \geqq 2$

$$\mu(\text{co}\,(P_n)) \leqq 2n-4$$

*Bizonyítás.* Legyen a $P_n = \{s_1, s_2, \ldots, s_n\}$ ponthalmaz konkáv burka co $(P_n) = \{s_{i_1}, s_{i_2}, \ldots, s_{i_m}\}$. Ez azt jelenti, hogy az algoritmus végrehajtása során $n-m$ darab pontot töröltünk. Jelöljük $n_b$-vel azon vizsgálatok számát, amelyek eredményeképpen balra léptünk, $n_j$-vel azon vizsgálatok számát, amelyek után jobbra kellett lépnünk. Mivel co $(P_n)$ legalább kételemű (azaz $m \geqq 2$ teljesül) és minden olyan vizsgálat, amely balra lépést eredményez, törléssel jár, érvényes a következő egyenlőtlenség:

$$n_b = n-m \leqq n-2.$$

Az is könnyen belátható, hogy $n_j \leqq n-2$, így a két eredményt összevetve a tétel állítását kapjuk:

$$\mu(\text{co}\,(P_n)) = n_j + n_b \leqq 2n-4.$$

*Megjegyzés.* Nyilvánvaló, hogy a műveletek száma legalább $n-2$. A tétel $n=2$ esetben pontos értéket ad, vagyis $\mu(\text{co}\,(P_2))=0$. (Két pont esetén nem kell vizsgálat.)

## 7. Megjegyzések

— A $(P)$ feladat általánosítható, ugyanis a feltételrendszerünk általánosabb célfüggvény mellett is megoldható. L. [2], [4].

— A kapott $LP$ feladat megoldására hatékony módszerek léteznek. L. [5].

## IRODALOM

[1] GRAHAM, R. L., "An efficient algorithm for determining convex hull of a finite planely set", *Information Processing Letters* **1** (1972) 132—133.

[2] LOOTSMA, F. A., *Numerical Methods for Nonlinear Optimization* (Academic Press, London, 1972).

[3] RAPCSÁK, T., „Lineáris programozási modell egy tereprendezési feladat megoldására", *Alkalmazott Matematikai Lapok* 7 (1981) 99—105.

[4] SIMMONS, D. M., *Nonlinear Programming for Operation Research* (Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1975.)

[5] ORCHARD—HAYS, W., *Advanced Linear Programming Computing Techniques* (McGraw-Hill Book Company, New York, 1968).

(Beérkezett: 1983. február 7.)

GERGÓ LAJOS
EÖTVÖS LORÁND TUDOMÁNYEGYETEM SZÁMÍTÓKÖZPONT
1117 BUDAPEST, BOGDÁNFY ÚT 10/B.

# SOLUTION FOR A CLASS OF PARAMETRIC OPTIMIZATION PROBLEMS

L. GERGÓ

In the present paper a class of one-parameter programming problems is considered, where some inequalities are to be satisfied on certain intervals of the parameter (see the problem $(P)$.

We will show that $(P)$ comes from a geometrical problem, then we prove the existence of the solution. Using the special properties of the underlying geometrical porblem, $(P)$ can be approximated with an $LP$ problem.

After showing an algorithm for determining the concave hull of finitely many points, we give an upper bound on the required number of operations of the algorithm.

*Alkalmazott Matematikai Lapok* **10** (1984) 73—85

# EGYDIMENZIÓS SZABÁSI FELADATOK MEGOLDÁSA OSZLOPGENERÁLÁSSAL

GALAMBOS GÁBOR ÉS IMREH BALÁZS

Szeged

A dolgozat az egydimenziós szabási feladat megoldásával foglalkozik. Elsőként ismertetjük a lineáris célfüggvénnyel rendelkező modell megoldására szolgáló eljárást, amelyet GILMORE és GOMORY [6] dolgozott ki. Ezt követően hiperbolikus célfüggvénnyel rendelkező modelleket vizsgálunk, és felhasználva a lineáris eset megoldásánál alkalmazott ötletet, a tekintett modellek megoldására szolgáló eljárásokat adunk meg. Ez utóbbi új eredmények ismertetésén kívül, a jelen dolgozattal szeretnénk hozzájárulni a tárgyalt eljárások minél szélesebb körben történő alkalmazásához is.

## 1. Bevezetés

Adott méretű rudak, gerendák, csövek nagyüzemi előállítása általában két lépésben történik. Az első lépésben nagy mennyiségben előállítanak különböző hosszúságú félkész termékeket. Ezeket a második lépésben kisebb darabokra vágják fel, amely során a darabok mennyiségét és méretét a megrendelések határozzák meg. A darabolási művelettel szemben támasztott természetes követelmény az, hogy ezt a leggazdaságosabban végezzék el. A leggazdaságosabb darabolás meghatározását leíró feladatokat *egydimenziós szabási feladatoknak* (a továbbiakban ESZF) nevezzük.

A vázolt problémával már KANTOROVICS [16] is foglalkozott, és a következő optimumszámítási modellt rendelte hozzá.

Legyenek adva korlátlan mennyiségben $L$ hosszúságú félésztermékek, ahol $L$ pozitív egész, és jelölje $c$ egy félkész termék előállítási költségét. Legyenek adva továbbá $l_1, \ldots, l_n$ ($l_i \leqq L$ és $l_i$ pozitív egész ($i=1, \ldots, n$)) hosszúságú megrendelések, amelyekből rendre $r_1, \ldots, r_n$ ($r_i$ pozitív egész ($i=1, \ldots, n$)) darabot kell kivágni a félésztermékek feldarabolásával. Az $\mathbf{A}_j=(a_{1j}, \ldots, a_{nj})^T$ vektort lehetséges szabásnak nevezzük, ha $a_{ij}$ nemnegatív egész ($i=1, \ldots, n$) és $\sum_{i=1}^{n} l_i a_{ij} \leqq L$. Jelölje $\mathbf{A}=(\mathbf{A}_1, \ldots, \mathbf{A}_p)$ az összes lehetséges szabásokból összeállított mátrixot. Legyen $\mathbf{R}=(r_1, \ldots, r_n)^T$ és $\mathbf{C}=(c, \ldots, c)$ $p$-dimenziós vektor. Akkor a fenti problémához az alábbi egészértékű programozási feladat rendelhető.

$$\begin{aligned} &\mathbf{AX} = \mathbf{R} \\ &\mathbf{X} \geqq \mathbf{O} \\ &\underline{\mathbf{X} \text{ egész}} \\ &\mathbf{CX} \to \min. \end{aligned} \tag{1.1}$$

Az (1.1) feladat megoldását illetően KANTOROVICS rámutatott, hogy konkrét szabási feladatok esetén a lehetséges szabások hatalmas száma miatt már az $\mathbf{A}$ mátrix

meghatározása is komoly nehézségekbe ütközik. Ennek kiküszöbölésére néhány szerző (például METZGER [17]) azt javasolta, hogy ne tekintsük az összes lehetséges szabást, hanem csak azokat, amelyeknek a szabási selejtje egy adott korlát alatt marad.

A fenti feladat megoldásával kapcsolatos nehézségeket alapvetően GILMORE és GOMORY [6] hidalta át. Első lépésként egyszerűsítették a modellt az egészértékűségi feltétel elhagyásával. Szerintük késztermékek ipari méretekben történő előállításához használt optimumszámítási modellben az egészértékűségi feladat optimális megoldásának valamilyen irányú kerekítése az (1.1) feladat optimális megoldásának egy jó közelítését adja. Ezt követően a folytonos feladat megoldására kidolgoztak egy olyan eljárást, amely a módosított szimplex módszeren alapul és nem szükséges hozzá az **A** mátrix meghatározása, hanem egy egyszerű feltételes szélsőértékfeladat megoldásának eredményeként adódik az **A** mátrix azon oszlopa, amelynek a bázisba történő bevonása a célfüggvényben javítást eredményezhet. (A bázis fogalmába beleértjük azt is, hogy az egyenletrendszer jobb oldalának a bázisra vonatkozó koordinátái nemnegatívak.) A vázolt eljárás *oszlopgenerálás módszere* néven ismeretes.

Az ESZF megoldására más eljárások is kidolgozást nyertek. Ezzel kapcsolatban csak utalnánk a [3], [5], [10], [15], [18] munkákra, mivel a jelen dolgozat az oszlopgenerálás módszeréhez kapcsolódik.

Az említett módszer kidolgozása után megkezdődött az eljárás kiterjeszthetőségének a vizsgálata. Egyrészt azt vizsgálták, hogy milyen, az (1.1) feladatnál bonyolultabb egydimenziós szabási feladatok megoldása vezethető vissza az oszlopgenerálás módszerére. Elsőként arra az esetre sikerült kiterjeszteni az eljárást, amikor a félkész termékek korlátozott mennyiségben állnak rendelkezésre. A korlátos modellt és a kapcsolódó eljárást a [7] dolgozat tartalmazza. További, más irányú kiterjesztések találhatók a [12], [13] dolgozatokban. Másrészt kísérletek történtek a módszer alkalmazására magasabb dimenziós szabási feladatok megoldására. Ilyen irányú eredmény található a [8] dolgozatban kétdimenziós *„guillotine"* szabások esetére.

A jelen dolgozatban további egydimenziós kiterjesztési lehetőséget mutatunk be. Elsőként röviden ismertetjük az oszlopgenerálás módszerét. Ezt követően megmutatjuk, hogy hiperbolikus célfüggvény esetén is alkalmazható az eljárás abban az esetben, ha a félkész termékek korlátlan mennyiségben állnak rendelkezésre. Végezetül a korlátos esetet tárgyaljuk, és erre vonatkozóan is megadunk egy, az oszlopgenerálás módszerén alapuló eljárást.

## 2. Az oszlopgenerálás módszere

Bővítsük a bevezetésben leírt szabási problémát azzal, hogy $m$ féle, rendre $L_1, \ldots, L_m$ hosszúságú félkész termékek állnak rendelkezésre korlátlan mennyiségben. Az általánosság megszorítása nélkül feltételezhetjük, hogy $L_i \geqq L_j$, ha $1 \leqq i < j \leqq m$. Továbbá vegyük észre, hogy a szabási problémának akkor és csak akkor van lehetséges megoldása, ha bármely $1 \leqq i \leqq n$ indexre van olyan $1 \leqq j \leqq m$, hogy $l_i \geqq L_j$. A továbbiakban ezt mindig feltételezzük. A problémához a következő modell rendelhető.

$$\begin{aligned} \mathbf{AX} &= \mathbf{R} \\ \mathbf{X} &\geqq \mathbf{O} \\ \hline \mathbf{CX} &\to \min, \end{aligned} \tag{2.1}$$

ahol $\mathbf{C}=(c_1, \ldots, c_p)$, és $c_i=c_j=f(k)$, ha az $i$-edik és $j$-edik szabásokat a $k$-adik félkész termékféleség két különböző darabjából vágtuk ki. Lényegében feltételezzük, hogy a költség független a konkrét szabástól, csak attól függ, hogy melyik félkész termékféleségből történik a szabás. Az ilyen költséget *szabásfüggetlen költségnek* nevezzük, és amint arra a fejezet végén rámutatunk, célszerű élesen elválasztani a szabásfüggő költségtől, mivel az utóbbi költség esetén az oszlopgenerálás módszere általában nem lesz érvényes.

Mielőtt rátérnénk a módszer tárgyalására, megadjuk, hogy az árnyékár fogalmát milyen értelemben használjuk. Ehhez legyen $\mathbf{B}$ a (2.1) feladat egy bázisa, és jelölje $c_{i_1}, \ldots, c_{i_n}$ rendre a bázisban szereplő oszlopvektorokhoz tartozó célfüggvényegyütthatókat. Legyen továbbá $\tilde{\mathbf{c}}=(c_{i_1}, \ldots, c_{i_n})$. Akkor a $\mathbf{B}$ *bázishoz tartozó árnyékár vektoron* a $\tilde{\mathbf{c}}\mathbf{B}^{-1}$ vektort értjük.

Az oszlopgenerálás módszere az alábbi két tételen alapul, amelyek más formában megtalálhatók GILMORE [9] összefoglaló munkájában.

2.1. TÉTEL. Legyen adva a (2.1) feladat egy $\mathbf{B}$ bázisa, és jelölje a $\mathbf{B}$ bázishoz tartozó árnyékár vektort $\mathbf{b}=(b_1, \ldots, b_n)$. Legyen továbbá $1 \leqq j \leqq m$ tetszőleges. Akkor az $M_j=\max\{\sum_{k=1}^{n} b_k v_k: \sum_{k=1}^{n} l_k v_k \leqq L_j$ és $v_k$ nemnegatív egész $(k=1, \ldots, n)\}$ hátizsák-feladat $\mathbf{V}_0$ optimális megoldása lehetséges szabás, és ha $M_j>f(j)$, akkor $\mathbf{V}_0$ nem szerepel a $\mathbf{B}$ bázisban, továbbá a bázisba történő bevonása a célfüggvényben csökkenést eredményezhet.

*Bizonyítás.* A definíciókból adódik, hogy ha $\mathbf{V}$ a $j$-edik félkész termékféleség egy lehetséges szabása, akkor $\mathbf{V}$ a fenti hátizsák-feladatnak egy lehetséges megoldása, és fordítva. Következésképpen $\mathbf{V}_0$ szerepel az $\mathbf{A}$ mátrixban úgy, hogy a megfelelő célfüggvényegyütthatója $f(j)$. Másrészt a módosított szimplex módszer (lásd pl. [2]) alapján, ha $\mathbf{V}_0$ szerepel a bázisban, akkor $\mathbf{b}\mathbf{V}_0=f(j)$, ami ellentmond az $M_j>f(j)$ feltételezésünknek. Így $\mathbf{V}_0$ nem szerepel a bázisban. Másrészt, ismét a módosított szimplex módszer alapján, $\mathbf{V}_0$ bevonása a bázisba javítást eredményezhet, ha $f(j)-\mathbf{b}\mathbf{V}_0<0$. Ez utóbbi feltétel viszont teljesül, mivel $M_j=\mathbf{b}\mathbf{V}_0$ és $M_j>f(j)$. Ezzel az állítást igazoltuk.

Felhasználva az előző tétel jelöléseit, a következő állítást mondhatjuk ki.

2.2. TÉTEL. A (2.1) feladat $\mathbf{B}$ bázisához tartozó bázismegoldás optimális megoldás, ha $f(j) \geqq M_j$ $(j=1, \ldots, m)$.

*Bizonyítás.* Legyen $1 \leqq j \leqq m$ tetszőleges, és jelölje $\mathbf{A}_q$ a $j$-edik félkész termékféleség tetszőleges lehetséges szabását. Akkor $\mathbf{A}_q$ lehetséges megoldása az előző tételben szereplő $M_j$ optimumú hátizsák-feladatnak, és így $\mathbf{b}\mathbf{A}_q \leqq M_j$. Másrészt $M_j \leqq f(j)$ és $f(j)=c_q$, amivel $c_q-\mathbf{b}\mathbf{A}_q \geqq 0$ adódik. Mivel $\mathbf{A}_q$ az $\mathbf{A}$ mátrix tetszőleges oszlopa, ezért a módosított szimplex módszer optimum-kritériumából következik az állítás.

A fenti két tétel alapján adódik az alábbi, a (2.1) feladat megoldására szolgáló eljárás.

*1. Előkészítő rész.* Egy $\mathbf{B}_0$ induló bázis meghatározása, a $\mathbf{B}_0$ ismeretében a $\mathbf{B}_0^{-1}$ mátrix és a $\mathbf{b}^{(0)}$ árnyékár vektor meghatározása.

*2. Iterációs rész (r-edik iteráció).* Az $M_j=\max\{\sum_{n=1}^{n} b_k^{(r)} v_k: \sum_{k=1}^{n} l_k v_k \leqq L_j$ és $v_k$

nemnegatív egész $(k=1, \ldots n)\}$ $(j=1, \ldots, m)$ hátizsák-feladatok megoldása. Ha $f(j) \geqq M_j$, $j=1, \ldots, m$ teljesül, akkor vége az eljárásnak, a $\mathbf{B}_r$ bázishoz tartozó bázismegoldás optimális megoldása a feladatnak. Ellenkező esetben legyen $1 \leqq j \leqq m$ olyan index, amelyre $f(j) < M_j$, és jelölje $\mathbf{V}^{(r)}$ az $M_j$-hez tartozó hátizsák-feladat optimális megoldását. Vonjuk be a $\mathbf{V}^{(r)}$ vektort a bázisba a szimplex módszernek megfelelően, azaz képezzük az $\mathbf{S}=\mathbf{B}_r^{-1}\mathbf{R}$, $\mathbf{T}=\mathbf{B}_r^{-1}\mathbf{V}^{(r)}$ vektorokat és a $\min \{s_i/t_i^{(r)}: t_i^{(r)}>0, 1 \leqq i \leqq n\}$ mennyiséget. A tekintett minimum létezik, ugyanis ellenkező esetben azt kapnánk, hogy a célfüggvény alulról nem korlátos a lehetséges megoldások halmazán. Másrészt könnyen belátható, hogy az adott feltételek mellett a (2.1) feladatnak létezik az optimális megoldása, ami ellentmond az előzőeknek. Következésképpen a tekintett minimum létezik. Hajtsuk végre a minimum által meghatározható generáló elemek valamelyikével a bázisváltoztatást. Ennek eredményeként adódik az új bázis inverze, jelölje ezt $\mathbf{B}_{r+1}^{-1}$. Határozzuk meg ezek után az új árnyékár vektort, jelölje ezt $\mathbf{b}_{r+1}$. Végül $r$ helyett az $r+1$ értékkel folytassuk az eljárást az iterációs résznél.

*1. Megjegyzés.* Mivel a generáló elemeket nem *l*-szabály alapján választjuk, ezért nem biztosított az eljárás végessége. Azonban a végesség elérhető a lexikografikus technika alkalmazásával módosított szimplex módszer esetén is. (Lásd: [19]).

*2. Megjegyzés.* A hátizsák-feladat megoldására számos eljárás kidolgozásra került. A [6] dolgozatban GILMORE és GOMORY adott meg egy, a dinamikus programozás elvén alapuló módszert, amelyet a későbbiek során a [9] dolgozatban a szerző továbbfejlesztett. Az utóbbi évtizedben további eljárások (lásd [4], [11], [14], [20]) kerültek kidolgozásra.

*3. Megjegyzés.* Az ismertetett eljárással kapcsolatban célszerű kihangsúlyozni, hogy szabásfüggetlen költségek esetére alkalmazható. Szabásfüggő költségeknél minden lehetséges szabáshoz tartozik egy, az illető szabástól függő célfüggvényegyüttható, ami azt eredményezi, hogy általános esetben az **A** mátrix meghatározása nem kerülhető el. Bizonyos speciális esetekben, amikor a célfüggvényegyütthatókra alkalmas megszorításokat teszünk, a kapott feladat megoldása visszavezethető az oszlopgenerálás módszerére. Ilyen jellegű eredmény található PIERCE [18] munkájában.

*4. Megjegyzés.* Mivel az **A** mátrix számos speciális oszlopot tartalmaz, ezért igen egyszerű egy $\mathbf{B}_0$ induló bázis meghatározása. Tekintsük például a következő $\mathbf{A}_i=(0, \ldots, 0, a_{ii}, 0, \ldots, 0)^T$ $(i=1, \ldots, n)$ vektorokat, ahol bármely $1 \leqq i \leqq n$ indexre $a_{ii}$ az $\mathbf{A}_i$ vektor *i*-edik komponense és $a_{ii}$ az $L_m/l_i$ hányados egész része. Nyilvánvalóan a tekintett vektorok bázist alkotnak, és a belőlük képezett mátrix inverze közvetlenül felírható.

## 3. Hiperbolikus célfüggvénnyel rendelkező ESZF megoldása az oszlopgenerálás módszerével

Az előzőekben a gazdaságosság mértékét a felhasznált félkész termékek előállítási költségeinek összegével azonosítottuk. Az ipari alkalmazásokban azonban más gazdasági szempontok is dominánsak lehetnek, amelyek a célfüggvény megváltozását eredményezik. Abban az esetben, ha a felhasznált félkész termékek egységnyi előállítási költségére eső selejtet akarjuk minimalizálni, akkor hiperbolikus célfüggvényt

kapunk. Felhasználva az (1.1), illetve (2.1) feladatoknál bevezetett jelöléseket, és feltéve, hogy $f(k)$ a $k$-adik félkész termékféleség egy darabjának az előállítási költségét jelöli, a problémához a következő optimumszámítási modell rendelhető:

$$\begin{array}{l} \mathbf{AX} = \mathbf{R} \\ \mathbf{X} \geqq \mathbf{O} \\ \hline \dfrac{\mathbf{SX}}{\mathbf{CX}} \to \min, \end{array} \tag{3.1}$$

ahol $\mathbf{S}=(s_1, \ldots, s_p)$, és $s_j=L_k-\sum_{i=1}^{n} l_i a_{ij}$, ha az $\mathbf{A}_j=(a_{1j}, \ldots, a_{nj})^T$ szabás a $k$-adik félkész termékféleség egy szabása, azaz $s_j$ megadja az $\mathbf{A}_j$ szabás által keletkező selejtet. A (3.1) feladattal kapcsolatban feltételezzük, hogy $c_i>0$ $(i=1, \ldots, p)$, ami nem jelent erős megszorítást, mivel a gyakorlati alkalmazásoknál ez rendre teljesül.

A következőkben megmutatjuk, hogy a (3.1) feladat megoldása visszavezethető az oszlopgenerálás módszerére, és az eljáráshoz nem szükséges az $\mathbf{A}$ mátrix és az $\mathbf{S}$ vektor meghatározása. Ezt követően a (3.1) feladat korlátos változatát vizsgáljuk, azaz feltételezzük, hogy a $j$-edik félkész termékféleségből $g_j$ ($g_j$ pozitív egész) darab használható fel bármely $1 \leqq j \leqq m$ indexre. Megmutatjuk, hogy ez az eset is visszavezethető az oszlopgenerálás módszerére.

*a) A félkész termékek korlátlan mennyiségben állnak rendelkezésre*

A címben szereplő esethez a (3.1) feladat rendelhető. CHARNES és COOPER [1] munkájából ismert, hogy tetszőleges $\mathbf{A}$ mátrix esetén a (3.1) feladat megoldása bizonyos feltételek mellett visszavezethető az alábbi lineáris programozási feladat megoldására.

$$\begin{array}{l} \mathbf{CY} \quad = 1 \\ \mathbf{AY}-\mathbf{R}t \quad = \mathbf{O} \\ \mathbf{Y} \geqq \mathbf{O}, \quad t > 0 \\ \hline \mathbf{SY} \to \min. \end{array} \tag{3.2}$$

A visszavezethetőség feltétele az, hogy a (3.1) feladat lehetséges megoldásainak halmaza korlátos legyen, és a (3.1) feladat tetszőleges $\mathbf{X}$ lehetséges megoldására $\mathbf{CX}>0$ teljesüljön. Ilyen feltételek mellett érvényes a következő állítás.

3.1. TÉTEL [1]. A (3.1) és (3.2) feladatoknak egyidejűleg létezik optimális megoldása, és ha $(\mathbf{Y}, t)$ a (3.2) feladat egy lehetséges megoldása, akkor $t>0$, továbbá ha $(\mathbf{Y}_0, t_0)$ a (3.2) feladat optimális megoldása, akkor $\mathbf{X}_0=\mathbf{Y}_0/t_0$ a (3.1) feladat optimális megoldása.

A fenti tételből közvetlenül adódik, hogy az adott feltételek mellett a (3.1) feladat optimális megoldásának meghatározásához elegendő a (3.2) feladat optimális megoldását meghatározni. Másrészt feltételeink szerint (szabási feladat esetén) az $\mathbf{A}$ mátrix elemei nemnegatív egészek. Ebből egyszerűen következik, hogy a (3.1) feladat lehetséges megoldásainak halmaza korlátos, és $\mathbf{0}$ nem lehetséges megoldás. Ez utóbbi

tényből $\mathbf{X}$ nemnegativitásával és a $c_i>0$ $(i=1, \ldots, p)$ feltevésünkkel adódik, hogy a (3.1) feladat bármely $\mathbf{X}$ lehetséges megoldására $\mathbf{CX}>0$ teljesül. Összegezve azt kaptuk, hogy szabási probléma esetén mindkét feltétel teljesül, így elegendő a (3.2) feladat megoldására szorítkozni. Ez utóbbi feladat megoldására szolgáló eljáráshoz fel fogjuk használni az alábbi állításokat.

3.2. TÉTEL. Legyen adva a (3.2) feladat egy $\mathbf{B}$ bázisa, és jelölje $\mathbf{b}=(b_0, b_1, \ldots, b_n)$ a $\mathbf{B}$ bázishoz tartozó árnyékár vektort. Legyen továbbá $1\leqq j\leqq m$ tetszőleges. Akkor az $M_j=\max\{\sum_{k=1}^{n}(l_k+b_k)v_k: \sum_{k=1}^{n} l_k v_k\leqq L_j,\ v_k$ nemnegatív egész $(k=1, \ldots, n)\}$ hátizsák-feladat $\mathbf{V}_0=(v_{01}, \ldots, v_{0n})^T$ optimális megoldásával képezett $\mathbf{Q}=(f(j), v_{01}, \ldots, v_{0n})^T$ vektor a (3.2) feladat együtthatómátrixának egy oszlopvektora, és ha $L_j-b_0f(j)<M_j$, akkor a $\mathbf{Q}$ vektor nem szerepel a bázisban, továbbá a bázisba történő bevonása a célfüggvényérték csökkenését eredményezheti.

*Bizonyítás.* A definíciókból következik, hogy ha $\mathbf{V}=(v_1, \ldots, v_n)^T$ a $j$-edik félkész termékféleség egy lehetséges szabása, akkor az $(f(j), v_1, \ldots, v_n)^T$ vektor a (3.2) feladat egy oszlopvektora, és fordítva. Következésképpen $\mathbf{Q}$ a (3.2) feladat egy oszlopvektora, és a megfelelő célfüggvényegyüttható $L_j-\sum_{k=1}^{n} l_k v_{0k}$. A módosított szimplex módszer alapján, ha $\mathbf{Q}$ szerepel a bázisban, akkor $\mathbf{bQ}=L_j-\sum_{k=1}^{n} l_k v_{0k}$. Rendezve ezt az egyenlőséget

$$\sum_{k=1}^{n}(l_k+b_k)v_{0k} = L_j-b_0f(j)$$

adódik, ami ellentmond az $L_j-b_0f(j)<M_j$ feltételnek. Így, ha $L_j-b_0f(j)<M_j$, akkor $\mathbf{Q}$ nem szerepel a bázisban. Másrészt $\mathbf{Q}$ bevonása a bázisba akkor eredményezhet a célfüggvényben javítást, ha $(L_j-\sum_{k=1}^{n} l_k v_{0k})-\mathbf{bQ}<0$. Rendezve ezt az egyenlőtlenséget, a javítás feltételeként az $L_j-b_0f(j)<\sum_{k=1}^{n}(l_k+b_k)v_{0k}$ egyenlőtlenség adódik, ami az $M_j=\sum_{k=1}^{n}(l_k+b_k)v_{0k}$ egyenlőséggel az állítás érvényességét eredményezi.

Felhasználva a fenti tétel jelöléseit, az alábbi állítást mondhatjuk ki.

3.3. TÉTEL. A (3.2) feladat $\mathbf{B}$ bázisához tartozó bázismegoldás optimális megoldás, ha $L_j-b_0f(j)\geqq M_j$ $(j=1, \ldots, m)$.

*Bizonyítás.* Jelölje $\mathbf{Q}$ a (3.2) feladat egy tetszőleges olyan oszlopvektorát, amely nem szerepel a bázisban. A 3.1. tételből tudjuk, hogy bármely $(\mathbf{Y}, t)$ lehetséges megoldásra $t>0$ teljesül. Következésképpen a $\mathbf{B}$ bázishoz tartozó bázismegoldásban a $t$ változó értéke pozitív, ami azt eredményezi, hogy a $(0, -r_1, \ldots, -r_n)^T$ vektor benne van a bázisban. Ebből azt kapjuk, hogy a $\mathbf{Q}^T=(f(j), a_{1r}, \ldots, a_{nr})$ valamely $1\leqq j\leqq m$, $1\leqq r\leqq p$ indexekre, és $(a_{1r}, \ldots, a_{nr})$ lehetséges megoldása az $M_j$-hez tartozó hátizsák-feladatnak. Ebből viszont $\sum_{k=1}^{n}(l_k+b_k)a_{kr}\leqq M_j$ adódik. Másrészt $M_j\leqq$

$\leqq L_j - b_0 f(j)$, és így $\sum_{k=1}^{n} (l_k + b_k) a_{kr} \leqq L_j - b_0 f(j)$. A kapott egyenlőtlenséget rendezve az $(L_j - \sum_{k=1}^{n} l_k a_{kr}) - \mathbf{bQ} \geqq 0$ egyenlőtlenséghez jutunk. Mivel a **Q**-hoz tartozó célfüggvényegyüttható éppen $L_j - \sum_{k=1}^{n} l_k a_{kr}$, és **Q** a (3.2) feladat tetszőleges, a bázisban nem szereplő oszlopvektora, ezért a módosított szimplex módszer optimumkritériumával adódik az állítás.

A fenti két tétel alapján a 2. részben megadott eljáráshoz hasonló eljárás adható meg a (3.2) feladat megoldására, és a 2. részben tett 1. megjegyzés erre az eljárásra is érvényes. Gyakorlati szempontból nem jelentéktelen, hogy a (3.2) feladat szerkezete lehetővé teszi egy $\mathbf{B}_0$ induló bázis, és az illető bázis inverzének egy egyszerű meghatározását. Valóban, könnyen belátható, hogy az $\mathbf{A}_0' = (0, -r_1, \ldots, -r_n)^T$ és $\mathbf{A}_i' = (f(m), 0, \ldots, 0, a_{ii}, 0, \ldots, 0)^T$ $(i = 1, \ldots, n)$ vektorok lineárisan függetlenek, ahol bármely $1 \leqq i \leqq n$ indexre $a_{ii}$ az $\mathbf{A}_i'$ vektor $i+1$-edik komponense, és $a_{ii}$ az $L_m/l_i$ hányados egész része. Másrészt egyszerű számítással ellenőrizhető, hogy az $\mathbf{A}_0', \mathbf{A}_1', \ldots, \mathbf{A}_n'$ vektorokból álló **B** mátrix inverze az alábbi mátrix

$$\left(\begin{array}{c|c} \delta & \mathbf{F} \\ \hline \mathbf{H} & \mathbf{U} \end{array}\right)$$

ahol

$$\delta = \left(f(m) \sum_{k=1}^{n} r_k / a_{kk}\right)^{-1}, \quad h_j = \delta r_j / a_{jj} \quad (j = 1, \ldots, n),$$

$$f_j = -\delta f(m)/a_{jj} \quad (j = 1, \ldots, n) \quad \text{és}$$

$$u_{ij} = \begin{cases} (1 - \delta r_i f(m)/a_{ii})/a_{ii}, & \text{ha} \quad i = j \\ -\delta r_i f(m)/(a_{ii} a_{jj}), & \text{ha} \quad i \neq j \end{cases} \quad (i = 1, \ldots, n; \quad j = 1, \ldots, n).$$

Végül vegyük észre, hogy a (3.2) feladat jobb oldali vektorának a fenti vektorrendszerre vonatkozó koordinátáit a $\mathbf{B}^{-1}$ mátrix első oszlopa tartalmazza. Mivel $\mathbf{B}^{-1}$ első oszlopának minden egyes eleme nemnegatív, ezért a tekintett vektorrendszer bázist alkot az általunk használt bázisfogalomnak megfelelően.

*b) A félkész termékek korlátos mennyiségben állnak rendelkezésre*

Minden egyes $1 \leqq i \leqq p$ indexre az $\mathbf{A}_i$ lehetséges szabáshoz rendeljük hozzá az $\mathbf{A}_i' = (a_{1i}, \ldots, a_{mi})^T$ vektort, ahol

$$a_{ji}' = \begin{cases} 1, & \text{ha az } \mathbf{A}_i \text{ szabást a } j\text{-edik félkész termékféleségből vágjuk ki,} \\ 0, & \text{különben.} \end{cases}$$

Legyen továbbá $\mathbf{G}=(g_1, \ldots, g_m)^T$. Akkor felhasználva az előzőek során bevezetett jelöléseket, a korlátos problémához a következő modell rendelhető.

$$\begin{aligned} &\mathbf{AX} = \mathbf{R} \\ &\mathbf{A'X} \leqq \mathbf{G} \\ &\mathbf{X} \geqq \mathbf{O} \\ &\hline \\ &\frac{\mathbf{SX}}{\mathbf{CX}} \to \min. \end{aligned} \tag{3.3}$$

Jelölje $\mathbf{E}_m$ az $m \times m$-es egységmátrixot, és legyen $\mathbf{U}=(u_1, \ldots, u_m)^T$. Egyszerűen belátható, hogy a (3.3) feladatnak és az alábbi (3.4) feladatnak egyidejűleg létezik optimális megoldása, és (3.4) optimális megoldásából közvetlenül származtatható (3.3) optimális megoldása.

$$\begin{aligned} &\mathbf{AX} = \mathbf{R} \\ &\mathbf{A'X}+\mathbf{E}_m\mathbf{U} = \mathbf{G} \\ &\mathbf{X} \geqq \mathbf{O}, \quad \mathbf{U} \geqq \mathbf{O} \\ &\hline \\ &\frac{\mathbf{SX}}{\mathbf{CX}} \to \min. \end{aligned} \tag{3.4}$$

A korábban tett feltételekből közvetlenül adódik, hogy a (3.4) feladat lehetséges megoldásainak halmaza korlátos, és tetszőleges $(\mathbf{X}, \mathbf{U})$ lehetséges megoldásra $\mathbf{CX}>0$. Ez a 3.1. tétel alapján azt eredményezi, hogy a (3.4), és így a (3.3) feladatok optimális megoldásának meghatározásához elegendő az alábbi (3.5) feladat optimális megoldását meghatározni.

$$\begin{aligned} &\mathbf{CY} = 1 \\ &\mathbf{AY} - \mathbf{R}t = \mathbf{O} \\ &\mathbf{A'Y}+\mathbf{E}_m\mathbf{W}-\mathbf{G}t = \mathbf{O} \\ &\mathbf{Y} \geqq \mathbf{O}, \quad \mathbf{W} \geqq \mathbf{O}, \quad t > 0 \\ &\hline \\ &\mathbf{SY} \to \min. \end{aligned} \tag{3.5}$$

A fenti (3.5) feladat megoldását is az oszlopgenerálás módszerére vezetjük vissza. Az eljáráshoz szükséges az alábbi két állítás, amelyek bizonyítása teljesen hasonló a 3.2. és 3.3. tételek bizonyításához, ezért a részletes bizonyítástól eltekintünk, csak kimondjuk a megfelelő állításokat.

3.4. TÉTEL. Legyen adva a (3.5) feladat egy $\mathbf{B}$ bázisa, és jelölje $\mathbf{b}=(b_0, b_1, \ldots, b_{n+m})$ a $\mathbf{B}$ bázishoz tartozó árnyékár vektort. Legyen továbbá $1 \leqq j \leqq m$ tetszőleges. Akkor érvényesek a következők:

(1) az $M_j=\max\left\{\sum_{k=1}^{n}(l_k+b_k)v_k : \sum_{k=1}^{n} l_k v_k \leqq L_j, v_k \text{ nemnegatív egész } (k=1,\ldots,n)\right\}$ hátizsák-feladat $\mathbf{V}_0=(v_{01}, \ldots, v_{0n})^T$ optimális megoldásával képezett $\mathbf{Q}=(f(j), v_{01}, \ldots, v_{0n}, 0, \ldots, 0, 1, 0, \ldots, 0)^T$ vektor, amelynek $n+j+1$-edik komponense

1, a (3.5) feladat egy oszlopvektora, és ha $M_j > L_j - b_0 f(j) - b_{n+j}$, akkor $\mathbf{Q}$ nem szerepel a bázisban, továbbá a bázisba történő bevonása a célfüggvényben csökkenést eredményezhet,

(2) ha $b_{n+j} > 0$, akkor a $(0, \ldots, 0, 1, 0, \ldots, 0)$ vektor, amelynek $n+j+1$-edik komponense 1, nem szerepel a bázisban, és a bázisba történő bevonása a célfüggvényben csökkenést eredményezhet.

A fenti tétel jelöléseit használva, érvényes a következő

3.5. TÉTEL. A (3.5) feladat $\mathbf{B}$ bázisához tartozó bázismegoldás a feladat optimális megoldása, ha $M_j \leqq L_j - b_0 f(j) - b_{n+j}$ $(j=1, \ldots, m)$ és $b_{n+i} \leqq 0$ $(i=1, \ldots, m)$.

A fenti két tétel alapján az előzőekben megadott eljárásokhoz hasonló eljárás adható meg, amelyre érvényes a 2. részben tett 1. megjegyzés.

Ellentétben a (3.2) feladattal a (3.5) feladatnak nem minden esetben létezik lehetséges megoldása, és abban az esetben, ha a (3.5) feladat kielégíthető, nehézkes az induló bázis meghatározása. Szerencsés módon erre is alkalmazható az oszlopgenerálás módszere. Ehhez jelölje $\mathbf{E}_n$ az $n \times n$-es egységmátrixot, és vegyük a $p_0$, $\mathbf{P} = (p_1, \ldots, p_n)^T$ mesterséges változókat. Ezek után tekintsük az alábbi feladatot.

$$
\begin{aligned}
&\mathbf{CY} \quad + p_0 \qquad\qquad\quad = 1 \\
&\mathbf{AY} - \mathbf{R}t \quad + \mathbf{E}_n \mathbf{P} \qquad = \mathbf{O} \\
&\mathbf{A}'\mathbf{Y} - \mathbf{G}t \qquad\quad + \mathbf{E}_m \mathbf{W} = \mathbf{O} \\
&\mathbf{Y} \geqq \mathbf{O}, \quad \mathbf{W} \geqq \mathbf{O}, \quad \mathbf{P} \geqq \mathbf{O}, \quad p_0 \geqq 0, \quad t > 0 \\
&p_0 + p_1 + \ldots + p_n \to \min.
\end{aligned} \tag{3.6}
$$

Ismeretes, hogy a (3.5) feladatnak pontosan akkor van lehetséges megoldása, ha a (3.6) feladatnak létezik optimális megoldása, és az optimum értéke 0. Másrészt a (3.6) feladatban a $p_0, p_1, \ldots, p_n, w_1, \ldots, w_m$ változókhoz tartozó oszlopvektorok bázist alkotnak. Így felhasználva az alábbi két tételt, az előzőekhez hasonló, az oszlopgenerálás elvén alapuló eljárás adódik a (3.6) feladat megoldására.

3.6. TÉTEL. Legyen adva a (3.6) feladat egy $\mathbf{B}$ bázisa, és jelölje $\mathbf{b} = (b_0, b_1, \ldots, b_{n+m})$ a $\mathbf{B}$ bázishoz tartozó árnyékár vektort. Legyen továbbá $1 \leqq j \leqq m$, $0 \leqq i \leqq n$ tetszőleges. Akkor érvényesek a következők:

(1) az $M_j = \max \left\{ \sum_{k=1}^{n} b_k v_k : \sum_{k=1}^{n} l_k v_k \leqq L_j, \; v_k \text{ nemnegatív egész } (k=1, \ldots, n) \right\}$ hátizsák-feladat $\mathbf{V}_0 = (v_{01}, \ldots, v_{0n})^T$ optimális megoldásával képzett $\mathbf{Q} = (f(j), v_{01}, \ldots, v_{0n}, 0, \ldots, 0, 1, 0, \ldots, 0)^T$ vektor, amelynek $n+j+1$-edik komponense 1, a (3.6) feladat egy oszlopvektora, és ha $f(j) b_0 + b_{n+j} + M_j > 0$, akkor $\mathbf{Q}$ nem szerepel a bázisban, továbbá bevonása a bázisba a célfüggvényben csökkenést eredményezhet,

(2) ha $\sum_{k=1}^{n} b_k r_k + \sum_{k=1}^{m} b_{n+k} g_k < 0$, akkor a $\mathbf{Q} = (0, -r_1, \ldots, -r_n, -g_1, \ldots, -g_m)^T$ vektor nem szerepel a bázisban, és a bázisba történő bevonása a célfüggvényben csökkenést eredményezhet,

(3) ha $b_{n+j} > 0$, akkor a $\mathbf{Q} = (0, \ldots, 0, 1, 0, \ldots, 0)^T$ vektor, amelynek $n+j+$

+1-edik komponense 1, nem szerepel a bázisban, és a bázisba történő bevonása a célfüggvényben csökkenést eredményezhet,

(4) ha $b_i>1$, akkor a $\mathbf{Q}=(0, \ldots, 0, 1, 0, \ldots, 0)^T$ vektor, amelynek $i+1$-edik komponense 1, nem szerepel a bázisban, és a bázisba történő bevonása a célfüggvényben csökkenést eredményezhet.

3.7. TÉTEL. A (3.6) feladat **B** bázisához tartozó bázismegoldás a feladat optimális megoldása, ha $f(j)b_0+b_{n+j}+M_j\leqq 0$ $(j=1, \ldots, m)$, $\sum_{k=1}^{n} b_k r_k+\sum_{k=1}^{m} b_{n+k}g_k\geqq 0$, $b_{n+j}\leqq 0$ $(j=1, \ldots, m)$ és $b_i\leqq 1$ $(i=0, \ldots, n)$.

A 3.6. és 3.7. tételek alapján adódó eljárásra szintén érvényes a 2. rész 1. megjegyzése. Ha az eljárás végén az optimum értéke pozitív, akkor a (3.5) feladatnak nincs lehetséges megoldása. Ellenkező esetben az optimum értéke 0, és ha a $p_0, p_1, \ldots, p_n$ változókhoz tartozó oszlopvektorok egyike sem szerepel az aktuális bázisban, akkor ez a bázis a (3.5) feladat egy induló bázisa. Ha az aktuális bázisban van olyan vektor, amely a $p_0, p_1, \ldots, p_n$ változók valamelyikéhez tartozik, akkor az illető vektort ki kell vonnunk a bázisból úgy, hogy egyrészt a bázisváltoztatás során az optimum értéke ne változzék, másrészt a bázisba bevont új vektor különbözzék a mesterséges változókhoz tartozó oszlopvektoroktól. A következőkben megmutatjuk, hogy ez lehetséges, és megadunk egy eljárást a bázisba bevonható vektorok meghatározására. Ennek alapján kivonva a bázisból a mesterséges változókhoz tartozó oszlopvektorokat, a (3.5) feladat egy induló bázisát kapjuk.

Most tegyük fel, hogy az aktuális bázis a **B** mátrixszal van megadva, és a **B** mátrix $\mathbf{b}_q$ $q$-adik oszlopvektora valamelyik mesterséges változóhoz tartozik. Mivel az optimum értéke 0, ezért a (3.6) feladat jobb oldalán álló vektornak a **B** bázisra vonatkozó $q$-adik koordinátája 0. Következésképpen $\mathbf{b}_q$ helyett bevonva a bázisba bármelyik olyan oszlopvektort, amelynek a **B** bázisra vonatkozó $q$-adik koordinátája 0-tól különböző, az optimum értéke a bázisváltoztatással nem változik. Így elegendő egy olyan oszlopvektort meghatározni, amely egyrészt nem mesterséges változóhoz tartozik, másrészt a **B** bázisra vonatkozó $q$-adik koordinátája 0-tól különböző. Ehhez jelölje $\mathbf{b}'_q=(b'_0, b'_1, \ldots, b'_{n+m})$ a $\mathbf{B}^{-1}$ mátrix $q$-adik sorát, és legyen $1\leqq j\leqq m$ tetszőleges. A $j$-edik félkész termékhez tartozó $(v_1, \ldots, v_n)^T$ szabás kielégíti a $\sum_{k=1}^{n} l_k v_k\leqq L_j$, $v_k$ nemnegatív egész $(k=1, \ldots, n)$ feltételt, és csak ezen feltételt kielégítő vektorok lesznek lehetséges szabások. Így a (3.6) feladatban a $j$-edik félkész termékhez azok és csak azok az $(f(j), v_1, \ldots, v_n, 0, \ldots, 0, 1, 0, \ldots, 0)^T$ alakú vektorok tartoznak, amelyeknek $n+j+1$-edik komponense 1, továbbá a $v_1, \ldots, v_n$ értékekre a $\sum_{k=1}^{n} l_k v_k\leqq L_j$, $v_k$ nemnegatív egész $(k=1, \ldots, n)$ feltétel teljesül. A tekintett vektoroknak a **B** bázisra vonatkozó $q$-adik koordinátája $b'_0 f(j)+\sum_{k=1}^{n} b'_k v_k+b'_{n+j}$. Tekintsük most a

$$d_j=\min\Big\{\sum_{k=1}^{n} b'_k v_k:\ \sum_{k=1}^{n} l_k v_k\leqq L_j,\ v_k \quad \text{nemnegatív egész} \quad (k=1, \ldots, n)\Big\}$$

$$D_j=\max\Big\{\sum_{k=1}^{n} b'_k v_k:\ \sum_{k=1}^{n} l_k v_k\leqq L_j,\ v_k \quad \text{nemnegatív egész} \quad (k=1, \ldots, n)\Big\}$$

hátizsák-feladatokat. A fentiek alapján adódik, hogy a $j$-edik félkész termékhez tartozó oszlopvektorok **B**-re vonatkozó $q$-adik koordinátái akkor és csak akkor egyenlőek rendre 0-val, ha $-b'_0 f(j) - b'_{n+j} = d_j = D_j$. Másrészt $L_i \leqq L_j$, ha $i \leqq j$, így $d_m \leqq d_{m-1} \leqq \ldots \leqq d_1$ és $D_1 \leqq D_2 \leqq \ldots \leqq D_m$. A kapott egyenlőtlenségekből a $d_1 \leqq D_1$ összefüggést felhasználva azt kapjuk, hogy abban az esetben, ha minden egyes félkész termékhez tartozó oszlopvektorok **B**-re vonatkozó $q$-adik koordinátái rendre 0-val egyenlők, akkor $d_m = D_m$. A fentiek alapján adódik az alábbi eljárás a bázisba bevonható oszlopvektor meghatározása.

1. Ha $\sum_{k=1}^{n} b'_k r_k + \sum_{k=1}^{m} b'_{n+k} g_k \neq 0$, akkor a $(0, -r_1, \ldots, r_n, -g_1, \ldots, -g_m)^T$ vektor bevonható a bázisba. Ellenkező esetben a 2. lépés következik.

2. Ha $b'_{n+j} \neq 0$ valamely $1 \leqq j \leqq m$ indexre, akkor $(0, \ldots, 0, 1, 0, \ldots)^T$ vektor, amelynek $n+j+1$-edik komponense 1, bevonható a bázisba. Ellenkező esetben a 3. lépés következik.

3. Oldjuk meg a $d_m$-nek és $D_m$-nek megfelelő hátizsák-feladatokat, és jelölje az optimumokat $d_m$ és $D_m$. Ha $d_m \neq D_m$, akkor a $b'_0 f(m) + d_m + b'_{n+m}$, $b'_0 f(m) + D_m + b'_{n+m}$ értékek különbözőek, így valamelyikük 0-tól különböző. Vegyük a 0-tól különböző értéknek megfelelő feladat $\overline{\mathbf{V}} = (\bar{v}_1, \ldots, \bar{v}_n)^T$ optimális megoldását. Akkor az $(f(m), \bar{v}_1, \ldots, \bar{v}_n, 0, \ldots, 0, 1)^T$ vektor bevonható a bázisba. Ha $d_m = D_m$, akkor a 4. lépés következik.

4. Tekintsük a $b'_0 f(j) + d_m + b'_{n+j}$ $(j = 1, \ldots, m)$ értékeket. Ha valamelyik 0-tól különböző, akkor véve a megfelelő hátizsák-feladat $\overline{\mathbf{V}} = (\bar{v}_1, \ldots, \bar{v}_n)^T$ optimális megoldását, az $(f(j), \bar{v}_1, \ldots, \bar{v}_m, 0, \ldots, 0, 1, 0, \ldots, 0)^T$ vektor, amelynek $n+j+1$-edik komponense 1, bevonható a bázisba.

Ha a megadott eljárás során nem adódik egyetlen, a bázisba bevonható vektor sem, akkor a (3.6) feladat minden egyes, nem mesterséges változóhoz tartozó oszlopvektorának a **B** bázisra vonatkozó $q$-adik koordinátája 0, ahol $q$ az előzőek során rögzített indexet jelöli. Mivel a (3.6) feladat jobb oldalán szereplő vektor **B**-re vonatkozó $q$-adik komponense is 0, ezért azt kapjuk, hogy a (3.5) feladat egyenletrendszerében a $q$-adik egyenlet nem független a többi egyenlettől, azaz a (3.5) feladat mátrixának a rangja nem nagyobb, mint $n+m$.
Másrészt egyszerű számolással ellenőrizhető, hogy a

$$(0, -r_1, \ldots, -r_n, -g_1, \ldots -g_m)^T, \quad \mathbf{Q}_i = (f(m), 0, \ldots, 0, 1, 0, \ldots, 0, 1)^T$$

$(i = 1, \ldots, n)$, az $i+1$-edik komponens 1, $\mathbf{Q}'_j = (f(j), 0, \ldots, 0, 1, 0, \ldots, 0)^T$ $(j = 1, \ldots, m)$, az $n+j+1$-edik komponens 1, vektorok a paraméterekre tett feltételeink mellett lineárisan függetlenek. Mivel a felsorolt vektorok rendre szerepelnek a (3.5) feladatban, azt kapjuk, hogy a (3.5) feladat mátrixának rangja $n+m+1$, ami ellentmondás. Következésképpen a fentiekben megadott eljárás mindig biztosítja a bázisba bevonható vektor meghatározását.

A (3.1) és (3.3) modellekkel kapcsolatosan megemlítjük, hogy a megadott eljárásokhoz hasonló eljárások származtathatók a fentiek alapján abban az esetben, ha a modellben szereplő **S** vektor helyett tetszőleges olyan vektort tekintünk, amelyre $s_i = s_j$ teljesül, ha az $i$-edik és a $j$-edik szabásokat ugyanazon félkész termékféleségből vágjuk ki. Ebben az esetben, hasonlóan az előzőekhez, egyszerűen adódnak a meg-

felelő tételek, illetve feltételek. A speciális S vektorral történt tárgyalást a gyakorlati interpretálhatóság indokolja.

Befejezésül kiemelnénk a dolgozatban ismertetett eljárások előnyös, illetve hátrányos tulajdonságait. Gyakorlati szempontból igen jelentős, hogy a lehetséges szabások száma már 3—4 különböző félkésztermék és 20—25 féle megrendelés esetén is igen magas lehet. Ennek következtében a vállalatoknál felmerülő ESZF-ek megoldása a szimplex módszer közvetlen alkalmazásával általában reménytelen a helyileg rendelkezésre álló számítástechnikai eszközökkel. Ezzel szemben az oszlopgenerálás módszere egészen kis kapacitású számítógépeken is igen jól alkalmazható. A feladat megoldásához szükséges tárigény — a program tárolásához szükséges területtől eltekintve — a különböző megrendelések számának négyzetével arányos. Hátrányt jelent kisgépes megoldásnál a futási idő növekedése. Az egyes generáló oszlopok meghatározása — különösen hosszú félkésztermék esetén — elég hosszadalmas lehet. Az utóbbi időveszteség minimális lehet vagy éppenséggel tényleges időnyereséget is eredményezhet egy helyi kisszámítógépes alkalmazásnál egy távollevő, nagy kapacitással rendelkező számítógépen történő megoldással szemben.

## IRODALOM


[1] CHARNES, A. and COOPER, W. W., "Programming with linear fractional functions", *Naval Research Logistics Quaterly* **9** (1962) 181—186.

[2] DANTZIG, G. B., *Linear Programming and Extensions* (Princeton University Press, Princeton, New Yersey, 1963).

[3] EILON, S. and CHRISTOFIDES, N., "The loading problem", *Management Science* **17** (1971) 259—268.

[4] FRIDMAN, F. A., "Solution of the Knapsack Problem with Special Supplementary Constraints", in: *Proceeding of the VI. Winter School on Math. and Related Questions, Drogobych* (1973).

[5] GALAMBOS, G., „A szabási feladat és az ehhez kapcsolódó lefedési problémák megoldása egzakt és heurisztikus módszerekkel", egyetemi doktori értekezés, JATE, Szeged, 1981.

[6] GILMORE, P. C. and GOMORY, R. E., "A linear programming approach to the cutting stock problem, Part I. " *Operations Research* **9** (1961) 849—859.

[7] GILMORE, P. C. and GOMORY, R. E., "A linear programming approach to the cutting stock problem, Part II", *Operations Research* **11** (1963) 863—888.

[8] GILMORE, P. C. and GOMORY, R. E., "Multistage cutting stock problems of two and more dimensions", *Operations Research* **13** (1965) 94—120.

[9] GILMORE, P. C., "Cutting stock, linear programming, dynamic programming and integer programming, some interconnections", *Annals of Discrete Mathematics* **4** (1979) 215—235.

[10] GOLDEN, B. L., "Approaches to the cutting stock problem", *AIEE Transactions* **8** (1976) 265—274.

[11] GVOZDEV, S. E., "A generalization of the knapsack problems", *Upravljajemüje Szisztemü* **15** (1976) 16—31.

[12] HAESSLER, R. E., "The dissagregation problem in the paper industry", in: *Production Planning and Scheduling Problems,* Ed. *Ritzmanl* (1979).

[13] HAESSLER, R. E., "Controlling cutting pattern changes in one-dimensional trim problems", *Operations Research* **23** (1975) 483—493.

[14] INGARGIOLA, G. P., "A general algorithm for one dimensional packing algorithms", *Operations Research* **25** (1977) 752—759.

[15] JOHNSON, D. S., DEMERS, A., ULLMAN, J. D., GAREY, M. R. and GRAHAM, R. L., "Worst case perfomance bounds for simple one-dimensional parcking algorithms", *SIAM Journal Computing* **3** (1974) 299—325.

[16] KANTOROVICS, Z. V., "Mathematical methods of organizing and planning production", *Management Science* **6** (1962) 366—422.

[17] METZGER, R. W., *Elementary Mathematical Programming* (John Willey & Sons Inc., New York, 1966).

[18] PIERCE, J. F., *Some Large-Scale Production Scheduling Problems in the Paper Industry* (Prentice Hall Ind., 1974).
[19] PRÉKOPA, A., *Lineáris programozás I.*, (Bolyai János Matematikai Társulat kivadványa, Budapest 1968).
[20] TENG, J. T., "Matrix algorithm for larger knapsack problems", *Bulletin Inst. Math. Acad. Sinica* **6** (1978) 197—201.

(Beérkezett: 1983. január 20.)

GALAMBOS GÁBOR
JATE KALMÁR LÁSZLÓ KIBERNETIKAI LABORATÓRIUM
6720 SZEGED, ÁRPÁD TÉR 2.

IMREH BALÁZS
JATE SZÁMÍTÁSTUDOMÁNYI TANSZÉK
6720 SZEGED, ARADI VÉRTANÚK TERE 1.

## SOLUTION OF ONE-DIMENSIONAL CUTTING STOCK PROBLEMS BY COLUMN-GENERATION

G. GALAMBOS and B. IMREH

In this paper we deal with the one-dimensional cutting stock problems. First we present the column-generation method wich was first applied by GILMORE and GOMORY. Afterwards, such cutting stock problems are considered which can be described by hyperbolic programming models. Using the column-generation method, we give algorithms to solve these hyperbolic problems both in the unbounded and bounded cases.

*Alkalmazott Matematikai Lapok* 10 (1984) 87—102

# KÉTLÉPCSŐS MATEMATIKAI MODELL ÉS INTERAKTÍV PROGRAMRENDSZER CSATORNA- ÉS SZENNYVÍZTISZTÍTÓ HÁLÓZATOK TERVEZÉSÉRE

KOVÁCS LÁSZLÓ BÉLA, BOROS ENDRE, INOTAY FERENC

Budapest

Ez a dolgozat egy új, kétlépcsős modellt mutat be, amely csatorna- és szennyvíztisztító rendszerek tervezésére alkalmas. Az első lépcsőben nagyszámú, ún. főgyűjtőrendszert generálunk, amelyek mindegyike mérnöki szempontból konzisztens és legalább egy szempontból jobb, mint a rendszer tetszőleges másik eleme. Paraméterek értékének megadásával több vagy kevesebb főgyűjtőrendszert generálhatunk, amelyek egy, a dolgozatban pontosan definiált értelemben lényegesen különböznek egymástól. A második lépcsőben egy 0—1 tiszta diszkrét programozási feladatot fogalmazunk meg és oldunk meg, amelynek az alapelemei a potenciális szennyvíztisztítók és főgyűjtőrendszerek. (Ezeket reprezentálják a 0—1 változók). Ezután a tervezést segítő interaktív számítógépes programrendszer fő funkcióit vázoljuk. Példaként a balatoni üdülőövezet egy régiójára végzett terveket hasonlítjuk össze. További kutatási irányokat, alkalmazásokat, valamint más szennyvíztisztítórendszer tervező modellekkel való összehasonlítást is megadunk.

## 1. Bevezetés

Az 1970-es évek közepétől a Balaton eutrofizálódása felgyorsult, s ha a jelenlegi tendencia fennmarad, úgy néhány éven belül a Balaton jelentős része fürdésre alkalmatlanná válik.

Az eutrofizálódási folyamat biológiailag igen bonyolult, vizsgálata évek óta folyik. Az eutrofizálódás egyik fő oka mindenesetre a vízbe jutó nagy mennyiségű szervesanyag. Ennek jelentős része a tisztítatlan vagy nem megfelelően tisztított szennyvízből ered.

1977-es adatok szerint a Balaton-parti közüzemi vízellátottság mértéke meghaladja az országos átlagot, a csatornázottság azonban ennek mélyen alatta marad. Ez számszerűen azt jelenti, hogy a közüzemi vízzel ellátott lakosoknak csak mintegy 10—12%-a volt csatornázva ebben az időben [1].

Az eltelt években a fenti arány bár jelentősen változott, jelenleg is még a keletkező szennyvizek 50—60%-a minden tisztítás nélkül kerül a Balatonba, de a tisztított szennyvizek nagyrésze is a telepek tisztítástechnológiája (mechanikai vagy biológiai tisztítás) és időszakos túlterhelése miatt a Balatonba befolyva a mikroorganizmusok tápanyagellátását szolgálja.

Mindez az eutrofizálódás egyik oka, de egyúttal a beavatkozás lehetőségét is kínálja. Ilyenek a csatornahálózat és a szennyvíztisztítók kapacitásának bővítése, speciális víztisztítási technológiák alkalmazása, továbbá a tápanyagok lebontására szolgáló kémiai tisztítási rendszerek bevezetése (pl. foszforeltávolítás).

A következő évtizedekben nagy összegű beruházásokkal kell mindezt megvalósítani. Egy 1975-ös terv 1990-ig a csatornázottságot a már említett 10—12%-ról 60% fölé kívánja emelni ([2]).

Mivel a korlátos pénzeszközök minél hatékonyabb felhasználása igen sürgető, fontos kérdéssé vált a tervezési folyamat tanulmányozása, gyorsítása.

A jelen dolgozatban egy olyan matematikai modellt és interaktív programrendszert ismertetünk, amelynek fő célja a hagyományos tervezés segítése és gyorsítása, elsősorban a lehetőségek részletesebb és gyorsabb elemzése révén.

Szennyvíztisztító rendszerek tervezése során fellépő feladatok matematikai és számítógépes megoldásával igen sokan foglalkoztak az elmúlt két évtizedben.

A dolgozatok jelentős része azonban csupán egy-egy kiragadott részfeladattal foglalkozik: csatornaszakaszok domborzattól függő optimális vonalvezetésének tervezése, optimális választás különböző szennyvíztisztítási technológiák közül stb. ([3], [8], [9], [10], [11]). A teljes hálózat kialakításával foglalkozó néhány dolgozat ([4], [6], [7]) pedig olyan pénzügyi és technológiai feltételezéseket használt, amelyek nem alkalmazhatók a Balaton esetében fellépő feladatra. Így vált szükségessé az itt ismertetendő kétfázisú matematikai modell kidolgozása.

Ez a modell elsősorban a tervezési folyamat döntési és döntéselőkészítési fázisában adhat hatékony segítséget. Az elkészült programrendszerben azonban a mérnöki számításokhoz és az adatok kezeléséhez is adottak az eszközök.

A rendszer elkészítése során célunk volt a számításigényes mérnöki-tervezési fázisok kiváltása, az adatok és eredmények egységes, gyors kezelése, valamint a matematikai modell felállítása és megoldása révén a különböző lehetséges szituációk elemzése. Mindehhez csupán azokat az adatokat tételeztük fel ismertnek, amelyek a hagyományos tervezési folyamat során is felhasználásra kerülnek.

A munka megkezdésekor, 1980-ban összegyűjtött adatokkal végzett kísérletek eredményei az akkori körülmények között jól értelmezhetők, tendenciájukban megfelelnek az elvárásoknak, néhány újszerű javaslatot is tartalmaznak. A mai tényleges döntéshozásban való alkalmazáshoz természetesen új, és a mienknél szélesebbkörű adatfelvételre volna szükség. A modell és a számítógépes rendszer rugalmassága számos, esetleg közben megváltozott körülmény vagy célkitűzés figyelembevételét is lehetővé teszi.

A következő fejezetekben először a hagyományos tervezési eljárást mutatjuk be, majd vázoljuk a fellépő feladatokat. A negyedik fejezetben a matematikai modellt, ezt követően pedig az elkészült programrendszert ismertetjük. A hatodik fejezet az alkalmazás módjairól és egy konkrét példa eredményeiről szól, végül pedig a továbbfejlesztés és a más területeken való alkalmazás lehetőségeiről beszélünk.

## 2. A tervezési-döntéshozási folyamat

A modellezés során fő célunk a jelenlegi tervezési folyamat bizonyos fázisainak hatékony segítése, gyorsítása volt. Ezen fázisok feltárásához szükséges volt a folyamat megismerése.

Mielőtt ennek ismertetéséhez kezdenénk hangsúlyoznunk kell, hogy az alábbiakban ismertetendő 11 részfeladat a valóságban nem egyetlen munkacsoporthoz vagy személyhez, de még csak nem is egyetlen intézményhez kötődik. Éppen ezért ezek sorrendje sem feltétlenül az itt közölt. Igen fontosnak tartjuk még megjegyezni azt a tényt, hogy a valódi tervezési folyamatban ugyan fellépnek a felsorolt mozzanatok, azonban időben igen hosszan, több iteráción keresztül, egy-egy mozzanat többször

is megismétlésre kerülhet. Az itt ismertetendő pontok tehát csupán egy egyszerűsített vázlatát alkotják a tényleges tervezés során fellépő különféle tevékenységeknek.

A szennyvíztisztító rendszerek hagyományos tervezése a következő főbb lépésekből áll:

1. Topográfia, azaz a hely- és vízrajzi viszonyok felderítése.
2. Szennyvízforrások felderítése, azaz
   — lakossűrűségek,
   — ivóvíz ellátottság,
   — egyéb ipari és mezőgazdasági források felderítése.

A fenti adatok jelenlegi és a távlatra előrebecsült értékeire is szükség van.

3. Régiók és határaik kijelölése. Ez azon körzetek kijelölését jelenti, amelyekben önállóan is megoldható a csatornázás és szennyvíztisztítás feladata.
4. A befogadó vizek áttekintése. Ezek elhelyezkedése, viszonya az adott területhez megszabja a telepíthető szennyvíztisztítók helyét, fokozatát.
5. A létező rendszer felmérése. A meglevő főgyűjtők (csatorna szakaszok) és szennyvíztisztítók helyének, terhelésének, kapacitásának, az alkalmazott technológiáknak az összegyűjtése.
6. Potenciális főgyűjtők (csatornaszakaszok) megadása. Hol lehetséges csatorna — gravitációs avagy nyomóvezeték — telepítése; mi a várható közvetlen szennyvízterhelése.
7. Potenciális szennyvíztisztítók kijelölése. Azaz hol lehet újat telepíteni, vagy hol lehet a már meglevőt bővíteni? Mik a lehetséges méretek, milyen technológiát lehet/kell alkalmazni?
8. Iszapkezelési technológiák; a melléktermékek hasznosítása, elszállítása.

A fenti, többnyire adatgyűjtő lépéseket követik a tulajdonképpeni tervezés lépései:

9. Néhány lehetséges rendszer tervezése. Azaz az érintett terület kijelölése; a várható terhelések alapján a szükséges méretek, kapacitások, valamint a költségek számítása. Mivel a részletes számítások igen időigényesek, így ebben a fázisban az eredmények nem részletesek, kerekített, becsült értékeket tartalmaznak. (Előtanulmány, tanulmányterv, program tanulmány, engedélyezési, döntési terv.)
10. A fenti néhány lehetőség közül a „legjobb" kiválasztása. A döntés a kiszámolt eredmények mellett néhány további objektív/szubjektív szempont alapján történik. (Beruházási program.)
11. Az előző két lépésben (esetleg több iterációt is végezve) végül is kiválasztják az egyetlen „legjobb" terv-variánst. Ennek részletekbe menő tervezése az utolsó mozzanat (kiviteli terv).

### 3. A fellépő feladatok matematikai leírása

Az előző szakaszban vázolt tervezési folyamat első nyolc lépése lényegében a későbbi döntési fázisokhoz szükséges adatok megadását írja le. A jelenlegi számítógépes rendszer ezeket a lépéseket csupán azzal segítheti, hogy egy erre felkészült adatkezelő rész révén a már felvitt és még hiányzó adatokat egyaránt gyorsan/könnyen áttekinthetővé teszi.

A döntéshozás kritikus lépései a 9—10 lépések. „Jobb" döntés eléréséhez mindenek előtt több alternatív lehetőség kidolgozására lenne szükség. Ennek jelenleg nemcsak a nagy idő és munkaigény az akadálya, hanem ezek után a több lehetőség áttekintése, a helyes választás válna nehézzé. Éppen ezért ezekben a lépésekben kaphat elsősorban hasznos szerepet a számítógép.

A fellépő és megoldandó feladat a következőképpen fogalmazható meg: Először is a tényleges lehetőségeket jól kiaknázó, nagyszámú részterv variánst dolgozunk ki, majd ezek közül választunk ki egy optimális együttest.

Az „optimális" a következőkben a minimális költségűt fogja jelenteni. Minden más elérendő célt feltételként fogalmazunk meg — azaz csak olyan tervezési lehetőségeket állítunk elő, amelyek egyéb céljainkat teljesítik.

A hatodik fejezetben beszélünk majd az általunk elképzelt alkalmazásokról részletesebben, de a fő felhasználási módnak az ismételt, különböző szituációknak megfelelő feltételek melletti futtatást képzeljük. Így mindenképpen fontos a viszonylagos gyorsaság.

Ugyancsak fontos, hogy a fellépő szempontok közül sok nehezen vagy egyáltalán nem számszerűsíthető, így követelmény az interaktivitás is. Ez pedig nemcsak a számítógépes programrendszerre jelent megszorítást, hanem a használt matematikai módszerekre is.

A rendszer és a modell leírásában fontos fogalmak: a csatornaszakasz vagy főgyűjtő; a szennyvíztisztító vagy röviden tisztító; a tervezési variáns vagy lehetséges terv; és a főgyűjtőrendszer. Ez utóbbit az alábbiakban fogjuk definiálni. Szerepe az adatok és a terv közötti „távolság" — mely bonyolultságban jelent lényeges eltérést — áthidalása, lehetővé teszi az előbb felsorolt szempontok kielégítését.

Matematikai fogalmak közül a gráf és a lineáris diszkrét-programozás fogalmait fogjuk használni.

Egy $G=(V, E)$ irányított gráfon egy $V$ csúcshalmazból és egy $E \subseteq V \times V$ élhalmazból álló párt értünk. Az élek és csúcsok jelölésére használni fogjuk az $E=E(G)$ és $V=V(G)$ jelöléseket.

Az alábbiakban ismertetjük az adatoknak megfeleltetett matematikai fogalmakat, definícióikat.

*Főgyűjtőn* két, a térképen is azonosítható pont között vezető, közel egyenes, szennyvíz szállító vezetéket értünk. Egy főgyűjtő lehet már létező vagy potenciális, azaz még nem létező, de építhető; valamint lehet gravitációs csatorna vagy nyomóvezeték. A gyakorlatban ez utóbbi különbség elsősorban a működés különbsége, de a modellben az eltérő telepítési és üzemelési költségek különböztetik meg őket.

Minden egyes főgyűjtőhöz az alábbi adatok tartoznak:

— kezdő és végpont azonosítója
— típus: gravitációs- vagy nyomóvezeték és létező vagy potenciális főgyűjtő
— hosszúság (m)
— tereplejtés (‰)
— közvetlen szennyvízterhelés ($m^3$/nap)
  külön téli, külön nyári napokra számított átlagok, és ezek jövőbeli becsült értékei.

A szennyvíztisztítók vagy röviden *tisztítók* ugyancsak lehetnek létezők vagy potenciálisak, de a már létező tisztító bővítését is potenciális tisztítóként tekintjük.

Minden egyes tisztítót az alábbi adatokkal jellemzünk:

— hely azonosítója,
— típus: létező vagy potenciális; előírt/megengedett technológiák;
— méret vagy kapacitás ($m^3$/nap), amely a tartós üzemben technológiájának megfelelő mértékben megtisztított szennyvíz napi összmennyisége.

Tekintsük most egy adott körzet adatait rögzítetteknek, és feleltessük meg a főgyűjtők által alkotott hálózatot egy irányított $G$ gráfnak. Tehát a $G$ gráf $E(G)$ élei megfelelnek a főgyűjtőknek, $V(G)$ csúcsai pedig a főgyűjtők találkozási/elágazási csomópontjainak.

Ekkor a már létező vagy potenciális tisztítók a $V$ halmaz egy $T \subseteq V$ részhalmazának felelnek meg.

Az élekhez, valamint csúcsokhoz rendelt adatokra, vagy később meghatározandó méretekre, költségekre függvényszerűen fogunk hivatkozni; például az $e=e_{ij} \in E(G)$ élnek megfelelő főgyűjtő mentén jelentkező közvetlen szennyvíz terhelést, (ill. annak téli és nyári napokra számított átlagait) $f^t(e)=f^t_{ij}$, ill. $f^n(e)=f^n_{ij}$ fogja jelölni.

Nevezzük a $G$ gráf egy $G^\circ \subseteq G$ részgráfját *lehetséges terv*nek, ha

$$1^\circ \quad V(G^\circ) \subseteq V(G), \qquad E(G^\circ) \subseteq E(G)$$

$$2^\circ \quad d_0^+(j) = \begin{cases} 0, & \text{ha} \quad j \in T \cap V(G^\circ) \\ 1, & \text{ha} \quad j \in V(G^\circ) \setminus T \end{cases} \qquad \text{minden} \quad j \in V(G^\circ) \quad \text{esetén,}$$

ahol $d_0^+(j)$ a $G^\circ$ gráfban a $j \in V(G^\circ)$ csúcsból kiinduló (irányított) élek számát jelenti.

Szavakban e második feltétel azt jelenti, hogy a $G^\circ$ gráf csúcsainak megfelelő csomópontokból a $G^\circ$ gráf éleinek megfelelő csatornákon befolyó szennyvizet egyetlen továbbvezető csatorna szállítja el, mindaddig amíg a szennyvíz meg nem érkezik egy szennyvíztisztítóhoz, ahonnan már nem vezetjük tovább, mint szennyvizet.

Világos, hogy a $G^\circ$ lehetséges terv kijelölése után az input adatok már egyértelműen meghatározzák, megvalósítás esetén, a hálózat elemeinek szükséges minimális méreteit. Ez azt jelenti, hogy

— minden $e \in E(G^\circ)$ élhez meghatározható a $d(e)$ csőátmérő (cm), amelyen az adott körülmények között (hossz, tereplejtés, típus, ...) lehetséges a várhatóan jelentkező szennyvízterhelés továbbítása,

— minden olyan $i \in V(G^\circ)$ csúcshoz, ahonnan az egyetlen kivezető $e_{ij} \in E(G^\circ)$ él nyomóvezetéknek felel meg, olyan teljesítményű ($m^3$/óra) átemelő berendezést tervezhetünk csak, amely képes a jelentkező szennyvízmennyiség továbbítására.

Mindkét alkalommal a napi csúcsterhelésre kell méretezni, aminek óránkénti mennyiségét a napi össz. mennyiség 1/14-ed részével becsüljük,

— minden $t \in V(G^\circ) \cap T$ csúcshoz meghatározható a telepítendő szennyvíztisztító szükséges minimális mérete ($m^3$/nap).

A $G^\circ$ lehetséges tervet *megfelelően méretezett*nek nevezzük, ha a hálózat elemeihez kapacitás (méret, teljesítmény) értékeket rendeltünk, és ezek nem kisebbek mint az előbb meghatározott minimális értékek, továbbá a kapacitás értékeknek megfelelő költségeket is meghatároztuk.

Egy megfelelően méretezett $G^\circ$ lehetséges tervet végül is néhány főbb paraméterrel jellemzünk:

$f(G^\circ)$ — az összegyűjtött és megtisztított szennyvízmennyiség nyári csúcsidőszakban ($m^3$/nap)

$c(G^\circ)$ — a beruházás, üzemeltetés és fenntartás összköltsége $r$ évre összesen (eFt), ahol $r$ a kérdéses tervidőszak hossza, rögzített tervezési paraméter*.

Továbbá ilyen paraméterek lehetnek a víztisztítás hatásfokát jellemző adatok, pl.,

$p(G^\circ)$ — a Balatonba kerülő foszfor mennyisége (kg/nap), tiszta foszforban számítva.

Ezek után egy adott régióban $(G=(V, E))$ és adott tervidőszakban $(r)$ jelentkező beruházás (döntési) feladatát a következőképpen fogalmazhatjuk meg:

Keressük azt a minimális $c(G^\circ)$ költségű, megfelelően méretezett $G^\circ \subseteq G$ lehetséges tervet, amelyre

$$\text{(P1)} \qquad \begin{aligned} f(G^\circ) &\geqq Nq \\ (p(G^\circ) &\leqq M), \end{aligned}$$

ahol $N$ a csatornával ellátandó lakosok száma, $q$ az ún. lakosegyenérték, amely a Balaton környékén = 0,2 ($m^3$/fő/nap), és $M$ az adott régióban még elfogadható foszforterhelés felső korlátja.

A P1 feladat a fentiek alapján még igen nehezen kezelhető, hiszen egy adott $G$ gráfban még igen nagy számban vannak kijelölhető $G^\circ$ lehetséges tervek. Ekkor nemcsak az így jelentkező nagy számításigény (méretezések, költségek) okoz nehézséget, hanem már az összes lényegesen különböző $G^\circ$ részgráf megkeresése is.

A $G^\circ$ lehetséges tervek direkt felsorolása helyett azok implicit kezelése szükséges. Ezt segíti a következő fogalom, és annak felhasználása.

Nevezzünk egy $F \subseteq G$ kapacitásokkal ellátott részgráfot *főgyűjtőrendszer*nek, ha

1° $F$ lehetséges tervet alkot,
2° $F$ összefüggő részgráf,
3° $F$, mint lehetséges terv, megfelelően méretezett.

Szavakban elmondva egy gyökeres irányított fát nevezünk főgyűjtőrendszernek, ha a gyökérpont egy szennyvíztisztítónak felel meg, az élek a gyökérpont „felé" vannak irányítva és erre a részgráfra mint lehetséges tervre definíció szerint elvégeztük a méretezés és költség számításokat.

Hasonlóan mint a lehetséges terv esetében most is értelmezhetők az $f(F), p(F), \ldots$ függvények. Könnyen láthatóan az $F$ főgyűjtőrendszerben pontosan egy $t \in V(F) \cap T$ tisztítónak megfelelő csúcs található. Jelöljük ekkor $\hat{c}(F)$-fel, hasonlóan mint $c(F)$ esetén, az $r$ év alatt felmerülő összes költséget, elhagyva a $t$ tisztítónál fellépő, a megtisztított szennyvíz mennyiségétől független költségeket.

Legyenek most $F_1, \ldots, F_k$ főgyűjtőrendszerek, melyekre

$$V(F_i) \cap V(F_j) \subseteq T, \quad \text{ha} \quad 1 \leqq i \neq j \leqq k.$$

Legyen ekkor $G^\circ = \bigcup_{i=1}^{k} F_i$, és a $T \cap V(G^\circ)$-beli csúcsokban végezzük el a méretezési és költségszámításokat.

* Pontosabban: a kiindulási évre diszkontált összes üzemeltetési költség az egy évi üzemköltségnek $\lambda$-szorosa, ahol $\lambda = 1 + \sigma + \ldots + \sigma^{r-1}$, $\sigma$ a diszkontfaktor. Ily módon az egyszeri beruházási és az évenként fellépő üzemeltetési költség összemérhetővé válik.

Könnyen láthatóan ekkor $G^\circ$ egy megfelelően méretezett lehetséges terv lesz, és minden megfelelően méretezett lehetséges tervet megkapunk ilyen módon.

Megfogalmazhatjuk ekkor a kétfázisú tervezési eljárást:

Az *első fázis* során előállítjuk főgyűjtőrendszerek egy $S=\{F_1, \ldots, F_m\}$ halmazát, amely a következő tulajdonságú: Tetszőleges $F$ főgyűjtőrendszerhez van olyan $F_i \in S$, amelyre a $V(F), f(F), \hat{c}(F)$ és $V(F_i), f(F_i), \hat{c}(F_i)$ jellemzők nem nagyon térnek el, azaz

$$|V(F) \setminus V(F_i)| + |V(F_i) \setminus V(F)| \leqq K1,$$

$$|f(F) - f(F_i)| \leqq K2,$$

$$|\hat{c}(F_i) - \hat{c}(F)| \leqq K3.$$

Továbbá $S$ elemei nem nagyon hasonlók, azaz

$$|V(F_i) \setminus V(F_j)| \geqq K4 \quad \text{vagy} \quad |V(F_j) \setminus V(F_i)| \geqq K4.$$

A $K1$, $K2$, $K3$, $K4$ korlátok mindenkor az adott régiónak és a kitűzött céloknak megfelelően adhatók meg.

A $K1$, $K2$, $K3$, $K4$ paraméterek megfelelő választásával elérhetjük, hogy a főgyűjtőrendszerek (azaz a részterv variánsok) száma se túl nagy, se túl kicsi ne legyen, ezen kívül lehetőleg a lényegesen különböző főgyűjtőrendszereket vizsgáljuk a továbbiakban.

A *második fázis* során az előállított $S$ halmaz elemeiből kialakítható $G^\circ$ megfelelően méretezett lehetséges tervek között keressük azt, amelyik a P1 feladat feltételeit minimális költséggel megvalósítja.

Mindkét fázis megoldásával a következő fejezetben foglalkozunk.

## 4. Kétlépcsős matematikai modell

### 4.1 *Az első fázis során fellépő feladat megoldása*

A feladat tehát egy megfelelő $S$ halmaz generálása. Ezt segíti a következő két észrevétel:

1. A $G$ gráfban a gravitációs főgyűjtőknek megfelelő élek elágazás- és hurokmentes részgráfot alkotnak.
2. Legyen $E^*$ egy tetszőleges részhalmaza a $G$ gráf gravitációs csatornáknak megfelelő éleinek, ekkor a

(P2) $$\min \{c(F) | F \text{ főgyűjtőrendszer és } E^* \subseteqq E(F)\}$$

feladat egyértelműen megoldható.

Az első következik abból, hogy a tereplejtés meghatározza a gravitációs továbbvezetés lehetséges irányát. Ha mégis volnának elágazó gravitációs főgyűjtők, akkor 0 költségű nyomóvezetékek közbeiktatásával biztosítható ez a tulajdonság.

A második észrevétel nyilvánvaló, és P2 megoldása megkapható egy egyszerű leszámlálás segítségével, amely szerencsésen gyorsítható.

Legyen ekkor $t \in T$ egy tisztítónak megfelelő kijelölt csúcs a $G$ gráfban. Legyen továbbá $G_t \subseteq G$ az a feszített részgráf, amelyre $V(G_t)$ az összes olyan $i \in V(G)$ csúcsot tartalmazza, ahonnan vezet irányított út a $G$ gráfban a $t$ csúcshoz.

Legyen most $E^*$ az összes gravitációs főgyűjtőnek megfelelő $G_t$-beli élek halmaza, és legyen $F_0$ a P2 feladat megoldása.

Legyen továbbá $F_{i+1}$ a következő P3 feladat megoldása ($i=0, 1, 2, \ldots$):

(P3)
$$\min q\frac{\hat{c}(F)}{f(F)} \quad \text{feltéve, hogy}$$
$$F \quad \text{főgyűjtőrendszer,}$$
$$F \subseteq F_i \quad \text{és}$$
$$\alpha_1 f(F_0) \leqq f(F_i) - f(F) \leqq \alpha_2 f(F_0)$$

Szavakban, ha már $F_i$-t meghatároztuk, úgy az $F_i$-ből ágak elhagyásával kapható $F$ főgyűjtőrendszerek közül válasszuk azt, amely $F_i$-től lényegesen különböző ($\alpha_1$), de nem tér el túlságosan ($\alpha_2$), és az (eFt/fő)-ben mért relatív költségessége a minimális.

Ha a (P3) feladatnak már nem volna megoldása, de az

$$\{F | F \subseteq F_i, \quad \alpha_2 f(F_0) \leqq f(F_i) - f(F)\}$$

halmaz még nem üres, akkor $\alpha_2$ értékét ideiglenesen megnövelhetjük a szükséges mértékben.

Legyen ekkor $S_t = \{F_i | f(F_i) > 0, \ i=0, 1, \ldots\}$.

Természetesen a már létező főgyűjtőkhöz tartozó élek elhagyását nem engedjük meg, ha ez előírás.

Itt pl., $\alpha_1 = 0{,}1$ és $\alpha_2 = 0{,}2$ választás esetén az $S_t$ halmaz kb. 5—10 főgyűjtőrendszert fog tartalmazni.

Legyen végül

$$S = \bigcup_{t \in T} S_t.$$

Megjegyezzük, hogy az $\alpha_1$ és $\alpha_2$ paraméterek megfelelő választásával az egyáltalán lehetséges főgyűjtőrendszerek egy „tetszőlegesen sűrű" $S$ reprezentáns halmazát generálhatjuk íly módon. Következésképpen a (P1) feladat optimumában szereplő főgyűjtőrendszereket is jól közelíthetjük $S$ elemeivel.

### 4.2 *A második fázis feladata*

Adott $S$ halmaz esetén a P1 feladat legjobb $S$-beli megoldásának megkeresése felírható egy 0—1 változós diszkrét programozási feladatként, nevezzük ezt a továbbiakban *globális modell*nek.

A felállítandó globális modell célja egy adott területre kiválasztani a lehetséges főgyűjtőrendszerek és szennyvíztisztítók közül azokat, amelyek egy minimális költségű lehetséges tervet alkotnak. A megkövetelt feltételek között van az adott területre előírt csatornázottság (százalékosan), és a megengedett maximális foszforterhelés. A további feltételek a mérnöki realizálhatóságot biztosítják, azaz település(rész)enként

legfeljebb egyetlen főgyűjtőrendszer készülhet el; minden összegyűjtött szennyvizet valamelyik betervezett tisztítóhoz kell szállítani; a betervezett tisztítók kapacitása legalább akkora legyen mint a terhelésük.

A modell változói a lehetséges főgyűjtőrendszereknek, illetve tisztítóknak felelnek meg; a változók a 0 vagy 1 értéket vehetik fel, aszerint, hogy az adott objektum (főgyűjtőrendszer vagy tisztító) kimarad vagy bekerül a megoldásba.

*Jelölések:*

*A körzet jellemzői:*

$S$ a megengedett főgyűjtőrendszerek halmaza,
$T$ a megengedett tisztítók halmaza,
$V$ a körzet településeinek halmaza,
$N$ csatornázandó lakosegységek előírt száma,
$M$ megengedett legnagyobb foszforterhelés (kg/nap)-ban,
$q$ lakosegyenérték, 0,2 ($m^3$/fő/nap),
$r$ tervidőszak hossza (év),
$\lambda = 1 + \sigma + \ldots + \sigma^{r-1}$, ahol $\sigma$ az ún. diszkontfaktor.

*A megengedett főgyűjtőrendszerek (rövidítve FGYR-ek) jellemzői:*

$i$ a figyelembe vett FGYR-ek indexei, ($i \in S$),
$x_i \in \{0, 1\}$ az $i$-edik FGYR-nek megfeleltetett változó,
$c_i$, $u_i$ beruházási és üzemköltségek ($c_i + \lambda u_i = \hat{c}(F_i)$), amelyek magukba foglalják az $F_i$ főgyűjtőrendszer által összegyűjtött szennyvizek megtisztításánál fellépő költségek mennyiségtől függő részét,
$f_i = f(F_i)$ szennyvízhozam ($m^3$/nap),
$p_i = p(F_i)$ foszforhozam (kg/nap).

*A lehetséges tisztítók és méreteik:*

$j \in T$ a figyelembe vett tisztítók indexei,
$k \in U_j$ a $j$-edik tisztító megengedett változatainak (kapacitás, technológia) indexei,
$y_{jk} \in \{0, 1\}$ a $j$-edik tisztító $k$-adik változatának megfeleltetett változó,
$d_{jk}$, $r_{jk}$ beruházási és az üzemelési költségeknek a ténylegesen megtisztított szennyvíz mennyiségétől független része (a mennyiségtől függő költségek $c_i$, ill. $u_i$-ben foglaltaknak,
$K_{jk}$ a $j$-edik tisztító $k$-adik változatának kapacitása ($m^3$/nap)-ban,
$\mu_j$ foszfor eltávolítás hatásfoka,
$g(\mu_j)$ foszfor eltávolítás egységköltsége.

*A hálózat jellemzői:*

$W_j \subset S$ azon FGYR-k halmaza, amelyek gyökere a $j \in T$ tisztító,
$S_v \subset S$ azon FGYR-k halmaza, amelyek elvezetik a $v \in V$ (rész) település szennyvizét.

*A matematikai modell:*

$$\min \sum_{i\in S}(c_i+\lambda u_i)x_i+\sum_{j\in T}\sum_{k\in U_j}(d_{jk}+\lambda_{jk})y_{kj}+\lambda\sum_{j\in T}g(\mu_j)\sum_{i\in W_j}p_ix_i$$

$$\sum_{i\in S} f_ix_i \geqq qN \tag{1}$$

$$\sum_{i\in S_v} x_i \leqq 1 \qquad (\forall v\in V)$$

$$\sum_{k\in U_j} y_{jk} \leqq 1 \qquad (\forall j\in T) \tag{3}$$

$$\sum_{i\in W_j} f_ix_i \leqq \sum_{k\in U_j} K_{jk}y_{jk} \qquad (\forall j\in T) \tag{4}$$

$$\sum_{j\in T}\mu_j\sum_{i\in W_j}p_ix_i \leqq M \tag{5}$$

$$x_i, y_{jk}\in\{0,1\} \qquad (\forall\ i,j,k) \tag{6}$$

*A célfüggvény és a feltételek jelentése:*

Miclőtt rátérnénk a célfüggvény ismertetésére, külön kell szólnunk a méretezési és költségszámításokról.

A főgyűjtőrendszerekhez tartozó objektumok (gravitációs- és nyomócsövek, átemelő berendezések) méretezését a főgyűjtőrendszer generálásakor végezzük, az $S$ halmazhoz már a megfelelő méretezésekkel ellátva csatoljuk. A generálás során elsődlegesen a főgyűjtőrendszer gráfszerű hálózatát alakítjuk ki, majd ennek alapján (és az inputként rendelkezésre álló adatok: csatorna hossz; várható közvetlen szennyvízterhelés télen és nyáron; talaj meredeksége alapján) először meghatározzuk minden egyes csatorna szakaszon a várhatóan átfolyó szennyvízmennyiséget, majd ennek alapján a szükséges átmérőt. Az átemelőberendezéseket a rendszer a szükséges helyekre beiktatja és a továbbítandó szennyvízmennyiség alapján a megfelelő teljesítményt meghatározza.

Mindezek a számítások előre megadott táblázatok alapján történnek, ezek adatainak ellenőrzése, esetleges felújítása a számítógépes rendszerbe beépített funkció. A táblázatok jelenlegi adatai az 1980-ig érvényes szabványoknak és a mérnöki gyakorlatnak megfelelőek.

Ezek után ugyancsak táblázatok alapján ([12]) történik a főgyűjtőrendszerekhez rendelt $c_i$ és $u_i$ költségtényezők számítása.

Az így kialakuló költségek és a további lehetséges változtatások alapján módosítjuk a főgyűjtőrendszer gráfját és az új hálózatra elvégezzük e számításokat újra.

A szennyvíztisztítók méretezése a modell felállításának pillanataiban történik. Ekkor a figyelembe vett főgyűjtőrendszerek $S$ halmaza és szennyvíztisztítók $T$ halmaza alapján kialakulnak a $j\in T$ tisztítókhoz vezető főgyűjtőrendszerek $W_j$ halmazai. Ekkor ezek szennyvízterhelései alapján (és a lehetséges szennyvíztisztító méretek alapján, amelyeket előre táblázatosan adnak meg), meghatározzuk azokat a lehetséges méreteket, amelyek egy-egy lehetséges megoldásban előfordulhatnak (hiszen a $W_j$ elemei közül általában csak néhány kerülhet bele egy megoldásba, így csupán ezek várható terhelésére kell méretezni). Mivel előre nem ismert, hogy mely lehetséges meg-

oldás lesz optimális, így minden tisztítót több lehetséges kapacitással is figyelembe veszünk a modellben ($U_j$ halmazok). Minden ilyen változatban már figyelembe vettük a továbbvezetési lehetőségek alapján szükséges technológiai típust is (azaz I., II., III. fokozatú tisztítás szükséges-e?)

A technológiai fokozatok és a méret alapján táblázat ([12]) segítségével határozzuk meg a modellben szereplő költségtényezőket.

A modellben szereplő költségek elsődlegesen kétfélék: Egyszeri költségek (beruházás) és évenként felmerülők (üzemelési, munkabérek, energia és üzemanyag költségek, vegyszerköltségek, amortizációk.) Ez utóbbiakat a kiindulási évre vetítjük a $\sigma$ diszkontálási faktor segítségével. Így egy $r$ éves periódusban az egy évben felmerülő költségek $\lambda$-szorosa lép fel ($\lambda = 1+\sigma+\ldots+\sigma^{r-1}$) a kiindulási évre vetítve. ($\sigma \simeq 0{,}8$ jelenleg használt érték).

Pusztán a modell megkívánta okokból a szennyvíztisztítóknál fellépő költségeket más szempont szerint is két részre bontottuk.

A fellépő költségek egy része független a ténylegesen megtisztított szennyvízmennyiségtől (beruházás, amortizáció, bizonyos üzemelési és munkabér, bizonyos energia költség). Ezeket tartalmazzák a $d_{jk}$, $r_{jk}$ együtthatók.

A költségek más része arányos a megtisztított szennyvíz mennyiségével (vegyszerköltség, energiaköltség). Ezen költségek arányos részét az adott szennyvíztisztítóhoz vezető főgyűjtőrendszereknél vettük figyelembe (ezeket tartalmazzák a $c_i$ és $u_i$ együtthatók).

E szétválasztásnak elsődlegesen a kétfázisú tervezési menet az oka.

Természetesen a modellben célfüggvényértékként fellépő összköltségben ezek a részköltségek már összegződnek, így a célfüggvény a valódi összköltséget tükrözi.

A modellben tetszőleges foszforeltávolítási technológia figyelembe vehető. A jelenleg működő számítógépes rendszerben az alumíniumszulfátos foszforeltávolítás költségeivel számoltunk. A figyelembe vett lehetséges hatásfok értékek a 75%, 85%, 95% voltak. A megfelelő változó költség döntőrészben a szükséges vegyszer árából tevődik ki, amelynek mennyisége könnyen számítható.

A célfüggvény három tagból áll.

Az első tag a főgyűjtőrendszerek $r$ évi összköltségét méri, azaz a beruházási költségeket, valamint $r$ évre az amortizációs, az átemelők üzemelési és az összegyűjtött szennyvíz tisztítási költségeit tartalmazza.

A második tag a víztisztítók beruházási költségeit, valamint $r$ évre a megtisztított szennyvíz mennyiségétől független változó költségeket (üzemelés, munkabér, ...) tartalmazza.

A harmadik tag a foszfor eltávolítás $r$ évi összköltségét (vegyszer+munkabér) tartalmazza.

Az (1) feltétel biztosítja, hogy a megoldások valóban legalább $qN(\mathrm{m}^3/\mathrm{nap})$ szennyvíz összegyűjtésére legyenek alkalmasak.

A (2) feltételcsoport biztosítja, hogy minden település(rész)t legfeljebb egyetlen FGYR érintsen. Itt, ha valamely település(rész)re egyenlőséget írunk elő, úgy ott minden megoldásban biztosan lesz betervezett FGYR. Mindazokon a település(rész)-eken, ahol már van meglevő hálózat, automatikusan egyenlőséget követelünk meg. Tehát a már meglevő csatornákat a készülő megoldásokban felhasználjuk.

A (3) feltételcsoport azt biztosítja, hogy minden tisztító legfeljebb egyetlen változatában kerülhessen a megoldásokba. A már létező, de nem megszüntethető (bővíthető) tisztítók esetében egyenlőséget követelünk meg.

A (4) feltételcsoport biztosítja, hogy bármely tisztító betervezett kapacitása elegendő legyen az odaszállított szennyvíz megtisztításához.

Az (5) feltétel azt biztosítja, hogy az előírt foszforszintet a megoldások ne lépjék át.

A (6) feltételek azt jelentik, hogy a figyelembe vett objektumokra — FGYR-k, tisztítók valamely változatukban — csak kétféle döntés hozható: megvalósuljon avagy sem.

## 5. Interaktív programrendszer

Az IBM 3031 számítógépen APL nyelven készült el a programrendszer. A rendszer célja a globális modell felállítása és megoldása, valamint az ehhez szükséges adatkezelés, az eredmények értékelésével és gyors megjelenítésével együtt.

A rendszer interaktivitása a következőket jelenti:

A főeljárás meghívása után a rendszer a terminálon keresztül kérdéseket/választási lehetőségeket ad a felhasználónak, ismételten, több lépcsőben. A megfelelő válaszokkal érhető el a kívánt működés. A kérdés-felelet sorozat hossza bármely működési cél esetében nem több mint 6. Minden pillanatban mód van egy korábbi kérdéshez visszatérni (pl., hogy módosítsuk válaszunkat, ...). A feltett kérdések többnyire néhány karakterben, de legfeljebb egyetlen sorban megválaszolhatók. A tévedések számát csökkentendő a válaszok egy részénél (ahol lehetséges) azok egyszerű ellenőrzése is megtörténik. Amennyiben valamely válaszhoz szükség van olyan információra, amely az adatbázisból nyerhető, úgy kívánság esetén az lekérdezhető a képernyőn keresztül.

Egy sorozatban a kérdések között csupán néhány másodpercnyi a várakozás.

A teljes rendszer mintegy 50 függvényeljárásból (kb. 2500 APL sor) áll, s ebből kb. 20 szervezi a rendszer interaktív működtetését.

Az eljárások 7 főbb funkción keresztül működtethetők (az első kérdés az ezek közüli választás), és három főbb csoportba sorolhatók.

Az alábbiakban, csoportosítva, a főbb funkciók rövid leírását adjuk.

### 5.1 *Adatbázis, adatkezelés:*

UPDATE — Az adatok felvitelére, vagy módosítására szolgál; természetesen nagyobb mennyiségű adat a szokásos batch üzemmódban is bevihető a rendszerbe.

REGIONS — Az adatbázisból nyerhető információk lekérdezésére szolgál. Kérés esetén a táblázatos, vagy egyszerű térképes választ a sornyomtatóra is elküldi.

RESULT — A nyert eredmények áttekintésére, értelmezésére, tömör vagy részletes kiíratására és néhány szempont szerinti elemzésére szolgál.

### 5.2 *Méretezési és költség számítások*

ENGMSS — Egy kijelölt $F$ főgyűjtőrendszerre végzi el a méretezési- és költségszámításokat.

FGYRGEN — A 4.1-ben leírtaknak megfelelően egy kijelölt $t \in T$ tisztítóhoz elkészíti a főgyűjtőrendszereknek egy $S_t$ sorozatát.

### 5.3 *A globális modell felállítása és megoldása*

PROBGEN A kijelölt és lekérdezett adatoknak megfelelően a globális modell együtthatóit kiszámítja és tárolja a megoldó algoritmusnak megfelelő formátumban.

PROBSOL A felállított feladat(ok) megoldását végzi. A jelenlegi változatban az IBM 3031-en meglevő MPSX programcsomag segítségével.

## 6. A rendszer alkalmazása, futtatási eredmények

A teljes Balatonpartra elkészített adatbázis és a működő rendszer néhány tipikus jellemzője:

— A hálózat összesen kb. 1000 főgyűjtőből és 50 tisztítóból áll. Mindezek 7 régióra oszthatók. A teljes hálózat automatikusan három független részre bomlik, északi, keleti, déli partra.

— Egy adott $t$ tisztítóhoz a megfelelő $S_t$ halmaz előállítása (FGYRGEN) mintegy 2—15 percet vett igénybe, az adott körzet méretétől függően.

— A legnagyobb felállított feladat kb. 800 változót és 90 feltételt tartalmazott; az egy régióra vonatkozó feladatok átlagosan 200—300 változót és 20—30 feltételt tartalmaznak.

— A feladatok generálása (PROBGEN) 2—3 percig tart; a megoldás (PROBSOL, MPSX) a csatornázottság megkövetelt értékétől függően 1—5 percig. (A 800 változós feladatot az MPSX magasabb csatornázottsági szint esetén nem tudta megoldani 15 percen belül.)

Az itt közölt idők nem CPU idők, hanem a terminál előtt eltöltött valódi időtartamokat jelentik, a gép átlagos terheltsége esetén, kivéve az MPSX futtatást.

Mindezek alapján úgy tűnik, igen sokféle kísérletet lehet gyorsan elvégezni segítségével, így valóban hatékony segítséget adhat a tervezéshez.

Ilyen lehetséges kísérletek a következők:

Egy adott körzetre (pl. egy régióra) különböző tervidőszakok $r=5, 10 \ldots, 30$ év) és különböző csatornázottsági követelmények (25—60%) mellett végezni futtatásokat. Az eredményekből átfogó képet nyerhetünk az adott körzet jellemzőiről, egy adott költségkeret lehetséges felhasználásáról, a különböző variánsok ütemezéséről, stb.

Egy adott körzetben különböző tisztítókat javasolva választ kaphatunk arra, hogy az adott területen milyen különbség van a néhány regionális, vagy több lokális tisztítóra épülő tervek műszaki, gazdasági paraméterei között. Esetleg mindez miképpen függ a megkövetelt csatornázottságtól.

Egy adott körzetben különböző településeken írhatjuk elő a kötelező csatornázást, és így egy hosszabb távú beruházás különféle ütemezései vethetők egybe.

Természetesen nagyobb feladatokat felállítva egyszerre több körzet is vizsgálható, s így esetleg újabb információt nyerhetünk a beruházások sorrendezéséről, a különböző technológiai változtatások (pl. foszforeltávolítás szintje) régiónként eltérő mértékéről.

A rendszer és a modell csekély módosításával még igen sok más kérdés vizsgálata válik lehetővé (pl. mellékgyűjtők kezelése, talajviszonyok figyelembe vétele, stb.).

Az eddig elmondottakat példával illusztráljuk a következőkben:

Az ún. balatoni V. régióra készítettünk néhány futást az 1980-as költség és az 1990-es (becsült) terhelés adatok alapján.

Az V. régió a Balaton északi partján Zánka—Szigliget között helyezkedik el. Az összesen 10 településből álló adatbázisnak megfelelő $G=(V, E)$ gráfban $|E|=130$, $|V|=100$. A hálózatban kb. 12 km a már meglevő csőhálózat hossza és kb. 3000 m³/nap az így összegyűjtött szennyvíz mennyisége. A potenciális főgyűjtők hossza kb. 80 km és ezek kb. 6000 m³/nap szennyvízmennyiséget gyűjthetnek össze az 1980-as adatok értelmében. A régió átlagos csatornázottsága mintegy 22%.

A régióban már létezik két tisztító (Révfülöp, 1200 m³/nap és Badacsony, 600 m³/nap). Lehetséges ezek bővítése, főként a révfülöpié, a badacsonyié csak igen kis mértékben; lehetséges néhány kis méretű lokális tisztító telepítése (Zánka, Szigliget, Badacsonyörs) vagy egy nagyobb regionális központi telep létrehozása (Badacsonytomaj közelében). $r=30$ évre és 25%—60% megkövetelt csatornázottság mellett (5%-onként növekvő lépcsőkben) készültek futások.

További követelés volt, hogy a tisztítókat mind III. fokozatra (mechanikai, biológiai tisztítás és foszforeltávolítás) kell tervezni, hiszen csak a Balaton a lehetséges befogadó, közvetve vagy közvetlenül.

A megoldások mindegyikében a révfülöpi és badacsonyi meglevő tisztítók bővítése, valamint két másik kis tisztító létesítése szerepel (Szigliget, Badacsonyörs). Csupán 55—60%-os megkövetelt csatornázottság esetén került be a megoldásba a lehetséges regionális tisztító (Badacsonytomaj).

A 30 évre számított összköltséget az 1. táblázat 2. oszlopa mutatja.

Összehasonlításként ugyanilyen csatornázottságok mellett, csak a regionális tisztítót megengedve készítettünk futásokat. Ez kb. 100—200 MFt többletköltséget jelent, amint az az 1. táblázat 3. oszlopából kiolvasható. De meg kell jegyezni, hogy a relatív nagyobb összköltség ellenére a regionális szennyvíztisztítótelep létrehozása és üzemeltetése a Balaton szempontjából sokkal biztonságosabb (egyszerűbb felügyelet, karbantartás), mint sok kis telep üzemeltetése.

1. TÁBLÁZAT

| Megkövetelt csatornázottság (%) | Összköltség ált. feltételek (MFt) | Összköltség csak regionális tisztítók (MFt) |
|---|---|---|
| 30 | 500,7 | 715,6 |
| 35 | 603,4 | 812,5 |
| 40 | 731,3 | 959,8 |
| 45 | 848,8 | 1050,6 |
| 50 | 1013,6 | 1165,8 |
| 55 | 1179,9 | 1293,1 |
| 60 | 1419,3 | 1520,2 |

## 7. Továbbfejlesztési lehetőségek

A már eddigiek mellett még sokrétű alkalmazási lehetőségek kínálkoznak a rendszer csekély változtatásával.

— Megcserélhető a célfüggvény és az (1) feltétel szerepe a modellben, és így egy adott pénzkeret leghatékonyabb felhasználására kaphatunk javaslatot.

— Lehetséges a kisebb körzetekben részletesebb adatrendszer készítése (mellékgyűjtők, talajviszonyok, stb.) és így egy kisebb területen hasonló, ámde részletesebb elemzéseket végezhetünk.

— Kicsit nagyobb változtatás árán lehetőség nyílna a tervek időbeni ütemezésére, most már pontosabb modellel.

A konkrét balatoni felhasználást illetően mindenképpen szükségesnek látszik a teljes balatoni vízgyűjtő terület vizsgálata. Ehhez alig kell változtatni a meglevő rendszeren, csupán a megfelelő adatokat kell beszerezni és a számítógépbe táplálni.

A meglevő módszerek és a rendszer egészen más területeken is felhasználható volna, pl.:

— ivóvíz szolgáltatás bővítése egy adott területen,
— elektromos hálózat fejlesztése,
— egy nagyváros telefon hálózatának felújítása, bővítése.

Különösen ez utóbbi területet érezzük fontosnak, hiszen a következő évtizedekben Budapesten jelentős összegű beruházásokkal kell a telefonhálózatot fejleszteni.

## IRODALOM

[1] A Balaton üdülőkörzet regionális rendezési terve, Városépítési Tudományos és Tervező Intézet, Budapest, 1980. március, Tsz.: 3754/78.
[2] Balaton térség szennyvízelvezetése, Tanulmány, VIZITERV, Budapest, 1975. Tsz.: 19 936.
[3] DAJANI, J. S., HASIT, Y. and MCCULLERS, S. D., "Mathematical programming in sewer network design", *Engineering Optimization* **3** (1977) 27—35.
[4] JARVIS, J. J., RARDIN, R. L., UNGER, V. E., MOORE, R. W. and SCHIMPELER, C. C., "Optimal design of regional wastewater systems: A fixed-charge network flow model", *Operations Research* **26** (1978) 538—550.
[5] KOVÁCS, L. B., DOBOLYI, E. és INOTAY, F., „A Balaton térség szennyvíztisztítási modelljének kiinduló pontjai és problémái", *MTA SZTAKI Working Paper* MO/17 1980.
[6] MAYS, L. W., WENZEL, H. G. and LIEBMAN, J. C. "Model for layout and design of sewer systems", *ASCE* **102** (WR2) (1976) 385—405.
[7] MAYS, L. W. and YEN, B. C. "Optimal cost design of branched sewer systems", *Water Resources Research* **11** (1975) 37—47.
[8] MERRIT, L. B. and BOGAN, R. H., "Computer-based optimal design of sewer systems", *ASCE*, **99** (EE1) (1973) 35—53.
[9] ORON, G., "An algorithm for optimizing nonlinear constrained zero-one problems to improve wastewater treatment", *Engineering Optimization* **4** (1979) 109—115.
[10] WALSH, S. and BROWN, L. C., "Least cost method for sewer design", *ASCE*, **99** (EE3), (1973), 333—345.
[11] WALTERS, G. A. and TEMPELMAN, A. B., "Non optimal dynamic programming algorithm in the design of minimum cost drainage systems", *Engineering Optimization* **4** (1979) 139—148.
[12] „Tájékoztató a közüzemi (települési) szennyvíztisztító telepek és a telepszerű lakásépítéshez tartozó víz-, csatornahálózatok beruházási költségeiről", OVH, Budapest, 1981. december.
[13] „Balaton térség szennyvízelvezetése", tanulmány, VIZITERV, 1975.

(Beérkezett: 1983. május 16.)

KOVÁCS LÁSZLÓ BÉLA ÉS BOROS ENDRE
MTA SZÁMÍTÁSTECHNIKAI ÉS AUTOMATIZÁLÁSI KUTATÓ INTÉZET
1111 BUDAPEST, KENDE U. 13—17.

INOTAY FERENC
VÍZGÉP
2040 BUDAÖRS, KOMÁROMI U. 22.

# A TWO-STAGE MATHEMATICAL MODEL AND INTERACTIVE PROGRAM SYSTEM FOR PLANNING NETWORKS OF SEWER SYSTEMS AND WASTE WATER TREATMENT PLANTS — WITH APPLICATION TO THE LAKE BALATON AREA

L. B. Kovács, E. Boros and F. Inotay

This paper describes a new, two-stage model for sewer system design. In the first stage a large number of so-called main sewer systems are constructed, each of which is consistent from the engineering viewpoint and none of which is uniformly inferior to any other. Parameters are provided to select more or less number of main sewer systems, which are substantially different in a sense well-defined in the paper. In the second stage a pure 0—1 integer programming problem is formulated and solved, the basic elements of which (represented by 0—1 variables) are the potential waste water treatment plants and main sewer systems. The corresponding interactive computer program system functions are outlined.

As an example, several different plans are compared for one of the regions of the Lake Balaton resort area. Further research directions, applications, and relations to other sewer system planning models are shortly discussed.

*Alkalmazott Matematikai Lapok* **10** (1984) 103—113

# NÉHÁNY ADALÉK A KVÁZIKONVEX FÜGGVÉNYEK ELMÉLETÉHEZ

KOMLÓSI SÁNDOR

Pécs

A dolgozatban a pontban való lokális kvázikonvexitás, lokális pszeudokonvexitás és lokális szigorú pszeudokonvexitás fogalmak kerülnek bevezetésre, kissé módosítva ezen terminusok korábban használatos jelentését. Vizsgáljuk a lokális általánosított konvexitási tulajdonságok kapcsolatait a megfelelő, halmazon értelmezett általánosított konvexitási tulajdonságokkal. Bevezetve a *kvázi-Hesse mátrix* fogalmát, segítségével jellemezzük kétszer differenciálható függvények esetében a lokális pszeudokonvexitás, lokális szigorú pszeudokonvexitás függvénytulajdonságokat. A *kvázi-Hesse mátrix* segítségével elegendő feltételt adunk feltételes lokális szélsőérték létezésére vonatkozóan.

## 1. Bevezetés

A matematikai programozás elméletében fontos szerepet játszanak a konvexitás különféle általánosításai (kvázikonvexitás, pszeudokonvexitás, stb.). Ezek segítségével a konvex programokra érvényes optimalitási feltételeket ki lehet terjeszteni matematikai programozási feladatoknak a konvex programok osztályánál lényegesen bővebb osztályaira. Éppen ezért fontos, hogy rendelkezzünk olyan módszerekkel, amelyek segítségével eldönthető, hogy adott függvény adott halmazon rendelkezik-e az általánosított konvexitás valamelyik válfajával.

Kétszer differenciálható függvények esetében több ilyen módszer is létezik. Az egyik módszer a *szegélyezett Hesse-mátrix* minorjainak vizsgálatán alapszik [1, 3, 4, 10]. Egy másik módszer az $\mathbf{u}^*\nabla^2 f(\mathbf{x})\mathbf{u}$ kvadratikus alaknak a $\nabla f(\mathbf{x})\mathbf{u}=0$ altérre való restrikciójának vizsgálatára épül [2, 6, 7, 9, 11, 18]. (A $\nabla^2 f(\mathbf{x})$ szimbólum az $f(\mathbf{x})$ függvény *Hesse-mátrixát*, a $\nabla f(\mathbf{x})$ szimbólum az $f(\mathbf{x})$ függvény gradiensét jelöli. Kényelmi okokból a $\nabla f(\mathbf{x})$ vektort sorvektornak tekintjük. A „*" szimbólum a transzponálás jele.)

Egy további módszer a

$$\mathbf{H}(\mathbf{x}; r(\mathbf{x})) = \nabla^2 f(\mathbf{x}) + r(\mathbf{x})\nabla f(\mathbf{x})^*\nabla f(\mathbf{x})$$

*kiterjesztett Hesse-mátrix* vizsgálatán alapszik [2, 3, 5, 11, 15, 21]. Az említett módszerek összehasonlítása megtalálható a [7, 9] cikkekben.

Dolgozatunkban egy új módszert adunk a pszeudokonvexitás (pszeudokonkavitás) és a szigorú pszeudokonvexitás (szigorú pszeudokonkavitás) vizsgálatára kétszer differenciálható függvények esetére. Módszerünk az $f(\mathbf{x})=f(\mathbf{x}_0)$ egyenletű nívófelületek lokális tulajdonságainak vizsgálatán alapszik. (Hasonló vizsgálódások találhatók a [11, 17, 18] cikkekben.) Ennek kapcsán különböző általánosított konvexitási tulajdonságok (a klasszikus analízis szellemében fogant) lokális értelmezésére van szükségünk.

A 2. részben a lokális és globális (halmazra vonatkozó) általánosított konvexitási tulajdonságok kapcsolatát vizsgáljuk.

A 3. részben szükséges és elegendő feltételt adunk arra, hogy egy $f(\mathbf{x})$ függvény egy $\mathbf{x}_0$ pontban lokálisan pszeudokonvex, illetve lokálisan szigorúan pszeudokonvex legyen. Bevezetve a *kvázi-Hesse mátrix* fogalmát, segítségével jellemezzük a pszeudokonvex, illetve szigorúan pszeudokonvex függvényeket. Ez a karakterizálás szoros analógiát mutat a konvexitás és szigorú konvexitás függvénytulajdonságok *Hesse-mátrixszal* való jellemzésével. Ez az oka, hogy a 3. részben bevezetett $\mathbf{Q}_i(\mathbf{x}_0)$ mátrix megnevezésére a *kvázi-Hesse mátrix* elnevezést javasoljuk.

A 4. részben megmutatjuk, hogy a lokális szigorú pszeudokonvexitás másodrendű feltétele bizonyos körülmények között feltételes lokális szélsőérték létezésének elgendő feltételéül szolgál. Ez az eredmény RAPCSÁK TAMÁSTÓL származik, mi csupán témánk tárgyalásába beleillő új bizonyítást adunk rá [17].

## 2. Lokális és globális általánosított konvexitás

Jelöljön (a dolgozat egészében) az $f(\mathbf{x})$ egy olyan valós értékű függvényt, amely értelmezett és *(Fréchet-féle értelemben)* differenciálható a $D \subset R^n$ nyílt halmazon.

Dolgozatunkban figyelmünket a kvázikonvex, pszeudokonvex és szigorúan pszeudokonvex függvények körére korlátozzuk. Ezen függvényosztályokra vonatkozó vizsgálódásoknak tekintélyes irodalma van. A rájuk vonatkozó alapvető eredmények megtalálhatók pl. a [3, 8, 9, 13, 14, 16] művekben.

Emlékeztetünk arra, hogy az $f(\mathbf{x})$ függvényt a $C \subset D$ konvex halmazon kvázikonvexnek, pszeudokonvexnek, illetve szigorúan pszeudokonvexnek nevezzük, ha a következő feltételek közül az adott felsorolásnak megfelelő feltétel teljesül:

$$\text{ha} \quad \mathbf{x}_1, \mathbf{x}_2 \in C,\ t \in [0,1] \quad \text{és} \quad f(\mathbf{x}_1) \leqq f(\mathbf{x}_2), \tag{2.1}$$
$$\text{akkor} \quad f(t\mathbf{x}_1 + (1-t)\mathbf{x}_2) \leqq f(\mathbf{x}_2);$$

$$\text{ha} \quad \mathbf{x}_1, \mathbf{x}_2 \in C \quad \text{és} \quad f(\mathbf{x}_1) < f(\mathbf{x}_2), \tag{2.2}$$
$$\text{akkor} \quad \nabla f(\mathbf{x}_2)(\mathbf{x}_1 - \mathbf{x}_2) < 0;$$

$$\text{ha} \quad \mathbf{x}_1, \mathbf{x}_2 \in C,\ \mathbf{x}_1 \neq \mathbf{x}_2 \quad \text{és} \quad f(\mathbf{x}_1) \leqq f(\mathbf{x}_2), \tag{2.3}$$
$$\text{akkor} \quad \nabla f(\mathbf{x}_2)(\mathbf{x}_1 - \mathbf{x}_2) < 0.$$

Jól ismert, hogy differenciálható $f(\mathbf{x})$ függvény esetében a (2.1) feltétel ekvivalens a következővel:

$$\text{ha} \quad \mathbf{x}_1, \mathbf{x}_2 \in C \quad \text{és} \quad f(\mathbf{x}_1) \leqq f(\mathbf{x}_2), \tag{2.4}$$
$$\text{akkor} \quad \nabla f(\mathbf{x}_2)(\mathbf{x}_1 - \mathbf{x}_2) \leqq 0.$$

Differenciálható függvény egy pontban való lokális kvázikonvexitását, illetve lokális pszeudokonvexitását először MARTOS BÉLA definiálta [14, Chapter 7.] a következő módon. Legyen $C \subset D$ nyílt halmaz, és $\mathbf{x}_0 \in C$. Az $f(\mathbf{x})$ függvényt az $\mathbf{x}_0$ pontban (a $C$ halmazra vonatkozóan) lokálisan kvázikonvexnek, illetve lokálisan pszeu-

dokonvexnek nevezi, ha a következő feltételek közül az adott sorrendnek megfelelő feltétel teljesül:

$$\text{ha} \quad \mathbf{x}\in C \quad \text{és} \quad f(\mathbf{x}) \leqq f(\mathbf{x}_0), \quad \text{akkor} \quad \nabla f(\mathbf{x}_0)(\mathbf{x}-\mathbf{x}_0) \leqq 0;$$

$$\text{ha} \quad \mathbf{x}\in C \quad \text{és} \quad f(\mathbf{x}) \leqq f(\mathbf{x}_0), \quad \text{akkor} \quad \nabla f(\mathbf{x}_0)(\mathbf{x}-\mathbf{x}_0) \leqq 0,$$

$$f(\mathbf{x}) < f(\mathbf{x}_0) \quad \text{esetén pedig} \quad \nabla f(\mathbf{x}_0)(\mathbf{x}-\mathbf{x}_0) < 0.$$

A MARTOS BÉLA által bevezetett fenti fogalmak (a lokális jelző ellenére) az adott függvénynek nem csupán lokális (az adott pont „közelében" való) viselkedését fejezik ki, ezért, ragaszkodva a lokális jelző klasszikus analízisben használatos jelentéséhez, a lokális általánosított konvexitási tulajdonságok következő értelmezését javasoljuk.

2.1. DEFINÍCIÓ. A differenciálható $f(\mathbf{x})$ függvényt az $\mathbf{x}_0\in C$ pontban lokálisan kvázikonvexnek, lokálisan pszeudokonvexnek, illetve lokálisan szigorúan pszeudokonvexnek nevezzük (a $C$ halmazra vonatkozóan), ha az $\mathbf{x}_0$ pontnak van olyan $G$ környezete, hogy a következő feltételek közül a felsorolás sorrendjének megfelelő feltétel teljesül:

(2.5)
$$\text{ha} \quad \mathbf{x}\in C\cap G \quad \text{és} \quad f(\mathbf{x}) \leqq f(\mathbf{x}_0),$$
$$\text{akkor} \quad \nabla f(\mathbf{x}_0)(\mathbf{x}-\mathbf{x}_0) \leqq 0;$$

(2.6)
$$\text{ha} \quad \mathbf{x}\in C\cap G \quad \text{és} \quad f(\mathbf{x}) \leqq f(\mathbf{x}_0),$$
$$\text{akkor} \quad \nabla f(\mathbf{x}_0)(\mathbf{x}-\mathbf{x}_0) \leqq 0,$$
$$f(\mathbf{x}) < f(\mathbf{x}_0) \quad \text{esetén pedig} \quad \nabla f(\mathbf{x}_0)(\mathbf{x}-\mathbf{x}_0) < 0;$$

(2.7)
$$\text{ha} \quad \mathbf{x}\in C\cap G,\ \mathbf{x}\neq\mathbf{x}_0 \quad \text{és} \quad f(\mathbf{x}) \leqq f(\mathbf{x}_0),$$
$$\text{akkor} \quad \nabla f(\mathbf{x}_0)(\mathbf{x}-\mathbf{x}_0) < 0.$$

Megjegyezzük, hogy a pontban való kvázikonvexitást, mint az adott pont „közelében" való függvényviselkedést KÉRI GERZSON is értelmezte és vizsgálta [12, 4. Definition], az általa megfogalmazott követelmény azonban eltér a (2.5) feltételtől.

*2.2. Megjegyzés.* Fenti definíciókból közvetlenül következik, hogy nyílt $C\subset D$ halmaz esetén $f(\mathbf{x})$ akkor és csak akkor lokálisan (szigorúan) pszeudokonvex az $\mathbf{x}_0\in C$ stacionárius pontban, amelyre tehát $\nabla f(\mathbf{x}_0)=0$, ha $\mathbf{x}_0$-ban $f(\mathbf{x})$-nek (szigorú) lokális minimuma van.

A lokális kvázikonvexitás és lokális pszeudokonvexitás általunk adott értelmezése mellett is érvényes RAPCSÁK TAMÁS következő tétele.

2.3. LEMMA. [18] Legyen $C\subset D$ nyílt halmaz. Ha $f(\mathbf{x})$ az $\mathbf{x}_0\in C$ pontban lokálisan kvázikonvex és $\nabla f(\mathbf{x}_0)\neq 0$, akkor $f(\mathbf{x})$ $\mathbf{x}_0$-ban lokálisan pszeudokonvex.

A továbbiakban a lokális és globális általánosított konvexitási tulajdonságok kapcsolatát vizsgáljuk.

2.4. TÉTEL. Ha az $f(\mathbf{x})$ függvény lokálisan (szigorúan) pszeudokonvex a nyílt konvex $C$ halmaz minden pontjában, akkor $f(\mathbf{x})$ $C$-n (szigorúan) pszeudokonvex.

*Bizonyítás.* Először a pszeudokonvexitásra vonatkozó állítást bizonyítjuk. Legyenek $\mathbf{x}_1, \mathbf{x}_2 \in C$ tetszőleges vektorok. Tegyük fel, hogy $f(\mathbf{x}_1) < f(\mathbf{x}_2)$. Azt kell megmutatnunk, hogy ekkor $\nabla f(\mathbf{x}_2)(\mathbf{x}_1 - \mathbf{x}_2) < 0$. Vezessük be a következő függvényeket:

$$\mathbf{x}(t) = (1-t)\cdot\mathbf{x}_2 + t\cdot\mathbf{x}_1, \quad t\in[0,1] \quad \text{és}$$
$$y(t) = f(\mathbf{x}(t)), \quad t\in[0,1].$$

Megmutatjuk, hogy az $y(t)$ függvény a [0, 1]-on egyedül a 0-ban veszi fel maximumát. Indirekt módon okoskodva tegyük fel, hogy a $t_0$ pozitív valós szám $(0 < t_0 < 1)$ a legutolsó maximumhelye $y(t)$-nek [0, 1]-en. Legyen $\mathbf{x}_0 = \mathbf{x}(t_0)$. Mivel

$$y'(t_0) = \nabla f(\mathbf{x}_0)(\mathbf{x}_1 - \mathbf{x}_2) = 0,$$

ezért minden $t\in[0,1]$-re

$$\nabla f(\mathbf{x}_0)(\mathbf{x}(t) - \mathbf{x}_0) = 0. \tag{2.8}$$

Feltevés szerint $f(\mathbf{x})$ lokálisan pszeudokonvex $\mathbf{x}_0$-ban, vagyis létezik $\mathbf{x}_0$-nak olyan $G$ környezete, amelyre teljesül a (2.6) feltétel. Ehhez a $G$-hez található olyan $\tau$ pozitív szám, hogy $t_0 < \tau < 1$ és $\mathbf{x}(\tau)\in G\cap C$. Nyilvánvaló, hogy $f(\mathbf{x}(\tau)) < f(\mathbf{x}_0)$, következésképpen $\nabla f(\mathbf{x}_0)(\mathbf{x}(\tau) - \mathbf{x}_0) < 0$. Ez utóbbi egyenlőtlenség viszont ellentmondásban van (2.8)-cal. Beláttuk tehát, hogy $y(t) < y(0)$ minden $t\in(0,1]$-re.

Feltevés szerint $f(\mathbf{x})$ lokálisan pszeudokonvex $\mathbf{x}_2$-ben, ennélfogva létezik olyan $\sigma\in(0,1)$, hogy $f(\mathbf{x}(\sigma)) < f(\mathbf{x}_2)$, és $\nabla f(\mathbf{x}_2)(\mathbf{x}(\sigma) - \mathbf{x}_2) = \sigma\cdot\nabla f(\mathbf{x}_2)(\mathbf{x}_1 - \mathbf{x}_2) < 0$. Ez utóbbi egyenlőtlenségből közvetlenül következik a bizonyítandó $\nabla f(\mathbf{x}_2)(\mathbf{x}_1 - \mathbf{x}_2) < 0$ egyenlőtlenség.

A szigorú pszeudokonvexitásra vonatkozó állítás bizonyítása teljesen hasonlóan történik.

Az $f(x) = -x^2$ egyváltozós függvény példája mutatja, hogy a lokális és a globális kvázikonvexitás között nem érvényes a 2.4. tétel analogonja. Ez a függvény ugyanis minden pontban lokálisan kvázikonvex, de egyetlen olyan konvex halmazon sem kvázikonvex, amely az $x_0 = 0$ pontot tartalmazza. Ha azonban bizonyos megszorításokat teszünk az $f(\mathbf{x})$ függvényre vonatkozóan, akkor már a 2.4. tétellel analóg tétel bizonyítható a kvázikonvex esetre. Ebből a célból felhasználjuk a [9] cikkben bevezetett következő fogalmat.

2.5. DEFINÍCIÓ. Legyen az egyváltozós $y(t)$ függvény értelmezett a nyílt $(a, b)$ intervallumon. A $t_0\in(a,b)$ pontot az $y(t)$ függvény *féligszigorú lokális maximumhelyének* nevezzük, ha $y(t)$-nek $t_0$-ban lokális maximuma van és léteznek olyan $t_1, t_2$ valós számok, hogy $a < t_1 < t_0 < t_2 < b$ és $y(t_1) < y(t_0)$ és $y(t_2) < y(t_0)$.

2.6. TÉTEL. Ha az $f(\mathbf{x})$ függvény lokálisan kvázikonvex a nyílt konvex $C$ halmaz minden pontjában és ha teljesül a következő feltétel:

(2.9) ha $\mathbf{z}\in C$ olyan, hogy $\nabla f(\mathbf{z}) = 0$, akkor bármely $\mathbf{x}\in C$-re az $y(t) = f(\mathbf{z} + t(\mathbf{x} - \mathbf{z}))$ függvénynek a $t = 0$ pont nem félig szigorú lokális maximumhelye,

akkor $f(\mathbf{x})$ globálisan kvázikonvex $C$-n.

*Bizonyítás.* Legyenek $\mathbf{x}_1, \mathbf{x}_2\in C$ tetszőlegesek. Tegyük fel, hogy $f(\mathbf{x}_1) \leqq f(\mathbf{x}_2)$. Tekintsük a 2.4. tétel bizonyításában definiált $\mathbf{x}(t)$, $y(t)$ függvényeket. Azt kell megmutatnunk, hogy $y(t) \leqq y(0)$ minden $t\in[0,1]$-re.

Tegyük fel ezzel ellentétben, hogy létezik olyan $t_0$ pozitív szám $(0 < t_0 < 1)$, amely olyan maximumhelye $y(t)$-nek, amelyre $y(t_0) > y(0)$. Feltehetjük, hogy $t_0$ az utolsó

maximumhely [0, 1]-en. Legyen $\mathbf{x}_0=\mathbf{x}(t_0)$. Ha $\nabla f(\mathbf{x}_0)\neq\mathbf{0}$, akkor a 2.3. lemma szerint $f(\mathbf{x})$ lokálisan pszeudokonvex $\mathbf{x}_0$-ban, másrészt viszont $\nabla f(\mathbf{x}_0)(\mathbf{x}(t)-\mathbf{x}_0)=0$ minden $t\in[0, 1]$-re. Ezek a tények ellentmondanak egymásnak. Ha $\nabla f(\mathbf{x}_0)=\mathbf{0}$, akkor a $t_0$-ra vonatkozó feltétel ellentmondásban van (2.9)-cel.

A (2.9) feltétel W. E. DIEWERT, M. AVRIEL és I. ZANG szerzőktől ered [9].

A továbbiakban a fenti eredmények néhány következményét fogalmazzuk meg. A 2.4. tétel és 2.3. lemma alapján nyilvánvaló a

2.7. KÖVETKEZMÉNY. Ha az $f(\mathbf{x})$ függvény lokálisan kvázikonvex a $C$ nyílt konvex halmaz minden pontjában és $\nabla f(\mathbf{x})\neq\mathbf{0}$ $C$-n, akkor $f(\mathbf{x})$ globálisan pszeudokonvex $C$-n.

A 2.4. tételből és a 2.2. megjegyzésből adódik a

2.8. KÖVETKEZMÉNY. Legyen $f(\mathbf{x})$ lokálisan kvázikonvex a $C$ nyílt konvex halmaz minden pontjában. $f(\mathbf{x})$ akkor és csak akkor pszeudokonvex $C$-n, ha minden olyan $\mathbf{z}\in C$, amelyre $\nabla f(\mathbf{z})=0$ teljesül, lokális minimumpontja $f(\mathbf{x})$-nek.

E legutóbbi következmény tartalmazza J. P. CROUZEIX és J. A. FERLAND egyik eredményét [7, Theorem 7.], mely szerint ha $f(\mathbf{x})$ kvázikonvex a $C$ nyílt konvex halmazon, akkor ahhoz, hogy $f(\mathbf{x})$ pszeudokonvex legyen $C$-n szükséges és elegendő, hogy minden olyan $\mathbf{z}\in C$, amelyre $\nabla f(\mathbf{z})=0$, lokális minimumpontja legyen $f(\mathbf{x})$-nek.

## 3. A lokális, illetve globális kvázikonvexitás szükséges és elegendő feltételei

További vizsgálódásainkban alapvető szerepet kap az $f(\mathbf{x})=f(\mathbf{x}_0)$ egyenletű nívófelületek lokális tulajdonságainak vizsgálata. Vizsgálódásunk alapvető segédeszköze az implicit-függvény tétel ([22], VI. fejezet, 3.). A továbbiakban, hogy alkalmazni tudjuk ezt a tételt, az $f(\mathbf{x})$ függvényről hallgatólagosan mindig feltesszük, hogy folytonosan differenciálható $D$-n.

Legyen $\mathbf{x}_0\in D$, amelyre $\nabla f(\mathbf{x}_0)\neq 0$. Jelöljön $i$ egy olyan indexet, amelyre $f'_{x_i}(\mathbf{x}_0)\neq\neq 0$. Vezessük be a következő jelöléseket:

$$\mathbf{u}=(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)^*, \quad v=x_i.$$

A továbbiakban az $\mathbf{x}=(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n)^*$ vektor és az $(\mathbf{u}, v)$ pár között nem teszünk különbséget.

Jelölje $\mathbf{I}$ az $n$-edrendű egységmátrixot. Hagyjuk el $\mathbf{I}$-ből az $i$-edik oszlopot és jelölje $\mathbf{P}_i$ az íly módon létrejött $n\times(n-1)$-es mátrixot. Ekkor $\mathbf{u}=\mathbf{P}_i^*\mathbf{x}$. Legyen $\nabla_{\mathbf{u}} f(\mathbf{x})=\nabla f(\mathbf{x})\mathbf{P}_i$.

Tekintsük az $f(\mathbf{u}, v)=f(\mathbf{x}_0)$ egyenletet, ahol $\mathbf{x}_0=(\mathbf{u}_0, v_0)$. Az implicit-függvény tétel szerint igazak a következő állítások:

$\mathbf{x}_0$-nak létezik olyan $G$ környezete, $\mathbf{u}_0$-nak létezik olyan $N$ környezete és létezik egyetlen olyan $N$-en definiált és ott folytonosan differenciálható $h_i(\mathbf{u})$ függvény, hogy

$(\mathbf{u}, v)\in G$ és $f(\mathbf{u}, v)=f(\mathbf{x}_0)$ akkor és csak akkor teljesül, ha $\mathbf{u}\in N$ és $v=h_i(\mathbf{u})$;

(3.1)

minden $\mathbf{u}\in N$-re

$$\nabla h_i(\mathbf{u}) = -\frac{\nabla_{\mathbf{u}} f(\mathbf{u}, h_i(\mathbf{u}))}{f'_v(\mathbf{u}, h_i(\mathbf{u}))}; \tag{3.2}$$

speciálisan

$$\nabla h_i(\mathbf{u}_0) = -\frac{\nabla_{\mathbf{u}} f(\mathbf{x}_0)}{f'_v(\mathbf{x}_0)}. \tag{3.3}$$

Nevezzük a $h_i(\mathbf{u})$ függvényt az $f(\mathbf{x})$ függvény $\mathbf{x}_0$ *pontbeli* $x_i$ *változóhoz asszociált lokális szintfelület-függvényének*.

Vezessük be az $\mathbf{u}$ változó következő függvényét:

$$H_i(\mathbf{x}_0; \mathbf{u}) = -f'_{x_i}(\mathbf{x}_0)\, h_i(\mathbf{u}).$$

A (3.2) összefüggés felhasználásával egyszerűen igazolható a következő állítás:

3.1. LEMMA. Minden $\mathbf{u}\in N$-re

$$\nabla H_i(\mathbf{x}_0; \mathbf{u}_0)(\mathbf{u}-\mathbf{u}_0) + H_i(\mathbf{x}_0; \mathbf{u}_0) - H_i(\mathbf{x}_0; \mathbf{u}) = \nabla f(\mathbf{x}_0)(\mathbf{x}-\mathbf{x}_0),$$

ahol $\mathbf{x}=(\mathbf{u}, h_i(\mathbf{u}))$.

A következő tétel képezi további vizsgálódásaink alapját.

3.2. TÉTEL. Legyen $f(\mathbf{x})$ $D$-n folytonosan differenciálható. Legyen $f'_{x_i}(\mathbf{x}_0)\neq 0$, ahol $\mathbf{x}_0=(\mathbf{u}_0, v_0)\in D$. Ahhoz, hogy $f(\mathbf{x})$ $\mathbf{x}_0$-ban lokálisan (szigorúan) pszeudokonvex legyen, szükséges és elegendő, hogy a $H_i(\mathbf{x}_0; \mathbf{u})$ függvény $\mathbf{u}_0$-ban lokálisan (szigorúan) konvex legyen.

*Bizonyítás.* A 2.3. lemma alapján a lokális pszeudokonvexitás vizsgálata helyett elegendő csupán a lokális kvázikonvexitást vizsgálni.

*Szükségesség:* Legyen $f(\mathbf{x})$ lokálisan kvázikonvex az $\mathbf{x}_0$ pontban. Az implicitfüggvény tétel szerint van $\mathbf{x}_0$-nak olyan $G$ és $\mathbf{u}_0$-nak olyan $N$ környezete, melyekre teljesül a (3.1) állítás. Az általánosság megszorítása nélkül feltehetjük, hogy $G$-n teljesül a (2.6) feltétel. Mivel minden $\mathbf{u}\in N$ esetén $f(\mathbf{u}, h_i(\mathbf{u}))=f(\mathbf{x}_0)$ és $\mathbf{x}=(\mathbf{u}, h_i(\mathbf{u}))\in G$, ezért a 3.1. lemma és a (2.6) feltétel alapján minden $\mathbf{u}\in N$-re igaz a következő egyenlőtlenség:

$$\nabla H_i(\mathbf{x}_0; \mathbf{u}_0)(\mathbf{u}-\mathbf{u}_0) + H_i(\mathbf{x}_0; \mathbf{u}_0) - H_i(\mathbf{x}_0; \mathbf{u}) = \nabla f(\mathbf{x}_0)(\mathbf{x}-\mathbf{x}_0) \leqq 0.$$

*Elegendőség.* Az általánosság megszorítása nélkül feltehetjük, hogy minden $\mathbf{u}\in N$-re

$$\nabla H_i(\mathbf{x}_0; \mathbf{u}_0)(\mathbf{u}-\mathbf{u}_0) + H_i(\mathbf{x}_0; \mathbf{u}_0) - H_i(\mathbf{x}_0; \mathbf{u}) \leqq 0.$$

Most megmutatjuk, hogy a (2.6) feltétel $f(\mathbf{x})$-re vonatkozóan teljesül a $G$ környezeten. Okoskodjunk indirekt módon: tegyük fel, hogy van olyan $\mathbf{x}_1\in G$, amelyre $f(\mathbf{x}_1)\leqq \leqq f(\mathbf{x}_0)$ és mégis $\nabla f(\mathbf{x}_0)(\mathbf{x}_1-\mathbf{x}_0)>0$.

Folytonossági okokból feltehetjük, hogy $f(\mathbf{x}_1)=f(\mathbf{x}_0)$. Legyen $\mathbf{x}_1=(\mathbf{u}_1, v_1)$. Nyilvánvaló, hogy $\mathbf{u}_1\in N$ és $v_1=h_i(\mathbf{u}_1)$. Tekintettel a 3.1. lemmára, igaz a következő egyenlőtlenség:

$$\nabla H_i(\mathbf{x}_0; \mathbf{u}_0)(\mathbf{u}_1-\mathbf{u}_0) + H_i(\mathbf{x}_0; \mathbf{u}_0) - H_i(\mathbf{x}_0; \mathbf{u}_1) = \nabla f(\mathbf{x}_0)(\mathbf{x}_1-\mathbf{x}_0) > 0.$$

Ez viszont ellentmond kiindulási feltevésünknek.

A lokális szigorú pszeudokonvexitásra vonatkozó állítás teljesen hasonlóan bizonyítható.

Az imént bizonyított tétel állításához hasonló állítások találhatók az [1, 11, 18] cikkekben.

Mostantól kezdve feltesszük, hogy $f(\mathbf{x})$ kétszer differenciálható $D$-n.

Legyen $\mathbf{x}_0 \in D$ és jelölje $\mathbf{e}_i$ az $i$-edik egységvektort. Tegyük fel, hogy $\nabla f(\mathbf{x}_0)\mathbf{e}_i \neq 0$. Vezessük be a következő mátrixokat:

$$\mathbf{K}_1(\mathbf{x}_0) = \nabla^2 f(\mathbf{x}_0);$$

$$\mathbf{K}_2^i(\mathbf{x}_0) = \frac{1}{\nabla f(\mathbf{x}_0)\mathbf{e}_i}\left(\nabla f(\mathbf{x}_0)^* \mathbf{e}_i^* \nabla^2 f(\mathbf{x}_0) + \nabla^2 f(\mathbf{x}_0)\mathbf{e}_i \nabla f(\mathbf{x}_0)\right);$$

$$\mathbf{K}_3^i(\mathbf{x}_0) = \frac{\mathbf{e}_i^* \nabla^2 f(\mathbf{x}_0)\mathbf{e}_i}{(\nabla f(\mathbf{x}_0)\mathbf{e}_i)^2} \nabla f(\mathbf{x}_0)^* \nabla f(\mathbf{x}_0).$$

3.3. DEFINÍCIÓ. Nevezzük a

$$\mathbf{Q}_i(\mathbf{x}_0) = \mathbf{P}_i^*\left(\mathbf{K}_1(\mathbf{x}_0) - \mathbf{K}_2^i(\mathbf{x}_0) + \mathbf{K}_3^i(\mathbf{x}_0)\right)\mathbf{P}_i$$

mátrixot az $f(\mathbf{x})$ függvény $\mathbf{x}_0$ pontbeli, az $x_i$ változóhoz asszociált *kvázi-Hesse mátrixának*. (A $\mathbf{P}_i$ mátrix jelen paragrafus elején van definiálva.) $\mathbf{Q}_i(\mathbf{x}_0)$ szimmetrikus, $(n-1)$-ed rendű mátrix. Kiszámítva a $H_i(\mathbf{x}_0\,;\,\mathbf{u})$ függvény $\mathbf{u}_0$ pontbeli *Hesse-féle mátrixát*, a következő eredményre jutunk.

3.4. LEMMA. $\mathbf{Q}_i(\mathbf{x}_0) = \nabla^2 H_i(\mathbf{x}_0\,;\,\mathbf{u}_0)$, ahol $\mathbf{x}_0 = (\mathbf{u}_0, v_0)$.

A továbbiakban szükségünk lesz az $\mathbf{u}_0$ pont $N$ környezetén értelmezett $H_i(\mathbf{x}_0\,;\,\mathbf{u})$ függvény és a $\mathbf{Q}_i(\mathbf{x})$ *kvázi-Hesse mátrix* közötti következő kapcsolat ismeretére.

3.5. LEMMA. Legyen $\mathbf{x}(\mathbf{u}) = (\mathbf{u}, h_i(\mathbf{u}))$. Ekkor minden $\mathbf{u} \in N$-re

$$\nabla^2 H_i(\mathbf{x}_0\,;\,\mathbf{u}) = \frac{f'_{x_i}(\mathbf{x}_0)}{f'_{x_i}(\mathbf{x}(\mathbf{u}))}\,\mathbf{Q}_i(\mathbf{x}(\mathbf{u})).$$

*Bizonyítás.* Legyen $\mathbf{u} \in N$ tetszőleges. Mivel $f'_{x_i}(\mathbf{x}(\mathbf{u})) \neq 0$, ezért az implicit-függvény tétel szerint van $\mathbf{x}(\mathbf{u})$-nak olyan $G_1$ és $\mathbf{u}$-nak olyan $N_1$ környezete és létezik egyetlen olyan $N_1$-en definiált $k_i(\mathbf{s})$ függvény, hogy $(\mathbf{s}, w) \in G_1$ és $f(\mathbf{s}, w) = f(\mathbf{x}(\mathbf{u})) = f(\mathbf{x}_0)$ akkor és csak akkor teljesül, ha $\mathbf{s} \in N_1$ és $w = k_i(\mathbf{s})$. Ebből nyilvánvaló, hogy a $h_i(\mathbf{s})$ és a $k_i(\mathbf{s})$ függvények egybeesnek az $N \cap N_1$ nyílt halmazon. Mivel $\mathbf{u} \in N \cap N_1$, ezért $\nabla^2 h_i(\mathbf{u}) = \nabla^2 k_i(\mathbf{u})$.

Másrészt viszont

$$\nabla^2 H_i(\mathbf{x}_0\,;\,\mathbf{u}) = -f'_{x_i}(\mathbf{x}_0)\nabla^2 h_i(\mathbf{u}) = -f'_{x_i}(\mathbf{x}_0)\nabla^2 k_i(\mathbf{u}).$$

A 3.4. lemma szerint pedig

$$\mathbf{Q}_i(\mathbf{x}(\mathbf{u})) = -f'_{x_i}(\mathbf{x}(\mathbf{u}))\nabla^2 k_i(\mathbf{u}).$$

A bizonyítandó lemma állítása közvetlenül következik e két egyenletből.

A következőkben a lokális (szigorú) pszeudokonvexitás, illetve a (szigorú) pszeudokonvexitás függvénytulajdonságokat jellemezzük a *kvázi-Hesse mátrix* segítségével.

3.6. TÉTEL. Legyen $f(\mathbf{x})$ kétszer differenciálható $D$-n és $f'_{x_i}(\mathbf{x}_0)\neq 0$, ahol $\mathbf{x}_0\in D$. Ha $f(\mathbf{x})$ lokálisan kvázikonvex $\mathbf{x}_0$-ban, akkor a $\mathbf{Q}_i(\mathbf{x}_0)$ *kvázi-Hesse mátrix* pozitív szemidefinit.

A tétel állítása a 3.2. tétel és a 3.4. lemma alapján nyilvánvaló.

3.7. TÉTEL. Legyen $f(\mathbf{x})$ kétszer differenciálható a $C$ nyílt konvex halmazon. Tegyük fel, hogy $f'_{x_i}(\mathbf{x})\neq 0$ $C$-n. Az $f(\mathbf{x})$ függvény akkor és csak akkor pszeudokonvex $C$-n, ha az $f(\mathbf{x})$ függvény $x_i$ változóhoz asszociált $\mathbf{Q}_i(\mathbf{x})$ *kvázi-Hesse mátrixa* minden $\mathbf{x}\in C$-re pozitív szemidefinit. Ha $\mathbf{Q}_i(\mathbf{x})$ minden $\mathbf{x}\in C$-re pozitív definit, akkor $f(\mathbf{x})$ szigorúan pszeudokonvex $C$-n.

*Bizonyítás.* A 3.5. lemma számbavételével adódik, hogy minden $\mathbf{x}\in C$-re, az $\mathbf{x}$ ponthoz tartozó $H_i(\mathbf{x};\mathbf{u})$ függvény az $\mathbf{u}$ változónak konvex, (illetve szigorúan konvex) függvénye. A 3.2. tétel alapján $f(\mathbf{x})$ lokálisan pszeudokonvex (illetve lokálisan szigorúan pszeudokonvex) minden $\mathbf{x}\in C$ pontban. Ebből a 2.4. tétel felhasználásával nyerjük, hogy $f(\mathbf{x})$ pszeudokonvex (illetve szigorúan pszeudokonvex) $C$-n.

Ha $f(\mathbf{x})$ pszeudokonvex $C$-n, akkor a 3.6. tétel szerint a $\mathbf{Q}_i(\mathbf{x})$ *kvázi-Hesse mátrix* minden $\mathbf{x}\in C$-re pozitív szemidefinit.

3.8. TÉTEL. Legyen $f(\mathbf{x})$ kétszer folytonosan differenciálható az $\mathbf{x}_0\in D$ pontban, ahol $f'_{x_i}(\mathbf{x}_0)\neq 0$. Ha $\mathbf{Q}_i(\mathbf{x}_0)$ pozitív definit, akkor $f(\mathbf{x})$ lokálisan szigorúan pszeudokonvex $\mathbf{x}_0$-ban.

*Bizonyítás.* Folytonossági okok miatt $\mathbf{x}_0$-nak van olyan $G$ konvex környezete, hogy minden $\mathbf{x}\in G$ esetén $f'_{x_i}(\mathbf{x})\neq 0$ és $\mathbf{Q}_i(\mathbf{x})$ pozitív definit. A 3.7. tétel szerint ekkor $f(\mathbf{x})$ szigorúan pszeudokonvex $G$-n, következésképpen lokálisan szigorúan pszeudokonvex $x_0$-ban.

Különböző változókhoz asszociált *kvázi-Hesse mátrixok* kapcsolatát vizsgálva gyengíteni tudjuk a 3.7. tétel feltételeit. Ebből a célból a $\mathbf{Q}_i(\mathbf{x})$ mátrix differenciálgeometriai jelentését vizsgáljuk.

Tekintsük az $f(\mathbf{x})=f(\mathbf{x}_0)$ nívófelületet az $\mathbf{x}_0$ pont közelében. Ennek egy $n-1$ dimenziós elemi felületként való megadását szolgáltatja a $h_i(\mathbf{u})$ függvény. (Feltételezzük, hogy $f'_{x_i}(\mathbf{x}_0)\neq 0$!) Jelölje $\mathbf{B}_i(\mathbf{x}_0)$ az $\{(\mathbf{u}, h_i(\mathbf{u}))\,|\,\mathbf{u}\in N\}$ elemi felület $\mathbf{x}_0=(\mathbf{u}_0, h_i(\mathbf{x}_0))$ pontjához tartozó második alapmennyiségeinek mátrixát. Ismeretes ([17], 115. old. (4.18)), hogy

$$\mathbf{B}_i(\mathbf{x}_0) = -\frac{1}{\|\nabla f(\mathbf{x}_0)\|}\,\mathbf{Q}_i(\mathbf{x}_0). \tag{3.4}$$

Tegyük fel, hogy $f'_{x_j}(\mathbf{x}_0)\neq 0$, ahol $i\neq j$. Jelölje $h_j(\mathbf{z})$, $\mathbf{z}\in L$ az $f(\mathbf{x})$ függvény $\mathbf{x}_0$ pontbeli, $x_j$ változóhoz asszociált, az implicit-függvény tétel által meghatározott lokális szintfelület-függvényét. Jelölje $\mathbf{B}_j(\mathbf{x}_0)$ a $\{(\mathbf{z}, h_j(\mathbf{z}))\,|\,\mathbf{z}\in L\}$ elemi felület $\mathbf{x}_0=(\mathbf{z}_0, h_j(\mathbf{z}_0))$ pontjához tartozó második alapmennyiségeinek mátrixát, amelyre ugyancsak érvényes a

$$\mathbf{B}_j(\mathbf{x}_0) = -\frac{1}{\|\nabla f(\mathbf{x}_0)\|}\,\mathbf{Q}_j(\mathbf{x}_0) \tag{3.5}$$

összefüggés. Ismeretes továbbá, hogy a $\mathbf{B}_i(\mathbf{x}_0)$ és $\mathbf{B}_j(\mathbf{x}_0)$ mátrixok hasonlóak, azaz van

olyan invertálható $A$ $(n-1)$-edrendű mátrix, hogy

$$\mathbf{B}_j(\mathbf{x}_0) = \mathbf{A}^{-1}\mathbf{B}_i(\mathbf{x}_0)\mathbf{A}. \tag{3.6}$$

A (3.4)—(3.6) összefüggésekből közvetlenül adódik a következő eredmény.

3.9. LEMMA. A $\mathbf{Q}_i(\mathbf{x}_0)$ és $\mathbf{Q}_j(\mathbf{x}_0)$ mátrixok hasonlóak.

Tekintettel a 3.9. lemmára, a 3.7. tétel bizonyításának csekély módosításával igazolhatók a következő állítások.

3.10. TÉTEL. Legyen $f(\mathbf{x})$ kétszer differenciálható a $C$ nyílt konvex halmazon. Az $f(\mathbf{x})$ függvény akkor és csak akkor pszeudokonvex $C$-n, ha teljesülnek a következő feltételek:

(i) ha $\mathbf{z}\in C$ és $\nabla f(\mathbf{z})=\mathbf{0}$, akkor $\mathbf{z}$-ben $f(x)$-nek lokális minimuma van,
(ii) minden $\mathbf{x}\in C$-hez, amelyre $\nabla f(\mathbf{x})\neq\mathbf{0}$, található olyan $i=i(\mathbf{x})$ index, hogy a $\mathbf{Q}_i(\mathbf{x})$ *kvázi-Hesse mátrix* pozitív szemidefinit.

3.11. TÉTEL. Legyen $f(\mathbf{x})$ kétszer differenciálható a $C$ nyílt konvex halmazon. Az $f(\mathbf{x})$ függvény szigorúan pszeudokonvex $C$-n, ha teljesülnek a következő feltételek:

(i) ha $\mathbf{z}\in C$ és $\nabla f(\mathbf{z})=\mathbf{0}$, akkor $\mathbf{z}$-ben $f(\mathbf{x})$-nek szigorú lokális minimuma van,
(ii) minden $\mathbf{x}\in C$-hez, amelyre $\nabla f(\mathbf{x})\neq\mathbf{0}$, található olyan $j=j(x)$ index, hogy a $\mathbf{Q}_j(\mathbf{x})$ *kvázi-Hesse mátrix* pozitív definit.

Végezetül szeretném megjegyezni, hogy analóg tételek fogalmazhatók meg kvázikonkáv, (szigorúan) pszeudokonkáv és pszeudomonoton függvények esetére is.

## 4. Feltételes lokális szélsőérték létezésének egy elegendő feltétele

Ennek a résznek a célja, hogy tárgyalásunkba beillő, új bizonyítást adjunk RAPCSÁK TAMÁS egy tételére ([17], 4.1. tétel következménye).

Tekintsük a továbbiakban a következő matematikai programozási feladatot:

$$\min f(\mathbf{x})$$

$$g_i(\mathbf{x}) \geqq 0, \quad i = 1, 2, \ldots, m, \tag{4.1}$$

$$\mathbf{x}\in D,$$

ahol az $f(\mathbf{x})$, $g_i(\mathbf{x})$ $i=1, 2, \ldots, m$ függvények értelmezettek és differenciálhatók a $D$ nyílt halmazon. Jelölje $D_0$ a (4.1) feladat megengedett megoldásainak halmazát.

4.1. DEFINÍCIÓ. ([14], Chapter 8., Definition 5.) Az $\mathbf{x}_0\in D_0$ pontot a (4.1) feladat KTL-stacionárius pontjának nevezzük, ha léteznek olyan $t_1, t_2, \ldots, t_m$ nemnegatív *Kuhn—Tucker—Lagrange multiplikátorok,* melyekre teljesülnek a következők:

$$\nabla f(\mathbf{x}_0) = \sum_{i=1}^{m} t_i \nabla g_i(\mathbf{x}_0), \tag{4.2}$$

$$t_i g_i(\mathbf{x}_0) = 0, \quad i = 1, 2, \ldots, m. \tag{4.3}$$

4.2. TÉTEL. Tekintsük a (4.1) feladatot. Tegyük fel, hogy az $\mathbf{x}_0 \in D_0$ pontban teljesülnek a következő feltételek:

(i) $\mathbf{x}_0$ KTL-stacionárius pontja a (4.1) feladatnak,
(ii) $\mathbf{x}_0$-nak van olyan $G$ környezete, hogy a $G \cap D_0$ halmaz csillagszerű az $\mathbf{x}_0$ pontban,
(iii) $f(\mathbf{x})$ $\mathbf{x}_0$-ban lokálisan (szigorúan) pszeudokonvex a $D_0$ halmazra vonatkozóan.

Ekkor $f(\mathbf{x})$-nek feltételes (szigorú) lokális minimuma van $\mathbf{x}_0$-ban $D_0$-ra vonatkozóan.

*Bizonyítás.* A lokális pszeudokonvexitás esetét vizsgálva, az általánosság megszorítása nélkül feltehetjük, hogy a (2.6) feltétel teljesül a $G \cap D_0$ halmazon. Okoskodjunk indirekt módon: tegyük fel, hogy van olyan $\mathbf{x}_1 \in G \cap D_0$, hogy $f(\mathbf{x}_1) < f(\mathbf{x}_0)$.

Ebből a (2.6) feltétel folytán a

$$\nabla f(\mathbf{x}_0)(\mathbf{x}_1 - \mathbf{x}_0) < 0 \tag{4.4}$$

egyenlőtlenség adódik. Tekintettel arra, hogy a $G \cap D_0$ halmaz csillagszerű $\mathbf{x}_0$-ban, következésképpen az $[\mathbf{x}_1, \mathbf{x}_0]$ szakasz $G \cap D_0$-ban fekszik. Bizonyítható, hogy minden $i = 1, 2, \ldots, m$-re

$$t_i \nabla g_i(\mathbf{x}_0)(\mathbf{x}_1 - \mathbf{x}_0) \geqq 0. \tag{4.5}$$

A (4.2) és (4.5) feltételekből azonban

$$\nabla f(\mathbf{x}_0)(\mathbf{x}_1 - \mathbf{x}_0) = \sum_{i=1}^{m} t_i \nabla g_i(\mathbf{x}_0)(\mathbf{x}_1 - \mathbf{x}_0) \geqq 0$$

következik, ami ellentmond (4.4)-nek. Eszerint tehát minden $\mathbf{x} \in G \cap D_0$-ra $f(\mathbf{x}) \geqq f(\mathbf{x}_0)$.

Az teljesen nyilvánvaló, hogy ha $G \cap D_0$-on a (2.6) feltétel helyett a nála erősebb (2.7) feltétel teljesül, akkor egyetlen $\mathbf{x} \in G \cap D_0$, $\mathbf{x} \neq \mathbf{x}_0$ vektorra sem lehet $f(\mathbf{x}) = f(\mathbf{x}_0)$.

Ebből a tételből és a 3.8. tételből közvetlenül kiadódik a következő, RAPCSÁK TAMÁS eredményével ([17], 3.4. tétel és 4.1. tétel következménye) lényegében megegyező állítás.

4.3. TÉTEL. Tekintsük a (4.1) feladatot. Tegyük fel, hogy $f(\mathbf{x})$ kétszer folytonosan differenciálható $D$-n és az $\mathbf{x}_0 \in D_0$ pontban teljesülnek a következő feltételek:

(i) $\mathbf{x}_0$ KTL-stacionárius pontja a (4.1) feladatnak,
(ii) $\mathbf{x}_0$-nak van olyan $G$ környezete, hogy a $G \cap D_0$ halmaz csillagszerű $\mathbf{x}_0$-ban,
(iii) van olyan $i$ index, hogy $f'_{x_i}(\mathbf{x}_0) \neq 0$ és a $\mathbf{Q}_i(\mathbf{x}_0)$ *kvázi-Hesse mátrix* pozitív definit.

Ekkor $f(\mathbf{x})$-nek feltételes szigorú lokális minimuma van $\mathbf{x}_0$-ban ($D_0$-ra vonatkozóan).

## IRODALOM

[1] ARROW, K. J. and ENTHOVEN, A. D., "Quasi-concave programming", *Econometrica* **29** (1961) 778—800.
[2] AVRIEL, M., "$r$-convex functions", *Mathematical Programming* **2** (1972) 309—323.
[3] AVRIEL, M., *Nonlinear Programming: Analysis and Methods*, (Prentice Hall, Englewood Cliffs, NJ,1976).

[4] AVRIEL, M. and SCHAIBLE, S., "Second order characterization of pseudoconvex functions", *Mathematical Programming* **14** (1978) 170—185.
[5] AVRIEL, M. and ZANG, I., "Generalized convex functions with applications to nonlinear programming", in: *Mathematical programs for activity analysis*, Ed. *P. van Moeseke* (North-Holland, Amsterdam, 1974) 23—33.
[6] CROUZEIX, J. P., "On second order conditions for quasiconvexity", *Mathematical Programming* **18** (1980) 349—352.
[7] CROUZEIX, J. P. and FERLAND, J. A., "Criteria for quasi-convexity and pseudo-convexity: relationships and comparisons", *Mathematical Programming* **23** (1982), 193—205.
[8] DIEWERT, W. E., "Alternative characterizations of six kinds of quasiconcavity in the nondifferentiable case with applications to nonsmooth programming", in: *Generalized Concavity in Optimization and Economics*, Eds.: *S. Schaible and W. T. Ziemba* (Academic Press, New York, 1981) 51—93.
[9] DIEWERT, W. E., AVRIEL, M. and ZANG, I., "Nine kinds of quasiconcavity and concavity", Discussion Paper 77—31, Department of Economics University of British Columbia, 1977.
[10] FERLAND, J. A., "Mathematical programming problems with quasi-convex objective functions", *Mathematical Programming* **3** (1972) 296—301.
[11] GERENCSÉR, L., "On a close relation between quasi-convex and convex functions and related investigations" *Math. Operationsforschung und Statistik* **4** (1973) 201—211.
[12] KÉRI, G., "An examination of nonnegativity and quasiconvexity conditions of quadratic forms on the nonnegative orthant", *Studia Sci. Math. Hungarica* **7** (1972) 11—20.
[13] MANGASARIAN, O. L., *Nonlinear Programming*, (McGraw-Hill, New York, 1969).
[14] MARTOS, B., *Nonlinear Programming: Theory and Methods*, (Akadémia Kiadó, Budapest, 1975).
[15] MEREAU, P. and PAQUET, J. G., "Second order conditions for pseudoconvex functions", *SIAM Journal on Applied Mathematics* **27** (1974) 131—137.
[16] PONSTEIN, J., "Seven kinds of convexity", *SIAM Review* **9** (1967) 115—119.
[17] RAPCSÁK, T., „Az optimalitás másodrendű feltételeiről", *Alk. Mat. Lapok* **4** (1978) 109—116.
[18] RAPCSÁK, T., „A SUMT-módszer alkalmazása nem konvex programozási feladatok esetén", *Alk. Mat. Lapok* **2** (1976) 427—437.
[19] SCHAIBLE, S., "Beiträge zur Quasi-Convexen Programmierung", Ph. D. Dissertation, Universität Köln (1971).
[20] SCHAIBLE, S., "Quasi-convex optimization in general real linear spaces", *Zeitschrift für Operations Research* **16** (1972) 205—213.
[21] SCHAIBLE, S., "Second-order characterization of pseudoconvex quadratic functions", *Journal of Optimization Theory and Applications* **21** (1977) 15—26.
[22] SZŐKEFALVI-NAGY, GY., GEHÉR, L. és NAGY, P., *Differenciálgeometria* (Műszaki Könyvkiadó, Budapest, 1979).

(Beérkezett: 1980. szeptember 22.)

(Átdolgozva beérkezett: 1983. január 21.)

KOMLÓSI SÁNDOR
PJPTE KÖZGAZDASÁGTUDOMÁNYI KAR MÓDSZERTANI TANSZÉK
7601 PÉCS, RÁKÓCZI ÚT 80.

## CONTRIBUTION TO THE THEORY OF QUASICONVEX FUNCTIONS

S. KOMLÓSI

In this paper definitions of different kinds of local generalized convexity are given and their relations to the global generalized convexity properties are investigated. The notion of *quasi-Hessian matrix* of a twice differentiable function is introduced and local pseudoconvexity, local strict pseudoconvexity, pseudoconvexity and strict pseudoconvexity are characterized by its help. A sufficient condition of strict local optimality for a mathematical programming problem based on the *quasi-Hessian* is presented.

*Alkalmazott Matematikai Lapok* **10** (1984) 115—123

# AZ ÍVKONVEXITÁSRÓL

RAPCSÁK TAMÁS

Budapest

A dolgozatban az ívkonvex függvények tulajdonságait vizsgáljuk, majd ennek alapján szükséges és elegendő feltételeket adunk arra, hogy egy függvény adott felület feletti lokális optimuma globális is legyen.

## 1. Bevezetés

A nemlineáris programozási feladatok egyik fontos osztályát a konvex programozási feladatok alkotják. A feladatosztály jelentőségét az adja, hogy itt bármely lokális optimum egyben globális is, ezért a számítógépes algoritmusok lényegesen hatékonyabbak, mint más esetekben. A matematikai programozás területén sokan foglalkoztak a konvexitás fogalmának az általánosításával [1], [4], [5]. Ebbe a csoportba tartozik az ívkonvexitás is, amelynek a definíciója az [1] könyvben található. Ebben a cikkben először az ívkonvex függvények tulajdonságaival foglalkozunk, majd egy felület adott vektormező irányában való konvexitását definiáljuk. Ezek az eredmények adnak lehetőséget arra, hogy bizonyos típusú, nem konvex programozási feladatok esetén is (pl. egyenlőség feltételekkel korlátozott nemlineáris programozási feladat, ahol a feltételek felületet határoznak meg) szükséges és elegendő feltételeket tudjunk adni arra, hogy bármely lokális optimum egyben globális is legyen. A dolgozatban szereplő definíciók és tételek a 2.1, 2.3 definíciók kivételével újak.

## 2. Az ívkonvexitás

Azt mondjuk, hogy egy $C \subset R^n$ halmaz konvex, ha bármely két pontjával együtt az összekötő szakaszt is tartalmazza. Ezt a definíciót általánosítani lehet, ha az összekötő szakasz helyett valamilyen folytonos összekötő ívet tekintünk. Ez az analízisben jól ismert alábbi definícióhoz vezet.

2.1. DEFINÍCIÓ ([1]). A $C \subset R^n$ halmazt ívösszefüggőnek nevezzük, ha bármely $\mathbf{x}_1 \in C$, $\mathbf{x}_2 \in C$ pontpár esetén van legalább egy a $[0, 1] \subset R$ intervallumon értelmezett, folytonos, vektorértékű $\mathbf{x}(t)$ függvény, amelyre

$$\mathbf{x}(t) \in C, \quad \forall t \in [0, 1] \quad \text{és} \quad \mathbf{x}(0) = \mathbf{x}_1, \quad \mathbf{x}(1) = \mathbf{x}_2. \tag{2.1}$$

Az $\mathbf{x}(t)$ függvényt összekötő ívnek is nevezzük.

Megjegyezzük, hogy $\mathbf{x}(t)$ általában függ az $\mathbf{x}_1$, $\mathbf{x}_2$ pontoktól, valamint a $C$ halmaztól és bármely $\mathbf{x}_1$, $\mathbf{x}_2$ pár esetén, amelyet egy ívösszefüggő halmazból veszünk, nemcsak egy összekötő ív lehetséges. Ha $\mathbf{x}_1 = \mathbf{x}_2 \in C$, akkor $\mathbf{x}(t) = \mathbf{x}_1$, $0 \leqq t \leqq 1$.

Nyilvánvaló, hogy egy konvex halmaz egyben ívösszefüggő is és

$$\mathbf{x}(t) = (1-t)\mathbf{x}_1 + t\mathbf{x}_2, \quad 0 \leqq t \leqq 1. \tag{2.2}$$

Jól ismert állítás, hogy egy függvény akkor és csak akkor konvex, ha az értelmezési tartományába eső bármely szakaszon konvex. Ez az állítás természetesen nem jelenti azt, hogy konvex függvények esetén a (2.2) helyett egy másik összekötő ívet tekintve, azon is teljesül a konvexitás [1].

2.2. DEFINÍCIÓ. Legyen $C$ egy ívösszefüggő halmaz. Az összekötő ívek egy részhalmazát a megengedett ívek egy családjának (vagy rövidebben megengedett íveknek) nevezzük, ha

a) bármely két $C$-beli pont esetén van a kitüntetett részhalmazhoz tartozó összekötő ív is,
b) a kitüntetett görbék bármely részgörbéje is kitüntetett.

2.3. DEFINÍCIÓ ([1]). Legyen $C \subset R^n$ egy ívösszefüggő halmaz és $\mathcal{M}$ a megengedett ívek egy családja. Az $f(\mathbf{x})$ skalár függvény, amely egy a $C$-t tartalmazó nyílt halmazon van értelmezve ívkonvex (az $\mathcal{M}$ halmazra nézve), ha bármely két $\mathbf{x}_1, \mathbf{x}_2 \in C$ pontot és egy megengedett ívet tekintve igaz a következő egyenlőtlenség:

$$f(\mathbf{x}(t)) \leqq (1-t)f(\mathbf{x}_1) + tf(\mathbf{x}_2), \quad 0 \leqq t \leqq 1. \tag{2.3}$$

2.4. LEMMA. Ha $f(\mathbf{x})$ ívkonvex a $C$ halmazon és $\mathbf{x}_0 \in C$ lokális minimum, akkor $\mathbf{x}_0$ egyben globális minimum pont is.

*Bizonyítás.* Tegyük fel, hogy nem igaz az állítás. Ekkor létezik $\mathbf{x}_1 \in C$ úgy, hogy $f(\mathbf{x}_0) > f(\mathbf{x}_1)$. Mivel $\mathbf{x}_0$ lokális minimum, ezért van olyan $V$ környezete, amelyben az $f(\mathbf{x}) \geqq f(\mathbf{x}_0)$, $\mathbf{x} \in C \cap V$ egyenlőtlenség teljesül.

Az ívkonvexitás miatt van legalább egy megengedett $\mathbf{x}(t)$ ív, amelyre

$$f(\mathbf{x}(t)) \leqq (1-t)f(\mathbf{x}_0) + tf(\mathbf{x}_1), \quad 0 \leqq t \leqq 1. \tag{2.4}$$

Ha a $\hat{t}$ értéket úgy választjuk, hogy $\hat{\mathbf{x}} = \mathbf{x}(\hat{t}) \in C \cap V$ legyen, akkor

$$f(\mathbf{x}_0) \leqq f(\hat{\mathbf{x}}) \leqq (1-\hat{t})f(\mathbf{x}_0) + \hat{t}f(\mathbf{x}_1) < f(\mathbf{x}_0). \tag{2.5}$$

Ez ellentmondás, így bebizonyítottuk az állítást.

Tegyük fel a továbbiakban, hogy $C \subset R^n$ egy nyílt, összefüggő halmaz. Ebből következik, hogy $C$ ívösszefüggő halmaz is [1].

2.5. LEMMA. Ha $f(\mathbf{x})$ differenciálható, ívkonvex függvény és a megengedett ívek differenciálhatóak, akkor bármely $\mathbf{x}_1, \mathbf{x}_2 \in C$ pont és $\mathbf{x}(t)$ $(\mathbf{x}(0) = \mathbf{x}_1, \mathbf{x}(1) = \mathbf{x}_2)$ megengedett ív esetén

$$f(\mathbf{x}_2) - f(\mathbf{x}_1) \geqq \nabla f(\mathbf{x}_1)\,\dot{\mathbf{x}}(0). \tag{2.6}$$

(Az $\dot{\mathbf{x}}(0)$ a görbe $t$ szerinti differenciálhányadosát jelenti a 0 pontban.)

*Bizonyítás.* Mivel $f(\mathbf{x})$ ívkonvex, ezért

$$f(\mathbf{x}(t)) \leqq (1-t)f(\mathbf{x}_1) + tf(\mathbf{x}_2), \quad 0 \leqq t \leqq 1. \tag{2.7}$$

Átrendezve az egyenlőtlenséget azt kapjuk, hogy

$$f(\mathbf{x}(t))-f(\mathbf{x}_1) \leqq t(f(\mathbf{x}_2)-f(\mathbf{x}_1)). \tag{2.8}$$

Ha $\mathbf{x}(t)\neq\mathbf{x}_1$, akkor $t>0$ és

$$\frac{f(\mathbf{x}(t))-f(\mathbf{x}_1)}{t} \leqq f(\mathbf{x}_2)-f(\mathbf{x}_1). \tag{2.9}$$

Határértéket véve kapjuk, hogy

$$\lim_{t\to 0}\frac{f(\mathbf{x}(t))-f(\mathbf{x}_1)}{t} = \frac{df(\mathbf{x}(t))}{dt}\bigg|_{t=0} \leqq f(\mathbf{x}_2)-f(\mathbf{x}_1). \tag{2.10}$$

Ebből következik, hogy

$$\nabla f(\mathbf{x}_1)\dot{\mathbf{x}}(0) \leqq f(\mathbf{x}_2)-f(\mathbf{x}_1), \tag{2.11}$$

ami az állítás.

2.6. LEMMA. Ha az $f(\mathbf{x})$ függvény, valamint a megengedett ívek differenciálhatók és bármely $\mathbf{x}_1, \mathbf{x}_2\in C$ és $\mathbf{x}(t)$ megengedett ív esetén

$$f(\mathbf{x}_2)-f(\mathbf{x}_1) \geqq \nabla f(\mathbf{x}_1)\dot{\mathbf{x}}(0), \tag{2.12}$$

akkor $f(\mathbf{x})$ ívkonvex.

*Bizonyítás.* Legyen $\mathbf{x}(t)$ $\mathbf{x}_1$ és $\mathbf{x}_2$ között egy megengedett ív és $\hat{\mathbf{x}}=\mathbf{x}(\hat{t})$ az ív egy tetszőleges belső pontja. Így $\hat{t}$ egy fix érték, amelyre $0<\hat{t}<1$.

Tekintsük az

$$\begin{aligned} \mathbf{x}_{\hat{\mathbf{x}}\mathbf{x}_2}(\lambda) &= \mathbf{x}(\hat{t}+\lambda(1-\hat{t})), \quad 0 \leqq \lambda \leqq 1 \\ \mathbf{x}_{\hat{\mathbf{x}}\mathbf{x}_1}(\mu) &= \mathbf{x}(\hat{t}-\mu\hat{t}), \qquad 0 \leqq \mu \leqq 1 \end{aligned} \tag{2.13}$$

görbéket. Mivel ezek is megengedett ívek, ezért

$$\begin{aligned} f(\mathbf{x}_2)-f(\hat{\mathbf{x}}) &\geqq \nabla f(\hat{\mathbf{x}})\dot{\mathbf{x}}_{\hat{\mathbf{x}}\mathbf{x}_2}(0) \\ f(\mathbf{x}_1)-f(\hat{\mathbf{x}}) &\geqq \nabla f(\hat{\mathbf{x}})\dot{\mathbf{x}}_{\hat{\mathbf{x}}\mathbf{x}_1}(0). \end{aligned} \tag{2.14}$$

A (2.14) egyenlőtlenségekből nyerjük, hogy

$$\begin{aligned} f(\mathbf{x}_2)-f(\hat{\mathbf{x}}) &\geqq \nabla f(\hat{\mathbf{x}})\dot{\mathbf{x}}(\hat{t})(1-\hat{t}) \\ f(\mathbf{x}_1)-f(\hat{\mathbf{x}}) &\geqq \nabla f(\hat{\mathbf{x}})\dot{\mathbf{x}}(\hat{t})(-\hat{t}). \end{aligned} \tag{2.15}$$

Ebből következik, hogy

$$\begin{aligned} \hat{t}f(\mathbf{x}_2)-\hat{t}f(\hat{\mathbf{x}}) &\geqq \nabla f(\hat{\mathbf{x}})\dot{\mathbf{x}}(\hat{t})\hat{t}(1-\hat{t}) \\ (1-\hat{t})f(\mathbf{x}_1)-(1-\hat{t})f(\hat{\mathbf{x}}) &\geqq -\nabla f(\hat{\mathbf{x}})\dot{\mathbf{x}}(\hat{t})(1-\hat{t})\hat{t}. \end{aligned} \tag{2.16}$$

Az egyenlőtlenségeket összeadva kapjuk, hogy

$$(1-\hat{t})f(\mathbf{x}_1)+\hat{t}f(\mathbf{x}_2) \geqq f(\hat{\mathbf{x}}), \tag{2.17}$$

ami éppen az állítás.

2.7. LEMMA. Ha az $f(\mathbf{x})$ függvény és a megengedett ívek kétszer folytonosan differenciálhatóak és $f(\mathbf{x})$ ívkonvex, akkor tetszőleges $\mathbf{x}(t)$ megengedett ívet tekintve az alábbi egyenlőtlenség teljesül.

$$\frac{d^2}{dt^2} f(\mathbf{x}(t)) \geqq 0, \quad 0 \leqq t \leqq 1. \tag{2.18}$$

*Bizonyítás.* A (2.18) egyenlőtlenséget elég a $t=0$ esetben bizonyítani, ugyanis az $\mathbf{x}(\hat{t}+\mu(1-\hat{t}))$ $(0 \leqq \mu \leqq 1)$ görbét tekintve az eredmény tetszőleges $0<\hat{t} \leqq 1$ esetre átvihető.

Az $f(\mathbf{x}(t))$ függvényt fejtsük a 0 pont körül *Maclaurin sorba.*

$$f(\mathbf{x}(t)) = f(\mathbf{x}(0)) + t\frac{d}{dt} f(\mathbf{x}(0)) + \frac{1}{2} t^2 \frac{d^2}{dt^2} f(\mathbf{x}(\xi)), \tag{2.19}$$

$$\xi = \theta t, \quad 0 < \theta < 1, \quad t > 0.$$

Ez nem más mint

$$f(\mathbf{x}(t)) = f(\mathbf{x}(0)) + t\nabla f(\mathbf{x}(0))\dot{\mathbf{x}}(0) + \frac{1}{2} t^2 \frac{d^2}{dt^2} f(\mathbf{x}(\xi)), \tag{2.20}$$

$$\xi = \theta t, \quad 0 < \theta < 1, \quad t > 0.$$

Ha a $t$ értékét rögzítjük, akkor az $\mathbf{x}(0)$ és az $\mathbf{x}(t)$ pont között az

$$\mathbf{x}(\mu t), \quad 0 \leqq \mu \leqq 1 \tag{2.21}$$

egy megengedett ív és erre a 2.5. lemma alapján az

$$f(\mathbf{x}(t)) - f(\mathbf{x}(0)) \geqq t\nabla f(\mathbf{x}_0)\dot{\mathbf{x}}(0) \tag{2.22}$$

egyenlőtlenség teljesül, ezért

$$\frac{d^2}{dt^2} f(\mathbf{x}(\xi)) \geqq 0, \ \xi = \theta t, \quad 0 < \theta < 1. \tag{2.23}$$

Mivel ez minden $t>0$ értékre igaz, ezért

$$\lim_{t \to 0} \frac{d^2}{dt^2} f(\mathbf{x}(\xi)) = \frac{d^2}{dt^2} f(\mathbf{x}(0)) \geqq 0, \tag{2.24}$$

ami az állítás.

2.8. LEMMA. Ha az $f(\mathbf{x})$ függvény és a megengedett ívek kétszer differenciálhatóak és bármely megengedett ívet tekintve

$$\frac{d^2}{dt^2} f(\mathbf{x}(t)) \geqq 0, \quad 0 \leqq t \leqq 1, \tag{2.25}$$

akkor $f(\mathbf{x})$ ívkonvex a $C$ halmazon.

*Bizonyítás.* A (2.19) egyenlőségből következik, hogy tetszőleges megengedett ívet tekintve

$$f(\mathbf{x}(t)) - f(\mathbf{x}(0)) \geqq t\nabla f(\mathbf{x}_0)\dot{\mathbf{x}}(0), \quad 0 \leqq t \leqq 1. \tag{2.26}$$

Ezért bármely $\mathbf{x}_1, \mathbf{x}_2 \in C$ és $\mathbf{x}(t)$ $(\mathbf{x}(0)=\mathbf{x}_1, \mathbf{x}(1)=\mathbf{x}_2)$ megengedett ív esetén

$$f(\mathbf{x}_2)-f(\mathbf{x}_1) \geqq \nabla f(\mathbf{x}_1)\dot{\mathbf{x}}(0). \tag{2.27}$$

Ez viszont a 2.6. lemma alapján éppen az ívkonvexitást jelenti.

Ha az összekötő görbék ívhosszparaméterben vannak megadva, akkor két tetszőlegesen választott $\mathbf{x}_1, \mathbf{x}_2 \in C$ pont között a megengedett ívek $\mathbf{x}(s)$, $\mathbf{x}(0)=\mathbf{x}_1$, $\mathbf{x}(s_0)=x_2$ alakúak. Bevezetve az $\mathbf{x}(s)=\mathbf{x}(\lambda s_0)=\tilde{\mathbf{x}}(\lambda)$, $0 \leqq \lambda \leqq 1$ jelölést, az előzőekben tárgyalt esetet kapjuk meg.

A továbbiakban az ívhossz szerinti differenciálást vesszővel jelöljük, azaz

$$\frac{d\mathbf{x}(s)}{ds} = \mathbf{x}'(s). \tag{2.28}$$

### 3. A függvények konvexitási tulajdonságának általánosítása felületekre

Ebben a részben azt vizsgáljuk meg, hogy mikor nevezhetünk egy $\mathbf{x}(\mathbf{u})$, $\mathbf{u} \in U_k$ $k$-dimenziós felületet valamilyen értelemben konvexnek. A differenciálgeometriában a konvexitás általánosítása a geodetikusan konvex felületekhez vezetett. Egy felületet akkor nevezünk geodetikusan konvexnek, ha bármely két pontját összekötő, minimális hosszúságú geodetikus is a felülethez tartozik [3], [8], [10]. A [6] cikk is foglalkozik azzal a kérdéssel, hogy egy $(n-1)$-dimenziós felület mikor konvex. A dolgozatban található megközelítés abban különbözik az előzőektől, hogy itt a felületen kívül egy vektormező is adott, amelyet a nemlineáris programozási feladatban a célfüggvény határoz meg.

Először szükségünk van a függvények konvexitás fogalmának differenciálgeometriai interpretációjára. Legyen $X_1 \subset R^n$ egy nyílt, konvex halmaz és $f(\mathbf{x})$ egy ezen értelmezett, kétszer folytonosan differenciálható konvex függvény. Jól ismert az az állítás, hogy $f(\mathbf{x})$ akkor és csak akkor konvex az $X_1$ halmazon, ha a *Hesse-mátrixa* minden pontban pozitív szemidefinit.

Tekintsük most az $R^{n+1}$ dimenziós térben az $y-f(\mathbf{x})=0$ nívófelületet. Ez a felület felírható az alábbi formában is.

$$\mathbf{x}(\mathbf{u}) = \begin{cases} x_1 = u_1 \\ x_2 = u_2 \\ \quad \vdots \\ x_n = u_n \\ x_{n+1} = f(\mathbf{u}) \end{cases}, \quad \mathbf{u} \in U_n = X_1 \subset R^n. \tag{3.1}$$

A (3.1) felület paramétervonalérintői a következők:

$$\frac{\partial \mathbf{x}}{\partial u_1} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \\ \dfrac{\partial f}{\partial u_1} \end{pmatrix}, \quad \frac{\partial \mathbf{x}}{\partial u_2} = \begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \\ \dfrac{\partial f}{\partial u_2} \end{pmatrix}, \ldots, \frac{\partial \mathbf{x}}{\partial u_n} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \\ \dfrac{\partial f}{\partial u_n} \end{pmatrix}. \tag{3.2}$$

Ezek a vektorok lineárisan függetlenek, ezért (3.1) differenciálgeometriai értelemben is felületet határoz meg. A (3.1) hiperfelület normálvektora egy tetszőleges pontban

$$\mathbf{n} = -(\nabla f(\mathbf{x}), 1). \tag{3.3}$$

Határozzuk meg a (3.1) felület második alapformájának a mátrixát, amit a továbbiakban **B** jelöl. Ez a definíció alapján a következő:

$$b_{ij} = \frac{\partial^2 \mathbf{x}}{\partial u_i \partial u_j} \cdot \frac{1}{\|\mathbf{n}\|} \mathbf{n}, \quad i, j = 1, \ldots, n, \tag{3.4}$$

ahol a $\| \ \|$ szimbólum a vektor normát jelenti.

Ebből következik, hogy

$$\mathbf{B} = \frac{1}{\|\mathbf{n}\|} \mathbf{H}, \tag{3.5}$$

ahol **H** az $f(\mathbf{x})$ függvény *Hesse-mátrixát* jelenti.

A konvexitás tehát az alaphalmazra jellemző globális tulajdonságokból (az $X_1$ konvex, nyílt halmaz) és a függvényre jellemző lokális tulajdonságokból tevődik össze (a *Hesse-mátrix* vagy a (3.1) hiperfelület második alapmennyiségeinek a mátrixa minden pontban pozitív szemidefinit).

Tekintsünk most egy $k$-dimenziós felületet. Ennél a felületnél a normálvektorok tere $(n-k)$-dimenziós és mivel a második alapformát minden normálvektor irányában lehet értelmezni, ezért a definícióban egy iránymezőt is meg kell adni.

3.1. DEFINÍCIÓ. Legyen adott az $\mathbf{x}(\mathbf{u})$, $\mathbf{u} \in U_k$ $k$-dimenziós felület és rendeljünk hozzá ehhez a felülethez egy $V$ vektormezőt. (A vektormező vektorai nem kell, hogy szükségképpen benne legyenek a normális térben.) Azt mondjuk, hogy az $\mathbf{x}(\mathbf{u})$ felület a $V$ vektormező irányában konvex (konkáv), ha a megfelelő második alapforma minden pontban nemnegatív (nempozitív).

A (3.1) felületnél a $V$ vektormező a $(-\nabla f(\mathbf{x}), 1)$ vektorokból állt.

A fenti definíció annyival általánosabb, mint a klasszikus konvexitás fogalom, hogy az alaphalmazra nem kell kikötni a konvexitást, a felület nem speciális nívófelület és a $V$ vektormező is tetszőleges lehet.

## 4. Függvények felületek feletti ívkonvexitásáról

A nemlineáris programozásban a konvex programozási feladatok legjobb tulajdonsága, hogy bármelyik lokális minimum egyben globális is, ezért a számítógépes algoritmusok lényegesen hatékonyabbak, mint más esetekben. Azonban ez a tulajdonság elvész, ha a nemlineáris feltételek között egyenlőség feltételek is vannak. Ebben a részben a célunk az, hogy ha a nemlineáris programozási feladat feltételei felületet határoznak meg, akkor szükséges és elegendő feltételeket tudjunk megadni arra nézve, hogy bármelyik lokális minimum egyben globális is legyen. Ehhez felhasználjuk az előző részek eredményeit.

Legyen adott az $\mathbf{x}(\mathbf{u}) \in R^n$, $\mathbf{u} \in U_k$ $k$-dimenziós felület és egy nyílt $C \subset R^n$ $(\mathbf{x}(\mathbf{u}) \in C, \ \mathbf{u} \in U_k)$ halmazon értelmezett kétszer folytonosan differenciálható $f(\mathbf{x})$ függvény. A differenciálgeometriai vizsgálatoknál a felület pontjaiban a lokális koor-

dinátarendszer az érintőteret kifeszítő és az arra ortogonális, úgynevezett normális teret kifeszítő vektorokból áll. Az általánosság megszorítása nélkül feltehetjük, hogy olyan lokális, ortonormált koordinátarendszert választunk, amelyben a $\nabla f(\mathbf{x})$ vektor normális irányú összetevője (amit $\nabla f(\mathbf{x})_N$ jelöl) az első normális irányú koordinátatengely irányába mutat. Így

$$\nabla f(\mathbf{x}) = \nabla f(\mathbf{x})_T + \nabla f(\mathbf{x})_N \tag{4.1}$$

ahol $\nabla f(\mathbf{x})_T$ a tangenciális irányú összetevőt jelenti.

4.1. TÉTEL. Legyenek az $\mathbf{x}(\mathbf{u})$ felületen a megengedett ívek az ívhosszparaméterben megadott geodetikusok. Akkor annak elegendő feltétele, hogy az $f(\mathbf{x})$ függvény az $\mathbf{x}(\mathbf{u})$ felületen ívkonvex legyen az, hogy az $f(\mathbf{x})$ függvény a $C$ halmazon és az $\mathbf{x}(\mathbf{u})$ felület a $\nabla f(\mathbf{x})_N$ vektormező irányában konvex legyen.

*Bizonyítás.* A 2.8. lemmát és az utána következő megjegyzést felhasználva elegendő azt bizonyítani, hogy bármely megengedett ívet (geodetikust) tekintve

$$\frac{d^2}{dt^2} f(\mathbf{x}(ts_0)) \geqq 0, \quad 0 \leqq t \leqq 1. \tag{4.2}$$

Mivel a megengedett ívek ívhosszparaméterben vannak megadva, azért

$$\frac{d}{dt} f(\mathbf{x}(ts_0)) = \nabla f(\mathbf{x}(s))\mathbf{x}'(s) \cdot s_0 \tag{4.3}$$

$$\frac{d^2}{dt^2} f(\mathbf{x}(ts_0)) = \mathbf{x}'(s)^T \nabla^2 f(\mathbf{x}(s))\mathbf{x}'(s)s_0^2 + \nabla f(\mathbf{x}(s))\mathbf{x}''(s)s_0^2.$$

Kihasználva azt, hogy felületen vagyunk az $\mathbf{x}(\mathbf{u}(s))$ görbe deriváltjaira a következő kifejezéseket nyerjük:

$$\frac{d\mathbf{x}(\mathbf{u}(s))}{ds} = \frac{\partial \mathbf{x}}{\partial u_i} \cdot u_i' \tag{4.4}$$

$$\frac{d^2\mathbf{x}(\mathbf{u}(s))}{ds} = \frac{\partial^2 \mathbf{x}}{\partial u_i \partial u_j} u_i' u_j' + \frac{\partial \mathbf{x}}{\partial u_i} u_i'',$$

ahol az *Einstein-féle konvenció* szerint az egy tagban kétszer előforduló azonos indexre a szummáció jel kiírása nélkül is szummációt értünk.

A *Gauss-féle egyenletek* szerint

$$\frac{\partial^2 \mathbf{x}}{\partial u_i \partial u_j} = \Gamma_{ij}^{\sigma} \frac{\partial \mathbf{x}}{\partial u_\sigma} + b_{ij}^{\gamma} \mathbf{n}_\gamma, \quad i, j = 1, \ldots, k \tag{4.5}$$

ahol a $\Gamma_{ij}^{\sigma}$ mennyiségek a *másodfajú Christoffel-szimbólumok*, a $b_{ij}^{\gamma}$ mennyiségek pedig a megfelelő normális irányba eső második alapforma mátrixának az elemei.

A (4.4) második egyenlőségében felhasználva a *Gauss-féle egyenleteket* azt kapjuk, hogy

$$\frac{d^2\mathbf{x}(\mathbf{u}(s))}{ds} = (\Gamma_{ij}^{\sigma} u_i' u_j' + u_\sigma'') \frac{\partial \mathbf{x}}{\partial u_\sigma} + b_{ij}^{\gamma} u_i' u_j' \mathbf{n}_\gamma. \tag{4.6}$$

A geodetikus vonal egyenlete ívhossz paraméterben

$$u''_\sigma = -\Gamma^\sigma_{ij} u'_i u'_j \tag{4.7}$$

és mivel $\mathbf{x}(\mathbf{u}(s))$ geodetikus, ezért a (4.6) egyenletben, a zárójelben levő kifejezés eltűnik.

Ebből következik, hogy

$$\nabla f(\mathbf{x}(s))\mathbf{x}''(s) = (\nabla f(\mathbf{x})_T + \nabla f(\mathbf{x})_N)\mathbf{x}''(s) = |\nabla f(\mathbf{x})_N| b^1_{ij} u'_i u'_j. \tag{4.8}$$

(A (4.8) egyenlőségben felhasználtuk, hogy $\nabla f(\mathbf{x})_T$ ortogonális minden normális irányú vektorra és $\nabla f(\mathbf{x})_N$ az első normális irányú koordinátatengely irányába esik.)

A (4.8) egyenlőségből következik, hogy a $\nabla f(\mathbf{x}(s))\mathbf{x}''(s)$ kifejezés akkor és csak akkor nemnegatív, ha az $\mathbf{x}(\mathbf{u})$ felület konvex a $\nabla f(\mathbf{x})_N$ vektormező irányában.

Az $f(\mathbf{x})$ függvény konvex, ezért a *Hesse-mátrixa* minden pontban pozitív szemidefinit, tehát (4.3) alapján valóban teljesülnek a (4.2) egyenlőtlenségek, azaz az $f(\mathbf{x})$ függvény az $\mathbf{x}(\mathbf{u})$ felületen ívkonvex.

4.2. KÖVETKEZMÉNY. A 4.1. tétel feltételei mellett az $\mathbf{x}(\mathbf{u})$ felületen az $f(\mathbf{x})$ függvény minden lokális minimuma egyben globális minimum is.

4.3. KÖVETKEZMÉNY. Az $f(\mathbf{x})$ függvény $\mathbf{x}(\mathbf{u})$ felület feletti ívkonvexitásának szükséges és elegendő feltétele, hogy minden pontban az

$$u'^T_i \frac{\partial \mathbf{x}^T}{\partial u_i} \nabla^2 f(\mathbf{x}(s)) \frac{\partial \mathbf{x}}{\partial u_i} u'_i + |\nabla f(\mathbf{x})_N| b^1_{ij} u'_i u'_j \geqq 0 \tag{4.9}$$

egyenlőtlenségek teljesüljenek.

*4.4. Megjegyzés.* Ha a 4.1. tételben és a 4.3. következményben az $\mathbf{x}(\mathbf{u})$ felület helyett euklideszi térbe beágyazott differenciálható sokaságot tekintünk, az állítás akkor is érvényben marad.

## IRODALOM


[1] AVRIEL, M., *Nonlinear Programming, Analysis and Methods* (Prentice-Hall, Inc. Englewood Cliffs, New Yersey, 1976).

[2] BAZARAA, M. S. and SHETTY, C. M., *Foundations of Optimization* (Springer-Verlag, Berlin, Heidelberg, New York, 1976).

[3] HICKS, N. J. *Notes on Differential Geometry* (D. Van Nostrand Company, Inc. Princeton, New Jersey, Toronto, New York, London, 1965).

[4] MANGASARIAN, O. L., *Nonlinear Programming* (McGraw-Hill Book Company, 1969).

[5] MARTOS, B., *Nonlinear Programming Theory and Methods* (Akadémiai Kiadó, Budapest, 1975).

[6] NOŽIČKA, F., "Spherical mappings of convex sets and convex parametrical optimization", *Survey of Mathematical Programming*, Proceedings of the 9th International Mathematical Programming Symposium, Edited by *A. Prékopa*, Akadémiai Kiadó, Budapest, 1979.

[7] RAPCSÁK, A. és TAMÁSSY, L., *Differenciálgeometria* (Tankönyvkiadó, Budapest, 1967).

[8] SPIVAK, M., *A Comprehensive Introduction to Differential Geometry* (Publish or Perish Inc. Berkeley 1979).

[9] SZŐKEFALVI-NAGY, GY., GEHÉR, L. és NAGY, P., *Differenciálgeometria* (Műszaki Könyvkiadó, Budapest, 1979).

[10] Сарафутдинов, В. А., "Теорема Погорелова—Клингенберла для многообразний, гомеаморфних $R^n$", Сибирский Математический Журнал (1977) 915—925.

(Beérkezett: 1983. április 20,)

RAPCSÁK TAMÁS
MTA SZÁMÍTÁSTECHNIKAI ÉS AUTOMATIZÁLÁSI KUTATÓ INTÉZET
1250 BUDAPEST, URI U. 49.

## ON THE ARCWISE CONVEXITY

T. RAPCSÁK

The paper deals with the characteristics of the arcwise convex functions, then gives on this basis necessary and sufficient conditions to ensure for any local minimum to be global.

*Alkalmazott Matematikai Lapok* **10** (1984) 125—148

# ÚJ NUMERIKUS MÓDSZER AZ ÁRAMLÁSTANILAG LINEÁRIS VEGYIPARI BERENDEZÉSEK SZIMULÁCIÓJÁHOZ

HALÁSZ GÁBOR

Veszprém

A hő- és anyagátadási folyamatok transzportegyenleteit *Hammerstein típusú integrálegyenlet* formájában adtuk meg. Összefoglaltuk a *Hammerstein típusú integrálegyenletek* egzisztencia és unicitás feltételeit, a vonatkozó irodalom áttekintésével. A *Hammerstein típusú integrálegyenlet-rendszer* megoldásához egy új, numerikusan stabilabb algoritmust és programot dolgoztunk ki. Leírtuk az algoritmus struktúráját és összehasonlítottuk a program működését más numerikus módszereken alapuló programok eredményeivel. Az összehasonlításhoz részben irodalomból ismert teszt feladatokat, részben két, a vegyipari szimulációs gyakorlatunkban előforduló példát használtunk.

## 1. Bevezetés

A vegyipari rendszerek (VIR) szimulációja több évtizedes múltra tekint vissza. A nemzetközi és a magyar irodalomban számos könyv [2, 17] jelent meg a VIR szimulációjáról annak metodikájáról. L. B. EVANS [11] a VIR szimulációjáról tartott *„state-of-art"* előadását így kezdte:

„Ma már egyetlen nagyobb vegyipari vagy petrolkémiai gyárat sem építenek meg anélkül, hogy előbb a folyamatot matematikai modellek segítségével ne szimulálnák."

A VIR-t matematikai szempontból viszonylag egyszerű, kevés paramétert tartalmazó bemenet-kimenet összefüggések segítségével számítják (algebrai egyenletrendszer). A számítástechnika mai lehetőségei mellett természetesen az egy-egy készüléken (reaktoron) belül lejátszódó fizikai, fizikai-kémiai folyamatok és a teljes VIR egyidejű vizsgálata gyakorlatilag nem lehetséges. Az egyedi modellezési egységek szimulációjánál ugyanis olyan bonyolult összefüggéseket vesznek figyelembe, melyek segítségével már pl. a reaktorban levő egy-egy katalizátor szemcsén, illetve katalizátor ágyban lejátszódó fizikai folyamatok is nyomon követhetők (nemlineáris parciális differenciálegyenlet-rendszer). Ugyanakkor a teljes VIR szimulációjánál egy-egy készülék működését általában azok hatásfokára utaló összefüggésekkel számítják.

A vegyipar szorpciós rendszereit vizsgálva merült fel annak igénye, hogy a több műveleti egységből álló VIR szimulációjánál célszerű lenne a modellek bonyolultságát tekintve egy középutat választani, azaz a néhány modellezési egységet tartalmazó rendszerek szimulációjánál az egyes modellezési egységeken belül kialakuló intenzív jellemzők eloszlásait (azaz matematikai szempontból a megoldás függvényeket) is számítani.

A VIR működését leíró egyenletek numerikus megoldásánál is, mint általában, a legnagyobb problémát a nemlinearitás miatt adódó iterációk jelentik. Az iterációk oka lehet:

— a fizikai-kémiai tulajdonságok közötti összefüggések nemlinearitása,
— a transzport egyenletek (mérlegek) nemlinearitása,
— a készülékek áramainak hálózatából eredő nemlinearitások, pl. visszacsatolások, szabályozások, stb.

A VIR számításánál így több egymásba ágyazott iterációs ciklusra van szükség, melynek következtében a számítási idő jelentősen megnő, valamint minden egyes iterációnál megszünhet a konvergencia, melynek következtében az egész szimuláció lehetetlenné válhat.

Konvergencia gyorsító eljárásokkal bizonyos esetekben a számítási időt sikerül jelentősen csökkenteni. Mások a hálózaton belüli iterációk számának minimalizálásával próbálták a feladatot könnyebbé tenni, a készülékek áramai hálózatának felszakítási helyének optimális megválasztásával minimalizálni lehet a hálózatból eredő iterációk számát [28].

BENEDEK [3] ezzel kapcsolatban távlati célt írt le: „A fizikai-kémiai tulajdonságok számításához, a készülékek modellezéséhez és a hálózat kezeléséhez a matematikai módszerek együttes átgondolása révén iterációmentes explicit megoldásokra kell törekedni."

VIRÁG [30] kutatásai nyomán lehetőség nyílott arra, hogy a hálózatból eredő iterációkat bizonyos esetekben explicit matematikai kifejezések megadásával elkerüljük.

A dolgozat célja a vonatkozó irodalom összefoglalása és a néhány műveleti, illetve a több modellezési egységből álló stacionárius működésű VIR szimulációjához kidolgozott új, kedvező tulajdonságú numerikus módszer ismertetése.

## 2. Előzmények

A bevezetésben a következő, a vegyiparban is használt fogalmak segítségével fejeztük ki magunkat: modellezési egység, készülék, vegyipari rendszer (VIR). E cikk keretei között a fenti szavak a következő jelentéssel bírnak.

A modellezési egység egy olyan térrész, melyet az jellemez, hogy bizonyos energia és tömegáramok belépnek oda és elhagyják azt. A térrésznek két kitüntetett pontja van, az áramok be- és/vagy kilépési helye. A fázisok (minden fázis formálisan vagy reálisan külön áram) közötti bizonyos áramok arányosak a fázisok közötti érintkezési felületekkel, a fázisok belsejében levő töménység-, hőmérséklet-, sebességkülönbséggel. A fázisok között így létrejövő anyag és/vagy entalpia cserét nevezzük átadásnak. A kitüntetett térrészen belül az áramok anyagot és/vagy entalpiát csak átadás formájában cserélhetnek.

A modellezési egységen belül érvényesek a megmaradási-, sztöchiometriai és termodinamikai törvények. Ezek a szabályok teszik lehetővé, hogy a bemenő anyag- és/vagy energia áramok birtokában kiszámítsuk a kimenő áramokat, a szabályokat a modellezési egység mérlegegyenletei reprezentálják, melyeket véges kifejezések formájában lehet megadni.

Egy vagy több modellezési egység ki- és belépő áramait meghatározott törvényszerűségek szerint összekötve felépíthetünk egy fizikailag létező objektumot, a készülék modelljét. A készülék szinonimájaként a berendezés fogalmát is használni fogjuk. A készülék a vegyi gyár olyan része, melyet valamilyen szempontból önállóan vizsgálunk.

A vegyi üzem ilyen készülékek halmazából áll. A vegyi üzem adott céllal összekapcsolt készülékeinek összességét nevezzük vegyipari rendszernek.

A differenciális mérlegegyenletek különböző formájú differenciálegyenletekre vezetnek. VIRÁG [30] a különböző típusú mérlegegyenleteket vizsgálva abból a feltevésből indult ki, hogy az áramlástanilag lineáris vegyipari berendezések olyan lineáris időinvariáns rendszerek, amelyek állapota a jellemző extenzív mennyiségek sűrűség eloszlásával jellemezhető. A funkcionálanalízis eszköztárát felhasználva sikerült olyan formalizmust találnia, amelyben számos, eddig különbözőnek tekintett modellt (diffúziós, recirkulációs, keveredési egységszám modellt, stb.) egységesen lehet kezelni. Ez matematikai szempontból azt jelenti, hogy az eddig formailag is eltérő leírásokat, azaz algebrai, első- és másodrendű differenciálegyenlet-rendszereket egyetlen típusú integrálegyenlet-rendszerrel lehet helyettesíteni.

Az általa bevezetett egységes matematikai modell formája stacionárius esetre vonatkoztatva:

(2.1) $$W = I_0 + \mathbf{KF} \cdot w$$

ahol

$$I_0, \quad w \in H$$

$$H := H_1 * H_2 * \ldots * H_j * \ldots * H_J.$$

$$H_j = L_2[0, L], \quad j = 1, 2, \ldots, J.$$

$$H \in [w_1(.), \ldots, w_j(.), \ldots, w_J(.)].$$

$$\mathbf{F}: H \mapsto H, \quad \mathbf{K}: H \mapsto H,$$

$$\mathbf{K} := K_{ij}; \; K_{ij}: H_j \mapsto H_i.$$

(2.2)
$$X(Z) = X_0 - \int_0^L K(Z, Y) \cdot F(Y, X_1(Y), \ldots, X_j(Y), \ldots, X_J(Y)) dY \quad j = 1, 2, \ldots, J$$

$Z$ — hosszkoordináta, $Z \in [0, L]$,
$Y$ — integrálási paraméter, $Y \in [0, L]$,
$L$ — modellezési egység hossza,
$X(Z)$ — az extenzív mennyiség sűrűség eloszlásának vektora,

$$F(Y, X(Y)) := F(Y, X_1(Y), X_2(Y), \ldots, X_J(Y))$$

a forrásfüggvény vektora feltételezzük, hogy olyan folytonos valós függvény, melyet az $X * Y$ vektortér felett értelmezünk és amelynek véges számú pontban az $X$ vektor szerinti deriváltjának elsőfajú szakadása lehet.

$X_0$ — a bemenő $X$ vektora (konstans)
$K(Z, Y)$ — mátrix, az áramok hálózatát és az egyes modellezési egységeken belül az áram hidrodinamikáját jellemző magfüggvény.

A szokásos VIR modellek [17] a (2.2) alakra hozhatók és a magfüggvények a következő tulajdonságokkal rendelkeznek:

$$K_{ii}(L, Y) = 1, \quad Y \in [0, L]; \quad 0 < i, j, \leqq J \quad \text{és} \tag{2.3}$$

$$K_{ij}(Z, Y) = 1, \quad \text{ha} \quad i > j, \quad Z, Y \in [0, L]$$

$$K_{ij}(Z, Y) \geqq 0, \quad i, j = 1, 2, \ldots, J, \tag{2.4}$$

$$\int_0^L \int_0^L U(Z) K(Z, Y) U(Y) dZ dY \geqq 0 \quad \forall U(Z) \in H. \tag{2.5}$$

Tehát a $K$ operátor a fent rögzített teljes téren értelmezett pozitív szemidefinit, korlátos, lineáris operátor. Az operátor linearitását kihasználva, a készülékek közötti áramok hálózatából származó iterációkat az integrál formalizmus keretében meg lehet szüntetni, mint ezt a dolgozatban megmutatjuk.

A fent leírt absztrakt konstrukció előnyeit a numerikus megoldás kidolgozásánál is kihasználtuk.

Az integrálegyenletek tipológiájában [22] a (2.2) egyenlet rendszert *Hammerstein típusúnak* nevezik, ha $I=1$, melynek tulajdonságai: a magfüggvény és a forrásfüggvény szétválasztható, a forrásfüggvény $X$-ben nemlineáris és az integrálási határok konstansok.

## 3. A Hammerstein típusú integrálegyenletek néhány egzisztencia és unicitási tétele, alkalmazásuk feltételei

Az ötvenes években matematikai kutatások homlokterébe kerültek a *Hammerstein típusú integrálegyenletek,* a megfelelő unicitási és egzisztencia tételek kidolgozása, ezen tételek közül néhány fontosabbat az alább ismertetünk.

1. E témában bibliának számít KRASZNOSZEL'SZKIJ [19] monográfiája, melyben a nemlineáris integrálegyenletekre topológiai módszereket alkalmazott.

A *Hammerstein egyenlet:* (itt $X$ — skalár, $I=1$)

$$X(Z) = \int_G K(Z, Y) \cdot F(Y, X(Y)) dY, \tag{3.1}$$

ahol $G$ — véges vagy végtelen dimenziójú halmaz.

Definiáljuk a következő **A** lineáris operátort:

$$\mathbf{A}X(Z) := \int_G k(Z, Y) X(Y) dY, \tag{3.2}$$

ekkor a (3.1)-nek megfelelő lineáris egyenlet:

$$X = \mathbf{A} \cdot X$$

$K(Z, Y)$ és $k(Z, Y)$ magok között a következő kapcsolat van:

Legyen a szimmetrikus $K(Z, Y)$ magnak véges számú negatív sajátértéke.

Ekkor a magot a következő alakban lehet felírni:

$$K(Z, Y) = -\sum_{i=1}^{n} \lambda_i \varphi_i(Z)\varphi_i(Y) + \sum_{i=n+1}^{\infty} \lambda_i \varphi_i(Z)\varphi_i(Y)$$

ahol $\lambda_1, \ldots, \lambda_n, \lambda_{n+1}, \ldots$ pozitív számok.

Legyen $k(Z, Y)$ pozitív definit mag és

$$k(Z, Y) = \sum_{i=1}^{n} \lambda_i \varphi_i(Z)\varphi_i(Y)$$

$\varphi_i(Z)$ $(i=1, 2, \ldots)$ ortonormált sajátfüggvény

$\lambda_i$ $(i=1, 2, \ldots)$ pozitív sajátérték.

Ekkor

$$k(Z, Y) = K(Z, Y) + 2 \cdot \sum_{i=1}^{n} \lambda_i \varphi_i(Z) \cdot \varphi_i(Y).$$

Legyen az **A** operátor olyan, hogy

$$\mathbf{A} = \mathbf{H} * \mathbf{H}^*, \tag{3.3}$$

ahol **H** — lineáris, teljesen folytonos operátor és a leképezés $L^2 \mapsto L^p$.

Tegyük fel, hogy a forrásfüggvény $(F)$ a következő feltételeknek tesz eleget:

$$\int_0^U F(Y, U)dU \geqq a \cdot U^2 - b(Y) \cdot |U|^{2-\gamma} - C(Y), \tag{3.4}$$

ahol

$$Y \in G; \quad -\infty \leqq U < \infty$$

és

$$0 < \gamma < 2; \; b(Y) \in L^{2/\gamma}; \; C(Y) \in L$$

és

$$a|\lambda_-| > 1,$$

ahol $\lambda_-$ a $K(Z, Y)$ legkisebb sajátértéke.

TÉTEL. Legyen $K(Z, Y)$ — olyan szimmetrikus, véges számú negatív sajátértékkel rendelkező mag, hogy az általa generált $k(Z, Y)$ pozitív definitív magnak megfelelő **A** operátort a (3.3) szerint fel lehet bontani. Legyen $F$ olyan, hogy $L^p \mapsto L^q$ és kielégíti a (3.4) feltételt, *akkor* a (3.1) legalább egy megoldással rendelkezik.

KRASZNOSZEL'SZKIJ az $F$-re vonatkozó feltételeket tovább szigorítva több mint egy, de *véges* számú megoldást is bizonyítani tudott.

KRASZNOSZEL'SZKIJ $F$-re vonatkozó feltételei számunkra nem konstruktívak.

2. TYIHONOV [29] a (3.4.) feltételrendszert némileg enyhítette (itt is egyetlen egyenletről van szó):

$$K(z, X(s)) := \int_0^b K(z, s, X(s))ds = u(z) \tag{3.5}$$

feladat korrekt kitűzésének feltételeit vizsgálta. Megmutatta azokat a feltételeket, amelyek mellett a (3.5) korrekt kitűzésű.

Ezek a feltételek a következők:

A) $K(z, X_1(s)) \neq K(z, X_2(s))$, ha $X_1(s) \not\equiv X_2(s)$,

B) $K(z, X(s))$ folytonos, $X(s) \in C$; $C \mapsto L$,

C) $K'_X$; $K''_{XX}$ folytonos,

D) ha $\int_a^b K'_X(z, s, X(s)W(s))ds = 0$, akkor $W(s) = 0$ azaz a lineáris egyenletnek csak triviális megoldása van.

Esetünkben a C. feltétel gyakran nem teljesül.

3. BAKER [4] a (3.4) egzisztencia feltételrendszerét némileg enyhítette. Állításait a következő módon fogalmazta meg:

Legyen $K(Z, Y)$ valós, szimmetrikus, folytonos és az összes sajátértékei negatívak, és

$$|F(Y, X)| \leqq C_1|X| + C_2,$$

ahol

$$C_1 \quad \text{és} \quad C_2 > 0 \quad \text{és} \quad 1/C_1 \geqq \varrho(K)$$

és $F(Y, X)$ rögzített $X$-nél nem csökken, *akkor* a (3.1.)-nek legalább egy folytonos megoldása van.

4. SPREKELS [27] munkájában — szempontunkból — új, hogy nem követeli meg a $K$ szimmetrikus voltát. Egy ún. *szuperlineáris Hammerstein egyenlettel* dolgozik:

$$\Pi(X(Z)) := \int_I K(Z, Y)F(Y, X(Y)) \cdot X(Y)^{1+\alpha} dY = X(Z)$$

$$z \in I; \quad I = [0, 1]; \quad \alpha > 0; \quad \dot{I} = (0, 1).$$

Feltételei:

$K(Z, Y)$ — folytonos $I \times I$-ban és pozitív $\dot{I} \times \dot{I}$-ban,
$F(Y, X)$ — folytonos $I \times R_+$-ban és nem csökken $X$-ben,
ha $X > 0$, akkor $F(Z, X) > 0$ majdnem minden $Z \in I$-re.

Ekkor a *Krasznoszel'szkij-tétel* segítségével az egzisztencia belátható. Egy ún. kúp iterációt használ a megoldáshoz. Legyen

$$\text{kúp} = \{X \in C_+ : aX(\bar{Z}) \leqq X \leqq b \cdot X(\bar{Z})\},$$

ahol $a, b$ — korlátos és mérhető függvények és $\bar{Z} \in T$ megfelelően van kiválasztva a $C(I)$ *Banach-téren* a természetes kúp, és

$$\bar{X} \in \prod_{n=1}^{\infty} \text{kúp}_n$$

elemet kell numerikusan megkeresni.

SPREKELS egy teljesen új iterációs sémát implementált a számítógépre, de ennek megvalósítása igen bonyolult volt, valamint az $a, b$ — megválasztása feladatfüggő.

5. DOLPH [9] összefoglaló cikkében egy olyan feltételrendszert ír le, amelyről nem lehetett triviálisan eldönteni, hogy az általunk használt integráloperátoroknál használható lesz-e. Így DOLPH unicitás tételét részletesen fogjuk ismertetni, majd egy két készülék-egységből álló összetett rendszer példáján megmutatjuk, hogy esetünkben a tétel feltételrendszere általában nem teljesül.

A következő meghatározásokat vezessük be:

Egy $G$: $H \mapsto H$ operátort akkor nevezünk monotonnak, ha az összes $x_1, x_2 \in H$-ra

$$\langle x_1 - x_2;\ Gx_1 - Gx_2 \rangle \geqq 0$$

*Szigorúan* monoton, ha $x_1 \neq x_2$-nél a ($\geqq$) jelet ($>$) jelre lehet a fenti egyenletben cserélni.

*Erősen* monoton, ha a fenti egyenlőtlenség jobb oldalára beírhatjuk a $c\|x_1 - x_2\|^2$ ($c$ — pozitív konstans) kifejezést. Ekkor *Dolph-tétele* szerint az $y + KFy = u$ egyenlet unicitásának elégséges feltétele, hogy $K$ és $F$ monoton legyen és *vagy* $K$ *vagy* $F$ szigorúan monoton legyen.

Tehát itt sem a $K$ szimmetrikus voltát, sem a véges számú sajátértékeket, sem egyéb — számunkra — nem konstruktív feltételt nem kell ellenőrizni.

A VIRÁG által kidolgozott magfüggvényeknél ellenőrizzük egy egyszerű példán a tétel feltételrendszerét:

Legyen két sorba kapcsolt készülék egységünk, melyek magfüggvényei rendre

$$K_1, K_2.$$

A sorba kapcsolás miatt az eredő magfüggvény [31]

$$\begin{bmatrix} K_1 & 0 \\ 0 & K_2 \end{bmatrix}$$

alakú.

A forrásfüggvény legyen az egyensúlyi (csillaggal jelölt) és az aktuális intenzív érték különbsége. A két készülék egység forrásfüggvénye rendre:

$$F_1 = c_1 - c_1^*,$$

$$F_2 = c_2 - c_2^*.$$

Végezzük el a *Dolph-tétel* ellenőrzését először $K$-ra. Legyen

$$x_1 = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} \quad \text{tetszőleges és} \quad x_2 = \begin{bmatrix} c_1^* \\ c_2^* \end{bmatrix},$$

$$Gx_1 - Gx_2 = \begin{bmatrix} K_1 & 0 \\ 0 & K_2 \end{bmatrix} \cdot \begin{bmatrix} c_1 - c_1^* \\ c_2 - c_2^* \end{bmatrix}.$$

A skalárszorzat:

$$\langle x_1 - x_2;\ Gx_1 - Gx_2 \rangle = [c_1 - c_1^*, c_2 - c_2^*] \begin{bmatrix} K_1 & 0 \\ 0 & K_2 \end{bmatrix} \begin{bmatrix} c_1 - c_1^* \\ c_2 - c_2^* \end{bmatrix} =$$

$$= \int_0^L (c_1(y) - c_1^*(y)) \cdot K_1(z, y) \cdot (c_1(y) - c_1^*(y)) dy +$$

$$+ \int_0^L (c_2(y) - c_2^*(y)) \cdot K_2(z, y) \cdot (c_2(y) - c_2^*(y)) dy.$$

Ez a kifejezés általános esetben (2.4) monotonitást biztosítja. Próbáljuk ki a forrásfüggvényre a DOLPH [9] feltételrendszerét, ha a forrás a legegyszerűbb alakú:

$$Fx_1 - Fx_2 = \begin{bmatrix} (c_1 - c_1^*) - (c_2 - c_2^*) \\ -(c_1 - c_1^*) + (c_2 - c_2^*) \end{bmatrix}$$

$$\langle x_1 - x_2;\ Fx_1 - Fx_2 \rangle = \int_0^L \{[c_1(y) - c_1^*(y)] - [c_2(y) - c_2^*(y)]\}^2 dy.$$

A monotonitás itt is teljesül, de a szigorú monotonitás feltétele már nem.

Az áttekintett tételek közül a 3. és 4. feltételrendszere esetünkben nem teljesül, mivel $K(Z, Y)$ — nem szimmetrikus és az egyenletünk nem szuperlineáris. Az 1. és 2. pont alatti tételek feltételrendszere számunkra nem konstruktív, mivel a gyakorlati feladatoknál $K$ és — $F$ alakja is változik.

A fenti okfejtésből természetesen nem az következik, hogy a VIRÁG által kidolgozott magfüggvények esetén semmilyen feladatnak sincs megoldása, mindössze arról lehet szó, hogy a megoldást még a legegyszerűbb esetben sem tudjuk előre garantálni.

*A Hammerstein típusú integrálegyenletek numerikus megoldása*

Célszerűnek tartottuk egy olyan módszer megkeresését, illetve kidolgozását, amelynek segítségével a *Hammerstein egyenlettípus* általánosan megoldható. Itt csak a stacionárius állapotra vonatkozó egyenletek megoldási módszereivel foglalkozunk.

A *Hammerstein típusú integrálegyenletek* a nemlineáris integrálegyenletek osztályába tartoznak, melyek megoldhatóságával, illetve megoldási módszereivel az ötvenes évektől foglalkoznak a matematikusok.

A *Hammerstein típusú integrálegyenletek* numerikus megoldásával viszonylag kevés szerző foglalkozott. Mint GOLDBERG [12] az integrálegyenletek megoldási módszeréről írt könyvében leszögezi: „A lineáris egyenletekkel ellentétben, minden egyes nemlinearitási típust külön-külön kell vizsgálni."

Rámutatott arra, hogy a *Hammerstein típusú egyenletek* a *Fredholm típusúaknak* lineáris megfelelői, melyeket itárációval lehet megoldani. Ha az $F$ forrásfüggvény elég sima, akkor a *Newton-módszert* ajánlja, egyébként ennek valamilyen módosított változatát.

Az integrálegyenletek numerikus megoldásával néhány monográfia is foglalkozik: DOLPH [9], DELVES [8]. Ezek főleg a *Fredholm típusú lineáris egyenletek* megoldásával, illetve a szinguláris magfüggvényekből eredő problémákkal foglalkoznak. A nemlineáris operátoregyenletek numerikus megoldása a tárgya RALL [23] könyvének. A *Hammerstein típusú feladat* kezelésére ő is a *Newton-módszert* írja le.

Függvények minimum helyeinek megkeresésére gyakran alkalmazzák a *Newton-módszert*. A *Newton-módszer* elmélete az $F(x)$ $n$-változós függvény minimalizálására jól ismert.

Ha $X^{(k)}$ az $X$ egzakt megoldás közelítése a $k$-adik iterációs lépésben, akkor definiáljuk úgy egy $Q^{(k)}(X)$ függvényt, hogy $F$ *Taylor-sorát* a kvadratikus tagig folytatjuk:

$$Q^{(k)}(x) = F^{(k)} + g^{(k)}(x - x^{(k)}) + \frac{1}{2}(x - x^{(k)})^T G^{(k)} \cdot (x - x^{(k)}),$$

ahol

$g(x) = \nabla \cdot F(x)$ — gradiens vektor,

$G(X) = \nabla^2 F(X) = \nabla \cdot \nabla \cdot F$ — *Hesse mátrix*,

felső index $T$ — transzponált vektor

Bizonyítható, hogy ha a *Hesse-mátrix* pozitív definit, akkor a minimumkeresési feladatnak egyetlen megoldása van. A *Newton-módszer* módosított változatainak egy része a *Hesse-mátrix* pozitív definitté tételére irányul [14]. A gradiens vektor segítségével számítható ki a szükséges korrekció iránya, ez több dimenziós esetben a *Jacobi-mátrix*. Mivel a *Jacobi-mátrix* számítása időigényes és numerikus hibáktól terhelt, ezért több módszer a *Jacobi-mátrix* helyettesítésével, illetve számítási stratégiájával foglalkozik [12, 14]. Sajnos — az általános tapasztalat szerint — a *Newton-iteráció* és annak módosított változatai is gyakran divergálnak, melyek oka, hogy bár a konvergencia sebesség gyors (másodrendű), de a konvergencia sugara gyakran a gyakorlat számára túlságosan kicsi [12], így ezt a módszert eredeti formájában szinte sehol sem használják.

DELVES [8] foglalta össze az integrálegyenletek numerikus megoldására készült algoritmusokat. Összefoglaló cikkében a C. N. R. S. számára írt, *Newton-módszeren* alapuló, egyetlen *Hammerstein-típusú integrálegyenletet* megoldó programra utal. A *Hammerstein-típusú integrálegyenlet-rendszert* megoldó programról nem tud.

A nemlineáris integrálegyenletek megoldásának egy másik útját kínálják a *variációs módszerek*, amelyek lényege a következő [6]: A megoldandó egyenlethez egy olyan funkcionált rendelünk hozzá, amelynek a szélsőérték helye megegyezik az egyenlet megoldásával. A funkcionál képzésére és a szélsőérték hatékony keresésére kész módszert nem lehet adni, mivel az a feladat jellegétől függ.

A harmadik eljárás az *invariáns beágyazás*, amelynek vegyipari szimulációs alkalmazásáról legutóbb SALGOVIČ [26] írt. Ezt akkor ajánlatos használni, ha a feladat kezdeti becslésére nem tudunk reális értéket kigondolni, azaz, ha a választott start értékek valószínűleg a konvergencia rádiuszon kívül lesznek. Ekkor a megoldandó egyenletet az egyenletek egyparaméteres seregébe ágyazzuk. Azaz előállítunk egy olyan paraméteres egyenletet (egyenletsereget), amelynek a paraméter egy rögzített értéke mellett ismerjük a megoldását, a paraméter egy másik rögzített értéke mellett az egyenlet az általunk megoldani kívánt egyenletbe megy át.

A megoldás paraméter-függését a közönséges differenciálegyenlet-rendszerek megoldásánál alkalmazott módszerekkel tudjuk meghatározni, és így az eredeti egyenlet megoldását a paraméterek fokozatos változtatásával kapjuk meg. A beágyazásos módszer számítási időigénye általában nagyságrenddel nagyobb, mint a *Newton-*, vagy a *variációs módszereké*, viszont numerikus stabilitása olyan nagy, hogy konvergenciája lényegében független a kezdeti becslés jóságától, ha a beágyazást szerencsésen végeztük el.

A numerikus stabilitás biztosítása a transzport egyenletek *peremérték* feladatának megoldásánál jelentős problémát okoz, még akkor is, ha ezek lineáris differenciálegyenletekké redukálhatók.

Jó példát mutatott erre BAHVALOV [6], aki a legegyszerűbb: $y'' - p \cdot y = 0$, $p = \text{const} > 0$ differenciálegyenlet peremérték feladatánál jelentkező számítási hibá-

ról megmutatta, hogy az arányos a következő összefüggéssel:

$$\frac{1}{h^2}\exp\left[\sqrt{p}\cdot X(1+O(\sigma))\right],$$

ahol $h$ — a lépésköz,

$X$ — integrálási hossz, $X=n\cdot h$,

$\sigma=\sqrt{p}\cdot X/n$.

Az intervallum végén tehát a lépésköztől függetlenül is katasztrofálisan nagy lehet a számítási hiba.

ABRAMOV [1] és szerzőtársai a közönséges lineáris differenciálegyenletek peremérték feladatának megoldását stabilabbá tevő eljárásokat foglalják össze. Ezeknek a módszereknek az az alapgondolatuk, hogy a peremérték feladatot bizonyos segédfüggvényekkel több kezdeti érték feladatra redukálják. Olyan segédfüggvények választását javasolják [1], hogy a peremérték feladat (tehát a teljes feladat) megoldása ne járjon nagy pontosságvesztéssel, azaz kíséreljünk meg a lehetőség szerint sima viselkedésű segédfüggvényeket használni, illetve a rájuk vonatkoztatott feladatot megoldani.

Az eddigi fejtegetések tisztán matematikai szempontok voltak. KUBIČEK [19] egy gyakorlati példán, a nem izotermikus, nem adiabatikus csőreaktor szimulációs modelljénél numerikusan kiszámolta, hogy bizonyos *Damköhler-szám* tartományban 5 stacionárius megoldás is van, tehát az unicitás valóban nem teljesül.

AMUNDSON, KUBIČEK és mások [18, 24] tárgyalták azokat a nehézségeket, melyek az $X(z)$ eloszlásgörbék (megoldások) igen meredek lefutásából erednek. A konkrét feladattól függően a differenciálegyenlet-rendszert vagy az egyik vagy a másik peremtől tartják célszerűnek integrálni, így közelítve a keresett peremekhez. Ha a differenciálegyenlet-rendszert a megfelelő peremfeltételekkel integrálegyenletekké tudjuk átalakítani, akkor ezzel a peremérték feladattal kapcsolatos nehézségeket többnyire sikerül megszüntetnünk. De, mint RALL [22] rámutatott, a többdimenziós feladatoknál a magfüggvény szingularitása problémát okozhat.

RAMKRISHNA és AMUNDSON [23] a stacionárius energia- és tömegátadási folyamatokat tanulmányozva mutattak rá, hogy elliptikus differenciálegyenletek bizonyos speciális peremfeltételek mellett szinguláris integrálegyenletekké írhatók át, más egyszerűbb peremek mellett pedig *Fredholm-típusú egyenletekké* redukálhatók. Ezeknek az egyenleteknek a megoldása esetenként egyszerűbb, a peremérték feladat speciális jellegétől független, általános módszerrel történhet.

A legégetőbb probléma, mint L. B. RALL megjegyezte [22]: „A nemlineáris integrálegyenletek megoldásánál a közelítő módszerek alkalmazását a nemlineáris egyenletrendszereket megoldó hatékony algoritmusok hiánya akadályozza".

*Összefoglalva* megállapíthatjuk, hogy a (2.2) egyenlet a (2.3—2.6) feltételeknek eleget tevő magfüggvényekkel az eddig ismert egzisztencia és unicitás feltételek egyikének sem tesz eleget.

Többen rámutattak arra, hogy bizonyos típusú peremérték feladatokat előnyösen lehet kezelni integrálegyenletként. Az általunk használt *Hammerstein-típusú integrálegyenlet-rendszert* megoldó algoritmust és programot tudomásunk szerint eddig még nem publikáltak.

Az előzőekben felsorolt problémák miatt egy új, a *Newton-iterációnál* numerikusan stabilabb, a feladathoz illeszkedő algoritmust dolgoztunk ki a *Hammerstein-típusú integrálegyenlet* megoldására.

## 4. Új numerikus módszer a Hammerstein-típusú integrálegyenlet-rendszer megoldására

Az ismeretlen $X$ vektor $j$-edik komponensére vonatkozó integrálegyenlet általános alakja az egyenlet részletezésével

(4.1)
$$X_j(z) = X_{0j} - \int_0^L K_j(z, y) \cdot F_j(y, X_1(y), X_2(y), \ldots, X_J(y)) dy, \quad j = 1, 2, 3, \ldots, J,$$

ahol $X_j(\,.\,)$ a keresett eloszlásfüggvény és

$$K_j(z, y) := B_j(z) \cdot K_j^*(z, y).$$

Az írásmód egyszerűsítésére vezessük be a következő jelölést:

(4.2) $$F_j(y, X(y)) := F_j(y, X_1(y), X_2(y), \ldots, X_J(y)), \quad j = 1, 2, \ldots, J.$$

A (4.1) egyenletben szereplő integrálást numerikusan a téglalapszabály szerint végezzük el. VIRÁG [30] a hibabecslések kapcsán megmutatta, hogy ez ekvivalens azzal, amikor az eredeti magfüggvény helyett egy alkalmas „lépcsős" magfüggvénnyel, azaz mérnöki szemlélettel egy módosított keveredési modellel dolgozunk. Az ezáltal okozott hibára becslést tudott adni. Az $X_j$ eloszlásfüggvényt a $z_n$ pontokban definiált ($n = 1, 2, \ldots N$) lépcsős függvények sorozatával közelítjük, amelyet iterációs módszerrel határozunk meg.

Az iteráció $m$-edik lépésében legyen a függvény: $X_j^m(\,.\,)$. A (4.1) egyenlet jobb és bal oldalának különbségét képezve egy $(z_n, X^m) \to \Delta(z_n, X^m)$ lépcsős függvényt kapunk,

(4.3) $$\Delta_j(z_n, X^m) = X_j^m(z_n) - X_{0j} + \int_0^L K_j(z_n, y) \cdot F_j(Y, X^m(y)) dy,$$

képezzük a $\Delta(\,.\,, X)$ függvény normáját, ahol $\Delta := (\Delta_1, \Delta_2, \ldots, \Delta_J)$

(4.4) $$I(X^m) := \|\Delta(.,X)\|^2 = \langle \Delta(.,X);\ \Delta(.,X) \rangle,$$

illetve diszkretizált formában

(4.5) $$I(X^m) = \sum_{j=1}^{J} \sum_{n=1}^{N} [\Delta(z_n, X^m)]^2, \quad \text{ha} \quad N \to \infty.$$

Az $I(X^m)$ funkcionál minimumhelye megegyezik a (4.1) egyenlet megoldásával. Ennek meghatározásához az $I$ funkcionál iránymenti deriváltját kell képezni.

(4.6) $$\frac{\partial I(X)}{\partial U} = \left.\frac{\partial I(X + a \cdot U)}{\partial a}\right|_{a=0} = 2\langle \Delta(.,X);\ U + \int_0^L K(.,y) \cdot \nabla \circ FU dy \rangle.$$

A (4.6) egyenletbeli skalárszorzat abban az $U$ irányban lesz maximális, azaz az $I$-funkcionál abban az $U$ irányban változik leggyorsabban, amelyben

(4.7) $$\Delta(.,X) = U + \int_0^L K(.,y) \cdot \nabla \circ FU dy.$$

Innen $U$-t kell meghatároznunk. A (4.7) másodfajú *Fredholm-típusú lineáris integrálegyenlet-rendszer* megoldását lineáris egyenletrendszer megoldására vezetjük vissza:

$$\Delta_j(z_n, X) = U_j(z_n) + \frac{L}{N} \cdot \sum_{i=1}^{N} K_j(z_n, y_i) \cdot \sum_{l=1}^{K} \frac{\partial F_j(y, X(y_i))}{\partial X_l} U_l(y_i). \tag{4.8}$$

A (4.8) egyenlet mátrix kifejtését az 1. ábrán láthatjuk.
A (4.8) egyenlet megoldásaként kapott $U$ függvény megmutatja az irányt, amely mentén az $X$-eloszlásfüggvényt módosítani kell ahhoz, hogy az $I$-funkcionál a leggyorsabban csökkenjen:

$$X^{m+1} = X^m + a \cdot U^m. \tag{4.9}$$

A módosítás mértékét, $a$-t úgy határozzuk meg, hogy az $I$-funkcionál az $U$ irányban minimális legyen. Az extrémum keresését egy egyszerűsített numerikus módszerrel végezzük.

Numerikus tapasztalataink (melyekről a cikk következő fejezeteiben lesz szó) és fizikai meggondolások alapján ([16]) az iterációt a bemeneti értékből számított egyensúlyi értéknek megfelelő konstans hőmérséklet és koncentráció eloszlásról indítjuk. A startnál a $[0, L]$ intervallumot csak néhány (3—6) pontra célszerű felosztani, így csökkentjük az egyenlet dimenzióját.

Az algoritmus főbb lépései (lásd 2. ábra):

*Első lépésben* a $z_n \to X_j^m(z_n)$ $(n=1, 2, \ldots N, j=1, 2, \ldots J)$ eloszlással meghatározzuk a (4.5) egyenletből $I(X_0^m)$-t.

*Második lépésben* az adott $F$ és $K$ és a számított $\Delta$ függvények birtokában a (4.8) egyenletet $U$-ra megoldjuk.

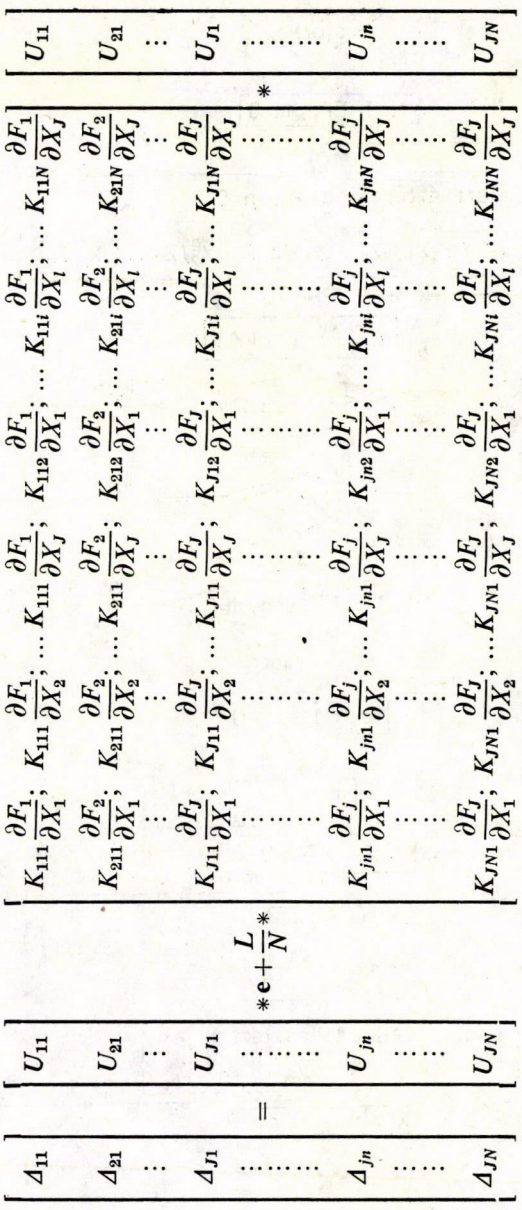
*Harmadik lépésben* a *Newton iterációhoz* hasonló gondolatmenet alapján $a$-t 1-nek választva a módosított $X^{m+1}$ eloszlásfüggvényt a (4.9) egyenlet alapján számítjuk.

*Negyedik lépésben* az $X^{m+1}$ birtokában a (4.5) egyenletből az $I(X^{m+1})$-et meghatározzuk. Ha $I(X^m) \leqq I(X^{m+1})$, akkor $a := a/2$ és a számítást a harmadik lépéstől folytatjuk, egyébként az egyes lépéstől addig, amíg $I(X) < \mathrm{EPS}/N$.

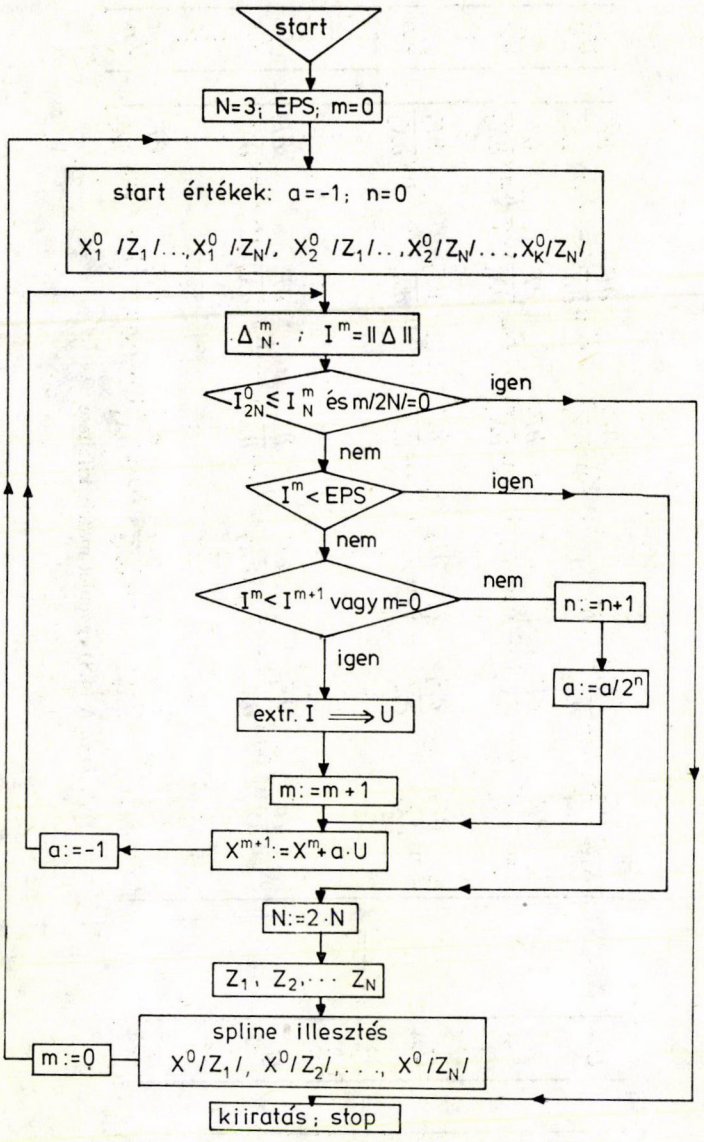
*Ötödik lépésben* $N$ diszkrét pontban, minden egyes komponensnél meghatározott diszkretizált eloszlásfüggvényre egy harmadrendű spline függvényt illesztünk, amely segítségével az új $N := 2 \cdot N$ számú pontban az $X$-függvényértékeket határozzuk meg.

Ha az új pontokban számított eloszlásfüggvényekre is igaz az $I(X) < \mathrm{EPS}/N$ feltétel, akkor a számítást befejezettnek tekintjük, egyébként a spline illesztéssel meghatározott függvényértékeket tekintve a start értékeknek, a számítást az első lépéstől folytatjuk.

Az algoritmust és a programot nem túl jelentős módosításokkal alkalmassá tettük több készülék-egységből álló tetszőleges hálózat szimulációjára is. A változtatás lényege abból ered, hogy a megoldó algoritmus szempontjából az áramok hálózata alapján egymás után megfelelő sorrendben „felfűzött" készülékek sorát egyetlen egységnek tekintjük, de a forrás- és magfüggvények számításánál lehetővé tettük, hogy minden egyes modellezési egységnek saját forrásfüggvénye és magfüggvénye lehessen.



*1. ábra.* A (4.8) egyenlet mátrix kifejtése




*2. ábra*

Hangsúlyozzuk, hogy mind a megoldandó egyenletrendszer mérete, mind az egyes fázisok áramlási viszonyaira vonatkozó információ, mind pedig a forrásfüggvény alakja csupán bemenő adata programunknak, a megoldó algoritmus ettől független.

## 5. Az általunk kidolgozott új algoritmus elvének összehasonlítása más numerikus eljárásokkal

Kihasználva azt, hogy az integrálegyenlet ekvivalens a megfelelő differenciálegyenlettel és annak peremfeltételeivel, az integrálegyenletes kezelésmód segítségével a peremek iterálásából eredő nehézségek elkerülhetők.

Algoritmusunkban minden egyes pont — függetlenül attól, hogy az az ismert bemeneti érték mellett van, vagy éppen a keresett kimeneti érték — egyenrangú!

A *Hammerstein-típusú integrálegyenletek* megoldására a *Newton-módszert* szokták alkalmazni [23]:

$$U^m(z) = X^{m+1}(z) - X^m(z), \tag{5.1}$$

$$\Delta^m(z) = U^m(z) + \int_0^L K(z, y) \cdot \nabla \circ F \cdot U^m. \tag{5.2}$$

A (4.7) és az (5.2) egyenletek formailag hasonlóak. Az (5.2) egyenletben $U^m$ a két iterációs lépés közötti különbséget jelenti, a (4.7) egyenletben pedig az $U$ az az irány, amely mentén $X$-eloszlást módosítva az $I$ funkcionál a leggyorsabban csökken. Innen látható, hogy ha (4.9)-ben $a=1$ választjuk, akkor módszerünk formailag a *Newton-módszerbe* megy át. Tapasztalataink szerint az $a=1$ értékkel azonban az iteráció gyakran divergál. Az ún. beágyazásos módszerek is hasonló iteráción alapulnak, ott azonban $a$-t „elég kicsinek" kell választanunk. Így általában ugyan biztosított a módszer konvergenciája, de a lépések száma, azaz az időigény jelentősen megnövekszik. Tapasztalataink szerint $a$-nak az általunk javasolt megválasztása a két módszer közötti ideális kompromisszumot jelenti.

A gradiens módszernél is a (4.6) egyenletből indulnak ki. Az $U$-irányt azonban nem a skalárszorzat jobb oldali tagjának inverzének segítségével, hanem a numerikusan lényegesen egyszerűbben meghatározható adjungált segítségével számítják. Tudomásunk szerint a módszer lineáris, elliptikus egyenleteknél nagyon jól alkalmazható, tapasztalataink szerint azonban az általunk vizsgált feladatoknál gyakran divergál. Ez nyilvánvalóan a nemlinearitás következménye.

## 6. Az integrálegyenletek numerikus megoldásán alapuló algoritmus és a program hatékonyságára vonatkozó tapasztalatok

Mikor valamely nemlineáris feladat megoldására módszert választunk, legalább három tényezőt kell mérlegelnünk: ezek a konvergencia biztosítása, a számítási idő, és a memória igény. A memória igény kis gépeknél, vagy igen nagy méretű feladatoknál meghatározó lehet. Esetünkben, amikor a hő- és anyagátadási folyamatokat olyan részletességgel szimuláljuk, hogy az intenzív jellemzők készüléken belüli eloszlását számítjuk, akkor a numerikus megoldás konvergenciájának és a reális számítási időnek a biztosítása jelent problémát.

Módszerünk memória igényét ($M$) és a számítási időt ($T$) a következő összefüggésekkel adhatjuk meg (mivel a számítási időt a mátrix invertálása limitálja):

$$M = (E.U.N)^2 \tag{6.1}$$

$$T = k \cdot M^{1,5} \cdot I,$$

ahol

$E$ = egyenletek száma = a számított intenzívek száma,
$U$ = a modellezési egységek száma,
$N$ = osztó pontok száma = a készüléken belül az eloszlásfüggvényeket hány pontban számítjuk,
$I$ = iterációk száma,
$k$ = a géptől függő faktor (esetünkben $\sim 10^{-5}$ s).

Ha pl. készülékenként 20 pontban számítunk 8 eloszlásfüggvényt, akkor 1 Mbyteos memória területen kb. 40 készüléket számíthatunk, viszont az IBM 3031-es számítógépen ha két egyenletet ($E=2$), öt modellezési egységet ($U=5$) egységként tíz pontban számítunk ($N=10$), akkor $M=(2\times5\times10)^2$, és egyetlen iterációs lépés ideje kb. 10 másodperc, ha a maximális méretű feladatot oldjuk meg, akkor 50 óra nagyságrendű lesz. Ebből is következik, hogy a számítást nem a memória igény, hanem a gépidő, illetve a konvergencia megfelelő sebességének elérése limitálja.

A fejezet további részében kijelöljük azt az alkalmazási területet, ahol — véleményünk szerint — az általunk ismertetett integrálegyenletek numerikus megoldásán alapuló programot a hő- és anyagátadási folyamatok szimulációjánál sikerrel lehet használni.

Algoritmusunk és a ráépülő programrendszer más numerikus eljárásokkal történő összehasonlítását az nehezítette, hogy eddig (beleértve a SZTAKI IBM 3031 programkönyvtárát is), nem találtunk kész programot a másodrendű differenciálegyenlet-rendszer *Danckwerts-féle peremekkel* való megoldására. Saját tapasztalataink egyrészt a lényegesen egyszerűbb, keveredés nélküli (elsőrendű differenciálegyenletrendszer), másrészt az ideálisan kevert (algebrai egyenletrendszer) esetekre vonatkoznak.

A nemlineáris algebrai egyenletrendszerrel való összehasonlításnál tesztfüggvényeket használtunk, az elsőrendű differenciálegyenlet-rendszer peremérték feladatának megoldását pedig reális problémákon végeztük el.

A *nemlineáris algebrai egyenleteket* megoldó szokásos algoritmusokat PALOSCHI eredményeivel [21] hasonlítottuk össze, aki az irodalomban ismert teszt feladatokat használta.

PALOSCHI többek között a következő három nemlineáris egyenletrendszer numerikus megoldását végezte el:

BUS 1

$$x_i = 10 - \sum_{j=1}^{10} x_j \quad i = 1, 2, \ldots, 9,$$

$$x_{10} = +1/(x_1 x_2 x_3 x_4 x_5 x_6 x_7 x_8 x_9),$$

$$x_i^0 = 0{,}5.$$

BUS 3

*(Freudenstein és Roth függvény)*

$$x_1 = 13 - ((-x_2 + 5) - 2)x_3,$$

$$x_2 = 29 - x_1/((x_2 + 1)x_2 - 14),$$

a) $x_0 = [15, -2]$,

b) $x_0 = [-5, 0]$,

c) $x_0 = [-5, 3]$.

BUS 9

*(Rosenbrock függvény)*

$$x_2 = x_1^2/10$$

$$x_1 = 1$$

$$x_0 = [-12, 1].$$

A fenti három egyenletrendszert numerikusan oldottuk meg programunk segítségével, ennek eredményeiről az 1. táblázat tájékoztat, mely alapján azt a következtetést lehet levonni, hogy programunk kevesebb iterációval oldja meg a vizsgált teszt-feladatokat, mint a *Broyden és Barnes módszerek*. A PALOSCHI által közölt és az általunk tapasztalt számítási idők nagyjából megegyeztek. Pontos időadatokat itt azért nem közöltünk, mivel programunk az IBM 3031-es, PALOSCHI programja pedig a CDC 6500-as számítógépen futott.

A *Davidenko módszer,* hasonlóan a saját programunkhoz, a függvények parciális deriváltjainak analitikus formáját használja fel az iránykeresésre. A *Davidenko módszernél* a számításokat az IBM programkönyvtár segítségével végeztük el és így a számítási időket is össze tudtuk hasonlítani a saját módszerünkével. Megállapítottuk, hogy a *Davidenko módszerrel* az IBM programmal számított feladatok CPU ideje lényegében megegyezett programunk futási idejével (az eltérés néhány 10% volt).

A kedvező eredmények véleményünk szerint annak köszönhetők, hogy a számításhoz szükséges deriváltakat programunk először analitikus formában állítja elő, majd a derivált függvények helyettesítési értékeivel számol, hasonlóan a *Davidenko módszerhez*. A program memória igénye, amely $n^2$-tel arányos, ahol $n$ az egyenletek száma, a *Broyden és Barnes módszerekének* néhányszorosa (1,5—3).

A következő részben két, a vegyipari gyakorlatból származó szimulációs példán hasonlítjuk össze a differenciálegyenletek hagyományos, illetve az integrálegyenletek új típusú numerikus megoldásának tapasztalatait. Igyekeztünk olyan példákat kiválasztani, amelyek szimulációja a szokásos, differenciálegyenlet-rendszerek megoldásán alapuló algoritmusokkal komoly nehézséget jelent, azaz vagy a számítási idő túlsá-

1. TÁBLÁZAT

Az iterációk számának összehasonlítása a nemlineáris algebrai egyenletek megoldásánál (* I. R. PALOSCHI adatai)

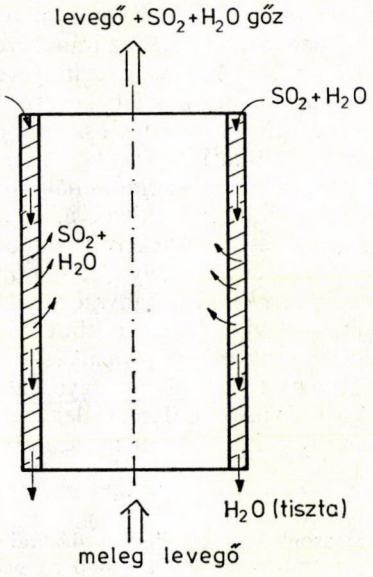
| Feladat | *Broyden** | *Barnes** | *Davidenko* | Saját |
|---|---|---|---|---|
| | Az iterációk száma | | | |
| BUS 1 | 25 | 25 | 10 | 12 |
| BUS 3 *a* | 65 | 19 | divergál | 6 |
| *b* | 303 | 50 | 10 | divergál |
| *c* | 14 | 13 | 8 | 5 |
| *d* | 56 | 16 | 6 | 6 |
| BUS 9 | 6 | 6 | 4 | 3 |

gosan nagy (órás nagyságrendű), vagy a konvergencia nem biztosított, és amelyek gyakorlati alkalmazhatóságához nem férhet kétség.

Ilyen feladatosztály a *nemlineáris differenciálegyenletek peremérték feladatai,* mivel nyilvánvaló, hogy a recirkulációmentes, két megegyező irányban áramló fázis közti hő- és anyagátadást leíró differenciálegyenlet-rendszer kezdeti érték feladatának megoldására a hagyományos módszer ajánlható, hiszen annak memóriaigénye a feladat méretével nem négyzetesen, hanem csak lineárisan nő.

CSAPÓ ZOLTÁN (Nehézvegyipari Kutató Intézet, Veszprém) programja az elsőrendű differenciálegyenlet-rendszer peremérték feladatát módosított *Newton-iterációval* oldja meg. Ez az eljárás olyan feladatok szimulációjára alkalmas, ahol az axiális visszakeveredés (effektív diffúzió) elhanyagolható. A következő feladatnál programját alkalmaztuk.

Első példában *esőfilmes készülékben* történő $SO_2$ deszorpcióját modelleztük úgy, hogy a meleg levegő hatására a deszorpcióval párhuzamosan jelentkező vízpárolgást is figyelembe vettük (3. ábra). A két fázisú, két komponensű ($SO_2$, $H_2O$) rendszerben a hő- és anyagátadási folyamatokat 6 egyenlet írta le, melyekben az egyensúlyi tényezők hőmérsékletfüggését az exponenciális alakú *Clapeyron-összefüggéssel* vettük figyelembe.




*3. ábra.* Elsőfilmes készülékben $SO_2$ deszorpciója és a $H_2O$ párolgása

Az anyagátadást leíró négy mérlegegyenlet:

$$V^F \varepsilon^F \frac{dc^F}{dz} = \Delta_i, \quad \text{ha} \quad i = 1, \quad \text{akkor} \quad i \quad \text{megfelel } SO_2\text{-nek},$$

$$V^G \varepsilon^G \frac{dc^G}{dz} = -\Delta_i, \quad \text{ha} \quad i = 2, \quad \text{akkor} \quad i \quad \text{megfelel } H_2O\text{-nak}.$$

A hajtóerő:

$$\Delta_i = \left(A_i \cdot \exp\left(B_i/T^G\right) \cdot c_i^F - c_i^G\right) \cdot \beta_i.$$

A hőátadást leíró két mérlegegyenlet:

$$V^G \varepsilon^G \varrho^G \mathbf{C}_p^G \cdot \frac{dT^G}{dz} = -\mathbf{C}_p^G (T^G - T^F) \cdot \Delta_1 - \alpha (T^G - T^F)$$

$$V^F \varepsilon^F \varrho^F \mathbf{C}_p^F \cdot \frac{dT^F}{dz} = \left(Q + \mathbf{C}_p^G (T^G - T^F)\right) \Delta_1 + \alpha (T^G - T^F),$$

ahol

$A$, $B$ — konstans,
$\mathbf{C}_p^G (T^G - T^F)$ — transzporthő,
$V$ — lineáris sebesség m/s,
$\mathbf{C}_p \cdot \varrho$ — fajhő sűrűség J/m³/K,
$\beta$ — anyagátadási tényező · fajlagos felülettel l/s,
$\alpha$ — hőátadási tényező · fajlagos felület (J/s/m³K),
$Q$ — latens hő J/kg,
$T$ — hőmérséklet K,
$c$ — koncentráció kg/m³,
$\varepsilon$ — térfogat kitöltési hányad;

felső index:

$G$ — gáz,
$F$ — folyadék;

alsó index:

$i=1$ $SO_2$,
$i=2$ $H_2O$,
$P$ — állandó nyomás.

A rögzített paraméterek értékei:

$$\varepsilon^F = 0{,}02, \quad \varepsilon^G = 0{,}98, \quad A_1 = 23{,}5, \quad A_2 = 31{,}3, \quad B_1 = 235,$$

$$B_2 = 4144, \quad \mathbf{C}_p^G = 0{,}238 \quad \mathbf{C}_p^G \varrho^G = 1400, \quad \mathbf{C}_p^F \varrho^F = 4{,}2 \cdot 10,$$

$$Q = 2 \cdot 10^6.$$

Mind az öt ($\alpha$, $\beta_1$, $\beta_2$, $V^G$, $V^F$) paramétert széles, de fizikailag reális intervallumban vizsgáltuk.

A változtatott paraméterek intervallumai:

$$50 \leqq \alpha \leqq 550, \quad 5 \cdot 10^{-4} \leqq \beta_1, \beta_2 \leqq 0{,}016, \quad 6 \leqq V^G/V^F \leqq 60.$$

Mivel több paraméter változását nehéz nyomonkövetni, ezért bevezetünk egy $0 \leqq r \leqq 1$ számot, melyet úgy határozunk meg, hogy az „$r$" növekedésével ebben a feladatban a numerikus megoldás nehézsége növekedjék. Jelölje az $i$-edik paraméter ($p_i$) maximális értékét $p_{i\max}$, minimális értékét $p_{i\min}$, ekkor az $r$-t a következő összefüggéssel de-

finiáljuk:

$$r = \frac{1}{I} \sum_{i=1}^{I} \frac{p_{im} - p_{i\min}}{p_{i\max} - p_{i\min}}$$

Az $r$ függvényében a számítási időt és az iterációs lépések számát a 2. táblázatban láthatjuk.

2. TÁBLÁZAT

Az iterációk száma és a számítási idő az $r$ függvényében

| $r$ | Iterációk száma | | Numerikus számítás ideje (sec) | |
|---|---|---|---|---|
| | differenciál- egyenlettel | integrál- | differenciál- egyenlettel | integrál- |
| 0 | 3 | 1 | 15 | 25 |
| 0,02 | 3 | 1 | 15 | 26 |
| 0,481 | 5 | 1 | 19 | 25 |
| 0,57 | 4 | 1 | 17 | 25 |
| 0,595 | 6 | 1 | 23 | 38 |
| 0,617[1] | 8 | 2 | 30 | 38 |
| 0,617[2] | divergál | 2 | — | 27 |
| 0,617[3] | divergál | 1 | — | 25 |
| 0,639 | divergál | 1 | — | 25 |
| 0,667 | divergál | 2 | — | 38 |
| 1,0 | divergál | 3 | — | 70 |

[1] $\beta_1 = 0{,}001$, $\beta_2 = 0{,}002$, [2] $\beta_1 = 0{,}002$, $\beta_2 = 0{,}001$, [3] $\beta_1 = 0{,}0015$, $\beta_2 = 0{,}0015$.

Az esőfilmes készülék számítást mind integrálegyenleteken, mind a differenciál egyenlet peremérték feladatát megoldó módszerrel 1% relatív hibával végeztük el. A differenciálegyenlet-rendszer peremérték feladatának megoldásánál az $r = 0{,}481$ futásnál kiválasztottuk a konvergencia szempontjából kedvező integrálási irányt, deriválási lépésközt és kezdeti becsléseket, mely értékekkel számítottuk a 2. táblázat adatait. A fenti kedvező becslések beállításához kb. 20 futásra (5 perc számítási idő) volt szükségünk. Az integrálegyenletekkel számolva a kezdeti becslés, mint egyetlen variálható paraméter a numerikus számítás ideje szempontjából lényegében közömbös volt.

A differenciálegyenleten alapuló megoldások számítási ideje, ha csak a sikeres futásokat tekintjük, kb. 40%-kal kisebb, mint az integrálegyenleteken alapuló számításoknál, az „$r$" az integrálegyenletre alapuló módszernél nagyobb volt (2. táblázat). Az „$r$", azaz a numerikus stabilitás növekedése az integrálegyenleteken alapuló megoldásoknál jelentős. A mintegy kétszeres növekedés azt jelenti, hogy pl. az egyik érzékeny paramétert, az anyagátadási tényező értékét huszszorosára lehetett növelni anélkül, hogy a számítás divergens lett volna.

Második példánkban egy állóréteges adszorpciós oszlopot modelleztünk. A modellel GYŐRI [15] doktori disszertációjában szereplő mérési adatokat kíséreltük meg interpretálni. Kísérleteinél GYŐRI állóréteges adszorberben szilikagélen vízgőzt kötött meg és az így kialakuló áttörési görbéket mérte.

Az állóréteges adszorberen, tehát szakaszos üzemű berendezésben végzett mérések eredményeit egy fiktív, a zónavándorlás sebességével megegyező nagyságú, de azzal ellentétes irányú rétegvándorlási sebesség bevezetésével áttranszformáltuk, így

időtől független, a megfelelő folyamatos üzemű berendezés stacionárius állapotát jellemző adatokat kaptunk. A zónavándorlás sebességét az együttes mérlegből számítottuk ki. Ezen átalakítások, illetve a következő modellek felállítása során alapfeltevésünk az volt, hogy a hő- és anyagátadási zónák kialakulnak, és sebességük megegyezik egymással.

Az izoterm és adiabatikus adszorpciós modellek által szolgáltatott koncentráció- és hőmérsékleteloszlás görbék, illetve a mért görbék között jelentős eltéréseket tapasztaltunk, ezért modelleinket [16] egyre finomítottuk és rendre kiszámítottuk az izoterm adiabatikus és a politróp modelleknek megfelelő görbéket.

A modellek felépítését korábbi munkánkban [16] már ismertettük.

Az adiabatikus és politróp modellek numerikus megoldása során problémát okozott az, hogy az anyagátadási tényező és a fajlagos felület szorzata nagyobb volt egynél. Ekkor a numerikus megoldás pontosságán belül teljes egyensúly áll be a fázisok között. Egyéb paraméter rendszer mellett az adiabatikus és politróp modelleknél 2—5 iterációs lépésen belül sikerült a megfelelő pontosságot elérni. Az izoterm modellnél 1—2 CPU s, az adiabatikus és politróp modellnél 3—5 CPU s, (5—8 Ft) volt egy-egy szimuláció ideje, ill. költsége az IBM 3031-es számítógépen.

A második példaként választott adszorpció szimulációját a differenciálegyenletrendszer segítségével reális paraméter rendszerek mellett nem tudtuk elvégezni. (Itt a paraméter rendszert számított és mért adatok illesztésével állítottuk be.) A kísérleti eredmények azt mutatták, hogy a hőmérséklet eloszlásnak maximuma van. Ezért egy kiegészítő programot kellett írnunk az integrálás iránya, a deriválás lépésköze és a kezdeti becslések egy hálón történő rendszeres változtatására annak reményében, hogy bizonyos start értékeknél az iteráció konvergálni fog, de a konvergenciát sem tudtuk biztosítani.

BALLA [5] a feladatot a szokásos *Newton-technikával* [2] linearizálta. A peremérték feladatot többféle faktorizációs módszerrel próbálta megoldani. A számításai eddig konvergencia problémák miatt nem vezettek eredményre.

PARLAGH hasonló algoritmuson alapuló programot készített, mint BALLA [3], de PARLAGH egy különlegesen pontos *Runge—Kutta formulát* alkalmazott, melyben csak a függvények növekményét veszi figyelembe. Így a használt duplapontos aritmetika kerekítési hibáit a négyszeres pontosság nagyságrendjébe tudta levinni. Az adszorpciós feladatot egy ESZR—32 számítógépen perces-tízperces futási idővel tudta megoldani.

Az általunk kidolgozott, az integrálegyenletek numerikus megoldásán alapuló módszert e feladatok osztályában akkor szükséges alkalmazni, mikor az eloszlásfüggvények alakja a differenciálegyenlet megoldása szempontjából nem kedvező.

Az integrálegyenleteken alapuló szimuláció nagy előnye, hogy segítségével nemcsak egyetlen készüléket, hanem készülékek tetszőleges hálózatát is szimulálni lehet. Módszerünket ebben az esetben is csak akkor ajánlhatjuk, ha a hálózatban levő visszacsatolások miatt (recirkuláció, ellenáram) a készülékről készülékre való számolás közvetlenül nem alkalmazható.

Módszerünkkel egyetlen algoritmussal, egyetlen programmal és így egyetlen input rendszerrel lehet olyan problémák együttesét szimulálni, amelyekben több, sokszor igen eltérő program felhasználására lenne szükség.

## 7. Összefoglalás

Dolgozatunkban az áramlástanilag lineáris, stacionárius működésű, egy dimenziós modellezési egységből álló vegyipari rendszer modellezésével és szimulációjával foglalkoztunk. A cikkben egy hatékény, új módszert írtunk le a néhány modellezési egységből álló vegyipari rendszer szimulációjához.

A téma irodalmának áttekintése alapján megállapítottuk, hogy a vegyipari rendszerek szimulációjánál használható *Hammerstein-típusú integrálegyenlet-rendszereknek* abban az osztályában, amelyeket VIRÁG definiált, egyetlen ismert unicitás, vagy egzisztencia tétel sem alkalmazható.

A *Hammerstein-típusú integrálegyenlet-rendszer* numerikus megoldására nem találtunk algoritmust, illetve programot. Célunk az volt, hogy az integrálegyenletek biztosította előnyöket megőrizve egy hatékonyan használható eszközt dolgozzunk ki a modellezési egységek modelljeinek numerikus megoldására.

A *Hammerstein-típusú nemlineáris integrálegyenlet-rendszert* a *Newton-* és *gradiens módszerek* előnyös ötvözetével oldottuk meg, melynek kidolgozásánál figyelembe vettük a Hammerstein egyenletek sajátosságait is.

Az algoritmus és a program működését több, már ismert megoldási módszerrel hasonlítottuk össze:

— Teszt feladatokként ismert nemlineáris algebrai egyenletrendszereket oldottunk meg módszerünkkel és a *Broyden, Bardes,* illetve a *Davidenko algoritmusokra* épülő programokkal. Az összehasonlítás eredményeként megállapítottuk, hogy programunk mind annak memória igényét tekintve, mind a futási időt összehasonlítva egyenértékű a fenti speciális programokkal.

— Egy filmkészülékben az $SO_2$ deszorpciójának és a víz párolgásának hőeffektusokat is tartalmazó modelljét kétféle módszerrel oldottuk meg. Megállapítottuk, hogy programunk konvergencia-rádiusza lényegesen meghaladja a differenciálegyenlet-rendszer peremérték feladatát Newton iterációval megoldó programét, de a paramétereknek abban a tartományában, ahol a hagyományos módszerrel is konvergens megoldást sikerült elérni a differenciálegyenlet-rendszer peremérték feladatát *módosított Newton-iterációval megoldó program* 20—80%-kal gyorsabban működött és memória igénye nagyságrenddel kisebb volt.

— A vízgőz szilikagélen történő adszorpcióját az adszorpcióshő figyelembevételével modelleztük. Ennek a példának kapcsán programunk hatékonyságát több, különböző, a differenciálegyenlet-rendszer peremérték feladatát megoldó algoritmussal hasonlítottuk össze. Az előző példánál említett, a peremérték feladatot *Newton* és *faktorizációs módszerrel* iteráló algoritmussal nem tudtunk konvergens megoldást biztosítani. PARLAGH integrálásnál jelentkező kerekítési hibákat nagyságrendekkel csökkentő programjával elfogadható számítási idővel oldotta meg a fenti feladatot.

### Köszönetnyilvánítás

Hálás köszönetemet fejezem ki DR. VIRÁG TIBORnak, témavezetőmnek, aki türelmes tanácsaival és útmutatásaival a munka elkészítése során mindig rendelkezésemre állt.

Köszönetemet fejezem ki DR. BLICKLE TIBORnak, aki lehetővé tette a cikk elkészítését.

Köszönetemet fejezem ki DR. BALLA KATALINnak, aki az elméleti részek megfogalmazásában volt segítségemre.

## IRODALOM


[1] ABRAMOV, A. A., BIRGER, E. S., KONYUKHOVA, N. B. and ULYANOVA, V. I., "On methods of numerical solution of boundary value problems for systems of linear ordinary differential equations", in: Colloquia Mathematica Societatis János Bolyai **22**, Numerical Methods, Keszthely (Hungary), 1977, 33—67.

[2] BENEDEK, P. és LÁSZLÓ, A., *A vegyészmérnöki tudomány alapjai* (Műszaki Könyvkiadó, Budapest, 1964).

[3] BENEDEK, P., „Vegyipari üzemek számítógépes szimulációja és tervezése", elemző tanulmány, OMFB, Budapest, 1977.

[4] BAKER, C. T. H., *The Numerical Treatment of Integral Equations* (Calderon Press, Oxford, 1977).

[5] BALLA, K., Személyes közlés, 1982.

[6] BAHVALOV, N. Sz., *A gépi matematika numerikus módszerei* (Műszaki Könyvkiadó, Budapest, 1977) 476—480.

[7] DELVES. L. M. and WALSH, J., *Numerical Solution of Integral Equations* (Calderon Press, Oxford, 1974).

[8] DELVES, L. M., ABD-ELAL, L. F. and HENDRY, J. A., "A set of modules for the solution of integral equations", *The Computer Journal* **24** (1981) 184—190.

[9] DOLPH, C. L. and MINTY, G. J., "On nonlinear integral equations of the Hammerstein type", in: Nonlinear Integral Equations, Ed. P. M. Anselone, Madison, The University of Wisconsin Press, 1964, 99—154.

[10] DONATI, G., MARINI, L. and MARZIANO, G. L., "A comprehensive approach to chemical engineering computational problems". Chem. Eng. Sci. **37** (1982) 1265—1282.

[11] EVANS, L. B., "Process flowsheeting: A state-of-art review", CHEMCOMP'82, Proceedings, Antwerpen, 1982, 1—12.

[12] GOLDBERG, M. A. (ed.) *Solution Methods for Integral Equations* (Plenum Press, New York and London, 1979).

[13] GOLDBERG, M. A., *A Survey of Numerical Methods for Integral Equations* 1—59.

[14] GOLDFELD, S. M., QUANDT, R. E. and TROTTER, H. F., *Econometrics* **34** (1966) 514.

[15] GYŐRI, I., Adiabatikus gáz adszorpció vizsgálata állóágyas adszorberben, Doktori disszertáció, Veszprém, 1976.

[16] HALÁSZ, G., BLICKLE, T., GYENIS, J., CSAPÓ, Z. and VIRÁG, T., "Unified simulation of sorption operations", in: Proceedings of the 5th International Congress in Scandinavia on Chemical Engineering, Copenhagen, 1980, 263—277.

[17] KAFAROV, V. V. and DOROHOV, I. N., *Sistemnij analiz processov himicheszkoj technologii* (Izdatyelsztvo Nauka, Moszkva, 1976).

[18] KUBIČEK, M., HOFMANN, H. and HLAVÁČEK, V., "Nonisothermal nonadiabatic tubular reactor, One dimensional model-detailed analysis", *Chemical Engineering Science* **34** (1979) 593—600.

[19] KRASZNOSZEL'SZKIJ, M. A., *Topologicseszkije metodi v teorii nelinejnih integral'nih uravnyényij* (Gosz. Izd. Tehnyiko-Teoreticseszkoj Literaturi, Moszkva, 1956).

[20] MARTIN, R. H., *Nonlinear Operators and Differential Equations in Banach Spaces* (John Wiley and Sons, New York, London, Sidney, Toronto, 1976).

[21] PALOSCHI, I. R., "A comparative study of algorithms of solving set of nonlinear equations", Report of Imperial College, London, 1980.

[22] RALL, B. L., *Computational Solution of Nonlinear Operational Equations* (John Wiley and Sons, New York, 1969).

[23] RAMKRISHNA, D. and AMUNDSON, N. R., "Boundary value problems in transport with mixed or oblique derivative boundary conditions I., Formulation of equivalent integral equations", *Chem. Eng. Sci.* **34** (1979) 301—308.

[24] RAYMON, R. L. and AMUNDSON, N. R., "Some observation on tubular reactor stability", *Canadian Journal on Chemical Engineering* **42** (1964) 173—177.

[25] RIDDEL, I. J. and DELVES, L. M., "The comparison of routines for solving Fredholm integral equations of the second kind", *The Computer Journal* **23** 274—285.

[26] SALGOVIČ, A., HLAVAČEK, V. and ILAVSKY, J. "Global simulation of countercurrent separation via one-parameter imbedding techniques", *Chem. Eng. Sci.* **36** (1981) 1599—1605.

[27] SPREKELS, J., "Finite dimensional cone iteration techniques for superlinear Hammerstein equations", *Num. Func. Anal. and Opt.* **1** (1979) 289—314.

[28] TARJÁN, K., GYENIS, J. és FRIEDLER, F., „Összetett vegyipari rendszerek műveleti egységei számítási sorrendjének meghatározása", *Magyar Kémikusok Lapja* **35** (1980) 490—494.

[29] TYIHONOV, A. N., "O resenii nelinejnih integral'nih uravnenij pervogo roda", *Dokladi Akademii Nauk SZSZSZR* **156** (1964) 1296—1299.
[30] VIRÁG, T., Az áramlástanilag lineáris vegyipari berendezések matematikai modelljeinek egységes kezelése, Kandidátusi értekezés, Budapest, 1979.

(Beérkezett: 1983. február 17.)

(Átdolgozva beérkezett: 1983. november 23.)

HALÁSZ GÁBOR
MTA MŰSZAKI KÉMIAI KUTATÓ INTÉZET
8200 VESZPRÉM, SCHÖNHERZ Z. U. 2.

## A NEW NUMERICAL METHOD FOR SIMULATION OF HYDRODYNAMICALLY LINEAR CHEMICAL EQUIPMENTS

G. HALÁSZ

The transport equations of heat and mass transfer processes may be formulated as *Hammerstein type integral equations*. The literature on the subject of uniqueness and existence conditions of Hammerstein type equations are summarized. A new, numerically stable algorithm has been elaborated to solve the *integral equation system of Hammerstein type*. This algorithm is compared with other numerical methods. Test problems from the literature and two examples from chemical industrial practice were used for the comparison.

*Alkalmazott Matematikai Lapok* **10** (1984) 149—172

# A HIDROSZTATIKUS CSŐVEZETÉKEK JELÁTVITELÉNEK PARCIÁLIS INTEGRO-DIFFERENCIÁLEGYENLET-RENDSZERÉRŐL

FÉNYES TAMÁS Budapest

HARKAY GÁBOR Budapest

A dolgozat módszert mutat be a hidrosztatikus csővezetékek dinamikai vizsgálatára. A súrlódásos folyadék (hidraulika olaj) lamináris áramlásából kiindulva, a csővezetéket rugalmasnak tekintve olyan általános egyenleteket ismertet a cikk, melyek alapján a csővezetékekben történő jelátvitel problémája elvileg teljes általánosságban megoldható, úgy a lökésszerű gerjesztések hatására fellépő tranziens folyamatok, mind a periódikus (színuszos) gerjesztések hatására fellépő állandósult állapotok esetén.

## 1. Bevezetés

A hidrosztatikus hajtás és irányítástechnikának (olajhidraulikának) az elmúlt 15 évben tapasztalt óriási fejlődése mindinkább szükségessé teszi a hidraulikus körfolyamatok statikus méretezésén és elemkiválasztásán túl az egyes elemek egymáshoz és a körfolyamhoz való dinamikus illesztésének a figyelembevételét is. Az olajhidraulikus rendszerekben alkalmazott csővezeték feladata, hogy az energiaközvetítő folyadékot az adott helyre (elembe, energiaátalakítóba) vezesse. A csővezetéknek biztonsággal el kell viselni a rendszerben fellépő dinamikus lengéseket, lökéseket, melyek nagysága sokszor többszöröse az üzemi nyomásnak.

A hidraulikus elemeket gyártó és rendszereket készítő üzemek tervezési segédletei alapján könnyen ki lehet választani az elemeket statikus jellemzők és névleges méretek alapján. Általában a gyakorlat azt mutatja, hogy a statikus méretezésen túl — a tervezésnél — nem nagy figyelmet szentelnek a csővezeték dinamikai hatásainak a figyelembevételére. A csővezetéket az energiaközvetítő folyadékkal együtt kell vizsgálni, mivel a hidraulikus körfolyamban együtt vannak jelen, hatásuk közös.

Színuszos gerjesztés esetén a gerjesztőfrekvenciának és az áramló folyadék viszkozitásának elsődleges hatása van a dinamikai tulajdonságokra. A nyomáslengés amplitúdója igen nagy lehet, ha a gerjesztő frekvencia közel esik, vagy megegyezik a csővezeték valamelyik rezonancia frekvenciájával. Ezért igen fontos a rezonancia frekvenciák ismerete, ezek azonban csak egy pontosan definiált általános érvényű matematikai modellből határozhatók meg.

A hazai és külföldi hidrosztatikus szakirodalomban több dolgozat foglalkozik a csővezetékek dinamikai problémáival, de ezekre jellemző, hogy csak speciális esetekben, meghatározott feltételek mellett (pl. meghatározott csővezetékhossz és gerjesztő frekvencia) érvényesek.

A szakirodalomból ki kell emelnünk LALLEMENT [7] dolgozatát, mely a hidrotatikus csővezetékek dinamikai vizsgálatát teljes általánosságban ismerteti. A szerző az alkalmazott matematikai analízis közlését mellőzi, tárgyalásmódja nem korrekt, eredményei a gyakorlat számára csak korlátozottan használhatók.

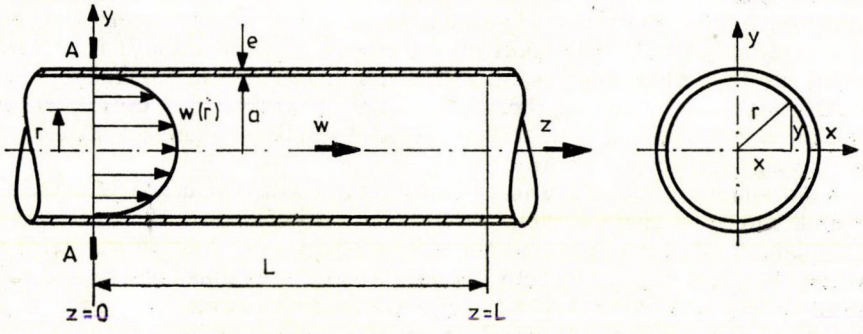
Cikkünkben a folyadék mechanikájának alapegyenleteiből kiindulva levezetjük a hidraulikus csővezeték jelátvitelének alapegyenleteit, majd ezek vizsgálatában a *Laplace-transzformációt* alkalmazva felírjuk az operátortartományban a csővezeték jelátvitelének legáltalánosabb összefüggéseit. Ezek alapján a gyakorlatilag legfontosabb tranziens folyamatokhoz tartozó időfüggvényeket *inverz Laplace-transzformációval* nyerjük.

Ezt követően a színuszos gerjesztések hatására fellépő állandósult állapot vizsgálatával foglalkozunk és meghatározzuk a csővezeték rezonanciagörbéit.

## 2. A hidrosztatikus csővezeték jelátviteli alap-egyenleteinek levezetése

Vizsgálataink alapjául tekintsük az alábbi 1. ábra koordinátarendszerében elhelyezett vízszintes csővezetéket.

Áramoljék a $p_0$ statikus nyomású folyadék (hidraulika olaj) kis sebességgel stacionáriusan az ábrán feltüntetett csővezetékben. Tegyük fel, hogy a csővezeték elején valamilyen zavarást alkalmazunk, pl. a csővezetéket az elején ($z=0$) nulla idő alatt elzárjuk. Ekkor a csővezeték $A-A$ keresztmetszetében megváltozik a sebesség és így a keresztmetszetben uralkodó nyomás is megváltozik és megváltoznak az áramlástani viszonyok a csővezeték végén is.




*1. ábra*

Általánosabban fogalmazva azt mondhatjuk, hogy feladatunk annak meghatározása, hogy ha a csővezeték elején ($z=0$) valamilyen adott zavarást alkalmazunk, milyen mértékben változtatja ez meg a csővezeték végén ($z=L$) a stacionárius áramlás hatására kialakult nyomás és tömegáram értékeket.

A továbbiak során az alábbi egyszerűsítő feltételezésekkel élünk:

— Az összenyomható folyadék newtoni és az áramlás lamináris, a nyomáshullám terjedési sebességéhez képest a folyadék sebessége kicsiny.
— A folyadék hőmérsékletváltozásától eltekintünk, vagyis a folyadék súrlódásából eredő hőfokváltozást elhanyagoljuk.
— A csővezeték alakváltozása tökéletesen rugalmas.
— Az áramlásnál a nehézségi gyorsulás hatása elhanyagolható.

Ha $p_0(z)$, $q_0$=konstans, $w_0(r)$ jelölik stacionárius állapotban a nyomás, tömegáram, és sebességi függvényeket, akkor a zavarás hatására kialakuló megfelelő függvények:

$$p(z, t) = p_0(z) + \Delta p(z, t); \quad q(z, t) = q_0 + \Delta q(z, t); \tag{2.1}$$

$$w(r, z, t) = w_0(r) + \Delta w(r, z, t).$$

Az 1. ábrán látható rugalmas csőfalra és a benne áramló, nyomás alatti valóságos munkafolyadékra a nehézségi gyorsulás elhanyagolásával az alábbi egyenletek írhatók fel ([1], [9]):

— az összenyomható közegre érvényes *Hooke-törvény:*

$$\frac{d\varrho}{\varrho} = \frac{dp}{E_0}, \tag{2.2}$$

— a rugalmas csővezeték alakváltozási egyenlete:

$$\frac{dA}{A} = \frac{2a}{E_{cs}e}\,dp, \tag{2.3}$$

— a kontinuitási egyenlet:

$$\frac{\partial(\varrho A)}{\partial t} = -\frac{\partial q}{\partial z}, \tag{2.4}$$

— a valóságos folyadék mozgásegyenlete *(Navier—Stokes egyenlet):*

$$\frac{\partial w}{\partial t} + w\frac{\partial w}{\partial z} = -\frac{1}{\varrho}\frac{\partial p}{\partial z} + \nu\left(\frac{\partial^2 w}{\partial x^2} + \frac{\partial^2 w}{\partial y^2} + \frac{\partial^2 w}{\partial z^2}\right). \tag{2.5}$$

Az egyenletekben szereplő (az előzőeken kívüli) mennyiségek:

$t$ idő,
$\nu$ a folyadék kinematikai viszkozitása,
$a$ a cső sugara,
$L$ a cső hossza,
$e$ a cső falvastagsága,
$A$ a cső keresztmetszete,
$\varrho$ a folyadék sűrűsége,
$E_0$ a folyadék rugalmassági modulusza,
$E_{cs}$ a csővezetők rugalmassági modulusza.

A csővezetékekben történő áramlások vizsgálatában a (2.5)-ben szereplő $w\frac{\partial w}{\partial z}$ nemlineáris tagot és a $\frac{\partial^2 w}{\partial z^2}$ kifejezést el szokták hanyagolni a többi mellett és így az ún. *linearizált Navier—Stokes egyenlet* adódik.

$$\frac{\partial w}{\partial t} = -\frac{1}{\varrho}\frac{\partial p}{\partial z} + \nu\left(\frac{\partial^2 w}{\partial x^2} + \frac{\partial^2 w}{\partial y^2}\right).$$

Az említett elhanyagolások jogosságát a gyakorlat igazolja. Különösen érvényes ez a hidrosztatikus jelátvitel esetén, ahol a csővezetékek hosszához képest a cső sugara kicsiny és az áramló folyadék sebessége is kicsiny. Az alábbiakban a *linearizált Navier—Stokes egyenlettel* dolgozunk.

(2.4) alapján írható, hogy

$$A\frac{\partial \varrho}{\partial t}+\varrho\frac{\partial A}{\partial t}=-\frac{\partial q}{\partial z}. \tag{2.6}$$

(2.2) és (2.3)-at (2.6)-ba helyettesítve adódik, hogy

$$\frac{A\varrho\left[\frac{1}{E_0}+\frac{2a}{E_{cs}e}\right]\partial p}{\partial t}=-\frac{\partial q}{\partial z}. \tag{2.7}$$

Itt a zárójelen belüli mennyiség az eredő (folyadék+csővezeték) rugalmassági modulusz reciproka.

$$\frac{1}{E_r}=\frac{1}{E_0}+\frac{2a}{E_{cs}e}. \tag{2.8}$$

(2.1) és (2.8) alapján a

$$\frac{A\varrho}{E_r}\cdot\frac{\partial \Delta p}{\partial t}+\frac{\partial \Delta q}{\partial z}=0. \tag{2.9}$$

(2.1)-ből és a *linearizált Navier—Stokes egyenletből* következik, hogy

$$\frac{\partial \Delta w}{\partial t}=-\frac{1}{\varrho}\frac{\partial p}{\partial z}+\nu\left(\frac{\partial^2 \Delta w}{\partial x^2}+\frac{\partial^2 \Delta w}{\partial y^2}\right). \tag{2.10}$$

Az $r=\sqrt{x^2+y^2}$ polárkoordináta bevezetésével a hengerszimmetria miatt (2.10) az alábbi alakra hozható.

$$\frac{\partial \Delta w}{\partial t}=-\frac{1}{\varrho}\frac{\partial p}{\partial z}+\nu\left(\frac{\partial^2 \Delta w}{\partial r^2}+\frac{1}{r}\frac{\partial w}{\partial r}\right). \tag{2.11}$$

A tömegáram és a folyadéksebesség között az alábbi definíciós egyenlet adja meg a kapcsolatot,

$$q=2\pi\varrho\int_0^a rw\,dr, \tag{2.12}$$

azaz a tömegáramot megkapjuk, ha a sűrűséget megszorozzuk a sebesség keresztmetszetre vonatkozó integráljával.

(2.1) és (2.12)-ből evidensen adódik, hogy

$$\Delta q=2\pi\varrho\int_0^a r\Delta w\,dr. \tag{2.13}$$

A (2.9), (2.11), (2.13) egyenletekből álló rendszert a csővezetékben történő hidraulikus jelátvitel egyenletrendszerének tekintjük. Írjuk le ezt mégegyszer úgy, hogy az egy-

szerűbb jelölés kedvéért — a félreértés veszélye nélkül — a $\Delta p$, $\Delta q$, $\Delta w$ helyett $p$, $q$, $w$ betűket írunk.

$$\frac{A\varrho}{E_r}\frac{\partial p}{\partial t}+\frac{\partial q}{\partial z}=0,$$

$$\frac{\partial w}{\partial t}=-\frac{1}{\varrho}\frac{\partial p}{\partial z}+\nu\left(\frac{\partial^2 w}{\partial r^2}+\frac{1}{r}\frac{\partial w}{\partial r}\right),\quad q=2\pi\varrho\int_0^a rw\,dr. \tag{2.14}$$

Még egyszer nyomatékosan hangsúlyozzuk, hogy (2.14)-ben az ismeretlen függvényekre bevezetett jelölések a már a zavarás hatására fellépő — a stacionárius folyamatra szuperponálódó — nyomás, tömegáram és sebesség függvényeket jelentik.

(2.14) a benne szereplő három ismeretlenre nézve egy parciális integro-differenciálegyenlet-rendszert képez, melyhez tartozó kiindulási értékek nullák.

$$\lim_{t\to-0}p(z,t)=\lim_{t\to-0}q(z,t)=\lim_{t\to-0}w(r,z,t)=0.$$

### 3. Laplace-transzformáció alkalmazása a hidrosztatikus csővezeték jelátviteli egyenletrendszerének vizsgálatában

Képezzük (2.14) egyenletek *Laplace-transzformáltját*. Figyelembe véve, hogy a kiindulási értékek nullák, a megfelelő *Laplace-transzformáltakat* nagy betűvel jelölve adódik, hogy

$$\frac{A\varrho}{E_r}sP(z,s)+\frac{dQ(z,s)}{dz}=0, \tag{3.1}$$

$$W''(r,s,z)+\frac{1}{r}W'(r,s,z)-\frac{s}{\nu}W(r,s,z)=\frac{1}{\varrho\nu}\frac{dP(z,s)}{dz}, \tag{3.2}$$

$$Q(z,s)=2\pi\varrho\int_0^a W(r,s,z)r\,dr, \tag{3.3}$$

ahol az $r$ szerinti deriválást vesszővel jelöljük. A sebesség ismerete a gyakorlatban érdektelen, annak $W$ *Laplace-transzformáltját* az alábbi módon lehet kiküszöbölni. (3.2) jobb oldala $r$-től nem függ. Egy partikuláris megoldás

$$-\frac{1}{s\varrho}\frac{dP}{dz}.$$

(3.2) azon megoldása most már, amely teljesíti azon feltételeket, hogy az $r=0$-ban korlátos és $r=a$-ban zérus (ahol a sebességnek el kell tűnnie) az alábbi alakú

$$W(r,s,z)=\frac{1}{s\varrho}\frac{dP}{dz}\frac{J_0\left(j\sqrt{\frac{s}{\nu}}r\right)}{J_0\left(j\sqrt{\frac{s}{\nu}}a\right)}-\frac{1}{s\varrho}\frac{dP}{dz}. \tag{3.4}$$

Itt a $J_0$ az *elsőfajú nulladrendű Bessel-függvényt*, $j$ az imaginárius egységet jelöli. A kapott megoldást (3.3)-ba helyettesítve és figyelembe véve a *Bessel-függvényekre* érvé-

nyes

$$\int_0^y xJ_0(x)dx = yJ_1(y)$$

összefüggést [4], melyben $J_1$ az *elsőrendű, elsőfajú Bessel-függvényt* jelöli, az integrálás könnyen elvégezhető és elemien adódik, hogy

$$Q(z,s) = \frac{-2\pi ja\sqrt{\nu}\,\frac{dP}{dz}\,J_1\left(j\sqrt{\frac{s}{\nu}}a\right)}{s^{3/2}J_0\left(j\sqrt{\frac{s}{\nu}}a\right)} - \frac{\pi a^2}{s}\frac{dP}{dz}. \tag{3.5}$$

Vegyük észre, hogy a (3.1) és (3.5) egyenletek az $r$ változót nem tartalmazzák, azok a $P$ és $Q$ transzformáltakra nézve egy $z$-re vonatkozó állandó együtthatójú differenciálegyenlet-rendszert alkotnak. Ennek általános megoldása könnyen felírható:

$$\begin{aligned} Q(z,s) = {} & \alpha(s)\,\mathrm{ch}\left\{\frac{s}{c}\left[1+\frac{2j\sqrt{\nu}\,J_1\left(j\sqrt{\frac{s}{\nu}}a\right)}{a\sqrt{s}\,J_0\left(j\sqrt{\frac{s}{\nu}}a\right)}\right]^{-1/2} z\right\} + \\ & + \beta(s)\,\mathrm{sh}\left\{\frac{s}{c}\left[1+\frac{2j\sqrt{\nu}\,J_1\left(j\sqrt{\frac{s}{\nu}}a\right)}{a\sqrt{s}\,J_0\left(j\sqrt{\frac{s}{\nu}}a\right)}\right]^{-1/2} (z-L)\right\}, \end{aligned} \tag{3.6}$$

$$\begin{aligned} P(z,s) = {} & -\frac{c}{A}\left[1+\frac{2j\sqrt{\nu}\,J_1\left(j\sqrt{\frac{s}{\nu}}a\right)}{a\sqrt{s}\,J_0\left(j\sqrt{\frac{s}{\nu}}a\right)}\right]^{-1/2} \alpha(s)\times \\ & \times\,\mathrm{sh}\left\{\frac{s}{c}\left[1+\frac{2j\sqrt{\nu}\,J_1\left(j\sqrt{\frac{s}{\nu}}a\right)}{a\sqrt{s}\,J_0\left(j\sqrt{\frac{s}{\nu}}a\right)}\right]^{-1/2} z\right\} - \\ & -\frac{c}{A}\left[1+\frac{2j\sqrt{\nu}\,J_1\left(j\sqrt{\frac{s}{\nu}}a\right)}{a\sqrt{s}\,J_0\left(j\sqrt{\frac{s}{\nu}}a\right)}\right]^{-1/2} \beta(s)\times \\ & \times\,\mathrm{ch}\left\{\frac{s}{c}\left[1+\frac{2j\sqrt{\nu}\,J_1\left(j\sqrt{\frac{s}{\nu}}a\right)}{a\sqrt{s}\,J_0\left(j\sqrt{\frac{s}{\nu}}a\right)}\right]^{-1/2} (z-L)\right\}. \end{aligned} \tag{3.7}$$

Fentiekben $c=\sqrt{\frac{E_r}{\varrho}}$ a nyomáshullám fázissebessége, $\alpha(s)$ és $\beta(s)$ integrálási állandók. Ezeket a hidraulikus jelátvitel peremfeltételei határozzák meg.

A csővezeték két végén levő nyomás és tömegáram *Laplace-transzformáltjai* $P(0, s)$, $Q(0, s)$, $P(L, s)$, $Q(L, s)$ közül kettő megadása egyértelműen meghatározza az $\alpha(s)$ és $\beta(s)$ operátorokat. Ha (3.6)-ban $z=L$, (3.7)-ben $z=0$ értéket helyettesítünk, úgy ezek közvetlenül felírhatók.

$$\alpha(s)=\frac{Q(L, s)}{\operatorname{ch}\left\{\frac{s}{c}\left[1+\frac{2j\sqrt{\nu}\,J_1\left(j\sqrt{\frac{s}{\nu}}a\right)}{a\sqrt{s}\,J_0\left(j\sqrt{\frac{s}{\nu}}a\right)}\right]^{-1/2}L\right\}},$$

$$\beta(s)=$$

$$=-\frac{-P(0, s)A}{c\left[1+\frac{2j\sqrt{\nu}}{a\sqrt{s}}\frac{J_1\left(j\sqrt{\frac{s}{\nu}}a\right)}{J_0\left(j\sqrt{\frac{s}{\nu}}a\right)}\right]^{-1/2}\operatorname{ch}\left\{\frac{s}{c}\left[1+\frac{2j\sqrt{\nu}}{a\sqrt{s}}\frac{J_1\left(j\sqrt{\frac{s}{\nu}}a\right)}{J_0\left(j\sqrt{\frac{s}{\nu}}a\right)}\right]^{-1/2}L\right\}}.$$

Visszatéve ezeket (3.6) és (3.7)-be megkapjuk a nyomás és tömegáram $z$ szerinti eloszlásának Laplace transzformáltjait. A hidraulikus jelátvitelnél a nyomás és tömegáram viselkedésének ismerete a cső két végpontja közötti szakaszban $(0<z<L)$ gyakorlatilag érdektelen, a folyamatokat csak a cső elején és végén vizsgáljuk. Ennek megfelelően az előbb meghatározott operátorok és (3.6), (3.7) alapján felírhatóak az alábbi összefüggések.

(3.8)
$$P(L, s)=\frac{P(0, s)}{\operatorname{ch}\left\{\frac{s}{c}\left[1+\frac{2j\sqrt{\nu}\,J_1\left(j\sqrt{\frac{s}{\nu}}a\right)}{a\sqrt{s}\,J_0\left(j\sqrt{\frac{s}{\nu}}a\right)}\right]^{-1/2}L\right\}}-$$

$$-Q(L, s)\frac{c}{A}\left[1+\frac{2j\sqrt{\nu}\,J_1\left(j\sqrt{\frac{s}{\nu}}a\right)}{a\sqrt{s}\,J_0\left(j\sqrt{\frac{s}{\nu}}a\right)}\right]^{-1/2}\times$$

$$\times\operatorname{th}\left\{\frac{s}{c}\left[1+\frac{2j\sqrt{\nu}\,J_1\left(j\sqrt{\frac{s}{\nu}}a\right)}{a\sqrt{s}\,J_0\left(j\sqrt{\frac{s}{\nu}}a\right)}\right]^{-1/2}L\right\},$$

$$(3.9)\qquad Q(0,s)=\frac{P(0,s)}{\left[1+\frac{2j\sqrt{\nu}J_1\left(j\sqrt{\frac{s}{\nu}}a\right)}{a\sqrt{s}J_0\left(j\sqrt{\frac{s}{\nu}}a\right)}\right]^{-1/2}}\frac{A}{c}\times$$

$$\times\operatorname{th}\left\{\frac{s}{c}\left[1+\frac{2j\sqrt{\nu}J_1\left(j\sqrt{\frac{s}{\nu}}a\right)}{a\sqrt{s}J_0\left(j\sqrt{\frac{s}{\nu}}a\right)}\right]^{-1/2}L\right\}+$$

$$+Q(L,s)\frac{1}{\operatorname{ch}\left\{\frac{s}{c}\left[1+\frac{2j\sqrt{\nu}J_1\left(j\sqrt{\frac{s}{\nu}}a\right)}{a\sqrt{s}J_0\left(j\sqrt{\frac{s}{\nu}}a\right)}\right]^{-1/2}L\right\}}.$$

(3.8) és (3.9) összefüggések megadják az operátortartományban a kapcsolatot a cső elején és végén fellépő nyomások, illetve tömegáramok között.

Vezessük most be dimenziónélküli mennyiségként a relatív időt $\bar{t}=\frac{t}{T}$, ahol $T=\frac{L}{c}$ és a hozzá tartozó $\bar{s}=Ts$ transzformációs változót, úgy fenti egyenletekben szereplő hiperbolikus függvények argumentuma az alábbi alakra transzformálódik:

$$(3.10)\qquad \bar{s}\left[1+\frac{2j\sqrt{\nu}J_1\left(j\sqrt{\frac{\bar{s}a^2c}{\nu L}}\right)}{\sqrt{\frac{a^2c\bar{s}}{L}}J_0\left(j\sqrt{\frac{\bar{s}a^2c}{\nu L}}\right)}\right]^{-1/2}.$$

Vezessük be továbbá a $\tau=\frac{L\nu}{ca^2}$ úgynevezett csillapítási (hasonlósági) tényezőt. A hiperbolikus függvények argumentuma

$$\bar{s}\varphi(\bar{s})$$

alakban írható, ahol

$$(3.11)\qquad \varphi(\bar{s})=\left[1+\frac{2jJ_1\left(j\sqrt{\frac{\bar{s}}{\tau}}\right)}{\sqrt{\frac{\bar{s}}{\tau}}J_0\left(j\sqrt{\frac{\bar{s}}{\tau}}\right)}\right]^{-1/2}.$$

Az alkalmazott jelölésekkel (3.8), (3.9) összefüggések az alábbi igen szemléletes alakban írhatók fel a *Laplace-transzformáció* hasonlósági tételének figyelembevételével.

$$
\begin{aligned}
P(L, \bar{s}) &= P(0, \bar{s}) \frac{1}{\operatorname{ch} \bar{s}\varphi(\bar{s})} - Q(L, \bar{s}) \frac{c}{A} \varphi(\bar{s}) \operatorname{th} \bar{s}\varphi(\bar{s}), \\
Q(0, \bar{s}) &= P(0, \bar{s}) \frac{A}{c\varphi(\bar{s})} \operatorname{th} \bar{s}\varphi(\bar{s}) + Q(L, \bar{s}) \frac{1}{\operatorname{ch} \bar{s}\varphi(\bar{s})},
\end{aligned}
\tag{3.12}
$$

ahol természetesen már a $\bar{t}$ relatív idő szerinti *Laplace-transzformáltak* szerepelnek. Írjuk fel (3.12) egyenletrendszert mátrixos formában

$$
\begin{bmatrix} P(L, \bar{s}) \\ \frac{c}{A} Q(0, \bar{s}) \end{bmatrix} = \begin{bmatrix} \frac{1}{\operatorname{ch} \bar{s}\varphi(\bar{s})} & -\varphi(\bar{s}) \operatorname{th} \bar{s}\varphi(\bar{s}) \\ \frac{1}{\varphi(\bar{s})} \operatorname{th} \bar{s}\varphi(\bar{s}) & \frac{1}{\operatorname{ch} \bar{s}\varphi(\bar{s})} \end{bmatrix} \begin{bmatrix} P(0, \bar{s}) \\ \frac{c}{A} Q(L, \bar{s}) \end{bmatrix}. \tag{3.13}
$$

A mátrixegyenlet bal oldali oszlopvektorában a csővégi nyomás és a cső elején érvényes tömegáram, a jobb oldali oszlopvektorában pedig a cső elején fellépő nyomás és a csővégi tömegáram szerepel. A négy függvény közül kettőt a peremfeltételekből kell meghatározni. A mátrix elemek (átviteli függvények) argumentumának bonyolultságán túl szembetűnik, hogy azok az $\bar{s}$ operátoron kívül csak a $\tau$ paramétertől függenek.

Vizsgáljuk most az irányítástechnikai hasonlóságot a $\bar{t}$ relatív időtartományban. Két csővezetéket akkor nevezünk irányítástechnikailag hasonlónak, ha a két csővezeték elején fellépő egyenlő bemenőjelek hatására a csővezetékek végén fellépő kimenőjelek megegyeznek. (3.13)-ból a hasonlóság szükséges és elégséges feltételei azonnal leolvashatók. Egyrészt látjuk, hogy teljesülnie kell a

$$
\frac{c_1}{A_1} = \frac{c_2}{A_2},
$$

vagy ami ugyanaz a

a)
$$
\frac{c_1}{a_1^2} = \frac{c_2}{a_2^2}
$$

feltételnek. Másrészt a két csővezetékre nézve a $\varphi(\bar{s})$ függvényeknek meg kell egyezniük. (3.11)-ből következik, hogy teljesülnie kell a $\tau_1 = \tau_2$, vagyis a

b)
$$
\frac{L_1 v_1}{c_1 a_1^2} = \frac{L_2 v_2}{c_2 a_2^2}
$$

feltételnek. Az a) és b) feltételek a relatív időtartományban az irányítástechnikai

hasonlóság szükséges és elégséges kritériumai. A (3.13)-ban szereplő átviteli mátrix elemeire vezessük be az alábbi jelöléseket.

$$(3.14) \qquad A(\bar{s}) = \frac{1}{\operatorname{ch} \bar{s}\varphi(\bar{s})}; \quad B(\bar{s}) = \varphi(\bar{s}) \operatorname{th} \bar{s}\varphi(\bar{s}); \quad C(\bar{s}) = \frac{1}{\varphi(\bar{s})} \operatorname{th} \bar{s}\varphi(\bar{s}).$$

Az $A(\bar{s})$, $B(\bar{s})$, $C(\bar{s})$ operátorok adják meg a rendszer válaszát *Dirac-δ gerjesztések* hatására. A $\varphi(\bar{s})$ bonyolultsága miatt az $A(\bar{s})$, $B(\bar{s})$, $C(\bar{s})$ operátorok időtartománybeli pontos invertálására nincs lehetőség. A $B(\bar{s})$ és $C(\bar{s})$ operátorok a kifejtési tétellel nem is invertálhatók, hiszen fizikailag könnyen beláthatóan *Dirac-komponenst* kell tartalmazniuk és látni fogjuk, hogy a *Dirac-komponenstől* eltekintve is a $B(\bar{s})$, $C(\bar{s})$ *inverz Laplace-transzformáltjai* a $\bar{t}=0$ környezetében nem korlátosak.

Matematikailag is egyszerűen belátható, hogy a $B(\bar{s})$ és $C(\bar{s})$ operátorok nem lehetnek klasszikus függvények *Laplace-transzformáltjai*. Ugyanis, ha azok lennének, akkor mint ismeretes fenn kellene állnia a

$$\lim_{\bar{s}\to\infty} B(\bar{s}) = \lim_{\bar{s}\to\infty} C(\bar{s}) = 0$$

összefüggéseknek, ha a valós tengely pozitív irányában tartunk a végtelenhez. Azonban könnyen belátható, hogy

$$\lim_{\bar{s}\to\infty} B(\bar{s}) = \lim_{\bar{s}\to\infty} \varphi(\bar{s}) \operatorname{th} \bar{s}\varphi(\bar{s}) = 1.$$

Ugyanis vezessük be (3.11)-ben a *módosított Bessel-függvényeket* [3]

$$(3.15) \qquad \varphi(\bar{s}) = \left[1 - \frac{2I_1\left(\sqrt{\frac{\bar{s}}{\tau}}\right)}{\sqrt{\frac{\bar{s}}{\tau}}\, I_0\left(\sqrt{\frac{\bar{s}}{\tau}}\right)}\right]^{-1/2}.$$

Nagy pozitív $\bar{s}$ értékekre a *Bessel-függvények* aszimptotikus sorából következik, hogy

$$I_1\left(\sqrt{\frac{\bar{s}}{\tau}}\right) \approx I_0\left(\sqrt{\frac{\bar{s}}{\tau}}\right)$$

és így

$$\lim_{\bar{s}\to\infty} \varphi(\bar{s}) = 1.$$

Vagyis $\lim_{\bar{s}\to\infty} B(\bar{s}) = 1$ és teljesen hasonlóan látható be, hogy

$$\lim_{\bar{s}\to\infty} C(\bar{s}) = 1.$$

A következőkben, a fellépő operátorok invertálására olyan közelítő módszert fogunk

alkalmazni, mely az időfüggvényeket elvben kis időértékekre, gyakorlatilag azonban a tranziens folyamatok lecsillapodásának időtartama alatt jól közelítik. A gyakorlatilag jó közelítés alatt azt értjük, hogy a számítások által nyert időfüggvények a mérési eredmények által kapott időfüggvényekkel közelítőleg megegyeznek. Az ismertetendő közelítő invertálás jogos voltát fizikailag az is indokolja, hogy a csillapított rendszer legnagyobb értékű nyomás és tömegáram lengései — melyek meghatározása műszaki szempontból a leglényegesebb — viszonylag rövid idő alatt lezajlanak. A Laplace-transzformáció elméletéből ismert, hogy az időfüggvény $\bar{t}=0$ környezetében való viselkedését a transzformált függvény nagy pozitív valósrészű $\bar{s}$ értékei determinálják. Ennek megfelelően ilyen $\bar{s}$ értékekre a mátrix elemei közelítőleg

$$A(\bar{s}) = \frac{1}{\operatorname{ch}\bar{s}\left(1-2\sqrt{\frac{\tau}{\bar{s}}}\right)^{-1/2}} \quad B(\bar{s}) = \left(1-2\sqrt{\frac{\tau}{\bar{s}}}\right)^{-1/2} \operatorname{th}\bar{s}\left(1-2\sqrt{\frac{\tau}{\bar{s}}}\right)^{-1/2}$$

(3.16)

$$C(\bar{s}) = \left(1-2\sqrt{\frac{\tau}{\bar{s}}}\right)^{+1/2} \operatorname{th}\bar{s}\left(1-2\sqrt{\frac{\tau}{\bar{s}}}\right)^{-1/2}.$$

A közelítő invertálás elvégezhetőségéhez még további egyszerűsítést hajtunk végre. Vesszük az

(3.17) $$\left(1-2\sqrt{\frac{\tau}{\bar{s}}}\right)^{-1/2}, \quad \left(1-2\sqrt{\frac{\tau}{\bar{s}}}\right)^{1/2}$$

kifejezések binomális sorának első három tagját

(3.18) $$\left(1-2\sqrt{\frac{\tau}{\bar{s}}}\right)^{-1/2} \approx 1+\sqrt{\frac{\tau}{\bar{s}}}+\frac{3\tau}{2\bar{s}}, \quad \left(1-2\sqrt{\frac{\tau}{\bar{s}}}\right)^{1/2} \approx 1-\sqrt{\frac{\tau}{\bar{s}}}-\frac{\tau}{2\bar{s}}.$$

Így az alábbi végső visszatranszformálandó függvényeket kapjuk.

(3.19) $$A(\bar{s}) = \frac{1}{\operatorname{ch}\left(\bar{s}+\sqrt{\tau\bar{s}}+\frac{3}{2}\tau\right)},$$

(3.20) $$B(\bar{s}) = \left(1+\sqrt{\frac{\tau}{\bar{s}}}+\frac{3\tau}{2\bar{s}}\right)\operatorname{th}\left(\bar{s}+\sqrt{\tau\bar{s}}+\frac{3}{2}\tau\right),$$

(3.21) $$C(\bar{s}) = \left(1-\sqrt{\frac{\tau}{\bar{s}}}-\frac{\tau}{2\bar{s}}\right)\operatorname{th}\left(\bar{s}+\sqrt{\tau\bar{s}}+\frac{3}{2}\tau\right).$$

Nyilvánvaló, hogy az alkalmazott közelítések során nyert kifejezések annál pontosabbak, minél kisebb a $\tau$ paraméter értéke, azaz a rendszer csillapítása. Az invertálást

WAGNER [8], [10] módszerével hajtjuk végre. A módszer lényege, hogy a hiperbolikus függvényeket exponenciális függvényekké alakítjuk át, majd az így kapott kifejezéseket sorbafejtve, az inverz transzformáció a végtelen sorok tagonkénti invertálásával egyszerűen elvégezhető. Végeredményben az időfüggvények olyan végtelen sorokkal állíthatók elő, melyek tetszőleges rögzített $\bar{t}_0$ időpontra véges sorokká redukálódnak és így nem túl nagy időintervallumban gyakorlatilag igen könnyen számíthatók.

Transzformáljuk először vissza a (3.19) operátort. Kapjuk, hogy

$$A(\bar{s}) = \frac{1}{\operatorname{ch}\left(\bar{s}+\sqrt{\tau\bar{s}}+\frac{3}{2}\tau\right)} = \frac{2}{e^{\left(\bar{s}+\sqrt{\tau\bar{s}}+\frac{3}{2}\tau\right)}+e^{-\left(\bar{s}+\sqrt{\tau\bar{s}}+\frac{3}{2}\tau\right)}} =$$

$$= \frac{2e^{-\left(\bar{s}+\sqrt{\tau\bar{s}}+\frac{3}{2}\tau\right)}}{1+e^{-2\left(\bar{s}+\sqrt{\tau\bar{s}}+\frac{3}{2}\tau\right)}} = 2\sum_{k=0}^{\infty}(-1)^k e^{-(2k+1)\left(\bar{s}+\sqrt{\tau\bar{s}}+\frac{3}{2}\tau\right)} = \tag{3.22}$$

$$= 2\sum_{k=0}^{\infty}(-1)^k e^{-(2k+1)\bar{s}} e^{-(2k+1)\sqrt{\tau\bar{s}}} e^{-(2k+1)\frac{3}{2}\tau}.$$

Bevezetve az

$$1(\bar{t}) = \begin{cases} 1, & \text{ha } \bar{t} > 0, \\ 0, & \text{ha } \bar{t} < 0 \end{cases}$$

egységugrásfüggvényt és figyelembevéve, hogy tetszőleges $a > 0$ mellett [3]

$$\mathcal{L}^{-1}\left[e^{-a\sqrt{\bar{s}}}\right] = \frac{a}{2\sqrt{\pi}\,\bar{t}^{3/2}} e^{-\frac{a^2}{4\bar{t}}},$$

az eltolási operátor ismert tulajdonsága alapján (3.22) tagonkénti visszatranszformálásával adódik, hogy

(3.23)

$$A(\bar{t}) = \sqrt{\frac{\tau}{\pi}}\sum_{k=0}^{\infty}(-1)^k(2k+1)e^{-(2k+1)\frac{3}{2}\tau}\,1[\bar{t}-(2k+1)]\,\frac{\exp\left\{-\frac{(2k+1)^2\tau}{4[\bar{t}-(2k+1)]}\right\}}{[\bar{t}-(2k+1)]^{3/2}}.$$

Bármilyen rögzített $\bar{t}_0$ időpontra a kapott sor véges számú tagot tartalmaz, mivel azon $k$ indexekre, melyekre

$$\bar{t}_0-(2k+1) < 0,$$

fennáll, hogy

$$1[\bar{t}_0-(2k+1)] = 0.$$

Hasonló elven transzformálható vissza $B(\bar{s})$ és $C(\bar{s})$ is. Adódik, hogy

$$B(\bar{s}) = \left(1+\sqrt{\frac{\tau}{\bar{s}}}+\frac{3\tau}{2\bar{s}}\right)\operatorname{th}\left(\bar{s}+\sqrt{\tau\bar{s}}+\frac{3}{2}\tau\right) =$$

$$= \left(1+\sqrt{\frac{\tau}{\bar{s}}}+\frac{3\tau}{2\bar{s}}\right)\frac{e^{\left(\bar{s}+\sqrt{\tau\bar{s}}+\frac{3}{2}\tau\right)}-e^{-\left(\bar{s}+\sqrt{\tau\bar{s}}+\frac{3}{2}\tau\right)}}{e^{\left(\bar{s}+\sqrt{\tau\bar{s}}+\frac{3}{2}\tau\right)}+e^{-\left(\bar{s}+\sqrt{\tau\bar{s}}+\frac{3}{2}\tau\right)}} =$$

$$= \left(1+\sqrt{\frac{\tau}{\bar{s}}}+\frac{3\tau}{2\bar{s}}\right)\left[\frac{1-e^{-2\left(\bar{s}+\sqrt{\tau\bar{s}}+\frac{3}{2}\tau\right)}}{1+e^{2\left(\bar{s}+\sqrt{\tau\bar{s}}+\frac{3}{2}\tau\right)}}\right] = \tag{3.24}$$

$$= \left(1+\sqrt{\frac{\tau}{\bar{s}}}+\frac{3\tau}{2\bar{s}}\right)\left[1-e^{-2\left(\bar{s}+\sqrt{\tau\bar{s}}+\frac{3}{2}\tau\right)}\right]\sum_{k=0}^{\infty}(-1)^k e^{-2k\left(\bar{s}+\sqrt{\tau\bar{s}}+\frac{3}{2}\tau\right)} =$$

$$= \left(1+\sqrt{\frac{\tau}{\bar{s}}}+\frac{3\tau}{2\bar{s}}\right)\left[\sum_{k=0}^{\infty}(-1)^k e^{-2k\left(\bar{s}+\sqrt{\tau\bar{s}}+\frac{3\tau}{2}\right)}-\sum_{k=0}^{\infty}(-1)^k e^{-2(k+1)\left(\bar{s}+\sqrt{\tau\bar{s}}+\frac{3\tau}{2}\right)}\right].$$

Ha a második szummában bevezetjük az $i=k+1$ új szummációs indexet, úgy

$$\sum_{k=0}^{\infty}(-1)^k e^{-2k\left(\bar{s}+\sqrt{\tau\bar{s}}+\frac{3}{2}\tau\right)}-\sum_{i=1}^{\infty}(-1)^{i-1}e^{-2i\left(\bar{s}+\sqrt{\tau\bar{s}}+\frac{3}{2}\tau\right)} =$$

$$= 1+2\sum_{k=1}^{\infty}(-1)^k e^{-2k\left(\bar{s}+\sqrt{\tau\bar{s}}+\frac{3}{2}\tau\right)}.$$

Tehát (3.24) az alábbi alakra hozható

$$B(\bar{s}) = 1+\sqrt{\frac{\tau}{\bar{s}}}+\frac{3\tau}{2\bar{s}}+\left(2+2\sqrt{\frac{\tau}{\bar{s}}}+\frac{3\tau}{\bar{s}}\right)\sum_{k=1}^{\infty}(-1)^k e^{-3k\tau}e^{-2k\bar{s}}e^{-2k\sqrt{\tau\bar{s}}} =$$

$$= 1+\sqrt{\frac{\tau}{\bar{s}}}+\frac{3\tau}{2\bar{s}}+\sum_{k=1}^{\infty}(-1)^k e^{-3k\tau}e^{-2k\bar{s}}\left[2e^{-2k\sqrt{\tau\bar{s}}}+2\sqrt{\tau}\frac{e^{-2k\sqrt{\tau\bar{s}}}}{\sqrt{\bar{s}}}+3\tau\frac{e^{-2k\sqrt{\tau\bar{s}}}}{\bar{s}}\right]. \tag{3.25}$$

A szumma előtt álló összeg elemien visszatranszformálható. Továbbá [3]-ból

$$\mathcal{L}^{-1}\left[\frac{e^{-2k\sqrt{\tau\bar{s}}}}{\sqrt{\bar{s}}}\right] = \frac{1}{\sqrt{\pi t}}e^{-\frac{k^2\tau}{t}},$$

$$\mathcal{L}^{-1}\left[\frac{e^{-2k\sqrt{\tau\bar{s}}}}{\bar{s}}\right] = \operatorname{erfc}\frac{k\sqrt{\tau}}{\sqrt{t}}, \quad \text{ahol} \quad \operatorname{erfc}x = \frac{2}{\sqrt{\pi}}\int_x^{\infty}e^{-u^2}du, \quad x \geqq 0.$$

Így $B(\bar{s})$ inverz transzformáltjára az alábbi kifejezést nyerjük

$$B(\bar{t}) = \delta(\bar{t}) + \sqrt{\frac{\tau}{\pi \bar{t}}} + \frac{3}{2}\tau + \sum_{k=1}^{\infty} (-1)^k e^{-3k\tau} \left[ 2\sqrt{\frac{\tau}{\pi}} k \frac{e^{-\frac{k^2\tau}{\bar{t}-2k}}}{(\bar{t}-2k)^{3/2}} + \right.$$

(3.26)

$$\left. + 2\sqrt{\frac{\tau}{\pi}} \frac{e^{-\frac{k^2\tau}{\bar{t}-2k}}}{\sqrt{\bar{t}-2k}} + 3\tau \operatorname{erfc} \frac{k\sqrt{\tau}}{\sqrt{\bar{t}-2k}} \right] 1(\bar{t}-2k).$$

Látható, hogy tetszőleges rögzített $\bar{t}_0$-ra (3.26) véges sorrá redukálódik. A erfc függvény kiszámítására jól kezelhető táblázat található [4]-ben. A $C(\bar{s})$ operátor teljesen hasonlóan transzformálható vissza. E. t nem részletezve, az alábbi végeredmény adódik

$$C(\bar{t}) = \delta(\bar{t}) - \sqrt{\frac{\tau}{\pi \bar{t}}} - \frac{\tau}{2} + \sum_{k=1}^{\infty} (-1)^k e^{-3\tau k} \times$$

(3.27)

$$\times \left[ 2\sqrt{\frac{\tau}{\pi}} k \frac{e^{-\frac{k^2\tau}{\bar{t}-2k}}}{(\bar{t}-2k)^{3/2}} - 2\sqrt{\frac{\tau}{\pi}} \frac{e^{-\frac{k^2\tau}{\bar{t}-2k}}}{\sqrt{\bar{t}-2k}} - \tau \operatorname{erfc} \frac{k\sqrt{\tau}}{\sqrt{\bar{t}-2k}} \right] 1(\bar{t}-2k).$$

Érdekes, hogy $B(\bar{t})$, $C(\bar{t})$ időfüggvények a *Dirac-$\delta(\bar{t})$ impulzusfüggvénytől* eltekintve sem korlátosak, mert $\frac{1}{\sqrt{\bar{t}}}$-vel arányos additív tagot tartalmaznak. Az $A(\bar{t})$, $B(\bar{t})$, $C(\bar{t})$ függvények igen nehezen ábrázolhatók (különösen kis $\tau$ értékekre), mert igen gyors változásúak. Ez fizikailag érthető, mert ezek a rendszer *Dirac-$\delta$ gerjesztésre* (vagy azzal arányos gerjesztésre) adott válaszai, az ún. súlyfüggvények.

Az irányítástechnika szerint [2], (3.13) mátrix egyenletnek az időtartományban az alábbi egyenletek felelnek meg:

$$p(L, t) = \int_0^{\bar{t}} A(\bar{t}-u) p(0, u) du - \frac{c}{A} \int_0^{\bar{t}} B(\bar{t}-u) q(L, u) du,$$

(3.28)

$$\frac{c}{A} q(0, \bar{t}) = \int_0^{\bar{t}} C(t-u) p(0, u) du + \frac{c}{A} \int_0^{\bar{t}} A(\bar{t}-u) q(L, u) du,$$

melyek a *Laplace-transzformáció* ismert konvolúció-tételének következményei. Az $A(\bar{t})$, $B(\bar{t})$, $C(\bar{t})$ függvények ismeretében tetszőleges bemenőjelek hatására a keletkező kimenőjelek (3.28)-ból meghatározhatók.

## 4. Az átmeneti jelenségek vizsgálata

A hirtelen terhelésváltozás hatására bekövetkező átmeneti vagy tranziens állapot vizsgálatához induljunk ki a

$$\begin{bmatrix} P(L, \bar{s}) \\ \frac{c}{A} Q(0, \bar{s}) \end{bmatrix} = \begin{bmatrix} A(\bar{s}) & -B(\bar{s}) \\ C(\bar{s}) & A(\bar{s}) \end{bmatrix} \begin{bmatrix} P(0, \bar{s}) \\ \frac{c}{A} Q(L, \bar{s}) \end{bmatrix} \tag{4.1}$$

mátrix egyenletből. A hidrosztatikus körfolyamokban a csővezeték elejéhez a szivattyú, a cső végéhez pedig sokszor valamilyen mennyiségszelep csatlakozik. Ez azt jelenti, hogy a csővezeték végén a $z=L$ helyen ilyenkor nem a nyomás, vagy tömegáram megadásával írjuk elő a peremfeltételt, hanem a két mennyiség között egy összefüggést írunk elő, melyet az operátortartományban az ún. *hidraulikai operátoros Ohm-törvény* fejez ki

$$P(L, \bar{s}) = Z(\bar{s}) \frac{c}{A} Q(L, \bar{s}). \tag{4.2}$$

A $Z(\bar{s})$ az ún. hidraulikus operátoros impedancia dimenzió nélküli mennyiség. A gyakorlatilag legfontosabb esetben az operátoros impedancia tiszta ohmos, ami azt jelenti, hogy $Z(\bar{s})=Z$ állandó, ($\bar{s}$-től független) és ekkor a cső végén a nyomás és tömegáram között egy adott lineáris összefüggés áll fenn. Nyitottvégű csővezetékeknél $Z=0$, lezártvégű csővezetékeknél $Z=\infty$. (4.1) és (4.2)-ből egyszerű számolással adódnak a csővezeték alábbi nyomás-nyomás, illetve nyomás-tömegáram átviteli függvényei:

$$\frac{P(L, \bar{s})}{P(0, \bar{s})} = \frac{A(\bar{s})}{1+\dfrac{B(\bar{s})}{Z(\bar{s})}}, \tag{4.3}$$

$$\frac{P(L, \bar{s})}{\dfrac{c}{A} Q(0, \bar{s})} = \frac{A(\bar{s})}{C(\bar{s})+\dfrac{1}{Z(\bar{s})}}. \tag{4.4}$$

A kapott kifejezések szemléletesen mutatják, hogy ha a cső elejére nyomás vagy tömegáram bemenőjelet adunk, akkor erre milyen módon reagál a cső, vagyis milyen lesz a cső végén a nyomás. Természetesen hasonlóan felírhatók a tömegáram-tömegáram, illetve tömegáram-nyomás átviteli függvények is, ezeknek azonban a gyakorlatban kisebb a jelentőségük.

A gyakorlatban előforduló átmeneti állapot legfontosabb esetei a hirtelen nyitás és zárás, melyeket az alábbiakban részletesen ismertetünk.

### I. *Hirtelen nyitás a cső elején*

Ekkor a csővezeték elejét hirtelen kinyitjuk, és állandó $p_0$ nyomáson tartjuk míg a vége el van zárva. Ekkor

$$P(0, \bar{s}) = \frac{p_0}{\bar{s}}, \quad Q(L, \bar{s}) = 0 \quad (Z = \infty)$$

(4.3)-ból

$$P(L, \bar{s}) = P(0, \bar{s}) A(\bar{s}) = \frac{p_0}{\bar{s}} A(\bar{s})$$

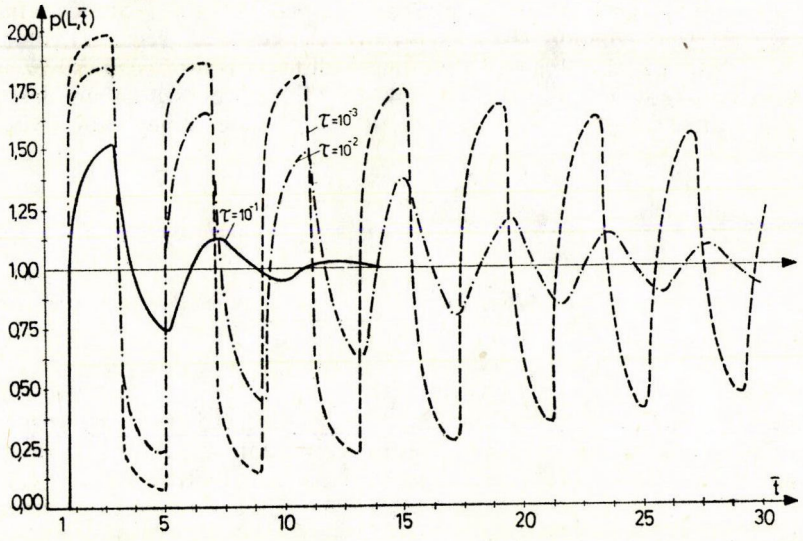
(3.19) és (3.22) figyelembevételével nyerjük, hogy

$$P(L, \bar{s}) = \frac{p_0}{\bar{s} \operatorname{ch}\left(\bar{s}+\sqrt{\tau\bar{s}}+\dfrac{3\tau}{2}\right)} = \frac{2p_0}{\bar{s}} \sum_{k=0}^{\infty} (-1)^k e^{-(2k+1)\bar{s}} e^{-(2k+1)\sqrt{\tau\bar{s}}} e^{-(2k+1)\frac{3\tau}{2}}. \tag{4.5}$$

Ennek inverz *Laplace-transzformáltja* az előző szakaszban ismertettek alapján

$$(4.6) \qquad p(L,\bar{t}) = 2p_0 \sum_{k=0}^{\infty} (-1)^k 1[\bar{t}-(2k+1)] e^{-(2k+1)\frac{3\tau}{2}} \operatorname{erfc} \frac{(2k+1)\sqrt{\tau}}{2\sqrt{\bar{t}-(2k+1)}}.$$

Ebből az összefüggésből $p_0=1$ bar esetén numerikusan kiszámoltuk a nyomást a $t$ relatív idő függvényében különböző $\tau$ értékek esetén. Ezt szemlélteti az alábbi 2. ábra.



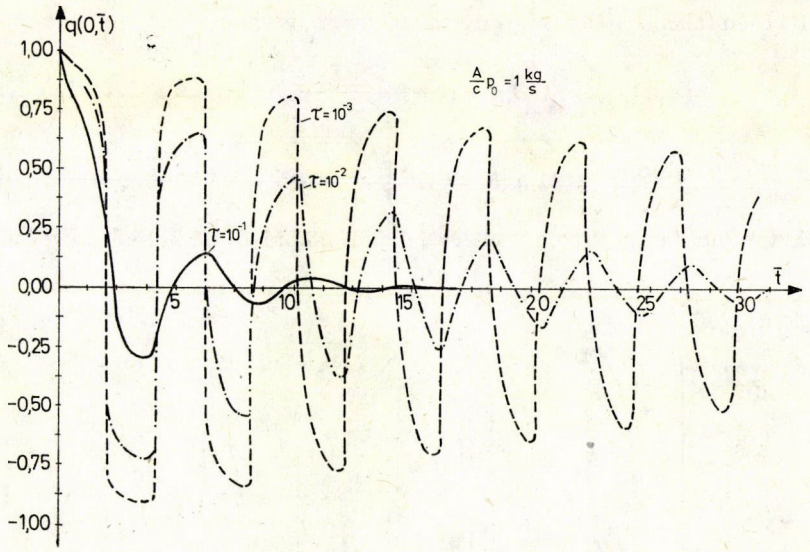

*2. ábra*

Meghatározhatjuk a cső elején fellépő tömegáram időfüggvényét is. A peremfeltételekből (4.1) alapján

$$Q(0,\bar{s}) = \frac{A}{c} P(0,\bar{s}) C(\bar{s}) = \frac{Ap_0}{c\bar{s}} C(\bar{s}).$$

Ennek visszatranszformálását ebben a dolgozatban nem közöljük, mert az teljesen analóg az alábbi részletesen ismertetett II. esettel. Az eredményül kapott $q(0,\bar{t})$ függvényt $\frac{A}{c}p_0=1$ kg/sec mellett numerikusan kiszámoltuk a $\bar{t}$ relatív idő függvényében különböző $\tau$ értékekre. Ezt szemlélteti a 3. ábra.
A függvények lefutásából megállapítható, hogy

$\tau \approx 10^{-1}$ esetén nagy,
$\tau \approx 10^{-2}$ esetén közepes,
$\tau \approx 10^{-3}$ esetén kis csillapítású csővezetékről beszélhetünk.




*3. ábra*

II. *Hirtelen zárás a cső végén*

Ebben az esetben a csővezeték elején alkalmazott állandó nyomás mellett a nyitott végű csővezetéket a végén hirtelen lezárjuk.

Ekkor

$$P(0, \bar{s}) = 0, \quad Q(L, \bar{s}) = \frac{-q_0}{\bar{s}}.$$

A peremfeltételek és (4.1)-ből adódik, hogy

$$P(L, \bar{s}) = \frac{c}{A} \frac{q_0}{\bar{s}} B(\bar{s}) \tag{4.7}$$

és (3.25) figyelembevételével

$$\frac{B(\bar{s})}{\bar{s}} = \frac{1}{\bar{s}} + \frac{\sqrt{\tau}}{\bar{s}^{3/2}} + \frac{3\tau}{2\bar{s}^2} + \sum_{k=1}^{\infty} (-1)^k e^{-2k\bar{s}} e^{-3\tau k} \left[ \frac{2e^{-2k\sqrt{\tau\bar{s}}}}{\bar{s}} + 2\sqrt{\tau} \frac{e^{-2k\sqrt{\tau\bar{s}}}}{\bar{s}^{3/2}} + \frac{3\tau e^{-2k\sqrt{\tau\bar{s}}}}{\bar{s}^2} \right]. \tag{4.8}$$

A szumma előtt álló tagok elemien visszatranszformálhatók, továbbá felhasználva az alábbi *Laplace-transzformációs* összefüggéseket [5],

$$\mathcal{L}^{-1}\left[\frac{e^{-\sqrt{a\bar{s}}}}{\bar{s}^{3/2}}\right] = 2\sqrt{\frac{\bar{t}}{\pi}} e^{-\frac{a}{4\bar{t}}} - \sqrt{a}\,\mathrm{erfc}\left(\frac{1}{2}\sqrt{\frac{a}{\bar{t}}}\right),$$

$$\mathcal{L}^{-1}\left[\frac{e^{-a\sqrt{\bar{s}}}}{\bar{s}^2}\right] = \left[\bar{t} + \frac{a^2}{2}\right] \mathrm{erfc}\,\frac{a}{2\sqrt{\bar{t}}} - a\sqrt{\frac{\bar{t}}{\pi}} e^{-\frac{a^2}{4\bar{t}}}.$$

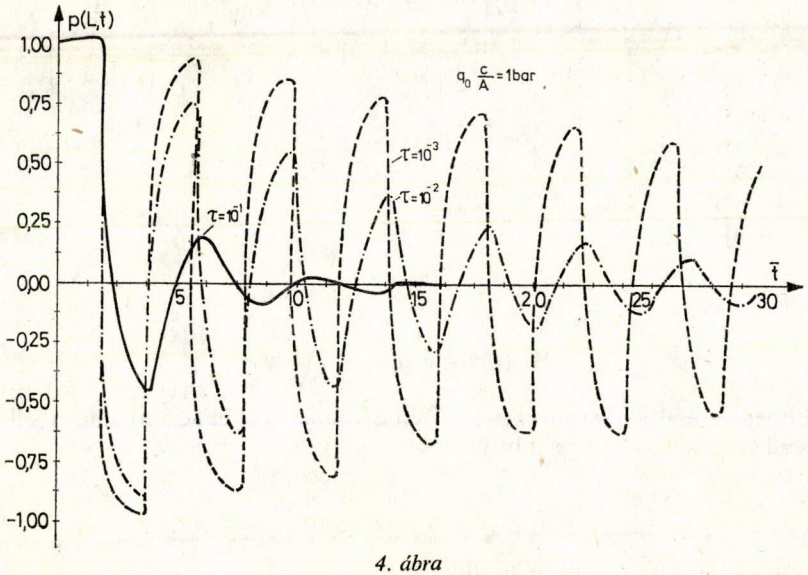
A cső végén fellépő nyomás időfüggvényére adódik, hogy

$$p(L,\bar{t}) = q_0 \frac{c}{A}\left\{1+2\sqrt{\frac{\tau\bar{t}}{\pi}}+\frac{3\tau\bar{t}}{2}+\sum_{k=1}^{\infty}(-1)^k e^{-3\tau k}1(\bar{t}-2k)\right.$$

(4.9)

$$\left.\left[[2+3\tau(\bar{t}-2k+2k^2\tau)-4k\tau]\operatorname{erfc} k\sqrt{\frac{\tau}{\bar{t}-2k}}+(4-6k\tau)\sqrt{\frac{\tau\bar{t}}{\pi}}\,e^{-\frac{k^2\tau}{\bar{t}-2k}}\right]\right\}.$$

Ezt numerikusan kiszámoltuk és a 4. ábrán ábrázoltuk különböző $\tau$ értékekre.




*4. ábra*

### III. *Hirtelen zárás a cső elején*

Az egyszerűség kedvéért feltételezzük, hogy a cső végén $Z=0$ impedancia van. Ekkor a peremfeltételek:

$$Q(0,\bar{s}) = -\frac{q_0}{\bar{s}}, \quad P(L,\bar{s}) = 0.$$

(4.1)-ből elemien adódik, hogy

$$P(0,\bar{s}) = \frac{c}{A}B(\bar{s})Q(0,\bar{s}) = -\frac{cq_0}{A\bar{s}}B(\bar{s}),$$

(4.10)

$$Q(L,\bar{s}) = Q(0,\bar{s})A(\bar{s}) = -\frac{q_0}{\bar{s}}A(\bar{s}).$$

Ezek inverzei konstans szorzótól eltekintve I. és II. esetekben már meg lettek határozva.

IV. *Hirtelen nyitás a cső végén*

$p_0$ állandó nyomásra kapcsolt lezártvégű csővezetéket a végén teljesen nyitjuk ($Z=0$). Ekkor a tömegáramok *Laplace-transzformáltjaira* (4.1)-ből a

$$P(0, \bar{s}) = 0, \quad P(L, \bar{s}) = -\frac{p_0}{\bar{s}}$$

peremfeltételek figyelembevételével kapjuk, hogy

$$Q(0, \bar{s}) = \frac{p_0 A A(\bar{s})}{c\bar{s}B(\bar{s})}, \quad Q(L, \bar{s}) = \frac{p_0 A}{\bar{s}cB(\bar{s})}.$$

Ezek visszatranszformálásával a kapott kifejezések bonyolultsága miatt e cikk kereteiben nem foglalkozunk.

Megjegyezzük, hogy különböző méretű csővezetékeknél a hirtelen nyitás-zárás eseteire kapott nyomásfüggvényeket a laboratóriumi mérésekkel összehasonlítottuk és azt tapasztaltuk, hogy az előzőek az utóbbiak igen jó közelítései.

## 5. Időben szinuszosan változó folyamatok vizsgálata

A jelátvitel — hidrosztatikus csővezetékeknél — gyakorlati szempontból legfontosabb feladata a színuszos gerjesztések hatására létrejövő folyamatok vizsgálata az állandósult állapotban.

Ez a (2.14) egyenletrendszerrel kapcsolatos kezdeti érték nélküli problémára vezet, ha az időfüggést a (2.14) egyenletrendszerben a villamosságtanban ismert módon az $e^{j\omega t}$ faktorral kiszeparáljuk. Ekkor (2.14) előírt peremfeltételeket kielégítő megoldása exakt meghatározható.

Az alábbiakban azonban nem lesz szükségünk arra, hogy vizsgálatainkat (2.14)-ből kiindulva hajtsuk végre. Ugyanis a *Laplace-transzformációval* kapott eredményekből az időben színuszosan változó folyamatokra vonatkozó eredmények közvetlenül adódnak.

Vizsgálódásunk kiinduló alapegyenlete a (3.13) átviteli mátrix egyenlet. Ha a mátrix elemeiben elvégezzük az $\bar{s}=j\bar{\omega}$ helyettesítést ahol $\omega$ a körfrekvencia és

$$\bar{\omega} = \omega T = \omega \frac{L}{c}$$

a relatív körfrekvencia, továbbá az oszlopvektorokban álló $P(0, \bar{s})$, $Q(L, \bar{s})$, $Q(0, \bar{s})$, $P(L, \bar{s})$ *Laplace-transzformáltak* helyére bevezetjük a

$$\bar{P}(0, j\bar{\omega}), \quad \bar{Q}(L, j\bar{\omega}), \quad \bar{Q}(0, j\bar{\omega}), \quad \bar{P}(L, j\bar{\omega})$$

állandósult állapotban fennálló színuszosan változó nyomások és tömegáramok komplex amplitúdóit, úgy az így kapott

(5.1)

$$\begin{bmatrix} \bar{P}(L, j\bar{\omega}) \\ \frac{c}{A}\bar{Q}(0, j\bar{\omega}) \end{bmatrix} = \begin{bmatrix} \frac{1}{\operatorname{ch} j\bar{\omega}\varphi(j\bar{\omega})} & -\varphi(j\bar{\omega}) \operatorname{th} j\bar{\omega}\varphi(j\omega) \\ \frac{1}{\varphi(j\bar{\omega})} \operatorname{th} j\bar{\omega}\varphi(j\bar{\omega}) & \frac{1}{\operatorname{ch} j\bar{\omega}\varphi(j\bar{\omega})} \end{bmatrix} \begin{bmatrix} \bar{P}(0, j\bar{\omega}) \\ \frac{c}{A}\bar{Q}(L, j\bar{\omega}) \end{bmatrix}$$

mátrixegyenlet a ki- és bemenőjelek komplex amplitúdói közti általános összefüggést fejezi ki. Ha még bevezetjük a $Z(\bar{s})$ operátoros impedancia helyére a $Z(j\bar{\omega})$ komplex impedanciát, úgy az egyes konkrét esetekre vonatkozó átviteli karakterisztikák egyszerűen adódnak.

(3.14) és (4.3)-ból kapjuk a nyomás-nyomás átviteli karakterisztikát:

$$\frac{\bar{P}(L, j\bar{\omega})}{\bar{P}(0, j\bar{\omega})} = \frac{1}{\operatorname{ch}[j\bar{\omega}\varphi(j\bar{\omega})] + \frac{\varphi(j\bar{\omega})}{Z(j\bar{\omega})} \operatorname{sh}[j\bar{\omega}\varphi(j\bar{\omega})]}. \tag{5.2}$$

(3.14) és (4.4)-ből kapjuk a nyomás-tömegáram átviteli karakterisztikát:

$$\frac{\bar{P}(L, j\bar{\omega})}{\frac{c}{A}\bar{Q}(0, j\bar{\omega})} = \frac{1}{\frac{1}{\varphi(j\bar{\omega})} \operatorname{sh}[j\bar{\omega}\varphi(j\bar{\omega})] + \frac{1}{Z(j\bar{\omega})} \operatorname{ch}[j\bar{\omega}\varphi(j\bar{\omega})]}. \tag{5.3}$$

ahol

$$\varphi(j\bar{\omega}) = \frac{1}{\sqrt{1 - \frac{2J_1\left(j^{3/2}\sqrt{\frac{\bar{\omega}}{\tau}}\right)}{j\sqrt{\frac{j\bar{\omega}}{\tau}} J_0\left(j^{3/2}\sqrt{\frac{\bar{\omega}}{\tau}}\right)}}} \tag{5.4}$$

természetesen egyszerűen felírhatók a gyakorlatilag kisebb jelentőségű egyéb átviteli karakterisztikák is. (5.2) alapján számítható az ún. színuszos nyomásgerjesztés esete, mikor a csővezeték elejére kapcsolt nyomás ismeretében a nyomást határozzuk meg a csővezeték végén. (5.3) alapján számítható az ún. színuszos tömegáramgerjesztés esete, mikor a csővezeték elejére betáplált tömegáram ismeretében a nyomást határozzuk meg a csővezeték végén.

A gyakorlatban az amplitúdókarakterisztikára van szükség, mely az átviteli karakterisztika abszolút értéke. Az (5.2), (5.3) kifejezések abszolút értéke könnyen meghatározható, mivel a bennük szereplő *komplex argomentumú Bessel-függvények* az ún. *Kelvin-függvények* valós és képzetes részének értékei grafikonból [4], vagy táblázatból [6] kivehetők.

A gyakorlatban igen fontos kis csillapítású vagy nagy frekvenciás határesetben (5.2), (5.3) kifejezések igen egyszerűen kezelhető alakra hozhatók. A *Kelvin-függvények* aszimptotikus viselkedésére jellemző, hogy ha $\sqrt{\frac{\bar{\omega}}{\tau}} \gg 1$, úgy

$$jJ_0\left(j^{3/2}\sqrt{\frac{\bar{\omega}}{\tau}}\right) \approx J_1\left(j^{3/2}\sqrt{\frac{\bar{\omega}}{\tau}}\right) \tag{5.5}$$

és jó közelítéssel

$$\varphi(j\bar{\omega}) = \frac{1}{\sqrt{1 - 2\sqrt{\frac{\tau}{j\bar{\omega}}}}}. \tag{5.6}$$

Ezt binomális sorba fejtve az első három tagig kapjuk, hogy

$$\varphi(j\bar{\omega}) = 1+\sqrt{\frac{\tau}{j\bar{\omega}}}+\frac{3\tau}{2j\bar{\omega}}. \tag{5.7}$$

$j\bar{\omega}$-val szorozva adódik, hogy

$$j\bar{\omega}\varphi(j\bar{\omega}) = j\bar{\omega}+\sqrt{j\bar{\omega}\tau}+\frac{3\tau}{2} = j\bar{\omega}+\sqrt{\frac{\tau\bar{\omega}}{2}}(1+j)+\frac{3\tau}{2} \tag{5.8}$$

(5.8) valós és képzetes része tehát

$$\alpha(\bar{\omega}) = \frac{3\tau}{2}+\sqrt{\frac{\tau\bar{\omega}}{2}},\quad \beta(\bar{\omega}) = \bar{\omega}+\sqrt{\frac{\tau\bar{\omega}}{2}}. \tag{5.9}$$

Tekintsük először a nyomásgerjesztés esetét. (5.2) jobb oldalán álló tört nevezőjére írható tehát, hogy

$$\operatorname{ch}(\alpha+j\beta)+\frac{1}{Z}\left(\frac{\beta}{\bar{\omega}}-j\frac{\alpha}{\bar{\omega}}\right)\operatorname{sh}(\alpha+j\beta) = \operatorname{ch}\alpha\cos\beta+j\operatorname{sh}\alpha\sin\beta+ \tag{5.10}$$

$$+\frac{1}{Z\bar{\omega}}(\beta-j\alpha)(\operatorname{sh}\alpha\cos\beta+j\operatorname{ch}\alpha\sin\beta).$$

Az egyszerűség kedvéért tekintsük azt a gyakorlatban legfontosabb esetet, mikor a $Z$ impedancia valós, azaz tiszta ohmos. Ekkor (5.10)-et valós és képzetes részre bontva az alábbi kifejezés adódik.

$$\operatorname{ch}\alpha\cos\beta+\frac{1}{Z\bar{\omega}}(\beta\operatorname{sh}\alpha\cos\beta+\alpha\operatorname{ch}\alpha\sin\beta)+j[\operatorname{sh}\alpha\sin\beta+ \tag{5.11}$$

$$+\frac{1}{Z\bar{\omega}}(\beta\operatorname{ch}\alpha\sin\beta-\alpha\operatorname{sh}\alpha\cos\beta)].$$

A nyomás-nyomás amplitúdókarakterisztika tehát (5.11) abszolút értékének reciproka. Hosszadalmas elemi számolással az alábbi végeredmény adódik:

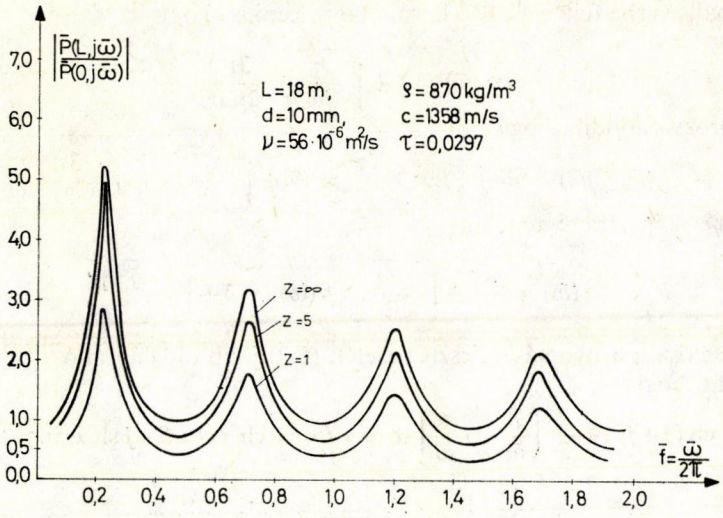
$$\left|\frac{\bar{P}(L,j\bar{\omega})}{\bar{P}(0,j\bar{\omega})}\right| = \left[\operatorname{ch}^2\alpha-\sin^2\beta+\frac{\beta}{Z\bar{\omega}}\operatorname{sh}2\alpha+\frac{\alpha}{Z\bar{\omega}}\sin 2\beta+\frac{\alpha^2+\beta^2}{Z^2\bar{\omega}^2}(\operatorname{sh}^2\alpha+\sin^2\beta)\right]^{-1/2} \tag{5.12}$$

(5.9) és (5.12) figyelembevételével a csővezeték nyomás-nyomás rezonancia görbéi igen egyszerűen meghatározhatók és a $Z$ lezáró impedancia paramétereként jellemző görbesereg rajzolható. Ezt egy adott elrendezés esetén az alábbi 5. ábra tünteti fel az $\bar{f}=\frac{\bar{\omega}}{2\pi}$ relatív frekvencia függvényében. A laboratóriumi mérések az elméleti rezonancia görbékkel való igen jó egyezést mutatták.

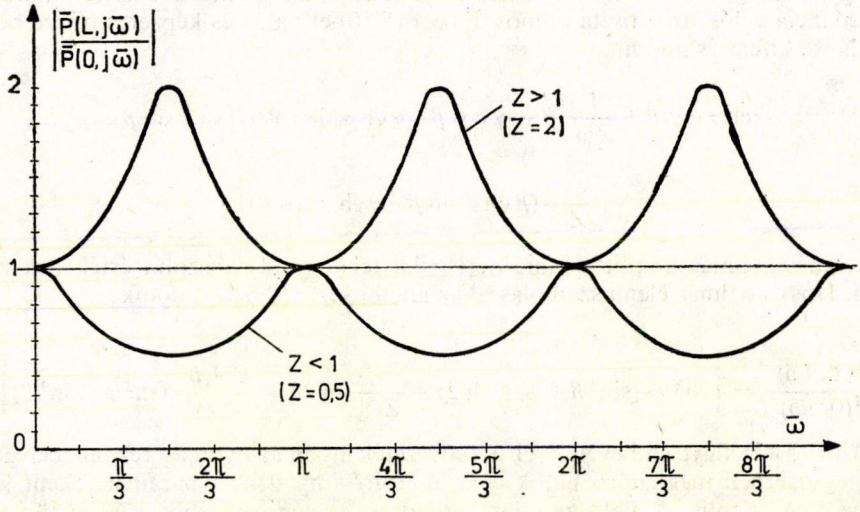
(5.12)-ből explicit kiszámíthatók az ideális, csillapítatlan rendszer sajátfrekvenciái. Ha $\tau=0$, akkor

$$\alpha = 0,$$

$$\beta = \bar{\omega},$$




*5. ábra*




*6. ábra*

így

$$\left|\frac{\bar{P}(L, j\bar{\omega})}{\bar{P}(0, j\bar{\omega})}\right| = \frac{1}{\sqrt{1-\sin^2\bar{\omega}+\frac{1}{Z^2}\sin^2\bar{\omega}}} = \frac{1}{\sqrt{1-\left[1-\frac{1}{Z^2}\right]\sin^2\bar{\omega}}}. \tag{5.13}$$

Ha $Z=1$, akkor (5.13) értéke 1. Ha $Z>1$, akkor az $\bar{\omega}_k$ rezonanciafrekvenciák azok a frekvenciák, melyek a gyök alatti kifejezést minimalizálják.

Tehát

$$\sin^2 \bar{\omega}_k = 1$$

és

$$\bar{\omega}_k = \left(\frac{2k+1}{2}\right)\pi, \quad k = 0, 1, 2, \ldots . \tag{5.14}$$

Ezeken a frekvenciákon a rezonanciagörbe a $Z$ értéket veszi fel. Ha $Z<1$, akkor hasonló meggondolások alapján

$$\sin^2 \bar{\omega}_k = 0$$

és

$$\bar{\omega}_k = k\pi, \quad k = 0, 1, 2, \ldots . \tag{5.15}$$

Ezeken a frekvenciákon a rezonanciagörbe az 1 értéket veszi fel. A csillapítatlan rendszer rezonanciagörbéit tünteti fel a 6. ábra.

Vizsgáljuk most a tömegáramgerjesztés esetét. Ekkor teljesen hasonló számolással kapjuk, hogy valós $Z$ felvételével a nyomás-tömegáram amplitúdókarakterisztika (rezonanciagörbe) az alábbi alakban írható

$$\left|\frac{\bar{P}(L, j\bar{\omega})}{\frac{c}{A}\bar{Q}(0, j\bar{\omega})}\right| = \left\{\frac{\bar{\omega}^2}{\alpha^2+\beta^2}(\mathrm{sh}^2\alpha+\sin^2\beta)+\frac{1}{Z^2}(\mathrm{ch}^2\alpha-\sin^2\beta)+\right. \tag{5.16}$$

$$\left.+\frac{\bar{\omega}}{(\alpha^2+\beta^2)Z}(\beta\,\mathrm{sh}\,2\alpha-\alpha\sin 2\beta)\right\}^{-1/2}$$

Az ideális csillapítatlan csővezetékre:

$$\left|\frac{\bar{P}(L, j\bar{\omega})}{\frac{c}{A}Q(0, j\bar{\omega})}\right| = \frac{1}{\sqrt{\sin^2\bar{\omega}+\frac{1}{Z^2}(1-\sin^2\bar{\omega})}} = \frac{Z}{\sqrt{1+(Z^2-1)\sin^2\bar{\omega}}}. \tag{5.17}$$

Ha $Z=1$, úgy a rezonanciagörbe most is az azonosan 1 értéket állítja elő. A rezonanciafrekvenciák:

$$\begin{aligned} &\bar{\omega}_k = k\pi, \quad \text{ha} \quad Z > 1 \\ &\qquad\qquad\qquad\qquad k = 0, 1, 2, \ldots, \\ &\bar{\omega}_k = \left(\frac{2k+1}{2}\right)\pi, \quad \text{ha} \quad Z < 1 \end{aligned} \tag{5.18}$$

A KIRK—YOUNG [6] táblázat alapján megállapítható, hogy ha

$$\sqrt{\frac{\bar{\omega}}{\tau}} \geqq 10$$

úgy (5.5) nagy pontossággal teljesül, és (5.6) binomális sorának az (5.7)-ben már figyelembe nem vett negyedik tagja abszolút értékben igen kicsiny az egységhez képest.

Igen lényeges, hogy kis csillapítás esetén ($\tau=10^{-3}$), a gyakorlatban alkalmazott méretű csővezetékekre, a probléma a gyakorlatban fellépő teljes frekvencia-tarto-

mányban ($1 < f < \infty$) nagyfrekvenciás határesetként tárgyalható, mert számításaink alapján kiderült, hogy az alkalmazott közelítések már $f = 1$ Hz-nél is igen jók.

Közepes és nagy csillapítás esetén az alacsony frekvenciák tartományában a pontos összefüggésekből kiindulva kell az amplitúdókarakterisztikákat meghatározni. A gyakorlatban alkalmazott csővezeték méreteknél, mint ahogy azt a számítások mutatják — a csillapítástól függően $f = 20$ Hz és $f = 30$ Hz közé esik az a frekvencia, melyre a fentebb tárgyalt közelítés még megengedett.

## IRODALOM

[1] BLAHÓ—GRUBER, *Folyadékok mechanikája* (Tankönyvkiadó, 1971).
[2] CSÁKI, F., *Szabályozások dinamikája* (Akadémiai Kiadó, 1974).
[3] DOETSCH, G., *Anleitung zum praktischen Gebrauch der Laplace Transformation* (orosznyelvű kiadás, Moszkva, 1965).
[4] EMDE—JAHNKE—LÖSCH, *Tafeln höherer Funktionen* (Teubner Verlag, 1960).
[5] GYITKIN—PRUDNYIKOV, *Szpravocsnyik po operácionnomu iszcsiszlenyiu* (Izdatyelsztvo vüszsaja skola, Moszkva, 1965).
[6] KIRK—YOUNG, *Bessel functions, Part IV. Kelvin functions* (orosznyelvű kiadás, Vücsiszlityelnüj centr, AN. SZSZSZSK, Moszkva, 1966).
[7] LALLEMENT, J., "Comportement dynamique des lignes hydrauliques Mechanique Materiaux Elektricite", Párizs, 1977. október.
[8] SIMONYI, K., *Elméleti villamosságtan,* (Tankönyvkiadó, 1981).
[9] SZENTMÁRTONY, T., *Folyadékok mechanikája I.* (Tankönyvkiadó, 1976).
[10] WAGNER, K., W., *Operatorenrechnung und Laplacesche Transformation* (Johann Ambrosius Barth Verlag, 1965).

(Beérkezett: 1983. február 3.)

FÉNYES TAMÁS
MTA MATEMATIKAI KUTATÓ INTÉZETE
1053 BUDAPEST, REÁLTANODA U. 13—15.

HARKAY GÁBOR
BÁNKI DONÁT GÉPIPARI MŰSZAKI FŐISKOLA
1081 BUDAPEST, NÉPSZÍNHÁZ U. 8.

## ÜBER DAS INTEGRO- DIFFERENTIALGLEICHUNGSSYSTEM DER SIGNALÜBERGABE IN DER HYDROSTATISCHEN ROHRLEITUNG

T. FÉNYES und G. HARKAY

Der Beitrag zeigt eine Methode auf die dynamische Prüfung der hydrostatischen Rohrleitung. Aus der laminarischen Strömung der wahrhaftigen Flüssigkeit (Hydrauliköl) ausgehend-die Rohrleitung elastisch gesehen- bespricht der Beitrag, solche allgemeine Gleichungen, die geeignet sind auf die Lösung der Signalübergabe in der Rohrleitung. Mit dieser Methode kann man sowohl die schlagartige Prozesse (transitorische Prozesse) als auch periodische Zustände (ständige sinus-Prozesse) in der Rohrleitung untersuchen.

*Alkalmazott Matematikai Lapok* **10** (1984) 173—176

# A NEWTON—KERNER-FÉLE POLINOM-GYÖKKERESŐ ELJÁRÁS EGY ÁLTALÁNOSÍTÁSA

VARGA GYULA

Budapest

A cikk általánosítja a *Kerner-féle polinom-gyökkereső eljárást*, hogy alkalmas legyen olyan polinomok összes gyökének egyidejű kiszámítására, amelyeknek egyik gyöke többszörös.

## 1. Bevezetés

Az összlépéses polinom-gyökkereső eljárásoknak nagy jelentőségük van a számítástechnikában. Egy polinom valamennyi gyökének egyidejű kiszámítása kiküszöböli az egymás után kiszámított, és a kerekítési hibák felhalmozódása miatt egyre pontatlanabb gyökök pontosításának szükségességét.

KERNER [1] olyan, a *Newton—Raphson-módszer* alkalmazásán alapuló eljárást adott, amely, kiindulva egy polinom összes gyökének valamely alkalmas közelítéséből, egyidejűleg számítja ki az adott polinom valamennyi gyökét. A *Vieta-féle gyökfüggvények* (a gyökök szimmetrikus függvényei) segítségével kapott iterációs eljárásának konvergenciája másodrendű. A szerző eljárását csupa egyszeres gyökökkel rendelkező polinomokra dolgozta ki.

Az alábbiakban, a *Vieta-féle gyökfüggvények* egy egyszerű tulajdonságát kihasználva, a fenti eljárásnak egy olyan általánosítását adjuk meg, amelynek segítségével olyan polinomok összes gyökének egyidejű kiszámítása is lehetséges, amelyeknek az egyszeres gyökök mellett egyetlen, ismert multiplicitású többszörös gyökük is van. Az általánosított gyökkereső eljárás másodrendű konvergenciáját is belátjuk.

## 2. Az eljárás leírása

Legyen

$$a(x) = \sum_{i=1}^{n+1} a_i x^{i-1} \tag{2.1}$$

$n$-edfokú komplex polinom, amelynek $m$ számú gyöke van ($m<n$), ezek közül $m-1$ egyszeres, és egy $k_m$-szeres ($n=m-1+k_m$). Legyen $a_{n+1}=1$, és $\mathbf{a}=(a_1, \ldots, a_n)^T$ a polinom együtthatóinak vektora.

Feladatul tűzzük ki a polinom összes gyökének egyidejű kiszámítását, vagyis a polinomnak

$$a(x) = (x-r_m^*)^{k_m} \prod_{i=1}^{m-1} (x-r_r^*) \tag{2.2}$$

alakban való tényezőkre bontását.

Legyen $\mathbf{r}^*=(r_1^*, \ldots, r_{m-1}^*)^T$ az egyszeres gyökök vektora, és legyen $\mathbf{r}=(r_1, \ldots, r_{m-1})^T$ ennek valamely közelítése. Kössük ki a továbbiakban a következő egyenlőség teljesülését (ez pontos gyökökre triviálisan igaz):

$$r_1+\ldots+r_{m-1}+k_m \cdot r_m = -a_n. \tag{2.3}$$

Akkor nyilvánvalóan

$$r_m = r_m(\mathbf{r}) = -\frac{1}{k_m}(a_n+r_1+\ldots+r_{m-1}). \tag{2.4}$$

Legyen az $\mathbf{r}$-hez tartozó közelítő polinom

$$b(\mathbf{r}, x) = \sum_{i=1}^{n+1} b_i(\mathbf{r}) x^{i-1} = (x-r_m)^{k_m} \prod_{i=1}^{m-1} (x-r_i), \tag{2.5}$$

ahol $b_{n+1}=1$, és $\mathbf{b}(\mathbf{r})=(b_1(\mathbf{r}), \ldots, b_n(\mathbf{r}))^T$ a közelítő polinom együtthatóinak vektora.

Kitűzött feladatunk értelmében meg kell oldanunk a

$$\mathbf{b}(\mathbf{r})-\mathbf{a} = \mathbf{0} \tag{2.6}$$

egyenletrendszert. Az egyenleteket az egyes komponensekre felírva kapjuk:

$$b_i(\mathbf{r})-a_i = 0, \quad (i = 1, \ldots, n), \tag{2.7}$$

amelyből láthatjuk, hogy egy $n$ egyenletből álló $m-1$ ismeretlenes túlhatározott nemlineáris egyenletrendszert kaptunk. Ez az egyenletrendszer egyértelműen megoldható, ha az $a(x)$ polinomnak $m-1$ egyszeres és egy $k_m$-szeres gyöke van, egyébként általában ellentmondásos. A (2.3)-ból látható, hogy az egyenletrendszer utolsó egyenlete triviálisan teljesül.

A (2.6) egyenletrendszert a *Newton—Raphson-módszer* egy módosított változatának segítségével oldjuk meg. Határozzuk meg az egyenletrendszer téglalapalakú *Jacobi-mátrixának* elemeit. Legyen

$$B_k(\mathbf{r}, x) = \frac{\partial b(\mathbf{r}, x)}{\partial r_k} = \sum_{i=1}^{n} \frac{\partial b_i(\mathbf{r})}{\partial r_k} x^{i-1} \quad (k = 1, \ldots, m-1), \tag{2.8}$$

akkor fennáll

$$B_k(\mathbf{r}, x) = -(x-r_m)^{k_m} \prod_{\substack{i=1 \\ i\neq k}}^{m-1} (x-r_i)+(x-r_m)^{k_m-1} \prod_{i=1}^{m-1} (x-r_i), \tag{2.9}$$

és látható, hogy az

$$\mathbf{U} = \left[\frac{\partial b_i(\mathbf{r})}{\partial r_k}\right] \quad (i = 1, \ldots, n), \quad (k = 1, \ldots, m-1),$$

*Jacobi-mátrix* elemei (2.9)-ből az $x=r_k$ helyettesítéssel adódnak a

$$B_k(\mathbf{r}, r_k) = -(r_k-r_m)^{k_m} \prod_{\substack{i=1 \\ i\neq k}}^{m-1} (r_k-r_i) \tag{2.10}$$

polinomnak mint $r_k$ polinomjának együtthatóiként.

Az $\mathbf{U}$ *Jacobi-mátrix* egy balinverzét egyszerűen megkaphatjuk az

$$\mathbf{U}^{-L} = [w_{k,i}], \quad w_{k,i} = \frac{r_k^{i-1}}{B_k(\mathbf{r}, r_k)} \quad (k = 1, \ldots, m-1), \quad (i = 1, \ldots, n) \tag{2.11}$$

alakban. A (2.6) egyenletrendszer megoldására szolgáló iterációs eljáráshoz a korrekciós tagot a

$$\Delta\mathbf{r} = \mathbf{U}^{-L}(\mathbf{a} - \mathbf{b}(\mathbf{r})) \tag{2.12}$$

képlettel adjuk meg. Ezt a $k$-adik komponensre felírva ($k=1, \ldots, m-1$), továbbá a képletben fellépő összegezéseket $a_{n+1}=b_{n+1}$ miatt $n$ helyett $n+1$-ig véve kapjuk az alábbi egyenlőséget:

$$\Delta r_k = \frac{\sum_{i=1}^{n+1} a_i r_k^{i-1} - \sum_{i=1}^{n+1} b_i(\mathbf{r}) r_k^{i-1}}{B_k(\mathbf{r}, r_k)} = \frac{a(r_k)}{B_k(\mathbf{r}, r_k)}. \tag{2.13}$$

Az egyenletrendszer megoldására szolgáló iterációs eljárás tehát, kiindulva valamely $\mathbf{r}^{(0)}$ közelítő vektorból, az alábbi képlettel adható meg:

$$r_k^{(l+1)} = r_k^{(l)} - \frac{a(r_k^{(l)})}{(r_k^{(l)} - r_m^{(l)})^{k_m} \prod_{\substack{i=1 \\ i \neq k}}^{m-1} (r_k^{(l)} - r_i^{(l)})} \quad (k = 1, \ldots, m-1), \quad (l = 0, 1, \ldots). \tag{14}$$

Az iterációból való kilépés feltételeként tekinthetjük a $\|\Delta\mathbf{r}\|_1 + |\Delta r_m| < m\varepsilon$ egyenlőtlenség teljesülését, ahol $\Delta r_m$ a (2.3) segítségével számítható ki.

Az iterációs eljárás konvergenciáját a komponensenként felírt

$$\varphi_k(\mathbf{r}) = r_k - \frac{a(r_k)}{(r_k - r_m)^{k_m} \prod_{\substack{i=1 \\ i \neq k}}^{m-1} (r_k - r_i)} \quad (k = 1, \ldots, m-1)$$

iterációs függvény $\mathbf{r}=\mathbf{r}^*$ helyen vett függvényértékének és első parciális deriváltjainak kiszámításával vizsgálhatjuk.

Az egyszerűen belátható

$$\varphi_k(\mathbf{r}^*) = r_k^*, \quad \left.\frac{\partial \varphi_k(\mathbf{r})}{\partial r_j}\right|_{\mathbf{r}=\mathbf{r}^*} = 0 \quad (k = 1, \ldots, m-1), \quad (j = 1, \ldots, m-1)$$

feltételek teljesüléséből következik, hogy ha $l \to \infty$, akkor $\mathbf{r}^{(l)} \to \mathbf{r}^*$ és vele együtt $r_m^{(l)} \to r_m^*$ másodrendben.

Azt az eredményt kaptuk tehát, hogy, bár az $a(x)$ polinom $k_m$-szeres gyökének közelítései ténylegesen nem vettek részt az iterációban, mégis megkaptuk a többivel együtt ezt a gyököt is a (2.4) felhasználásával.

### 3. Teszteredmények

A fentiekben leírt eljárást az alábbi valós együtthatós polinomokra próbáltuk ki az MTA CDC 3300 számítógépén:

(1) $$a(x) = (x-2)(x+3)(x-5)(x-4)^3, \quad E = 10^{-6},$$

az egyszeres gyökök kezdő közelítései:

$$1{,}8 \quad -2{,}9 \quad 5{,}2$$

továbbá (2.3) alapján a háromszoros gyök kezdő közelítése: 3,96̇.

Az iteráció során kapott korrekciók (a háromszoros gyök közelítéseinek korrekcióit (2.3) alapján számítottuk ki):

| $\Delta r_1$ | $\Delta r_2$ | $\Delta r_3$ | $\Delta r_4$ |
|---|---|---|---|
| 0,201 250 13 | −0,103 169 11 | −0,175 522 29 | 0,025 813 75 |
| −0,001 253 32 | 0,003 168 85 | −0,023 946 98 | 0,007 343 81 |
| 0,000 003 19 | 0,000 000 26 | −0,000 530 45 | 0,000 175 67 |
| 0,0 | 0,0 | −0,000 000 32 | 0,000 000 11 |

(2) $$a(x) = (x-1{,}3)(x+9{,}6)(x-71)(x+0{,}07)(x+12{,}5)^5, \quad \varepsilon = 10^{-6},$$

az egyszeres gyökök kezdő közelítései:

$$1{,}0 \quad -10{,}0 \quad 70{,}2 \quad -0{,}12$$

továbbá (2.3) alapján az ötszörös gyök kezdő közelítése: −12,19.

Az iteráció során kapott korrekciók (az ötszörös gyök közelítéseinek korrekcióit (2.3) alapján számítottuk ki):

| $\Delta r_1$ | $\Delta r_2$ | $\Delta r_3$ | $\Delta r_4$ | $\Delta r_5$ |
|---|---|---|---|---|
| 0,313 788 48 | 0,808 589 64 | 0,907 007 86 | 0,069 834 73 | −0,399 844 14 |
| −0,014 046 814 | −0,356 086 99 | −0,007 009 13 | −0,019 785 51 | 0,079 389 95 |
| 0,000 279 96 | −0,051 584 00 | 0,000 001 28 | −0,000 049 30 | 0,010 270 41 |
| −0,000 000 30 | −0,000 918 36 | 0,0 | 0,000 000 08 | 0,000 183 72 |

### IRODALOM

[1] KERNER, I. O., „Ein Gesamtschrittverfahren zur Berechnung der Nullstellen von Polynomen", *Numerische Mathematik* **8** (1966) 290—294.

(Beérkezett: 1983. március 2.)

VARGA GYULA
MTA SZÁMÍTÁSTECHNIKAI ÉS AUTOMATIZÁLÁSI KUTATÓ INTÉZET
1250 BUDAPEST, URI U. 49.

### ON A GENERALIZATION OF THE NEWTON—KERNER PROCEDURE

GY. VARGA

The paper gives a generalization of the *Newton—Kerner procedure* to make it suitable to the simultaneous computation of all zeros of polynomials which have a unique multiple zero too.

*Alkalmazott Matematikai Lapok* **10** (1984) 177—183

# PÁRHUZAMOS ALGORITMUS POLINOMOK MÁSODFOKÚ TÉNYEZŐKRE BONTÁSÁRA

VARGA GYULA

Budapest

A cikk egy párhuzamos eljárást ad meg valós együtthatós párosfokszámú polinomok másodfokú tényezőkre bontására valamennyi másodfokú tényező együtthatóinak egyidejű kiszámításával a *Newton—Kerner-eljárás* általánosításaként. A polinomnak egy többszörös másodfokú tényezője is lehet. Az eljárás konvergenciája másodrendű.

## 1. Bevezetés

Valós és komplex együtthatós polinomok faktorizálása rendszerint úgy történik, hogy a polinom egy vagy két gyökéhez tartozó első- vagy másodfokú gyöktényező kiszámítása után ezzel a tényezővel a polinomot elosztjuk, s a kapott alacsonyabbfokú polinomot hasonlóképpen bontjuk tovább. A *Bairstow-eljárás* [1] és általánosításai [2], [3], [4] egyszeres vagy ismert multiplicitású többszörös valós vagy konjugált komplex gyökpárokra illetőleg ismert multiplicitású többszörös valós gyökökre végzik el a fenti feladatot, de az algoritmusok, amelyek a *Newton—Raphson-eljáráson* alapulnak, polinomosztásokat végeznek, és így a kiszámított gyöktényezővel együtt a hányadospolinomot is megadják, valós polinomokra valós aritmetikával. Az ilyen szukcesszív eljárások hátránya, hogy az egymás után kiszámított gyöktényezők a kerekítési hibák felhalmozódása miatt egyre pontatlanabbakká válhatnak, és valamennyit javítani kell az eredeti polinom segítségével. KERNER [5] algoritmusa, amely ugyancsak a *Newton—Raphson-eljáráson* alapul, egyszerre végzi a csupa egyszeres gyökökkel rendelkező valós együtthatós polinomok összes gyökeinek kiszámítását csupa valós gyökök esetén valós, komplex gyökök esetén komplex aritmetikával. Ez egy fajta párhuzamos algoritmus, amely sikerrel alkalmazható, ha valamennyi gyök megfelelő közelítését ismerjük. Az eljárás a fent említett hátrányt kiküszöböli. Általánosításai közül megemlítjük a [6]-ot és [7]-et, amelyek ugyancsak az [5] célkitűzését valósítják meg egy, illetve két tetszőleges multiplicitású többszörös gyökkel is rendelkező polinomok esetére, ezenkívül [8]-at, amely WEIERSTRASS elgondolása alapján egy páros fokszámú polinom másodfokú tényezőkre való egyidejű felbontását végzi el. A felbontás egyes másodfokú tényezőinek lehet kétszeres gyökük, de különböző másodfokú tényezőknek közös gyökük nem lehet.

Az alábbiakban a *Kerner-féle,* illetve a [8]-ban leírt módszer általánosítását végezzük el a következő célkitűzés megvalósításával: Bontsuk fel másodfokú tényezőkre a párosfokszámú valós együtthatós, csupa egyszeres valós vagy konjugált komplex gyökpárokkal, ill. kétszeres valós gyökökkel és egy ismert multiplicitású többszörös valós vagy konjugált komplex gyökpárral rendelkező polinomokat a másodfokú tényezők együtthatóinak (amelyeknek valamilyen közelítését ismerjük) egyidejű kiszámításával ugyancsak a *Newton—Raphson-eljárás* alkalmazásával. Ez az általánosítás szintén párhuzamos algoritmust fog eredményezni.

## 2. A faktorizálási eljárás leírása

Legyen

$$A(x) = \sum_{i=1}^{2n+1} a_i x^{i-1} \tag{2.1}$$

$2n$-edfokú valós együtthatós polinom, $a_{2n+1}=1$ és $a_1 \neq 0$, legyen továbbá a polinom együtthatóiból képezett vektor

$$\mathbf{a} = (a_1, \ldots, a_{2n})^T.$$

Fenti célkitűzésünk szerint a polinomot

$$A(x) = (x^2 + r^*_{2m}x + r^*_{2m-1})^{k_m} \prod_{i=1}^{m-1} (x^2 + r^*_{2i}x + r^*_{2i-1}) \tag{2.2}$$

alakban akarjuk felbontani. A tényezők fokszámainak összeadásával adódik a $2n = 2m-2+2k_m$ egyenlőség.

Legyen

$$\mathbf{r}^* = (r^*_1, \ldots, r^*_{2m-2})^T$$

a pontos egyszeres másodfokú tényezők együtthatóinak vektora. A polinom (2.1) és (2.2) alakú előállításából $x$ megfelelő hatványainak összehasonlításával adódnak az alábbi egyenlőségek:

$$\sum_{i=1}^{m-1} r^*_{2i} + k_m r^*_{2m} = a_{2n}$$

$$\sum_{i=1}^{m-1} r^*_{2i-1} + k_m r^*_{2m-1} - \frac{k_m(k_m+1)}{2} r^{*2}_{2m} + k_m r^*_{2m} a_{2n} + \sum_{i=1}^{m-1} \left( r^*_{2i} \sum_{j=i+1}^{m-1} r^*_{2j} \right) = a_{2n-1}. \tag{2.3}$$

Ezekből $r^*_{2m}$ és $r^*_{2m-1}$ explicite is kifejezhető $r^*_{2m} = r^*_{2m}(\mathbf{r}^*)$, ill. $r^*_{2m-1} = r^*_{2m-1}(\mathbf{r}^*)$ alakban.

Legyen az $\mathbf{r}^*$ vektor valamely közelítése

$$\mathbf{r} = (r_1, \ldots, r_{2m-2})^T.$$

Kössük ki a (2.3) egyenletek teljesülését az $\mathbf{r}$ közelítő vektorra valamint $r_{2m-1}$-re és $r_{2m}$-re is.

Legyen az $A(x)$ polinom $\mathbf{r}$-hez tartozó közelítése

$$B(\mathbf{r}, x) = \sum_{i=1}^{2n-1} b_i(\mathbf{r}) x^{i-1}, \quad b_{2n+1} = 1, \tag{2.4}$$

a közelítő polinom együtthatóinak vektora

$$\mathbf{b}(\mathbf{r}) = (b_1(\mathbf{r}), \ldots, b_{2n}(\mathbf{r}))^T, \tag{2.5}$$

és legyen a közelítő polinom szorzatalakja

$$B(\mathbf{r}, x) = (x^2 + r_{2m}x + r_{2m-1})^{k_m} \prod_{i=1}^{m-1} (x^2 + r_{2i}x + r_{2i-1}). \tag{2.6}$$

Kitűzött feladatunk szerint meg kell oldanunk a

$$\mathbf{b}(\mathbf{r})-\mathbf{a}=\mathbf{0}, \tag{2.7}$$

vagy komponensenként felírva a

$$b_i(\mathbf{r})-a_i=0, \quad (i=1, 2, \ldots, 2n) \tag{2.8}$$

nemlineáris egyenletrendszert. A fenti kikötésünk szerint az utolsó két egyenlet automatikusan teljesül. A felírt $2m-2$ ismeretlenes egyenletrendszer $2n$ egyenletből áll, tehát általában ellentmondásos, és csak akkor oldható meg egyértelműen, ha az $A(x)$ polinomnak létezik (2.2) alakú felbontása.

Az egyenletrendszert a *Newton—Raphson módszer* egy speciális változatával oldjuk meg. Számítsuk ki a $2n\times(2m-2)$-es $\mathbf{U}$ *Jacobi mátrix* elemeit. Deriváljuk a $B(\mathbf{r}, x)$ polinomot rendre $r_{2k-1}$ és $r_{2k}$ szerint $(k=1, \ldots, m-1)$:

$$\frac{\partial B(\mathbf{r}, x)}{\partial r_{2k-1}}=\sum_{i=1}^{2n+1}\frac{\partial b_i(\mathbf{r})}{\partial r_{2k-1}}x^{i-1}=$$

$$=k_m(x^2+r_{2m}x+r_{2m-1})^{k_m-1}\cdot\frac{\partial r_{2m-1}}{\partial r_{2k-1}}\cdot\prod_{i=1}^{m-1}(x^2+r_{2i}x+r_{2i-1})+ \tag{2.9}$$

$$+(x^2+r_{2m}x+r_{2m-1})^{k_m}\cdot\prod_{\substack{i=1\\ i\neq k}}^{m-1}(x^2+r_{2i}x+r_{2i-1}),$$

$$\frac{\partial B(\mathbf{r}, x)}{\partial r_{2k}}=\sum_{i=1}^{2n+1}\frac{\partial b_i(\mathbf{r})}{\partial r_{2k}}x^{i-1}=$$

$$=k_m(x^2+r_{2m}x+r_{2m-1})^{k_m-1}\left(\frac{\partial r_{2m}}{\partial r_{2k}}x+\frac{\partial r_{2m-1}}{\partial r_{2k}}\right)\cdot\prod_{i=1}^{m-1}(x^2+r_{2i}x+r_{2i-1})+ \tag{2.10}$$

$$+(x^2+r_{2m}x+r_{2m-1})^{k_m}x\cdot\prod_{\substack{i=1\\ i\neq k}}^{m-1}(x^2+r_{2i}x+r_{2i-1}).$$

Legyenek $s_k$ és $t_k$ az $x^2+r_{2k}x+r_{2k-1}$ tényező zérushelyei, helyettesítsük be valamelyiket (2.9)-be és (2.10)-be.

Legyen

$$B_k(s_k)=\sum_{i=1}^{2n+1}\frac{\partial b_i(\mathbf{r})}{\partial r_{2k-1}}s_k^{i-1}=(s_k^2+r_{2m}s_k+r_{2m-1})^{k_m}\prod_{\substack{i=1\\ i\neq k}}^{m-1}(s_k^2+r_{2i}s_k+r_{2i-1}), \tag{2.11}$$

akkor

$$\sum_{i=1}^{2n+1}\frac{\partial b_i(\mathbf{r})}{\partial r_{2k}}s_k^{i-1}=(s_k^2+r_{2m}s_k+r_{2m-1})^{k_m}s_k\prod_{\substack{i=1\\ i\neq k}}^{m-1}(s_k^2+r_{2i}s_k+r_{2i-1})=s_kB_k(s_k), \tag{2.12}$$

és látható, hogy a *Jacobi-mátrix* $2k-1$-edik, ill. $2k$-adik oszlopának elemei éppen a (2.11)-gyel definiált $B_k(x)$, ill. az $xB_k(x)$ polinom együtthatói. A téglalapalakú $\mathbf{U}$

*Jacobi-mátrix* közönséges értelemben vett inverzéről nem beszélhetünk, de kiszámíthatjuk egy balinverzét.

Legyenek a $\mathbf{V}(2m-2)\times 2n$-es mátrix elemei az alábbiak:

$$\begin{aligned} V_{2k-1,j} &= s_k^{j-1} \\ V_{2k,j} &= t_k^{j-1} \end{aligned} \qquad (k = 1, \ldots, m-1;\ j = 1, \ldots, 2n), \tag{2.13}$$

akkor fennáll a $\mathbf{VU}=\mathbf{W}$ mátrixegyenlőség, ahol

$$\mathbf{W} = \begin{bmatrix} \mathbf{W}_1 & & 0 \\ & \ddots & \\ 0 & & \mathbf{W}_{m-1} \end{bmatrix}, \tag{2.14}$$

és a főátló mentén elhelyezkedő

$$\mathbf{W}_k = \begin{bmatrix} B_k(s_k) & s_k B_k(s_k) \\ B_k(t_k) & t_k B_k(t_k) \end{bmatrix} \quad (k = 1, \ldots, m-1)$$

blokkok $t_k \neq s_k$ esetén nemszingulárisak, tehát $\mathbf{W}$ invertálható, és $\mathbf{U}$ egy balinverzét

$$\mathbf{U}^{-L} = \mathbf{W}^{-1}\mathbf{V} \tag{2.15}$$

alakban kaphatjuk meg, ahol $\mathbf{W}^{-1}$

$$\mathbf{W}_k^{-1} = \frac{1}{t_k - s_k} \begin{bmatrix} \dfrac{t_k}{B_k(s_k)} & -\dfrac{s_k}{B_k(t_k)} \\ -\dfrac{1}{B_k(s_k)} & \dfrac{1}{B_k(t_k)} \end{bmatrix} \quad (k = 1, \ldots, m-1)$$

alakú blokkokból áll.

Az iterációs eljáráshoz a korrekciót a

$$\Delta\mathbf{r} = \mathbf{U}^{-L}(\mathbf{a} - \mathbf{b}(\mathbf{r})) \tag{2.16}$$

képlettel megadva, $\Delta r_{2k-1}$-re és $\Delta r_{2k}$-ra a [8] iterációs képletével alakilag egyező, de tőle tartalmilag különböző eredményt kaptunk, mert itt a (2.11) alapján

$$B_k(x) = (x^2 + r_{2m}x + r_{2m-1})^{k_m} \prod_{\substack{i=1 \\ i \neq k}}^{m-1} (x^2 + r_{2i}x + r_{2i-1}). \tag{2.17}$$

A (2.16) részletes felírásával az alábbi iterációs képleteket kapjuk:

$$\begin{aligned} \Delta r_{2k-1} &= \frac{1}{t_k - s_k}\left[\frac{A(s_k)t_k}{B_k(s_k)} - \frac{A(t_k)s_k}{B_k(t_k)}\right] \\ \Delta r_{2k} &= \frac{1}{t_k - s_k}\left[\frac{A(t_k)}{B_k(t_k)} - \frac{A(s_k)}{B_k(s_k)}\right] \end{aligned} \qquad (k = 1, \ldots, m-1). \tag{2.18}$$

Az új közelítéseket, kiindulva valamely alkalmas $\mathbf{r}^{(0)}$ vektorból, az $\mathbf{r}^{(l+1)} = \mathbf{r}^{(l)} + \Delta\mathbf{r}$ képlettel kaphatjuk meg, az iteráció végrehajtásához szükséges, de abban

ténylegesen részt nem vevő $r_{2m-1}$ és $r_{2m}$ mennyiségek aktuális közelítő értékeit pedig, kikötésünknek megfelelően, az explicit alakban is felírható (2.3) képletek adják.

Az iterációból való kilépés feltételeként tekinthetjük pl. a $\|\Delta\mathbf{r}\|_1+|r_{2m-1}^{(l+1)}-r_{2m-1}^{(l)}|+|r_{2m}^{(l+1)}-r_{2m}^{(l)}|<2m\varepsilon$ egyenlőtlenség teljesülését.

Az eljárás konvergenciájának rendjét a $\varphi(\mathbf{r})=\mathbf{r}+\Delta\mathbf{r}$ iterációs függvény $\mathbf{r}^*$-beli viselkedésével vizsgálhatjuk. A

$$\varphi(\mathbf{r}^*)=\mathbf{r}^*$$

és az egymástól különböző $s_k^*$ és $t_k^*$ (az $x^2+r_{2k}^*x+r_{2k-1}^*$ tényező gyökei) esetére közvetlenül, egybeeső $s_k^*$ és $t_k^*$ esetére pedig határátmenet útján belátható

$$\left.\frac{\partial\varphi_i(\mathbf{r})}{\partial r_j}\right|_{\mathbf{r}=\mathbf{r}^*}=0 \quad (i=1,\ldots,2m-2),\quad (j=1,\ldots,2m-2)$$

egyenlőségek teljesülése miatt a konvergencia másodrendű, ha $B_k(s_k^*)$ és $B_k(t_k^*)$ zérustól különbözők, vagyis az egyes másodfokú tényezőknek nincs közös gyökük.

## 3. Megjegyzések

1. Az előző szakasz végén a konvergencia rendjére vonatkozó megállapítást ki kell egészítenünk. A (2.18) képletek közvetlenül csak akkor alkalmazhatók, ha az aktuális közelítő másodfokú tényező gyökei különbözők. Bár $t_k\to s_k$ esetén (2.15)-nek nincs véges határértéke, (2.16)-nak van, és $t_k=s_k$ esetén a (2.18) képletek helyett az alábbi

$$(3.1)\qquad \Delta r_{2k-1}=-s_k^2\frac{d}{dx}\left(\frac{A(x)}{xB_k(x)}\right)\Bigg|_{x=s_k},\quad \Delta r_{2k}=\frac{d}{dx}\left(\frac{A(x)}{B_k(x)}\right)\Bigg|_{x=s_k}$$

képleteket használva, a kétszeres gyököt tartalmazó másodfokú tényezőt is megkaphatjuk.

2. Kezdő közelítéseket természetesen csak az egyszeres másodfokú tényezők együtthatóira kell megadnunk, de az eljárás sikeres befejezése után a többszörös másodfokú tényező együtthatóit is eredményül kapjuk.

3. Az eljárás az egyes másodfokú tényezők együtthatóinak korrekcióit egymástól függetlenül számítja ki az illető másodfokú tényező diszkriminánsának előjelétől függően valós vagy komplex aritmetikával, ezenkívül az egyes másodfokú tényezők gyökeit is megadja.

## 4. Teszteredmények

Az eljárás FORTRAN szubrutinját az MTA CDC 3300 számítógépén próbáltuk ki az alábbi polinomok faktorizálásával:

1. $f(x)=x^{10}-4x^8-10x^7-3x^6-14x^5-46x^4-72x^3+4x^2+72x+72.$

A polinomnak két egyszeres és egy háromszoros másodfokú tényezőjéről tudunk. A kezdő közelítés az egyszeres másodfokú tényezők együtthatóihoz:

$$\mathbf{r}^{(0)}=(3{,}1;\ -3{,}8;\ 2{,}89;\ -2{,}1)^T,\quad \varepsilon=10^{-6}.$$

A háromszoros másodfokú tényező együtthatóinak számított kezdő közelítése $1{,}9\dot{6}$ és $1{,}74\dot{5}$. Az iterációs közelítő értékek az egyes tényezők együtthatóihoz:

| | | |
|---|---|---|
| 2,892 412 18 | 3,078 354 33 | 2,014 304 78 |
| −3,999 585 18 | −2,004 039 84 | 2,001 208 34 |
| 2,993 632 17 | 3,003 071 27 | 2,000 022 82 |
| −3,998 386 56 | −2,001 612 49 | 1,999 999 68 |
| 2,999 994 47 | 2,999 998 92 | 2,000 000 86 |
| −3,999 998 01 | −2,000 001 98 | 2,000 000 00 |
| 3,000 000 00 | 3,000 000 00 | 2,000 000 00 |
| −4,000 000 00 | −2,000 000 00 | 2,000 000 00. |

2. $f(x) = x^{10} - x^9 + x^8 + 22x^7 - 20x^6 + 16x^5 + 160x^4 - 128x^3 + 64x^2 + 384x - 256.$

A polinomnak két egyszeres és egy háromszoros másodfokú tényezőjéről tudunk. Az egyik egyszeres másodfokú tényező teljes négyzet. A kezdő közelítés az egyszeres másodfokú tényezők együtthatóihoz:

$$\mathbf{r}^{(0)} = (-1{,}2;\ 0{,}9;\ 3{,}8;\ 3{,}7)^T, \quad \varepsilon = 10^{-6}.$$

A háromszoros másodfokú tényező együtthatóinak számított kezdő közelítése $3{,}45\dot{8}$ és $-1{,}8\dot{6}$. Az iterációs közelítő értékek az egyes tényezők együtthatóihoz:

| | | |
|---|---|---|
| −0,876 376 32 | 3,948 493 70 | 4,099 395 74 |
| 1,089 064 49 | 4,016 194 53 | −2,035 086 34 |
| −0,985 346 83 | 4,028 260 97 | 4,003 698 46 |
| 0,993 609 35 | 4,012 180 22 | −2,001 929 85 |
| −0,999 561 15 | 4,001 297 15 | 4,000 030 31 |
| 0,999 417 12 | 4,005 958 7 | −2,000 004 33 |
| 0,999 999 11 | 4,000 002 97 | 4,000 000 12 |
| 0,999 998 60 | 4,000 001 40 | −2,000 000 00 |
| -1,000 000 00 | 4,000 000 00 | 4,000 000 00 |
| 1,000 000 00 | 4,000 000 00 | −2,000 000 00. |

## IRODALOM

[1] RALSTON, A., *Bevezetés a numerikus analizisbe* (Műszaki Könyvkiadó, Budapest, 1969).
[2] VARGA, GY., „Kétszeres valós gyökökkel rendelkező valós együtthatós polinomok faktorizálása", *Alk. Mat. Lapok* **4** (1978).
[3] VARGA, GY., „Többszörös valós gyökökkel rendelkező valós együtthatós polinomok faktorizálása", *Alk. Mat. Lapok* **6** (1980).

[4] VARGA, GY., „Többszörös gyökpárokkal rendelkező valós együtthatós polinomok faktorizálása", *Alk. Mat. Lapok* 7 (1981).
[5] KERNER, I. O., „Ein Gesamtschrittverfahren zur Berechnung der Nullstellen von Polynomen", *Numer. Math.* **8** (1966) 290—294.
[6] VARGA, GY., „A Newton—Kerner-féle polinom-gyökkereső eljárás egy általánosítása", *Alk. Mat. Lapok* **10** (1984) 173—176.
[7] VARGA, GY., "On a generalization of the Newton—Kerner procedure for simultaneous calculation of all zeros of polynomials", Working Paper, MTA SZTAKI, Budapest.
[8] FILIPPI, S., „Ein verallgemeinertes Bairstow—Verfahren zur gleichzeitigen Ermittlung aller Nullstellen eines Polynoms", *Beiträge Numer. Math.* **4** (1975) 83—93.

(Beérkezett: 1983. április 20.)

VARGA GYULA
MTA SZÁMÍTÁSTECHNIKAI ÉS AUTOMATIZÁLÁSI KUTATÓ INTÉZET
1250 BUDAPEST, URI U. 49.

## ON A PARALLEL ALGORITHM FOR DECOMPOSITION OF POLYNOMIALS INTO QUADRATIC FACTORS

GY. VARGA

The paper gives a procedure based on the *Newton—Raphson method* for simultaneous decomposition of polynomials of even degree with real coefficients having simple real or conjugate complex and double real roots respectively and a real or conjugate complex pair of roots of known multiplicity into quadratic factors. The procedure is the generalization of the *Newton—Kerner method* and its convergence is quadratic.

*Alkalmazott Matematikai Lapok* **10** (1984) 185—201

# AFFIN PROJEKCIÓK VÉGTELEN SZORZATAI NUMERIKUS SZEMPONTBÓL*

STACHÓ LÁSZLÓ

Szeged

A cikk a lineáris funkcionál egyenletrendszerek megoldására szolgáló *Kaczmarz típusú eljárások* viselkedésével foglalkozik *Hilbert térben*, a megoldással nem rendelkező rendszerekre való különös tekintettel. A klasszikus esetben, véges dimenziós egyenletrendszernél, éles becslést nyújt a *randomizált Kaczmarz iterációk* konvergencia-sebességére.

## 1. Bevezetés

Amikor az elektronikus számítógépek megjelenése lehetővé tette a kézi számolással már kezelhetetlenül nagy

$$\begin{matrix} a_{11}x_1+\ldots+a_{1n}x_n = b_1 \\ \vdots \qquad\qquad \vdots \qquad \vdots \\ a_{m1}x_1+\ldots+a_{mn}x_n = b_m \end{matrix} \tag{1.1}$$

lineáris egyenletrendszerek effektív megoldását, a futtatási tapasztalatok kezdettől fogva igen negatívnak bizonyultak a klasszikus algebrai megoldási módszereket illetően. NEUMANN JÁNOS, W. BERGMANN és D. MONTGOMERY [11] már 1946-ban külön munkát szenteltek az (1.1) típusú egyenletrendszerek akkor ismert numerikus megoldási eljárásainak a vizsgálatára, és kimutatták az eliminációs módszerek igen erős instabilitását: A kerekítési hibák 20 ismeretlentől kezdve már használhatatlanná tehetik a *Gauss-eliminációt*. A figyelem ettől kezdve az iterációs eljárásokra irányult, mégpedig elsősorban az (1.1) olyan speciális eseteire, amelyeknél az $(a_{ij})$ mátrixra súlyos korlátozó feltételek (pl. főátló-dominancia) teljesülnek. A jelen dolgozat célja az (1.1) rendszernek a *Kaczmarz projekcióiból* alkotott konvex kombinációk (nem feltétlenül stacionárius) iterálásával való megoldási módszereinek a vizsgálata és ezek néhány általánosítása végtelen dimenzióra.

$\mathbf{R}^n$-nel jelölve az $\mathbf{y}=(y_1, \ldots, y_n)$ alakú valós $n$-esek vektorterét az $\langle \mathbf{x}, \mathbf{y}\rangle = \sum_{k=1}^{n} x_k y_k$ skaláris szorzattal ellátva, (1.1) így írható:

$$\langle \mathbf{a}_k, \mathbf{x}\rangle = b_k \quad (k = 1, \ldots, m). \tag{1.2}$$

---

* A dolgozat annak a tanulmánynak egy átdolgozása, amelyet a szerző az *Április 4. Gépipari Művek* (Kiskunfélegyháza, Csanyi u. 2.) számára készített az *Április 4. Gépipari Művek* és a *József Attila Tudomány Egyetem Analizis Tanszéke* között 1980. december 5-én létrejött kutatási-fejlesztési szerződés 3. pontjának keretében.

Vagyis az (1.1) feladat megoldása abból áll, hogy meg kell keresnünk az

$$M_k = \{\mathbf{x}: \langle \mathbf{a}_k, \mathbf{x}\rangle = b_k\} \quad (k = 1, \ldots, m)$$

hipersíkok metszetét. 1937-ben KACZMARZ [10] a következő eljárást javasolta:

Kiindulva tetszőleges $\mathbf{x}_0 \in \mathbf{R}^n$ pontból, képezzük annak $\mathbf{x}_1$ merőleges vetületét $M_1$-re. Majd $\mathbf{x}_1$-et vetítjük $M_2$-re, így kapjuk $\mathbf{x}_2$-t. Ezt $M_3$-ra vetítjük, stb., — az $M_m$-re való vetítés után újra $M_1$-re való vetítéssel folytatjuk az $(\mathbf{x}_i)$ sorozatot. Azaz

$$\mathbf{x}_{i+1} = F_{1+\mathrm{mod}_m i}(\mathbf{x}_i) \quad (i = 0, 1, 2, \ldots), \tag{1.3}$$

ahol $F_1, \ldots, F_m$ rendre az $M_1, \ldots, M_m$ hipersíkokra való merőleges vetítést jelölik. Amennyiben $\bigcap_{k=1}^{m} M_k$ egyetlen pont, akkor az $(\mathbf{x}_i)$ sorozat geometriai rendben konvergál hozzá, az (1.1) egyetlen $\mathbf{x}^*$ megoldásához, tehát valamely $\mathbf{x}_0$-tól függő $q(\mathbf{x}_0) < 1$ konstanssal

$$\|\mathbf{x}_i - \mathbf{x}^*\| \leq q(\mathbf{x}_0)^i \quad (i = 1, 2, \ldots).$$

(Itt $\|\mathbf{v}\| = \langle \mathbf{v}, \mathbf{v}\rangle^{1/2}$ a $\mathbf{v}$ vektor hossza).

Később [3], [14] a *Kaczmarz eljárással* rokon további geometriai iterációs módszereket publikáltak. Azonban az általunk ismert irodalom nem foglalkozik a konvergencia-rátájuk pontos becslésével. A jelen dolgozat 3. részében kitérünk erre a problémára egy általánosabb szituációban, és — alapvető szempontok szerint — tovább nem javítható becslést adunk az $\mathbf{a}_k$ vektorok néhány paramétere segítségével.

Ez a következő tényeken alapszik: $F_k(\mathbf{x}) = \mathbf{x} + \mu_k(\mathbf{x})\mathbf{a}_k$, ahol $\mu_k(\mathbf{x}) = (b_k - \langle \mathbf{x}, \mathbf{a}_k\rangle)/\|\mathbf{a}_k\|^2$. Így ha $\bar{Q}$ jelöli egy $Q$ mennyiségnek a numerikus végrehajtás során kapott értékét, akkor $\bar{F}_k(\mathbf{x}) - F_k(\mathbf{x}) = (\bar{\mu}_k(\mathbf{x}) - \mu_k(\mathbf{x}))\mathbf{a}_k = \delta_k(\mathbf{x})\mathbf{u}_k$ alakú, ahol $\mathbf{u}_k$ az $\mathbf{a}_k$ irányú egységvektort jelöli. Ugyanakkor a 3.7 lemmából könnyen következik, hogy ha pl. $m = n$ és az (1.1) rendszer determinánsa nem nulla, akkor megadható olyan $M(= M(\mathbf{u}_1, \ldots, \mathbf{u}_n))$ korlát, amelyre tetszőleges $\mathbf{x}_0 \in \mathbf{R}^N$ és $\varepsilon > 0$ mellett az $\{F(\mathbf{x}_0): F \in \Sigma\}$ ponthalmaz átmérője kisebb $M \cdot \varepsilon \cdot (1 + \|\mathbf{x}_0\|)$-nál, ahol $\Sigma$ az összes $\mathbf{x} \mapsto F_k(\mathbf{x}) + \delta\mathbf{u}_k$ $(k = 1, \ldots, n;\ |\delta| \leq \varepsilon)$ alakú leképezések által generált félcsoport. Azaz, ha a kerekítési hibát egy $\varepsilon$ érték alatt tudjuk tartani az $\mathbf{x}_0$-ból kiinduló *Kaczmarz iteráció* során, akkor a kapott $\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2, \ldots$ sorozat átmérője $M \cdot \varepsilon \cdot (1 + \|\mathbf{x}_0\|)$ alatt marad. (Az $M$ konstans — durván szólva — annál kisebb, minél nagyobb a különböző $\mathbf{u}_k$ vektorok közti minimális szög.)

Az (1.2) probléma felvethető minden változtatás nélkül $\mathbf{R}^n$ helyett tetszőleges *H Hilbert tér* vektoraira (a $H$-beli skalárszorzatot szintén $\langle\ ,\ \rangle$-vel jelölve). Sőt a végtelen dimenziós kontextus hívja fel a figyelmet az (1.2) alábbi átfogalmazásának fontosságára

$$\mathbf{x} \in M_k \quad (k = 1, \ldots, m), \tag{1.4}$$

ahol $M_1, \ldots, M_n$ tetszőleges affin alterei $H$-nak (azaz olyan zárt halmazok, melyek bármely két pontjukkal együtt a rajtuk áthaladó egyenest is tartalmazzák).

Véges dimenziós tér esetén az (1.4) feladat mindig átírható (1.2) alakúra, hiszen egy $r$ dimenziós affin altér $\mathbf{R}^n$-ben mindig előáll $(n-r)$ hipersík metszeteként. Ezzel szemben végtelen dimenzióban mindig vannak olyan affin alterek, amelyek csak végtelen sok hipersík metszeteként jönnek létre. A tisztán geometriai (1.4) feladatkitű-

zést célszerűbb a következő funkcionális alakban is kimondani:

$$F_k(\mathbf{x}) = \mathbf{x} \quad (k = 1, \ldots, m), \tag{1.5}$$

ahol $F_k$ az $M_k$-ra való ortogonális projekció.

A klasszikus mechanika számos hullámegyenlete olyan extrémum feladat, amely ekvivalens valamilyen (1.5) típusú fixpontproblémával végtelen dimenziós *Hilbert térben* [2], [4]. Az operátor-algebrák elmélete is felvetett hasonló kérdéseket: fontos szerepet játszik NEUMANN JÁNOS [13] cikkében az az észrevétel, hogy amennyiben $L_1$, $L_2$ alterei $H$-nak és $P_k$ az $L_k$-ra való ortogonális projekció ($k=1, 2$), akkor minden $\mathbf{x}\in H$-ra a $P_1\mathbf{x}$, $P_2P_1\mathbf{x}$, $P_1P_2P_1\mathbf{x}$, ... sorozat konvergál $\mathbf{x}$-nek az $L_1\cap L_2$-re való $P\mathbf{x}$ merőleges vetületéhez. Figyelemre méltó tény, hogy itt a konvergencia már nem szükségképpen geometriai rendű, szemben a véges dimenziós esettel. Ez a nehézség magyarázza, hogy a *Kaczmarz-tétel* végtelen dimenziós analogonját, amely kimondja, hogy amennyiben $M_1\cap\ldots\cap M_m\neq\emptyset$, akkor minden $\mathbf{x}_0\in H$-ra az (1.3) pontsorozat konvergál $\mathbf{x}_0$ vetületéhez $M_1\cap\ldots\cap M_m$-re, csak 1958-ban bizonyította be BROWDER [2], HALPERIN [8] majd APOSTOL [1] kissé később más megközelítéssel mély operátorelméleti általánosításokat adtak BROWDER e tételére.

A szerző köszönettel tartozik POGÁNY CSABÁNAK, aki a problémakörre felhívta a figyelmét.

## 2. Problémafelvetés

Az (1.4) problémát az eddigi irodalom szinte kizárólag csupán abban a speciális esetében tárgyalja, amikor $M_1\cap\ldots\cap M_m\neq\emptyset$. A numerikus matematika szempontjából ugyanakkor nagy jelentőségű lenne az (1.4) különböző megoldási algoritmusainak teljes viselkedésvizsgálata éppen az $M_1\cap\ldots\cap M_m=\emptyset$ esetben, mivel még a klaszszikus (1.1) feladathoz sem rendelkezünk pillanatnyilag hatékony előrejelző eljárással a megoldás létezésére vonatkozóan. Másrészt az utóbbi időben kibontakozó döntéselmélet gyakran vet fel olyan problémákat, amelyek (1.4) alakra hozhatók, de ahol (szemben pl. a legtöbb fizikai alkalmazással) a megoldás létezése, ill. egyértelműsége a priori nem ismert, esetleg éppen ez a fő kérdés. A döntéselmélet az (1.4) típusú rendszerekhez általánosított megoldás gyanánt megkonstruál olyan $\varphi$ operátorokat, amelyek az alaptér altér $m$-eseihez úgy rendelnek vektorokat, hogy $M_1\cap\ldots\cap M_m\neq\emptyset$ esetén az $\mathbf{x}=\varphi(M_1, \ldots, M_m)$ választás mindig kielégítse (1.4)-et (vö. [15]). Az ilyen döntés-operátorok iteratív kiszámítási módszereinek, (ill. célszerű új operátoroknak) a kidolgozásához lényeges a véges affin altérrendszerekre való merőleges vetítések konvex lineáris kombinációiból alkotott iteratív limeszek vizsgálata.

Ezzel kapcsolatban a következő alapvető kérdések tehetők fel:

1. Adott $\mathbf{x}\in H$ pont mellett hogyan jellemezhető az összes lehetséges

$$F_{i_1}(\mathbf{x}),\ F_{i_2}F_{i_1}(\mathbf{x}),\ F_{i_3}F_{i_2}F_{i_1}(\mathbf{x}),\ldots$$

alakú pontsorozatok torlódási pontjaiból alkotott halmaz, ahol $1\leqq i_k\leqq m$ ($k=1,2,\ldots$) és az ($i_k$) indexsorozatban az $1, \ldots, m$ számok mindegyike végtelen sokszor előfordul?

2. Speciálisan, igaz-e, hogy amennyiben $\{\mathbf{0}\}=M_1\cap\ldots\cap M_m$, akkor minden $\mathbf{x}\in H$ kiindulópontra és minden az $1, \ldots, m$ számok mindegyikét végtelen sokszor

tartalmazó $i_1, i_2, \ldots$ indexsorozatra

$$F_{i_n}F_{i_{n-1}}\ldots F_{i_1}(\mathbf{x}) \to \mathbf{0} \quad (n \to \infty)?$$

3. Hogyan jellemezhető az összes lehetséges

$$G_1(\mathbf{x}), \quad G_2G_1(\mathbf{x}), \quad G_3G_2G_1(\mathbf{x}), \ldots$$

alakú sorozatok torlódási pontjainak halmaza, ahol $\mathbf{x}\in H$ tetszőleges pont, a $G_1, G_2, \ldots$ leképezések az $F_1, \ldots, F_m$ affin operátorok véges kompozícióiból álló véges konvex lineáris kombinációk úgy, hogy az $F_j$ operátorok mindegyike előfordul végtelen sok $G_k$ kifejezésében és a $\{G_1, G_2, \ldots\}$ család véges sok különböző elemből áll csak.

A numerikus matematika szempontjából különösen érdekesek az alábbi aspektusai az 1, 2, 3 problémáknak:

4. Mik a torlódási pontjai egy 3-beli $G_k\ldots G_1(\mathbf{x})$ $(k=1, 2, \ldots)$ sorozatnak, hogy ha a $G_1, G_2, \ldots$ operátorsorozat periodikus (azaz $G_{k+N}=G_k$ valamely $N$-re)?

5. A $G_1=G_2=\ldots=G$ esetben melyek azok az $\mathbf{x}$ pontok, amelyekre a $G^k(\mathbf{x})$ $(k=1, 2, \ldots)$ sorozat geometriai rendben konvergál, s hogyan becsülhető a konvergencia sebessége ekkor?

6. Ha $\{0\}=M_1\cap\ldots\cap M_m$ és $H=\mathbf{R}^n$, adható-e 1-nél kisebb közös felső korlát az összes olyan $F_{i_n}\ldots F_{i_2}F_{i_1}$ alakú leképezések lineáris operátor normájára, amelyeknél az $i_1, \ldots, i_n$ indexek között az $1, \ldots, m$ számok mindegyike legalább egyszer előfordul?

A felsorolt kérdések egyikét sem tárgyalta az eddigi szakirodalom a teljes általánosságában. A kimerítő válaszhoz legközelebb 4-nél jutottak, ahol is APOSTOL [1] egy tétele segítségével lényeges haladást (és véges dimenzióra végleges eredményt) sikerült elérni. Figyelemre méltó, hogy technikailag milyen problematikus még annak a plauzibilis ténynek a bizonyítása is, hogy az 1. tárgyát képező torlódási ponthalmaz mindig korlátos, ha $H=\mathbf{R}^n$. A jelen dolgozat — bár más sorrendben — érinti mind a hat kérdést, azonban (első sorban a végtelen dimenziós esetben) számos problémát hagy nyitottan, s újabbakat is felvet. Ezek mindegyike érdemesnek tűnik további elméleti vizsgálatokra. Fő eredményünk a 6. kérdésre adott pozitív válasz, olyan bizonyítással, amely pontos becslést is szolgáltat, s így lehetőséget nyújt 5., ill. 2. teljes megválaszolásához a $H=\mathbf{R}^n$ esetben 1. és 3-mal kapcsolatban azt sikerült megállapítanunk, hogy mely irányokban korlátosak a kérdéses torlódási pont halmazok véges dimenzióban.

## 3. Affin alterekre való vetítések random iteráltjai $\mathbf{R}^N$-ben

Legyenek $M_1, \ldots, M_m$ rögzített affin alterei $\mathbf{R}^N$-nek. Az $M_k$-ra való (merőleges) vetítést jelöljük $F_k$-val. Bevezetjük továbbá a $\mathbf{b}_k=F_k(\mathbf{0})$, $L_k=M_k-\mathbf{b}_k$, $P_k$: $\mathbf{x}\mapsto$ $\mapsto F_k(\mathbf{x})-\mathbf{b}_k$ jelöléseket. Azaz $L_k$ az $M_k$ halmaz $\mathbf{0}$-ba való párhuzamos eltolásából kapott altér, $P_k$ pedig az $L_k$-ra való ortogonális projekció lineáris operátora.

Tekintsük az $\mathbf{y}\in\mathbf{R}^N$ vektorok

$$\mathbf{y}=\mathbf{y}^{(1)}+\mathbf{y}^{(2)}, \quad \mathbf{y}^{(1)}\perp L, \mathbf{y}^{(2)}\in L, \quad \text{ahol} \quad L=L_1\cap\ldots\cap L_m$$

direkt felbontását. Vegyük észre, hogy $F_k^{(s)}(\mathbf{x})=(F_k(\mathbf{x}))^{(s)}$-et írva

$$F_k^{(1)}(\mathbf{x}) = F_k(\mathbf{x}^{(1)}) = F_k^{(1)}(\mathbf{x}^{(1)}),\ F_k^{(2)}(\mathbf{x}) = \mathbf{x}^{(2)},\ b_k^{(1)} = \mathbf{b}_k,\ \mathbf{b}_k^{(2)} = \mathbf{0}$$

tetszőleges $k=1, \ldots, m$ és $\mathbf{x}\in\mathbf{R}^N$ mellett.

Innen azonnal adódik, hogy bármely véges $i_1, \ldots, i_n$ indexsorozatra

$$F_{i_n}F_{i_{n-1}}\ldots F_{i_1}(\mathbf{x}) = F_{i_n}^{(1)}F_{i_{n-1}}^{(1)}\ldots F_{i_1}^{(1)}(\mathbf{x}^{(1)})+\mathbf{x}^{(2)}. \tag{3.1}$$

Így az általánosság lényeges megszorítása nélkül maradhatunk az $F_{i_n}^{(1)}\ldots F_{i_1}^{(1)}$ alakú szorzatok vizsgálatánál, vagy ami ezzel ekvivalens, feltehetjük, hogy

$$L_1\cap\ldots\cap L_m = \{\mathbf{0}\}. \tag{3.2}$$

Az is látszik azonnal, hogy

$$\begin{aligned} F_{i_n}F_{i_{n-1}}\ldots F_{i_1}(\mathbf{x}) &= P_{i_n}(P_{i_{n-1}}+(\ldots+(P_{i_1}\mathbf{x}+\mathbf{b}_{i_1})\ldots)+b_{i_{n-1}})+\mathbf{b}_{i_n} = \\ &= P_{i_n}\ldots P_{i_1}\mathbf{x}+P_{i_n}\ldots P_{i_2}\mathbf{b}_{i_1}+P_{i_n}\ldots P_{i_3}\mathbf{b}_{i_2}+\ldots+P_{i_n}\mathbf{b}_{i_{n-1}}+\mathbf{b}_{i_n} = \\ &= P_{i_n}\ldots P_{i_1}\mathbf{x}+F_{i_n}\ldots F_{i_1}(\mathbf{0}). \end{aligned} \tag{3.3}$$

Ezt a fejezetet első sorban az

$$S = \bigcup_{\substack{(i_1,i_2\ldots)\in J\\ \mathbf{x}\in|\mathbf{R}^N}} \{\text{az } [F_{i_n}\ldots F_{i_1}(\mathbf{x})]_{n=1}^{\infty} \text{ sorozat torlódási pontjai}\}$$

halmaz vizsgálatának szenteljük, ahol a továbbiakban $I$ azon indexsorozatok családja, amelyek az $1, \ldots, m$ számokból állnak, és ezek mindegyikét végtelen sokszor felveszik.

3.1. TÉTEL. Minden $(i_1, i_2, \ldots)\in I$ indexsorozatra fennáll (3.2) esetén

$$P_{i_n}\ldots P_{i_1}\to 0 \quad (n\to\infty). \tag{3.4}$$

3.2. KÖVETKEZMÉNY. (3.2) esetén

$$S = \bigcup_{(i_1,\ldots)\in I} \{\text{az } [F_{i_n}\ldots F_{i_1}(\mathbf{0})]_{n=1}^{\infty} \text{ sorozat torlódási pontjai}\}.$$

*Bizonyítás.* (3.3)-ból azonnal látszik, hogy ha a tétel igaz, akkor $\mathbf{x}\in\mathbf{R}^N$ és $(i_1, i_2, \ldots)\in I$ esetén az $[F_{i_n}\ldots F_{i_1}(\mathbf{x})]_{n=1}^{\infty}$ torlódási pontjai azonosak az $[F_{i_n}\ldots F_{i_1}(\mathbf{0})]_{n=1}^{\infty}$ sorozatéival, ami bizonyítja (3.2)-t.

A tétel bizonyításához belátjuk a következő, jóval erősebb állítást:

3.3. TÉTEL. Ha (3.2) áll, akkor létezik olyan $q<1$ konstans, hogy

$$\|P_{j_n}\ldots P_{j_1}\| \leqq q \quad \text{valahányszor} \quad \{j_1, \ldots, j_n\} = \{1, \ldots, m\}. \tag{3.5}$$

A (3.5) becslés valóban adja (3.4)-et: Ha ugyanis (3.5) teljesül, akkor

$$\|P_{i_k}\ldots P_{i_1}\| \leqq q^{\nu(k)} \quad (k = 1, 2, \ldots), \tag{3.6}$$

ahol $\nu(k)$ az a legnagyobb $r$ szám, amelyhez választható olyan $0=n_0<n_1<\ldots$ $\ldots<n_r=k$ alakú beosztása az $\{1, \ldots, k\}$ indexsorozatnak, hogy $\{i_j\colon n_{t-1}<j<n_t\}=$ $=\{1, \ldots, m\}$ legyen $t=1, \ldots, r$ mellett. Márpedig $(i_1, i_2, \ldots)\in I$ esetén $\nu(k)\to\infty$ $(k\to\infty)$.

*3.3 Bizonyítása* $m$ szerinti teljes indukcióval történik.

Az $m=1$ esetben a $q=0$ választás triviálisan megfelel.

Tegyük fel ezután, hogy tetszőleges $L_1', \ldots, L_{m-1}' \subset \mathbf{R}^N$ alterek mellett $L_1' \cap \ldots \cap L_{m-1}' = \{\mathbf{0}\}$ esetén van olyan $q_{m-1}(L_1', \ldots, L_{m-1}') < 1$ szám, hogy

$$\|P_{i_{n'}}' \ldots P_{i_1}'\| \leqq q_{m-1}(L_1', \ldots, L_{m-1}') \quad \text{valahányszor} \quad \{i_1, \ldots, i_{n'}\} = \{1, \ldots, m-1\} \tag{3.7}$$

ahol $P_k'$ az $L_k'$-ra való merőleges vetítést jelöli.

Legyenek $L_1, \ldots, L_m$ olyan alterei $\mathbf{R}^N$-nek, amelyek metszete az origó, és legyen $j_1, \ldots, j_n$ egy olyan véges indexsorozat, melyre $\{j_1, \ldots, j_n\} = \{1, \ldots, m\}$. Tekintsük az $n' = \max\{k: \{j_1, \ldots, j_k\} \neq \{1, \ldots, m\}\}$ indexet. Mivel az $L_k$-ra való $P_k$ projekció mindig kontrakció,

$$\|P_{j_n} \ldots P_{j_1}\| \leqq \|P_{j_{n'+1}} P_{j_{n'}} \ldots P_{j_1}\| \tag{3.8}$$

hiszen $n' < n$ szükségképpen. Mivel $\{j_1, \ldots, j_{n'+1}\} = \{1, \ldots, m\}$ az általánosság megszorítása nélkül vehetjük, hogy $\{j_1, \ldots, j_{n'}\} = \{1, \ldots, m\}$ és $j_{n'+1} = m$. Ezt feltéve, vezessük be az $L' = L_1 \cap \ldots \cap L_{m-1}$, ill. $L_k' = \{\mathbf{x} \in L_k: \mathbf{x} \in L'\}$ $(k=1, \ldots, m-1)$ altereket és a rájuk való $P'$, ill. $P_1', \ldots, P_{m-1}'$ projekciókat. Ekkor $0 = P'P_k' = P_k'P'$, $P_k = P' + P_k'$ $(k=1, \ldots, m-1)$ és így

$$P_{j_{n'}} \ldots P_{j_1} = P' + P_{j_{n'}} \ldots P_{j_1}'. \tag{3.9}$$

Teljesül továbbá $L_1' \cap \ldots \cap L_{m-1}' = \{\mathbf{0}\}$. Tehát (3.7) szerint

$$\|P_{j_{n'}}' \ldots P_{j_1}'\| \leqq q_{m-1}(L_1', \ldots, L_{m-1}') = q < 1.$$

Rögzítsünk egy tetszőleges $\mathbf{x} \in \mathbf{R}^N$ egységvektort, és jelöljük $\mathbf{y}$-nal a $P_{j_{n'}} \ldots P_{j_1}\mathbf{x}$ vektor vetületét a $P'\mathbf{x}$ és $P_{j_{n'+1}} \ldots P_{j_1}\mathbf{x}$ vektorok által kifeszített 2 dimenziós $K$ altérre. Vegyük észre, hogy $\mathbf{x}' = P'\mathbf{x}$, ill. $\mathbf{x}'' = (1-P')\mathbf{x}$ -et írva,

$$P_{j_{n'}} \ldots P_{j_1}\mathbf{x} = \mathbf{x}' + P_{j_{n'}}' \ldots P_{j_1}'\mathbf{x}'', \tag{3.10}$$

és így

$$P_{\mathbf{y}}' = \mathbf{x}' = P'(P_{j_{n'}} \ldots P_{j_1}\mathbf{x}) \quad \text{és} \quad P_{j_{n'+1}}\mathbf{y} = P_{j_{n'+1}} \ldots P_{j_1}\mathbf{x} = P_m(P_{j_{n'}} \ldots P_{j_1}\mathbf{x}).$$

Vagyis $\mathbf{y} - \mathbf{x}' \perp \mathbf{x}$ és $\mathbf{y} - P_m P_{j_{n'}} \ldots P_{j_1}\mathbf{x} \perp P_m P_{j_{n'}} \ldots P_{j_1}\mathbf{x}$. Legyen $\varrho = \|\mathbf{y} - \mathbf{x}'\|$, $\xi = \|\mathbf{x}'\|$ és vegyünk fel $K$-ban egy olyan $\mathbf{e}_1$, $\mathbf{e}_2$ ortonormált bázist, melyre $\mathbf{x} = \xi\mathbf{e}_1$ és $\mathbf{y} - \mathbf{x}' = \varrho\mathbf{e}_2$. Jelölje $\mathbf{e}$ a $P_m P_{j_{n'}} \ldots P_{j_1}\mathbf{x}$ irányába mutató egységvektort. Ekkor nyilván

$$\begin{aligned} \|P_{j_{n'+1}} \ldots P_{j_1}\mathbf{x}\| &= \|P_m\mathbf{y}\| = \langle P_m\mathbf{y}, \mathbf{e}\rangle = \langle \mathbf{y}, P_m\mathbf{e}\rangle = \langle \mathbf{y}, \mathbf{e}\rangle = \\ &= \langle \xi\mathbf{e}_1 + \varrho\mathbf{e}_2, \langle \mathbf{e}, \mathbf{e}_1\rangle\mathbf{e}_1 + \langle \mathbf{e}, \mathbf{e}_2\rangle\mathbf{e}_2\rangle = \xi\langle \mathbf{e}, \mathbf{e}_1\rangle + \varrho\langle \mathbf{e}, \mathbf{e}_2\rangle \leqq \\ &\leqq \xi|\langle \mathbf{e}, \mathbf{e}_1\rangle| + \varrho|\langle \mathbf{e}, \mathbf{e}_2\rangle|. \end{aligned}$$

Mivel $\mathbf{e}$ egységvektor, $|\langle \mathbf{e}, \mathbf{e}_1\rangle|^2 + |\langle \mathbf{e}, \mathbf{e}_2\rangle|^2 = 1$. Azaz valamely $t \in [0, \pi/2]$-re $|\langle \mathbf{e}, \mathbf{e}_1\rangle| = \cos t$ és $|\langle \mathbf{e}, \mathbf{e}_2\rangle| = \sin t$. A $t$ szögre vonatkozóan a következő (pontos) alsó becsléssel rendelkezünk:

$$\cos t \leqq \sup\{|\langle \mathbf{f}, \mathbf{e}_1\rangle|: \|\mathbf{f}\| = 1, \mathbf{f} \in L'\} \leqq \sup\{|\langle \mathbf{f}, \mathbf{g}\rangle|: \|\mathbf{f}\| = \|\mathbf{g}\| = 1, \mathbf{f} \in L', \mathbf{g} \in L_m\}$$

(ugyanis könnyen konstruálható tetszőleges $\mathbf{y} \in L_m$ egységvektorhoz $\mathbf{x}$ úgy, hogy $P_{j_{n'}} \ldots P_{j_1}\mathbf{x}/\|P_{j_{n'}} \ldots P_{j_1}\| = P_m P_{j_{n'}} \ldots P_{j_1}\mathbf{x}/\|P_m P_{j_{n'}} \ldots P_{j_1}\| = \mathbf{g}$ legyen). Tehát geometriai-

lag interpretálva,

$$t \leqq \alpha_m = \sphericalangle(L_m, L_1 \cap \ldots \cap L_{m-1})$$

ahol

$$\alpha_m = \arccos \sup\{|\langle \mathbf{f}, \mathbf{g}\rangle| : \|\mathbf{f}\| = \|\mathbf{g}\| = 1,\ \mathbf{f} \in L',\ \mathbf{g} \in L_m\}$$

az $L_m$ és $L'$ alterek által bezárt szög. Mivel most $L_m \cap L' = \{\mathbf{0}\}$, az egységgömb kompaktsága miatt $\alpha_m > 0$. Becsüljük ezután $\varrho$-t és $\xi$-t. $Q$-val jelölve a $K$-ra való vetítést,

$$\varrho = \|\mathbf{y} - \mathbf{x}'\| = \|QP_{j_{n'}} \ldots P_{j_1}\mathbf{x} - Q\mathbf{x}'\| = \|Q(P_{j_{n'}} \ldots P_{j_1}\mathbf{x} - \mathbf{x}')\| \leqq$$

$$\leqq \|P_{j_{n'}} \ldots P_{j_1}\mathbf{x} - \mathbf{x}'\| = ((3.10)\ \text{szerint}) = \|P'_{j_{n'}} \ldots P'_{j_1}\mathbf{x}''\| \leqq$$

$$\leqq \|P'_{j_{n'}} \ldots P'_{j_1}\| \cdot \|\mathbf{x}''\| \leqq q \cdot \|\mathbf{x}''\|.$$

Másfelöl $\xi = \|\mathbf{x}'\|$ és $\|\mathbf{x}'\|^2 + \|\mathbf{x}''\|^2 = \|\mathbf{x}\|^2 = 1$. Így valamely $s \in [0, \pi/2]$ és $\varepsilon \in [0, q]$ mellett

$$\xi = \cos s, \quad \varrho = \varepsilon \cdot \sin s \tag{3.11}$$

írható. Így $\|P_{j_n} \ldots P_{j_1}\|$ ($\leqq \|P_{j_{n'+1}} \ldots P_{j_1}\|$)-ra a következő felső becslést kapjuk:

$$\|P_{j_n} \ldots P_{j_1}\| \leqq \max\{\cos s \cdot \cos t + \varepsilon \cdot \sin s \cdot \sin t : t \in [\alpha_m, \pi/2],\ s \in [0, \pi/2],\ \varepsilon \in [0, q]\}.$$

Ez tovább nem javítható, hiszen tetszőleges $s \in [0, \pi/2]$, $\varepsilon \in [0, q]$ párhoz nyilván választható olyan $j_1, \ldots, j_n$ indexsorozat és $\mathbf{x} \in \mathbf{R}^N$, hogy (3.11) teljesüljön. Tehát

$$\|P_{j_n} \ldots P_{j_1}\| \leqq \max\{\cos s \cdot \cos t + q \cdot \sin s \cdot \sin t : \alpha_m \leqq t \leqq \pi/2,\ 0 \leqq \varepsilon \leqq \pi/2\}.$$

Elemi számolás mutatja, hogy itt a jobb oldal értéke mennyi:

$$\|P_{j_n} \ldots P_{j_1}\| \leqq \frac{\cos^2 \alpha_m + q \cdot \sin^2 \alpha_m}{\sqrt{\cos^2 \alpha_m + q^2 \sin^2 \alpha_m}} = \left\langle \begin{pmatrix} \cos \alpha_m \\ q \sin \alpha_m \end{pmatrix}^0, \begin{pmatrix} \cos \alpha_m \\ \sin \alpha_m \end{pmatrix} \right\rangle < 1,$$

ami bizonyítja a tételt. $\left(\text{Itt } \begin{pmatrix} \alpha \\ \beta \end{pmatrix}^0 = \frac{1}{\sqrt{\alpha^2 + \beta^2}} \begin{pmatrix} \alpha \\ \beta \end{pmatrix}\right)$.

A bizonyítás alapján az alábbi rekurzív formulához jutunk a

$$q_m(L_1, \ldots, L_m) = \sup\{\|P_{j_n} \ldots P_{j_1}\| : \{j_1, \ldots, j_n\} = \{1, \ldots, m\}\}$$

konstansokkal kapcsolatban:

3.4. KÖVETKEZMÉNY.

$$q_m(L_1, \ldots, L_m) = \max_{k=1,\ldots,m} \frac{\cos^2 \beta_k + \varepsilon_k \sin^2 \beta_k}{\sqrt{\cos^2 \beta_k + \varepsilon_k^2 \cdot \sin^2 \beta_k}},$$

ahol

$$\beta_k = \sphericalangle\Big(L_k, \bigcap_{j \neq k} L_j^k\Big), \quad \varepsilon_k = q_{m-1}(L_1^k, \ldots, L_{k-1}^k, L_{k+1}^k, \ldots, L_m^k)$$

az $L_j^k = \{\mathbf{z} \in L_j : \mathbf{z} \perp \bigcap_{r \neq k, j} L_r\}$ alterek mellett.

*3.5 Megjegyzés.* Régóta ismeretes az a sejtés, hogy a (3.4) formula (3.1 tétel) végtelen dimenziós *Hilbert tér* esetén is áll, ha a konvergenciát nem az operátornorma szerinti, hanem az erős topológiában vesszük.

A fejezet lezárásaképpen rátérünk az $S$ halmaz korlátossági tulajdonságainak vizsgálatára.

3.6. TÉTEL. Ha $L_1 \cap \ldots \cap L_m = \{0\}$, akkor az $S$ halmaz korlátos.

*Bizonyítás.* (3.2) értelmében elegendő belátnunk, hogy a

$$T = \{F_{i_n} \ldots F_{i_1}(0):\ n = 1, 2, \ldots \quad \text{és} \quad \{i_1, \ldots, i_n\} = \{1, \ldots, m\}\}$$

halmaz korlátos. Tudjuk, hogy $\mathbf{R}^N$-ben egy affin altér mindig előáll véges sok hipersík metszeteként, amelyek normálisai páronként merőlegesek egymásra. Ezért $T$ korlátossága következik, ha megmutatjuk, hogy a

$$T' = \{G_{i_n} \ldots G_{i_1}(0):\ n = 1, 2, \ldots \quad \text{és} \quad \{i_1, \ldots, i_n\} = \{1, \ldots, \mu\}\}$$

halmaz korlátos, ha $G_1, \ldots, G_\mu$ adott hipersíkokra való merőleges vetítések $\mathbf{R}^N$-ben. E tény bizonyításának a kulcslépése a következő önmagában is érdekes konstrukció:

3.7. LEMMA. Legyenek $\mathbf{u}_1, \ldots, \mathbf{u}_\mu$ páronként különböző egységvektorok $\mathbf{R}^N$-ben, és jelölje $Q_i$ az $\mathbf{u}_i$-ra merőleges altérre való ortogonális projekciót (azaz $Q_i\mathbf{x} = \mathbf{x} - \langle \mathbf{x}, \mathbf{u}_i \rangle \mathbf{u}_i$ $i = 1, \ldots, \mu$). Ekkor létezik olyan nem-üres $A$ részhalmaza az $\mathbf{R}^N$ tér $B$ nyitott egységgömbjének és létezik $\varepsilon > 0$ úgy, hogy

$$(3.12) \qquad (Q_i A) + (-\varepsilon, \varepsilon) \cdot \mathbf{u}_i \left(= \{Q_i\mathbf{x} + \xi_i : |\xi| < \varepsilon,\ \mathbf{x} \in A\}\right) \subset A \quad i = 1, \ldots, \mu.$$

*Bizonyítás.* Feltehetjük, hogy az $\mathbf{u}_1, \ldots, \mathbf{u}_\mu$ vektorok által kifeszített tér maga $\mathbf{R}^N$ (egyébként ugyanis kiegészítjük az $\{\mathbf{u}_i\}_1^\mu$ rendszert további vektorokkal). Legyen ekkor $k = 0, \ldots, N-1$ mellett $\mathcal{U}_k$ az

$$\mathcal{U}_k = \{\{\lambda_1 \mathbf{u}_{j_1} + \ldots + \lambda_k \mathbf{u}_{j_k} : \lambda_1, \ldots, \lambda_k \in \mathbf{R}\} : \mathbf{u}_{j_1}, \ldots, \mathbf{u}_{j_k} \quad \text{lineárisan függetlenek}\}$$

altércsalád. (Tehát minden $U \in \mathcal{U}_k$ altér $k$ dimenziós.)

Rekurzióval megkonstruálunk egy olyan csökkenő $\delta_k, \varepsilon_k$ $k = 1, \ldots, N-1$ számpár sorozatot, amelyre az

$$(3.13) \qquad A = \bigcap_{k=1}^{N-1} \bigcap_{U = \mathcal{U}_{N-k}} \{\mathbf{x} \in B : \operatorname{dist}(\mathbf{x}, U) < 1 - \delta_k\}$$

választás $\varepsilon = \varepsilon_{N-1}$ mellett megfelel, mint látni fogjuk. (Itt $\operatorname{dist}(\mathbf{x}, U) = \min\{\|\mathbf{x} - \mathbf{y}\| : \mathbf{y} \in U\}$ az $\mathbf{x}$ pont távolsága az $U$ altértől.)

Legyen $\delta_1, \varepsilon_1 > 0$ olyan kicsiny, hogy az

$$C(U) = \{\mathbf{x} \in \bar{B} : \operatorname{dist}(\mathbf{x}, U) \geqq 1 - \delta_1\}$$

gömbsüvegpárok $U \in \mathcal{U}_{N-1}$ mellett páronként diszjunktak legyenek és teljesüljön

$$\{\langle \mathbf{y}, \mathbf{u}_i \rangle : \mathbf{y} \in C(U)\} \cap [-\varepsilon, \varepsilon] = \emptyset \quad \text{valahányszor} \quad \mathbf{u}_i \notin U \quad (U \in \mathcal{U}_{N-1}).$$

Miután $\delta_{k-1}, \varepsilon_{k-1}$-et megkonstruáltuk ($k \geqq 2$), úgy választjuk meg $\delta_k, \varepsilon_k$-t, hogy $0 < \delta_k \leqq \delta_{k-1}$, $0 < \varepsilon_k \leqq \varepsilon_{k-1}$ és

$$(3.14) \qquad d(U, \mathbf{u}, \varepsilon_k, \delta_k) \leqq \delta_{k-1}$$

legyen minden $U \in \mathcal{U}_{N-(k-1)}$, $\mathbf{u} \in \{\mathbf{u}_1, \ldots, \mathbf{u}_\mu\}$ és $\mathbf{u} \notin U$ mellett, ahol

$$d(U, \mathbf{u}, \varepsilon, \delta) = 1 - \operatorname{dist}\left(U + \mathbf{R}\mathbf{u}, \{\mathbf{x} : |\langle \mathbf{x}, \mathbf{u} \rangle| \leqq \varepsilon, \operatorname{dist}(\mathbf{x}, U) \leqq 1 - \delta, \|\mathbf{x}\| \leqq 1\}\right).$$

Igazoljuk, hogy (3.14) teljesíthető. Elegendő belátni, hogy

$$\lim_{\varepsilon,\delta\to 0+} d(U, \mathbf{u}, \varepsilon, \delta)=0, \tag{3.15}$$

ha $U$ egy tetszőlegesen rögzített $k(2\leqq k\leqq N-2)$ dimenziós altér és $\mathbf{u}$ egy nem benne fekvő egységvektor.

Legyen $U$, $\mathbf{u}$ adottak. Jelölje $\mathbf{u}_1$ az $\mathbf{u}$-projekció$_U\mathbf{u}$ vektor irányába mutató egységvektort, és legyen $\mathbf{u}_2$ a projekció$_U\mathbf{u}$ vektor irány egységvektora ha $\mathbf{u}\not\perp U$ (egyébként $\mathbf{u}_2=\mathbf{0}$ legyen). Most $\vartheta=\arccos\langle \mathbf{u}, \mathbf{u}_1\rangle$ mellett $\mathbf{u}=\cos\vartheta\cdot\mathbf{u}_1+\sin\vartheta\cdot\mathbf{u}_2$, $\vartheta\in[0, \pi/2)$. Vezessük be továbbá az $U'=\{\mathbf{u}'\in U: \|\mathbf{u}'\|=1, \mathbf{u}'\perp\mathbf{u}_1, \mathbf{u}_2\}$, ill. $V=\{\mathbf{v}: \mathbf{v}\perp\mathbf{u}, \mathbf{v}\perp U, \|\mathbf{v}\|=1\}$ irányvektor családokat. Vegyük észre, hogy $V\neq\emptyset$ mivel $\dim\{\mathbf{v}: \mathbf{v}\perp\mathbf{u}, U\}=N-(1+\dim U)=N-(1+k)\geqq 1$.

Tetszőleges $\mathbf{x}\in\mathbf{R}^N$-re, véve az

$$\mathbf{x}=\eta\mathbf{v}+\xi_1\mathbf{u}_1+\xi_2\mathbf{u}_2+\xi'\mathbf{u}, \quad \mathbf{v}\in V, \ \mathbf{u}'\in U' \ \xi_2=0, \quad \text{ha} \quad \mathbf{u}_2=\mathbf{0}, \ \eta\geqq 0$$

ortogonális felbontást, kapjuk, hogy

$$\|\mathbf{x}\| = (\eta^2+\xi_1^2+\xi_2^2+\xi'^2)^{1/2},$$

$$\operatorname{dist}(\mathbf{x}, U) = \|\eta\mathbf{v}+\xi_1\mathbf{u}_1\| = (\eta^2+\xi_1^2)^{1/2},$$

$$\langle \mathbf{x}, \mathbf{y}\rangle = \xi_1\cos\vartheta+\xi_2\sin\vartheta.$$

Tehát

$$d(U, \mathbf{u}, \varepsilon, \delta) = 1-\inf\{\eta\geqq 0: \exists\xi_1, \xi_2, \xi' \quad \eta^2+\xi_1^2+\xi_2^2+\xi'^2\leqq 1,$$

$$\eta^2+\xi_1^2\geqq(1-\delta)^2, \ |\xi_1\cos\vartheta+\xi_2\sin\vartheta|\leqq\varepsilon\} =$$

$$= 1-\inf\{\eta: \exists\xi_1, \xi_2(\eta, \xi_1, \xi_2)\in E(\varepsilon, \delta)\},$$

ahol

$$E(\varepsilon, \delta) = \{(\eta, \xi_1, \xi_2)\in\mathbf{R}^3: \eta\geqq 0, \ \eta^2+\xi_1^2+\xi_2^2\leqq 1, \ \eta^2+\xi_1^2\geqq(1-\delta)^2,$$

$$|\xi_1\cos\vartheta+\xi_2\sin\vartheta|\leqq\varepsilon\}.$$

Nyilván, $1>\varepsilon, \delta>0$ esetén $E(\varepsilon, \delta)\neq\emptyset$ és tetszőleges $(\eta, \xi_1, \xi_2)\in E(\varepsilon, \delta)$-ra $1\geqq\geqq(\eta^2+\xi_1^2)+\xi_2^2\geqq(1-\delta)^2+\xi_2^2$, ahonnan

$$\xi_2^2\leqq 1-(1-\delta)^2=\delta^2+2\delta,$$

$$|\xi_1|\leqq\frac{\varepsilon+|\xi_2|\sin\vartheta}{\cos\vartheta}\leqq\frac{\varepsilon+(\delta^2+2\delta)^{1/2}\sin\vartheta}{\cos\vartheta}$$

$$\eta^2\geqq(1-\delta)^2-\xi_1^2\geqq(1-\delta)^2-\left[\frac{\varepsilon+(\delta^2+2\delta)^{1/2}\sin\vartheta}{\cos\vartheta}\right]^2.$$

Így $d(U, \mathbf{u}, \varepsilon, \delta)\leqq 1-\left[(1-\delta)^2+\left[\frac{\varepsilon+(\delta^2+2\delta)^{1/2}\cdot\sin\vartheta}{\cos\vartheta}\right]^2\right]^{1/2}$, ami bizonyítja (3.15)-öt.

Verifikáljuk, hogy a (3.13)-ban definiált $A$ halmaz rendelkezik a (3.12) tulajdonsággal:

Tegyük fel (3.12)-vel ellentétben, hogy valamelyik $i$ index mellett volna olyan $\mathbf{a}\in A$ pont és $\varepsilon'\in(-\varepsilon, \varepsilon)$ szám, hogy $Q_i\mathbf{a}+\varepsilon'\mathbf{u}_i\notin A$. Mivel az $A$ halmaz nyitott, most

úgy is választható $\varepsilon'$, hogy az

$$\mathbf{x} = Q_i\mathbf{a} + \varepsilon'\mathbf{u}_i$$

pont határpontja legyen $A$-nak. Most $\|\mathbf{x}\| \leqq 1$ (hiszen $A \subset B$), és $|\langle \mathbf{x}, \mathbf{u}_i\rangle| = |\varepsilon'| \leqq \varepsilon$. Mivel pedig $\mathbf{x} \notin A$, létezik olyan minimális $k$ index és $U \in \mathscr{U}_{N-k}$, hogy $\operatorname{dist}(\mathbf{x}, U) \geqq \geqq 1 - \delta_k$.

Vegyük észre, hogy $\mathbf{u}_i \in U$ nem lehet. Ugyanis $\mathbf{u}_i \in U$-ból következne a $\operatorname{dist}(\mathbf{x}, U) = \operatorname{dist}(Q_i\mathbf{a} + \varepsilon'\mathbf{u}_i, U) = \operatorname{dist}(\mathbf{a}, U) < 1 - \delta_k$ ellentmondás.

$k = 1$ nem lehet, mert $k = 1$ esetén $U \in \mathscr{U}_{N-1}$ volna és $\mathbf{x}$ benne feküdne a $C(U)$ gömbsüvegpárban. De ekkor $\varepsilon_1$ definíciója szerint $|\langle \mathbf{x}, \mathbf{u}_i\rangle| > \varepsilon_1 \geqq \varepsilon_{N-1} = \varepsilon$ volna, mivel $\mathbf{u}_i \notin U$.

Tehát $\mathbf{u}_i \notin U$ és $k > 1$.

Mivel $\mathbf{u}_i \notin U$ és $U \in \mathscr{U}_{N-k}$, az $U^* = \{\mathbf{y} + \lambda\mathbf{u}_i : \mathbf{y} \in U, \lambda \in \mathbf{R}\}$ altér $\mathscr{U}_{N-(k-1)}$-be tartozik. Így a $k$ index minimalitása miatt $\operatorname{dist}(\mathbf{x}, U^*) < 1 - \delta_{k-1}$. Csakhogy ekkor $d(U, \mathbf{u}_i, \varepsilon_k, \delta_k) \leqq \delta_{k-1}$ is áll az $\varepsilon_k, \delta_k$ pár definíciója szerint. Innen pedig a $\operatorname{dist}(U^*, \mathbf{x}) \geqq \geqq \operatorname{dist}\left(U^*, \{\mathbf{x}' : |\langle \mathbf{x}', \mathbf{u}_i\rangle| \leqq \varepsilon_k,\ \operatorname{dist}(\mathbf{x}', U) \geqq 1 - \delta_k,\ \|\mathbf{x}'\| \leqq 1\}\right) \geqq 1 - \delta_{k-1}$ ellentmondásra jutunk.

Ezzel a 3.7 bizonyítása teljes.

3.7-ből a $T'$ halmaz korlátossága azonnal következik: Jelölje $\mathbf{u}_i$ a $H_i = \{G_i(\mathbf{x}) : \mathbf{x} \in \mathbf{R}^N\}$ hipersík egy normálvektorát $(i = 1, \ldots, \mu)$ és legyen $d = \max_{i=1,\ldots,\mu} \operatorname{dist}(\mathbf{0}, H_i)$. Konstruáljuk meg a lemma szerint az $A$ alakzatot és $\varepsilon$-t az $\mathbf{u}_1, \ldots, \mathbf{u}_\mu$ egységvektorokhoz, és legyen $A' = \frac{d}{\varepsilon} A \left(= \left\{\frac{d}{\varepsilon}\mathbf{x} : \mathbf{x} \in A\right\}\right)$. Ekkor $G_i A' \subset A'$, $i = 1, \ldots, \mu$. Mivel $\mathbf{0} \in A' \subset \frac{d}{\varepsilon} B$, a $T'$ halmaz része a $\frac{d}{\varepsilon} B$ gömbnek.

A 3.6 tételből már könnyen következik, hogy az $F_1, \ldots, F_m$ affin projekciók konvex kombinációiból alkotott random iteráltak torlódási pontjai mind beleesnek az $S$ halmaz konvex burkának a lezártjába.

Legyenek $\tilde{F}_1, \ldots, \tilde{F}_r$ konvex kombinációi $F_1, \ldots, F_m$-nek, nevezetesen

$$\tilde{F}_i = \sum_{j=1}^{\nu_i} \lambda_{ij} F_{t_{ij}(k_{ij})} \ldots F_{t_{ij}(1)}, \quad (i = 1, \ldots, r),$$

ahol $\sum_{j=1}^{\nu_i} \lambda_{ij} = 1$; $\lambda_{i1}, \ldots, \lambda_{i\nu_i} > 0$ és

$$\bigcup_{i=1}^{r} \bigcup_{j=1}^{\nu_i} \bigcup_{s=1}^{k_{ij}} \{t_{ij}(s)\} = \{1, \ldots, m\}.$$

Legyen ekkor

$$S^* = \bigcup_{\substack{(i_1, i_2, \ldots) \in I \\ \exists \mathbf{x} \in \mathbf{R}^N}} \{[\tilde{F}_{i_n} \ldots \tilde{F}_{i_1}(\mathbf{x})]_{n=1}^{\infty} \text{ torlódási pontjai}\}$$

ahol $I$ azon indexsorozatok családja, amelyek az $1, \ldots, r$ számokból állnak és ezért mindegyikét végtelen sokszor felveszik.

3.9. KÖVETKEZMÉNY. $S^* \subset \overline{\operatorname{co}}\, S$.

*Bizonyítás.* (3.1)-ből azonnal adódik, hogy az általánosság megszorítása nélkül feltehetjük, hogy $L_1 \cap \ldots \cap L_m = \{\mathbf{0}\}$. Ezután azt kell megmutatnunk, hogy ha $\Phi$ egy

tetszőleges valós lineáris funkcionál $\mathbf{R}^N$-en és

$$\gamma = \{\Phi(\mathbf{Z}): \mathbf{Z} \in \bigcup_{(i_1,\ldots)\in I} \{[F_{i_n} \ldots F_{i_1}(\mathbf{0})]_{n=1}^{\infty} \text{ torlódási pontjai}\}\}$$

(vö. a 3.2 következménnyel), akkor minden $(i_1, i_2, \ldots)\in \tilde{I}$ sorozatra és $\mathbf{x}\in\mathbf{R}^N$ pontra fennáll

$$\overline{\lim_{n\to\infty}} \Phi(\tilde{F}_{i_n} \ldots \tilde{F}_{i_1}(\mathbf{x})) \leqq \gamma. \tag{3.15}$$

Ehhez legyen $\mathbf{x}\in\mathbf{R}^N$ és $(i_1, \ldots)\in\tilde{I}$ tetszőlegesen rögzítve. Jelöljük $\mu_n$-nel azt a valószínűségi mértéket, amely az $\{1, \ldots, \nu_{i_n}\}$ halmazon van értelmezve úgy, hogy $\mu_n(\{j\})=\lambda_{i_n j}$ $(j=1, \ldots, \nu_{i_n})$. Legyen $\mu$ a $\mu = \prod_{n=1}^{\infty} \mu_n$ szorzatmérték. Definíció szerint (l. [7]), a $\mu$ mérték tartóhalmaza a $\Pi = \prod_{n=1}^{\infty} \{1, \ldots, \nu_{i_n}\}$ szorzathalmaz, vagyis az összes olyan $(j_1, j_2, \ldots)$ indexsorozatok halmaza, amelyekre $1 \leqq j_n \leqq \nu_{i_n}$ $(n=1, 2, \ldots)$. Ekkor

$$\Phi(\tilde{F}_{i_n} \ldots \tilde{F}_{i_1}(\mathbf{x})) = \sum_{j_n=1}^{\nu_{i_n}} \ldots \sum_{j_1=1}^{\nu_{i_1}} \lambda_{i_n j_n} \ldots \lambda_{i_1 j_1} \Phi[F_{t_{i_n j_n}(k_{i_n j_n})} \ldots F_{t_{i_1 j_1}(1)}(\mathbf{x})] = \int_{\Pi} \varphi_n d\mu, \tag{3.16}$$

ahol

$$\varphi(j_1, j_2, \ldots) = \Phi[(F_{t_{i_n j_n}(k_{i_n j_n})} \ldots F_{t_{i_n j_n}(1)}) \ldots F_{t_{i_1 j_1}(k_{i_1 j_1})} \ldots F_{t_{i_1 j_1}(1)})(\mathbf{x})].$$

Legyen $J = \{(j_1, j_2, \ldots)\in\Pi$: $t_{i_1 j_1}(1), \ldots, t_{i_1 j_1}(k_{i_1 j_1}), t_{i_2 j_2}(1), \ldots, t_{i_2 j_2}(k_{i_2 j_2}), \ldots$-ben az $1, \ldots, m$ számok valamelyike véges sokszor fordul elő$\}$. Vegyük észre, hogy $\mu(J)=0$. A 3.6 tétel szerint így

$$\overline{\lim_{n\to\infty}} \varphi_n \leqq \gamma \quad \mu\text{-majdnem mindenütt} \quad (\Pi \setminus J \text{ fölött}). \tag{3.17}$$

Másrészt a 3.7 lemma következtében az $\{F_{t_r} \ldots F_{t_1}(\mathbf{x})$: $r=1, 2, \ldots; 1 \leqq t_1, \ldots, t_r \leqq m\}$ halmaz belefoglalható egy, az $F_k$ leképezések mindegyikére invariáns és $\mathbf{x}$-et is tartalmazó, konvex korlátos halmazba, így korlátos. Vagyis

$$\sup\{|\Phi(F_{t_r}, \ldots, F_{t_1}(\mathbf{x}))|: r = 1, 2, \ldots; 1 \leqq t_1, \ldots, t_r \leqq m\} < \infty.$$

Innen $\sup_{n,(n_1,\ldots)} |\varphi_n(j_1, \ldots,)| < \infty$. Így (3.17) és a jól ismert *Fatou-lemma* (l. [17]) alapján $\overline{\lim_{n\to\infty}} \int_{\Pi} \varphi_n d\mu \leqq \int_{\Pi} \gamma d\mu = \gamma$, ami (3.15) és (3.16) szerint ekvivalens a bizonyítandóval.

## 4. Affin kontrakciók stacionárius iterációja Hilbert térben

Az (1.4) probléma megoldására kidolgozott numerikus eljárások [3], [10], [14]-ben a megoldást mindig az $F_1, \ldots, F_m$ affin operátorok segítségével megalkotott újabb $F_1', \ldots, F_{m'}'$ affin kifejezések végtelen szorzataként adják, azaz

$$\lim_{n\to\infty} F'_{1+\mathrm{mod}_{m'}(n)} F'_{1+\mathrm{mod}_{m'}(n-1)}, \cdots, F'_{1+\mathrm{mod}_{m'}(0)}(\mathbf{x})$$

alakban, amennyiben $M_1 \cap \ldots \cap M_m \neq \emptyset$. Bevezetve az $F = F'_m \ldots F'_1$ és $\mathbf{x}_k = F'_k \ldots F'_1(\mathbf{x})$ $(k = 1, \ldots, m')$ jelöléseket, az $[F'_{1+\mathrm{mod}_{m'}(n)} \ldots F'_{1+\mathrm{mod}_{m'}(0)}(\mathbf{x})]_{n=1}^{\infty}$ pontsorozat felbontható az $[F^n(\mathbf{x}'_k)]_{n=1}^{\infty}$ $(k = 1, \ldots, m')$ sorozatok egyesítésére. Vagyis az említett típusú módszerek általános vizsgálatakor szorítkozhatunk egyetlen affin operátor hatványainak az elemzésére. Könnyen látható, hogy a [3], [10], [14], [15]-beli eljárásokban fellépő $F'_k$ affin operátorok mind kontrakciók.

Ebben a részben $H$ végig egy tetszőlegesen rögzített *Hilbert teret* fog jelölni, $F$ pedig egy $H \to H$ affin operátort; $F(\mathbf{x}) = A\mathbf{x} + \mathbf{b}$, ahol $A$ az $F$ lineáris része $(\mathbf{b} = F(\mathbf{0}))$. Érdeklődésünk az $[F^n(\mathbf{x})]_{n=1}^{\infty}$ alakú pontsorozatok viselkedésére irányul, elsősorban is arra az esetre, amelyben az $F$ operátor affin merőleges vetítések véges konvex kombinációja.

4.1. LEMMA. Ha az $F$ leképezés lineáris része hatvány-korlátos, azaz a $\mu = \sup_n \|A^n\|$ mennyiség véges, akkor az $M = \{\mathbf{x}: [F^n(\mathbf{x})]_{n=1}^{\infty}$ konvergens$\}$ ponthalmaz a $H$ tér egy affin altere; $\mathbf{x} \in M$-re $\lim_{n\to\infty} F^n(\mathbf{x})$ az $F$ fixpontja.

*Bizonyítás.* Ha $\mathbf{x} \in M$, akkor $F(\lim_{n\to\infty} F^n(\mathbf{x})) = \lim_{n\to\infty} F^{n+1}(\mathbf{x}) = \lim_{n\to\infty} F^n(\mathbf{x})$ az $F$ operátor folytonossága miatt.

Ha tehát $M \neq \emptyset$, akkor $F$-nek van legalább egy $\mathbf{x}_0 (\in M)$ fixpontja. Elég tehát belátnunk, hogy az $\{\mathbf{x} - \mathbf{x}_0: \mathbf{x} \in M\}$ halmaz $H$ egy (lineáris) altere. Vegyük észre, hogy $F(\mathbf{x}) = A\mathbf{x} + \mathbf{b} = A(\mathbf{x} - \mathbf{x}_0) + A\mathbf{x}_0 + \mathbf{b} = A(\mathbf{x} - \mathbf{x}_0) + \mathbf{x}_0$ ahonnan

$$F^n(\mathbf{x}) = A^n(\mathbf{x} - \mathbf{x}_0) + \mathbf{x}_0 \quad (n = 0, 1, 2, \ldots). \tag{4.1}$$

Így $M = \mathbf{x}_0 + H_0$, ahol $H_0 = \{\mathbf{z}: [A^n\mathbf{z}]_{n=0}^{\infty}$ konvergens$\}$. Nyilván $\alpha_1\mathbf{z}_1 + \alpha_2\mathbf{z}_2 \in H_0$ valahányszor $\mathbf{z}_1, \mathbf{z}_2 \in H_0$ és $\alpha_1, \alpha_2 \in \mathbf{R}$. Vagyis csak $H_0$ zártságát kell megmutatnunk. Tegyük fel, hogy $H_0 \ni \mathbf{z}_1, \mathbf{z}_2, \ldots \to \mathbf{z}$. Egy alkalmas részsorozatra áttérve vehető, hogy a $\mathbf{h}_k = \mathbf{z}_k - \mathbf{z}_{k-1}$ $(\mathbf{z}_0 = \mathbf{0})$ különbségvektorok normájára már $\|\mathbf{h}_k\| \leqq 2^{-k}$ $(k = 1, 2, \ldots)$. Tekintsük az $[A^n\mathbf{z}]_{n=0}^{\infty}$ sorozatot. Legyen $\varepsilon > 0$ tetszőlegesen adott. Rögzítsünk egy olyan $N$ számot, melyre $\sum_{k \geqq N} \mu \cdot 2^{-k} < \varepsilon/2$. Mivel $m, n \to \infty$ mellett $A^n\mathbf{h}_k - A^m\mathbf{h}_k \to \mathbf{0}$, a $k = 1, 2, \ldots, N$ indexek mindegyikéhez létezik $\nu_k$ úgy, hogy $\|A^n\mathbf{h}_k - A^m\mathbf{h}_k\| < \varepsilon/(2N)$ valahányszor $n, m > \nu_k$. Ezért

$$\|A^n\mathbf{z} - A^m\mathbf{z}\| \leqq \sum_{k=1}^{N} \|A^n\mathbf{h}_k - A^m\mathbf{h}_k\| + \sum_{k>N} \|A^n - A^m\| \cdot \|\mathbf{h}_k\| \leqq N \cdot \frac{\varepsilon}{2N} + \sum_{k>N} 2\mu \cdot 2^{-k} < \varepsilon.$$

Tehát az $[A^n\mathbf{z}]_{n=0}^{\infty}$ sorozat *Cauchy tulajdonságú*, ahonnan $\mathbf{z} \in H_0$.

4.2. KÖVETKEZMÉNY. Létezik (egy és csak) olyan $Q$ lineáris projekció operátor a $H_0$ altéren, hogy minden $\mathbf{x} \in M$-re $\lim_{n\to\infty} F^n(\mathbf{x}) = (1 - Q)\mathbf{x}_0 + Q\mathbf{x}$. Fennáll $QH_0 = \{F$ fixpontjai$\} - \mathbf{x}_0$ és $\|Q\| \leqq \mu$. Speciálisan, ha $F$ kontrakció, akkor a $Q$ operátor ortogonális projekció, azaz $\lim_{n\to\infty} F^n(\mathbf{x}) = [\mathbf{x}$-nek $\{F$ fixpontjai$\}$-ra való merőleges vetülete$]$ $(\mathbf{x} \in M)$.

*Bizonyítás.* A $Q$: $H_0 \ni \mathbf{z} \mapsto \lim_{n\to\infty} A^n\mathbf{z}$ leképezés linearitása és a $\|Q\| \leqq \mu$, $Q^2 = Q$ relációk nyilvánvalók. Ha speciálisan $\|A\| \leqq 1$, akkor $\|Q\| \leqq 1$. De a *Hilbert térbeli* kontrakciók elméletéből (vö. [18]) jól ismert, hogy egy kontraktív lineáris projekció

szükségképpen ortogonális. Végül (4.1) szerint $F^n(\mathbf{x})=A^n(\mathbf{x}-\mathbf{x}_0)+\mathbf{x}_0 \to Q(\mathbf{x}-\mathbf{x}_0)+$ $+\mathbf{x}_0=(1-Q)\mathbf{x}_0+Q\mathbf{x}$ $(\mathbf{x}\in H)$.

4.3. LEMMA. Ha az $F$ affin operátor kontraktív (azaz $\|A\|\leqq 1$), akkor minden $\mathbf{x}\in H$ mellett

$$\frac{1}{n}F^n(\mathbf{x}) \to P^{(1)}\mathbf{b} \quad \text{és} \quad \frac{1}{n}\sum_{k=1}^{n}\left(F^k(\mathbf{x})-F^k(\mathbf{0})\right) \to P^{(1)}\mathbf{x} \quad (n\to\infty),$$

ahol $P^{(1)}$ jelöli a $\{\mathbf{z}\in H\colon A\mathbf{z}=\mathbf{z}\}$ sajátaltérre való ortogonális projekciót.

*Bizonyítás.* Az állítás azonnali következménye az ergodikus tételnek (l. [17]) és az $F^n(\mathbf{x})=A^n\mathbf{x}+A^{n-1}\mathbf{b}+A^{n-2}\mathbf{b}+\ldots+A\mathbf{b}+\mathbf{b}$ összefüggésnek.

4.4. PROPOZÍCIÓ. Ha $F$ kontraktív és az $F(\mathbf{x})=\mathbf{x}$ fixpont-egyenletnek van megoldása, akkor $P^{(1)}\mathbf{b}=\mathbf{0}$ (4.3 jelöléseivel) és bármely $\mathbf{x}\in H$-ra az $\frac{1}{n}\sum_{k=1}^{n}F^k(\mathbf{x})$ $(n=1, 2, \ldots)$ sorozat konvergens és a limesze $F$-nek egy fixpontja.

*Bizonyítás.* Tudjuk, hogy $\|A\|\leqq 1$ esetén az $A$ operátor önmagába képezi a $H^{(1)}=\{\mathbf{z}\colon A\mathbf{z}=\mathbf{z}\}$ sajátalteret és annak $H^{(0)}$ ortokomplementerét. Így $F(\mathbf{x}_0)=\mathbf{x}_0$-ból most $P^{(1)}\mathbf{x}_0=P^{(1)}F(\mathbf{x}_0)=P^{(1)}A\mathbf{x}_0+P^{(1)}\mathbf{b}=P^{(1)}\mathbf{x}_0+P^{(1)}\mathbf{b}$, azaz $P^{(1)}\mathbf{b}=\mathbf{0}$ adódik. (4.1)-et alkalmazva kapjuk, hogy

$$\frac{1}{n}\sum_{k=1}^{n}F^k(\mathbf{x}) = \frac{1}{n}\sum_{k=1}^{n}A^k(\mathbf{x}-\mathbf{x}_0)+\mathbf{x}_0.$$

Az ergodikus tétel szerint $\frac{1}{n}\sum_{k=1}^{n}A^k(\mathbf{x}-\mathbf{x}_0)\to P^{(1)}(\mathbf{x}-\mathbf{x}_0)$, azaz $\frac{1}{n}\sum_{k=1}^{n}F^k(\mathbf{x})\to$ $\to P^{(1)}\mathbf{x}+P^{(0)}\mathbf{x}_0 (n\to\infty)$, ahol $P^{(0)}$ a $H^{(0)}$-ra való merőleges vetítés. Azonban

$$F(P^{(1)}\mathbf{x}+P^{(0)}\mathbf{x}_0) = AP^{(1)}\mathbf{x}+AP^{(0)}\mathbf{x}_0+\mathbf{b} = P^{(1)}\mathbf{x}+A(\mathbf{x}_0-P^{(1)}\mathbf{x}_0)+\mathbf{b} =$$
$$= P^{(1)}\mathbf{x}+F(\mathbf{x}_0)-P^{(1)}\mathbf{x}_0 = P^{(1)}\mathbf{x}+\mathbf{x}_0-P^{(1)}\mathbf{x}_0 = P^{(1)}\mathbf{x}+P^{(0)}\mathbf{x}_0.$$

*4.5. Megjegyzés.* Ha a $H$ tér véges dimenziós és $P^{(1)}\mathbf{b}=\mathbf{0}$, akkor a *Cramer szabályból* azonnal adódik, hogy az $F(\mathbf{x})=\mathbf{x}$ fixpont-egyenlet $H^{(0)}$-ban megoldható. Így a 4.4 propozíció sugall egy numerikus eljárást a fixpont megkeresésére. Az ergodikus közepek konvergencia-sebessége azonban csak logaritmikus rendű.

Az előbbiekben kifejtett egyszerű általános meggondolások után most annak a számunkra fontos speciális esetnek a vizsgálatára térünk, amelynél az $F$ operátor véges sok $F_1, \ldots, F_m$ affin merőleges vetítés adott kompozícióinak konvex kombinációja. Ekkor, $P_k$-val jelölve a továbbiakban az $F_k$ leképezés lineáris részét, az $A$ operátor ($F$ lineáris része) a $P_1, \ldots, P_m$ ortogonális projekciókból alkotott véges szorzatok valamely véges konvex kombinációja. Azaz $S$-sel jelölve az $F_1, \ldots, F_m$, ill. $S'$-vel a $P_1, \ldots, P_m$ leképezések által generált kontrakció-félcsoportot, $F\in\operatorname{co} S$ és $A\in\operatorname{co} S'$.

I. HALPERIN [8] és C. APOSTOL [1] munkáira megy vissza a lineáris operátorok következő strukturális tulajdonsága alapvető voltának a felismerése a stacionárius iterációk szempontjából:

4.6. DEFINÍCIÓ. Mondjuk azt, hogy egy $H$-beli $B$ lineáris kontrakció operátor kvázi-projekció, ha

$$\lim_{\varepsilon\to 0+} \sup\{\|\mathbf{x}-B\mathbf{x}\| : \|\mathbf{x}\| \leqq 1,\ \ \|\mathbf{x}\|-\|B\mathbf{x}\| \leqq \varepsilon\} = 0. \tag{4.2}$$

4.7. TÉTEL. A co $S'$-beli operátorok mind kvázi-projekciók.

*Bizonyítás.* Legyen $\bar{H}$ a $H$-tér komplexifikációja (azaz $\bar{H}=\{\mathbf{f}+i\mathbf{g}\colon \mathbf{f},\mathbf{g}\in H\}$ az $\langle \mathbf{f}+i\mathbf{g}, \mathbf{f}'+i\mathbf{g}'\rangle = \langle \mathbf{f},\mathbf{f}'\rangle + i\langle \mathbf{g},\mathbf{f}'\rangle - i\langle \mathbf{f},\mathbf{g}'\rangle + \langle \mathbf{g},\mathbf{g}'\rangle$ skalárszorzattal ellátva) és legyen $Q$ a $H$ tér kvázi-projekcióinak a családja. Tetszőleges $\bar{H}$ fölötti lineáris $T$ operátornál jelölje $\varphi_T$ a

$$\varphi_T(\varepsilon) = \sup\{\|\mathbf{x}-T\mathbf{x}\| :\ \mathbf{x}\in\bar{H},\ \|\mathbf{x}\| \leqq 1,\ \|\mathbf{x}\|-\|T\mathbf{x}\| \leqq \varepsilon\} \quad (\varepsilon > 0)$$

modulus-függvényt. $T\in Q$ pontosan ha $\lim_{\varepsilon\to 0+} \varphi_T(\varepsilon)=0$.

A tétel azonnal következik az alábbi három lemmából:

4.8. LEMMA. Ha $T$ egy $\bar{H}$-beli pozitív kontrakció, akkor $T\in Q$. Mindig áll ilyenkor $\varphi_T(\varepsilon)\leqq\sqrt{2\varepsilon}$.

4.9. LEMMA. (C. APOSTOL). Ha $T_1, T_2\in Q$, akkor $T_1T_2\in Q$.

4.10. LEMMA. Ha $\lambda_1, \lambda_2\geqq 0$, $\lambda_1+\lambda_2=1$ és $T_1, T_2\in Q$, akkor $\lambda_1T_1+\lambda_2T_2\in Q$.

*A lemmák bizonyításai.*

4.8. LEMMA. A spektrál-előállítási tétel (l. [17]) szerint most

$$T = \int_0^1 \lambda dP_\lambda$$

írható valamely $[P_\lambda]_{\lambda\in[0,1]}$ projekció-sereggel. Tegyük fel, hogy $0<\varepsilon<1$, $\|\mathbf{x}\|\leqq 1$ és $\|\mathbf{x}\|-\|T\mathbf{x}\|\leqq\varepsilon$. Ekkor

$$\|\mathbf{x}\|-\|T\mathbf{x}\| = \frac{\|\mathbf{x}\|^2-\|T\mathbf{x}\|^2}{\|\mathbf{x}\|+\|T\mathbf{x}\|},$$

ahonnan

$$\frac{\|\mathbf{x}\|^2-\|T\mathbf{x}\|^2}{2\|\mathbf{x}\|} \leqq \|\mathbf{x}\|-\|T\mathbf{x}\| \leqq \varepsilon.$$

Jól ismert, hogy

$$\|\mathbf{x}\|^2 = \int_0^1 d\|P_\lambda\mathbf{x}\|^2,\quad \|T\mathbf{x}\|^2 = \int_0^1 \lambda^2 d\|P_\lambda\mathbf{x}\|^2,\quad \|\mathbf{x}-T\mathbf{x}\|^2 = \int_0^1 (1-\lambda)^2 d\|P_\lambda\mathbf{x}\|^2.$$

Mivel $(1-\lambda)^2\leqq 1-\lambda^2$ minden $\lambda\in[0,1]$ -re, így

$$\|\mathbf{x}-T\mathbf{x}\|^2 = \int_0^1 (1-\lambda)^2 d\|P_\lambda\mathbf{x}\|^2 \leqq \int_0^1 (1-\lambda^2) d\|P_\lambda\mathbf{x}\|^2 = \|\mathbf{x}\|^2-\|T\mathbf{x}\|^2.$$

Vagyis most $\|\mathbf{x}-T\mathbf{x}\|^2\leqq\|\mathbf{x}\|^2-\|T\mathbf{x}\|^2\leqq 2\varepsilon\|\mathbf{x}\|$, azaz $\varphi_T(\varepsilon)\leqq\sqrt{2\varepsilon}$.

4.9. LEMMA. Legyen $\varepsilon \geqq \|\mathbf{x}\| - \|T_1 T_2 \mathbf{x}\|$, $\|\mathbf{x}\| \leqq 1$. Ekkor $\varepsilon \geqq \|T_2\mathbf{x}\| - \|T_1 T_2 \mathbf{x}\|$ és $\|T_2\mathbf{x}\| \leqq 1$, ill. $\varepsilon \geqq \|\mathbf{x}\| - \|T_2\mathbf{x}\|$.

Innen $\varphi_{T_1}$ és $\varphi_{T_2}$ definíciója szerint

$$\|\mathbf{x} - T_1 T_2 \mathbf{x}\| \leqq \|\mathbf{x} - T_2\mathbf{x}\| + \|T_2\mathbf{x} - T_1 T_2 \mathbf{x}\| \leqq \varphi_{T_2}(\varepsilon) + \varphi_{T_1}(\varepsilon),$$

azaz $\varphi_{T_1 T_2} \leqq \varphi_{T_1} + \varphi_{T_2}$.

4.10. LEMMA. Legyen $\|\mathbf{x}\| - \|(\lambda_1 T_1 + \lambda_2 T_2)\mathbf{x}\| \leqq \varepsilon$, $\|\mathbf{x}\| \leqq 1$. Bevezetve az $\varepsilon_k = \|\mathbf{x}\| - \|T_k\mathbf{x}\|$ $(k = 1, 2)$ jelölést,

$$\lambda_1\varepsilon_1 + \lambda_2\varepsilon_2 = \|\mathbf{x}\| - (\lambda_1\|T_1\mathbf{x}\| + \lambda_2\|T_2\mathbf{x}\|) \leqq \|\mathbf{x}\| - \|(\lambda_1 T_1 + \lambda_2 T_2)\mathbf{x}\| \leqq \varepsilon.$$

Mivel $\varepsilon_1, \varepsilon_2 \geqq 0$, innen $\varepsilon_k \leqq \varepsilon/\lambda_k$ $(k = 1, 2)$. Tehát

$$\|\mathbf{x} - (\lambda_1 T_1 + \lambda_2 T_2)\mathbf{x}\| \leqq \lambda_1\|\mathbf{x} - T_1\mathbf{x}\| \leqq \lambda_1\varphi_{T_1}(\varepsilon_1) + \lambda_2\varphi_{T_2}(\varepsilon_2) \leqq$$
$$\leqq \lambda_1\varphi_{T_1}(\varepsilon/\lambda_1) + \lambda_2\varphi_{T_2}(\varepsilon/\lambda_2).$$

A kváziprojekciók iteráltjaira teljesül a

4.11. TÉTEL. (C. APOSTOL [1]). Ha $T$ egy kvázi-projekciója $\overline{H}$-nak, akkor $P^{(1)}$-gyel jelölve az $\{\mathbf{x} \in \overline{H}\colon T\mathbf{x} = \mathbf{x}\}$ altérre való ortogonális projekciót,

$$T^n\mathbf{x} \to P^{(1)}\mathbf{x} \quad (n \to \infty) \quad \text{minden} \quad \mathbf{x} \in \overline{H}\text{-ra}.$$

Az $F$ affin operátorra APOSTOL tételéből a következő a konklúziónk:

4.12. TÉTEL. Legyen $A \in \operatorname{co} S'$, és tegyük fel, hogy $A \notin \operatorname{co} S_1'$ valahányszor $S_1$ egy olyan leképezés félcsoportot jelöl, amit a $\{P_1, \ldots, P_m\}$ rendszer egy valódi részhalmaza generál. Ekkor létezik olyan $M'$ sűrű lineáris alsokasága az $L^\perp$ altérnek (ahol $L = L_1 \cap \ldots \cap L_m$, $L_k = P_k H$ mint a 3. fejezetben), melynél minden $\mathbf{b} \in M'$-re az $F_\mathbf{b}\colon \mathbf{x} \mapsto A\mathbf{x} + \mathbf{b}$ affin operátor iteráltjaival alkotott $[F_\mathbf{b}^n(\mathbf{x}_0)]_{n=1}^\infty$ sorozat konvergens tetszőleges $\mathbf{x}_0$ kezdőpont mellett és limesze az

$$F_\mathbf{b}(\mathbf{z}) = \mathbf{z}, \quad P\mathbf{z} = P\mathbf{x}_0$$

fixpont-egyenletrendszer egyetlen megoldása (itt $P$ az $L$-re való ortogonális projekciót jelöli).

*Bizonyítás.* Tekintsük az $F(\mathbf{z}) = \mathbf{z}$ fixpont egyenletet. Ennek akkor és csak akkor van megoldása, ha $\mathbf{b} \in (1 - A)H$. Megmutatjuk, hogy az $(1 - A)H$ alsokaság sűrű $L^\perp$-ben:

Tegyük fel, hogy $L \cap (1 - A)H \subset \{\mathbf{x}\colon \mathbf{x} \perp \mathbf{u}\}$ ahol $\mathbf{u}$ nem-zéró vektor $L^\perp$-ben. Mivel $PP_k = P_k P = P$ $(k = 1, \ldots, m)$, fennáll $AP = PA = P$ és $(1 - P)A = A(1 - P) = A - P = (1 - P)A(1 - P)$. Így $AL^\perp \subset L^\perp$ és $AL = L$, ahonnan $L^\perp \cap (1 - A)H = (1 - A)L^\perp = (1 - A)(1 - P)H = (1 - P)(1 - A)H$. Tehát $0 = \langle (1 - P)(1 - A)\mathbf{h}, \mathbf{u}\rangle = \langle (1 - A)\mathbf{h}, \mathbf{u}\rangle = \langle \mathbf{h}, (1 - A)^*\mathbf{u}\rangle$ minden $\mathbf{h} \in H$ mellett, azaz $A^*\mathbf{u} = \mathbf{u}$. Csakhogy ekkor $A\mathbf{u} = \mathbf{u}$ is, hiszen $\|A\| \leqq 1$ (vö. [18]). Ez pedig azt jelenti, hogy $A = \sum_j \lambda_j R_j$-t írva, ahol $\lambda_j \geqq 0$, $\sum_j \lambda_j = 1$ és $R_j \in S'$, fennáll $\lambda_j\|R_j\mathbf{u}\| = \lambda_j\|\mathbf{u}\|$ mindegyik $j$ indexre (hiszen $\|R_j\mathbf{u}\| \leqq \|u\|$ mindig). Tekintve, hogy alkalmas $i_{j1}, i_{j2}, \ldots, i_{jn_j}$ indexekkel $R_j =$

$= P_{i_{jn_j}} \dots P_{i_{j1}}$ írható, ez csak úgy lehet, ha $\mathbf{u} \in \bigcap_{k=1} L_{i_{jk}}$ minden $j$-re. Tehát $\mathbf{u} \in \bigcap_j \bigcap_{k=1} L_{i_{jk}}$. Nyilván $A \in \operatorname{co}\left[\bigcap_j \{P_{i_{jk}}: 1 \leqq k \leqq n_j\}\right.$ generátuma$]$. Így a tételbeli feltevés szerint $\bigcup_j \{P_{i_{jk}}: 1 \leqq k \leqq n_j\} = \{P_1, \dots, P_m\}$, azaz $\bigcap_j \bigcap_{k=1}^{n_j} L_{i_{jk}} = L$. Vagyis $\mathbf{u} \in L$, ellentmondásban azzal, hogy $\mathbf{u} \in L^\perp$ is.

Következésképpen $\mathbf{b}$ az $L^\perp$ altér egy sűrű lineáris alsokaságából ($(1-A)H \cap L^\perp$-ból) tetszőlegesen választható úgy, hogy az $A\mathbf{z}+\mathbf{b}=\mathbf{z}$ fixpont-egyenlet megoldható legyen. Egyben az is látszik, hogy az $A$ leképezés $L^\perp$-en injektív. Így $A\mathbf{z}_1+\mathbf{b}=\mathbf{z}_1$ és $A\mathbf{z}_2+\mathbf{b}=\mathbf{z}_2$ esetén $\mathbf{z}_1=\mathbf{z}_2$ szükségképpen.

Ha $A\mathbf{z}+\mathbf{b}=\mathbf{z}$, akkor $F_b^n(\mathbf{x}_0)=A^n(\mathbf{x}_0-\mathbf{z})+\mathbf{z}$ $(n=1, 2, \dots)$. Ekkor $PF_b^n(\mathbf{x}_0)= PA^n(\mathbf{x}_0-\mathbf{z})+P\mathbf{z}=P(\mathbf{x}_0-\mathbf{z})+P\mathbf{z}=P\mathbf{x}_0$ mindig. Végül a 4.7. és 4.11. tételek mutatják, hogy ilyenkor $F_b^n(\mathbf{x}_0) \to \mathbf{z}$ $(n \to \infty)$.

4.13. KÖVETKEZMÉNY. Ha $\dim H < \infty$, akkor $M' = H$.

*4.14. Megjegyzés.* Véges dimenzióban az $[F_b^n(\mathbf{x})]_{n=1}^\infty$ sorozatok konvergencia sebességére fennáll

$$\|F_b^n(\mathbf{x}) - \lim_{\nu\to\infty} F_b^\nu(\mathbf{x})\| = \left\|A^n\left(\mathbf{x} - \lim_{\nu\to\infty} F_b^\nu(\mathbf{x})\right)\right\| = \left\|A^n(1-P)\left(\mathbf{x} - \lim_{\nu\to\infty} F_b^\nu(\mathbf{x})\right)\right\| =$$

$$= \left\|(A|L^\perp)^n\left(\mathbf{x} - \lim_{\nu\to\infty} F_b^\nu(\mathbf{x})\right)\right\| \leqq \|A|L^\perp\|^n \cdot \left\|\mathbf{x} - \lim_{\nu\to\infty} F_b^\nu(\mathbf{x})\right\| = \operatorname{const}(\mathbf{x}) \cdot q(\mathbf{x})^n.$$

A 4.12. tétel bizonyításából kiderül, hogy $\mathbf{u} \in L^\perp$ mellett $1 = \|\mathbf{u}\| = \|A\mathbf{u}\|$ nem állhat, ha $A \in \operatorname{co} S'$. Így az egységgömb kompaktsága miatt $\|A|L^\perp\| = \sup\{\|A\mathbf{u}\|: \mathbf{u} \in L^\perp, \|\mathbf{u}\| = 1\} < 1$ ilyenkor. A 3.4. következmény segítségével $\|A|L^\perp\|$ az alábbi módon becsülhető felülről.

Tekintsük az $A$ operátort az $A = \sum_j \lambda_j R_j$ (ahol $R_j = P_{i_{jn_j}} \dots P_{i_{j1}}$) alakban, és legyen $Q_j$ az $\{z \in L^\perp: R_j\mathbf{z} = \mathbf{z}\}$ altérre való ortogonális projekció. A 3.4. következmény pontos becslést ad az $R_j' = (1-Q_j)R_k$ operátorok normájára (mindig $\|R_j'\| < 1$). Tegyük fel, hogy innen kaptuk az $\|R_j'\| \leqq q_j (< 1)$ egyenlőtlenségeket. Ekkor

(4.3)
$$\|A|L^\perp\| \leqq \max\left\{\left\|\sum_j \lambda_j(Q_j\mathbf{u} + \|(1-Q_j)\mathbf{u}\|\,\mathbf{v}_j)\right\|: \|\mathbf{u}\| \leqq 1,\ \mathbf{u} \in L^\perp,\ Q_j\mathbf{v}_j = 0,\ \|v_j\| \leqq q_j\right\}.$$

Jól ismert (az ún. *beng-beng elv*), hogy itt a maximum $\|\mathbf{v}_j\| = q_j$ mellett vétetik fel. A háromszög-egyenlőtlenség segítségével (4.3) jobb oldala még elég finoman növelhető az egyszerűbb kezelésű

$$\|A|L^\perp\| \leqq \max\left\{\sum_j \lambda_j \sqrt{\|Q_j\mathbf{u}\|^2 + q_j^2 \cdot (1 - \|Q_j\mathbf{u}\|^2)}: \|\mathbf{u}\| = 1,\ \mathbf{u} \in L^\perp\right\} \tag{4.4}$$

alakig. (A jobb oldal értéke mindig kisebb 1-nél, mivel $Q_jH \subset L^\perp$ minden $j$ indexre és $\bigcap_j Q_jH = \{\mathbf{0}\}$.

## IRODALOM

[1] APOSTOL, C., "Products of contractions in Hilbert space", *Acta Sci. Math.* **33** (1972) 91—94.
[2] BROWDER, F., "On some approximation methods for solutions of Dirichlet problems," *J. Math. Mech.* **7** (1958) 69—70.
[3] CIMMINO, G., "Calcolo approssimato per le soluzioni dei sistemi di equazioni lineari", *Ric. Sci. Progr. Tecn. Naz.* **1** (1938) 326—333.
[4] COURANT, R. und HILBERT, D., *Methoden der Mathematischen Physik* (Springer, Berlin, 1937).
[5] HALMOS, P., *Finite Dimensional Vector Spaces* (Van Nostrand, Princeton, 1958).
[6] HALMOS, P., *Introduction to Hilbert Space* (Chelsea, New York, 1951).
[7] HALMOS, P., *Measure Theory* (Van Nostrand, New York, 1950).
[8] HALPERIN, I., "The product of projection operators", *Acta Sci. Math.* **23** (1962) 96—99.
[9] HESTENS, M. and STIEFEL, E., "Methods of conjugate gradients for solving linear systems", *J. Res. Bur. Standards* **49** (1952) 409—436.
[10] KACZMARZ, S., „Angenäherte Auflösungen von Systemen linearer Gleichungen", *Bull. Intern. Acad. Sci. Polon.* **A** (1937) 355—357.
[11] BERGMANN, V. and von NEUMANN, J., *Solution of linear systems of high order* (Report for Bureau of Ordnance, Washington, 1946).
[12] von NEUMANN, J., *Mathematische Grundlagen der Quantenmechanik* (Springer, Berlin, 1932).
[13] von NEUMANN, J., "On rings of operators", *Annals of Math.* **50** (1949) 481—485.
[14] POGÁNY, Cs., „Lineáris egyenletrendszerek megoldása geometriai közelítő módszerekkel", *MTA III. Osztálya Közleményei* **17** (1967) 151—160.
[15] POGÁNY, Cs., „Ellentmondó feltételrendszerek kezeléséről II.", *MTA III. Osztálya Közleményei* **23** (1974), 197—202.
[16] RIESZ, F. und SZ.-NAGY, B., „Über Kontraktionen des Hilbertschen Raumes", *Acta Sci. Math.* **10** (1941—43) 202—205.
[17] RIESZ, F.—SZ.-NAGY, B., *Lecons d'analyse fontionnelle* (Akadémiai Kiadó, Budapest, 1955).
[18] SZ.-NAGY, B.—FOIAS, C., *Analyse harmonique des opérateurs de l'espace de Hilbert* (Akadémia, Kiadó, Budapest, 1967).

(Beérkezett: 1983. február 26.)

STACHÓ LÁSZLÓ
JATE BOLYAI INTÉZET
6720 SZEGED, ARADI VÉRTANÚK TERE 1.

## INFINITE PRODUCTS OF AFFINE PROJECTIONS FROM THE NUMERICAL POINT OF VIEW

L. STACHÓ

The behaviour of *Kaczmarz type methods* for solving systems of linear functional equations in *Hilbert space* is investigated when the existence of solutions is not provided. In the classical finite dimensional case a sharp estimate of the convergence rate is given for *random Kaczmarz iterations.*

A kiadásért felelős az Akadémiai Kiadó és Nyomda főigazgatója
Műszaki szerkesztő: Sándor István
A kézirat nyomdába érkezett: 1984. I. 18. — Terjedelem: 17,85 (A/5) ív
84-413 — Szegedi Nyomda, Szeged — F. v.: Dobó József igazgató

# ÚTMUTATÁS A SZERZŐKNEK

Az Alkalmazott Matematikai Lapok csak magyar nyelvű dolgozatokat közöl. A kéziratok gépelését olyan formában kérjük, hogy minden gépelt oldal 25, egyenként átlag 50 betűhelyes sort tartalmazzon. A közlésre szánt dolgozatokat három példányban kell beküldeni.

A kéziratok szerkezeti felépítésének a következő követelményeket kell kielégíteni. A fejlécnek tartalmaznia kell a dolgozat címét, a szerző teljes nevét, valamint annak a városnak a nevét, ahol a szerző dolgozik. A fejléc után egy, képletet nem tartalmazó, legfeljebb 200 szóból álló kivonatot kell minden esetben megadni. A dolgozatot címmel ellátott szakaszokra kell bontani, és az egyes szakaszokat arab sorszámmal kell ellátni. Az esetleges bevezetésnek mindig az első szakaszt kell alkotnia. Az irodalomjegyzék mindig az utolsó szakasz kell hogy legyen, és azt nem kell sorszámmal ellátni. Az irodalomjegyzék után, a kézirat befejezéseképpen fel kell tüntetni a szerző teljes nevét és a munkahelye (illetve lakása) pontos postai címét. A dolgozatban előforduló képleteket szakaszonként újrakezdődően, a képlet előtt két zárójel közé írt kettős számozással kell azonosítani. Természetesen nem szükséges minden képletet számozással ellátni. Az esetleges definíciókat és tételeket (segédtételeket és lemmákat) ugyancsak szakaszonként újrakezdődő, kettős számozással kell ellátni. Kérjük a szerzőket, hogy ezeket, valamint a tételek bizonyítását a szövegben kellő módon emeljék ki. Minden dolgozathoz csatolni kell egy angol, német, francia vagy orosz nyelvű, külön oldalra gépelt összefoglalót. Amennyiben lehetséges, kérjük a nyomtatás számára különösen nehézkes matematikai jelölések használatának az elkerülését.

A dolgozat ábráit és az esetleges lábjegyzeteket a dolgozat végén, különálló lapokon kérjük beküldeni. Mind az ábrákat, mind a lábjegyzeteket a dolgozat szakaszokra bontásától független, folytatólagos arab sorszámozással kell ellátni. Az ábrák elhelyezését a dolgozat megfelelő helyén, széljegyzetként feltüntetett, ábraazonosító sorszámokkal kell megadni. A lábjegyzetekre a dolgozaton belül az azonosító sorszám felső indexkénti használatával lehet hivatkozni.

Az irodalmi hivatkozások formája a következő. Minden hivatkozást fel kell sorolni a dolgozat végén található irodalomjegyzékben, a szerzők, illetve társszerzők esetén az első szerző neve szerinti alfabetikus sorrendben úgy, hogy külön, de folytatólagos sorszámozású listát alkossanak a latin és a cirill betűs nevű szerzők műveire vonatkozó hivatkozások, és mindkét részben a megfelelő alfabetikus sorrend legyen kialakítva. A folyóiratban megjelent cikkekre [1], a könyvekre [5], a kötetben megjelent dolgozatokra [4], a disszertációkra [3] és a gépi program leírásokra [2] a következő minta szerint kell hivatkozni:


[1] Farkas, J., »Über die Theorie der einfachen Ungleichungen«, *Journal für die reine und angewandte Mathematik* **124** (1902) 1—27.

[2] Kéri, G., „DUALSIMP", rutin a CDC 3300-as gépekre (Magyar Tudományos Akadémia Számítástechnikai és Automatizálási Kutató Intézete, CDC 3300 felhasználói ismertetők 2. 1973. május) 19—20.

[3] Prékopa, A., „Sztohasztikus rendszerek optimalizálási problémáiról", doktori értekezés. Magyar Tudományos Akadémia, Budapest, 1970.

[4] Prabhu, N. U., "Recent research on the ruin problem of collective risk theory", in: *Inventory Control and Water Storage* Ed. A. Prékopa (János Bolyai Mathematical Society and North-Holland Publishing Company, Amsterdam—London, 1973) 221—228.

[5] Zoutendijk, G., *Methods of Feasible Directions* (Elsevier Publishing Company, Amsterdam and New York, 1960).


A dolgozatok szövegében az irodalmi hivatkozás számait szögletes zárójelben kell megadni, mint például [5] vagy [4, 76—78]. A szerzők a dolgozatukról 100 darab különlenyomatot kapnak, ezek költsége — nyomott oldalanként 25 forint — a szerzői díjat terheli.

Ára: 50 Ft

INDEX: 26498
ISSN 0133—3399

## TARTALOMJEGYZÉK



## INDEX

# Alkalmazott matematikai lapok

1984/3-4

A MAGYAR TUDOMÁNYOS AKADÉMIA
MATEMATIKAI ÉS FIZIKAI TUDOMÁNYOK
OSZTÁLYÁNAK KÖZLEMÉNYEI

10.

KÖTET

AKADÉMIAI KIADÓ, BUDAPEST

# ALKALMAZOTT MATEMATIKAI LAPOK

## A MAGYAR TUDOMÁNYOS AKADÉMIA MATEMATIKAI ÉS FIZIKAI TUDOMÁNYOK OSZTÁLYÁNAK KÖZLEMÉNYEI





X. kötet 3—4. szám

Szerkesztőség: 1502 Budapest XI., Kende u. 13—17.
Kiadóhivatal: 1055 Budapest V., Alkotmány u. 21.


Az Alkalmazott Matematikai Lapok változó terjedelmű füzetekben jelenik meg, és olyan eredeti tudományos cikkeket publikál, amelyek a gyakorlatban, vagy más tudományokban közvetlenül felhasználható új matematikai eredményt tartalmaznak, illetve már ismert, de színvonalas matematikai apparátus újszerű és jelentős alkalmazását mutatják be. A folyóirat közöl cikk formájában megírt, új tudományos eredménynek számító programokat, és olyan, külföldi folyóiratban már publikált dolgozatokat, amelyek magyar nyelven történő megjelentetése elősegítheti az elért eredmények minél előbbi, széles körű hazai felhasználását.

A folyóirat feladata a Magyar Tudományos Akadémia III. (Matematikai és Fizikai) Osztályának munkájára vonatkozó közlemények, könyvismertetések stb. publikálása is.

A kéziratok a főszerkesztőhöz, vagy a szerkesztő bizottság bármely tagjához beküldhetők. A főszerkesztő címe:

Prékopa András, főszerkesztő
1502 Budapest, Kende u. 13—17.

Közlésre el nem fogadott kéziratokat a szerkesztőség lehetőleg visszajuttat a szerzőhöz, de a beküldött kéziratok megőrzéséért vagy továbbításáért felelősséget nem vállal.

Az Alkalmazott Matematikai Lapok előfizetési ára kötetenként 100 forint. Belföldi megrendelések az Akadémiai Kiadó, 1055 Budapest V., Alkotmány u. 21. címen (pénzforgalmi jelzőszám 215—11 488), külföldi megrendelések a Kultúra Külkereskedelmi Vállalat, H-1389 Budapest, Pf. 149. címen (pénzforgalmi jelzőszám 218—10 990) lehetségesek.

A Magyar Tudományos Akadémia III. (Matematikai és Fizikai) Osztálya a következő idegen nyelvű folyóiratokat adja ki:

1. Acta Mathematica Hungaricae,
2. Acta Physica Hungaricae,
3. Studia Scientiarum Mathematicarum Hungarica.

*Alkalmazott Matematikai Lapok* **10** (1984) 203—229

# STABILIS EGYÜTTÉLÉS ÉS BIFURKÁCIÓK A POPULÁCIÓDINAMIKÁBAN

FARKAS MIKLÓS

Budapest

Vizsgáljuk a klasszikus *Volterra-féle ragadozó-zsákmány modellt*, illetve két ragadozó versenyét egyetlen táplálékért. Megmutatjuk a klasszikus modellek fogyatékosságait és módosítjuk, finomítjuk a modelleket. A ragadozó-zsákmány modellbe korlátozó tényezőt és késleltetést vezetünk be. Megmutatjuk, hogy a késleltetés növekedésével a rendszer stabilis egyensúlyi helyzete elveszíti stabilitását, és bizonyos a paraméterekre kirótt feltételek mellett a rendszer stabilisan rezegni kezd. A ragadozók versengését abban az esetben tárgyaljuk részletesen, amikor a zsákmány is önreprodukáló. A zsákmányra nézve korlátozó tényezőt vezetünk be, a ragadozási rátát, illetve a ragadozók szaporodási rátáját a *Holling-féle függvénnyel* adjuk meg. Vizsgáljuk a versengő kizárás elvének érvényesülését, illetve a bőség paradoxonát. Megmutatjuk, hogy a környezet fenntartóképességének növekedése a stabilis egyensúlyi helyzetek bifurkációjára vezet. Bevezetjük a zipzárbifurkáció fogalmát, amely bizonyos feltételek mellett jelentkezik az *r-stratéga* és a *K-stratéga* versenyében.

## 1. A klasszikus ragadozó-zsákmány modell

Az első világháború alatt, a háborús cselekmények következtében mintegy négy éven át erősen korlátozódott a halászat az *Adriai tengeren*. A háború befejezése után újra kezdődő halászatokon a halászok azt tapasztalták, hogy igen nagy mértékben elszaporodtak a ragadozó halak a tengerben, és egyidejűleg erősen lecsökkent bizonyos értékes növényevő halfajok száma. Az olasz halbiológusok a statisztikai adatok elemzésére, a halpopuláció összetételében végbement változások magyarázatára és a követendő, helyes halászati stratégia kidolgozására VITO VOLTERRÁT, a kiváló olasz matematikust kérték fel. E munka eredményeképpen született a *Volterra-féle ragadozó-zsákmány modell* (lásd [17]), amelyet a következőkben ismertetünk. Ettől függetlenül lásd [11].

Induljunk ki abból, hogy egy adott ökológiai környezetben együtt él egy 1-essel jelölt „zsákmányul szolgáló" és egy 2-essel jelölt „ragadozó" faj. A zsákmányul szolgáló fajt, mint növényevőt képzeljük el és feltételezzük, hogy számára korlátlan mértékben található táplálék; a ragadozó fajról feltételezzük, hogy az adott és zártnak képzelt környezetben az előbbi zsákmányul szolgáló faj képezi az egyetlen táplálékforrást. Jelöljük $N_1$-gyel, illetve $N_2$-vel a zsákmány, illetve a ragadozó mennyiségét (számát). A feladat $N_1$ és $N_2$ dinamikájának, vagyis annak vizsgálata, hogyan alakulnak ezek az értékek az időben. Bár ebben az interpretációban $N_1$ és $N_2$ darabszámok, értéküket olyan nagynak feltételezzük, hogy viszonylag rövid időtartamok alatti megváltozásuk (állandó külső körülmények mellett) folytonosnak legyen tekinthető. Más szóval feltételezzük, hogy $N_i$: $R \mapsto R^+$ $(i=1,2)$ a $t$ időnek folytonosan differenciálható függvénye.

Ha ragadozó nincs jelen, vagyis $N_2=0$, akkor a legkézenfekvőbb feltevés az, hogy állandó környezeti körülmények között az 1-es faj *természetes szaporodási rátája* (számának időegység alatti egy főre eső megváltozása) állandó: $\dot{N}_1/N_1=\varepsilon_1>0$, ahol $\dot{N}_1=dN_1/dt$ az idő szerinti derivált és $\varepsilon_1$ állandó. Ez a feltevés természetesen az

$$\dot{N}_1 = \varepsilon_1 N_1 \tag{1.1}$$

differenciálegyenletre vezet, amelynek tetszőleges $N_1(0)$ kezdeti értékhez tartozó megoldása

$$N_1(t) = N_1(0)e^{\varepsilon_1 t}.$$

Erre a megoldásra alapozta THOMAS MALTHUS a 18. és 19. század fordulóján népesedési elméletét, amelyet azóta is bírálnak részben jogosan. Valóban nem tűnik túlzottan realisztikusnak az a modell, amely szerint „a magára hagyott faj" létszáma az időben exponenciálisan nő és végtelenhez tart.

Hasonló módon, ha a zsákmány a $t=0$ időpillanatban végleg eltűnik, vagyis $N_1=0$, akkor a legegyszerűbb feltevés az, hogy a 2-es faj táplálék hiányában állandó mortalitási rátával kihal, vagyis $\dot{N}_2/N_2=-\varepsilon_2$, ahol $\varepsilon_2>0$ állandó. Ezek szerint az $N_2$ függvény az

$$\dot{N}_2 = -\varepsilon_2 N_2 \tag{1.2}$$

differenciálegyenletnek tesz eleget, ahonnan

$$N_2(t) = N_2(0)e^{-\varepsilon_2 t}$$

következik.

Ha a két faj együttesen van jelen az ökológiai környezetben, akkor a legegyszerűbb feltevés nyilván az, hogy az egyik faj természetes szaporodási (mortalitási) rátája a másik faj számától függeni fog és ez a függés lineáris. Ennek a feltevésnek következtében a független (1.1) és (1.2) egyenletekből álló rendszer csatolt differenciálegyenlet rendszerré válik:

$$\dot{N}_1 = (\varepsilon_1-\gamma_1 N_2)N_1, \quad \dot{N}_2 = (-\varepsilon_2+\gamma_2 N_1)N_2. \tag{1.3}$$

Itt a $\gamma_1>0$ állandó a *„ragadozási ráta"* (egy ragadozó ennyivel csökkenti a zsákmányfaj szaporodási rátáját), a $\gamma_2>0$ állandó a *„zsákmánynak ragadozóvá való feldolgozódási rátája"* (egy zsákmány ennyivel növeli a ragadozó szaporodási rátáját.)

Az (1.3) rendszert nevezzük *Lotka—Volterra ragadozó-zsákmány modellnek.* A rendszer viszonylag egyszerűen vizsgálható az $N_1\geqq 0$, $N_2\geqq 0$ kvadránsban. Leglényegesebb tulajdonságai a következők. Két egyensúlyi helyzete van, az $(N_1, N_2)=(0, 0)$, illetve az $(N_1, N_2)=(\varepsilon_2/\gamma_2, \varepsilon_1/\gamma_1)$ pont. Ezek közül az első instabilis, a második centrum, vagyis *Ljapunov-értelemben stabilis,* de nem aszimptotikusan stabilis.

Az $(\varepsilon_2/\gamma_2, \varepsilon_1/\gamma_1)$ egyensúlyi helyzetben linearizált rendszernek e pont természetesen ugyancsak centruma, és a linearizált rendszer periodikus megoldásainak periódusa $\tau_0=2\pi/\sqrt{\varepsilon_1\varepsilon_2}$.

Behelyettesítéssel könnyen meggyőződhetünk arról, hogy az (1.3) rendszernek egy első integrálja

$$F(N_1, N_2) = \gamma_2 N_1+\gamma_1 N_2-\varepsilon_2 \ln N_1-\varepsilon_1 \ln N_2,$$

vagyis ez a függvény a rendszer pályagörbéi mentén állandó. A pályagörbék egyenletét tehát a

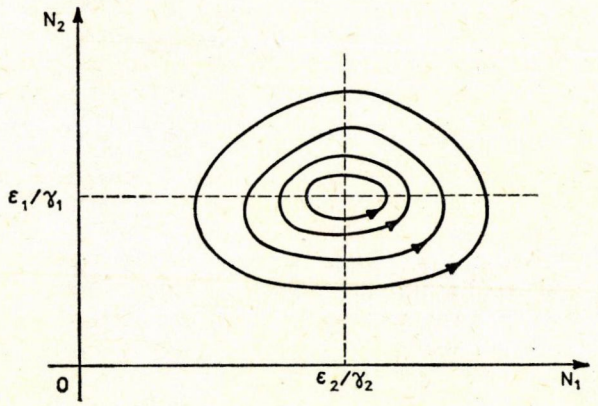
$$\gamma_2 N_1+\gamma_1 N_2-\varepsilon_2 \ln N_1-\varepsilon_1 \ln N_2 = c \tag{1.4}$$

alakban írhatjuk, ahol $c$ tetszőleges állandó. Az $F$ függvény abszolút minimumát az $(\varepsilon_2/\gamma_2, \varepsilon_1/\gamma_1)$ egyensúlyi helyzetben veszi fel (az $N_1$, $N_2$ sík első kvadránsában), tehát nem üres pálygörbét csak akkor kapunk, ha $c \geqq F(\varepsilon_2/\gamma_2, \varepsilon_2/\gamma_1)$. Az (1.4) görbék, az $F$ első integrál szintvonalai mind zártak. Ez azt jelenti, hogy az (1.3) rendszer összes megoldása periodikus. Ha az $(N_1(0), N_2(0))$ kezdeti érték elég közel van $(\varepsilon_2/\gamma_2, \varepsilon_1/\gamma_1)$-hez, akkor a megoldás $\tau$ periódusa közelítőleg $\tau_0=2\pi/\sqrt{\varepsilon_1\varepsilon_2}$. Osszuk el az (1.3) rendszer első egyenletét $N_1$-gyel, a másodikat $N_2$-vel és integráljuk az így kapott egyenleteket egy periódus hosszon, vagyis 0-tól $\tau$-ig. Felhasználva azt, hogy a megoldás $N_i(t)$ koordinátafüggvényei $(i=1, 2)$ periodikusak $\tau$ periódussal a következő eredményt kapjuk

$$\frac{1}{\tau}\int_0^\tau N_1(t)\,dt = \frac{\varepsilon_2}{\gamma_2}, \quad \frac{1}{\tau}\int_0^\tau N_2(t)\,dt = \frac{\varepsilon_1}{\gamma_1}, \tag{1.5}$$

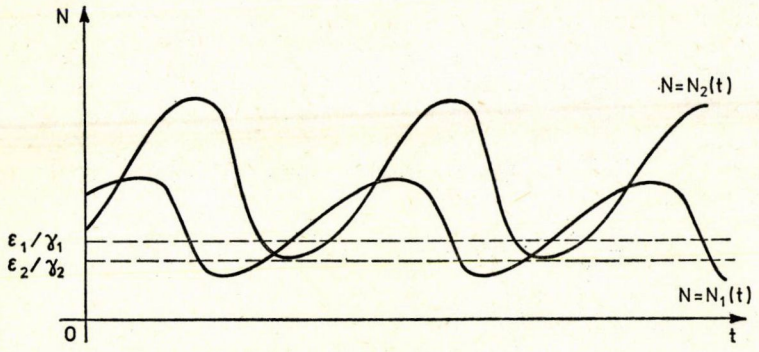
vagyis *mind a ragadozó, mind pedig a zsákmány integrálközepe, átlagos mennyisége egy periódusnyi időtartam alatt a (pozitív) kezdeti értéktől független állandó.*

Az (1.3) rendszer vizsgálata tehát azt mutatja, hogy bármilyen pozitív kezdeti értékpár mellett mind a zsákmány, mind pedig a ragadozó mennyisége az időnek periodikus függvénye ugyanazzal a periódussal; a zsákmány maximumhelyei, illetve a ragadozó maximumhelyei az időtengelyen elválasztják egymást és természetesen ugyanez a helyzet a minimumhelyekkel is. A zsákmány átlagos mennyisége $\varepsilon_2/\gamma_2$, a ragadozó átlagos mennyisége $\varepsilon_1/\gamma_1$, és ezek az értékek függetlenek a kezdeti értéktől. Ha a kezdeti értékeket kicsit megváltoztatjuk, akkor a pálya kissé deformálódik, a periódusidő kissé megváltozik, de az átlagértékek nem változnak. Ha a kezdeti értékek közel vannak az $(\varepsilon_2/\gamma_2, \varepsilon_1/\gamma_1)$ egyensúlyi helyzethez, akkor az oszcilláció kicsi, a megoldás koordinátafüggvényei közelítőleg állandók, ha a kezdeti értékek távol




*1. ábra.*
A *Lotka—Volterra modell* pályagörbéi

vannak az egyensúlyi helyzettől, akkor az oszcilláció erős, nagy maximumok és kis minimumok váltogatják egymást. Azokban a megoldásokban, amelyek túl közel kerülnek az $N_1$, illetve az $N_2$ tengelyhez, nem lehet megbízni, mivel ilyen esetben a modellben elhanyagolt statisztikus ingadozások a túlságosan kis pozitív mennyiségben jelenlevő faj számát zérusra csökkenthetik. Ha $N_i$ egyszer zérussá vált, akkor onnan már „nem támad fel". (Lásd az 1. és 2. ábrát.)




*2. ábra.*
A *Lotka—Volterra modell* megoldásának koordinátafüggvényei és ezek átlagértékei

Tételezzük fel most, hogy állandó, rendszeres „*lehalászás*" folyik, melynek során mindkét faj mennyiségét csökkentjük a meglevő mennyiségekkel arányos módon. Ez azt jelenti, hogy a zsákmány, ill. a ragadozó mennyiségének időegység alatti megváltozását $\varepsilon N_1$-gyel, ill. $\varepsilon N_2$-vel csökkentjük, vagyis (1.3) első egyenletének jobb oldalából $\varepsilon N_1$-et, második egyenletének jobb oldalából pedig $\varepsilon N_2$-t vonunk le. Ily módon az eredeti rendszerrel analóg felépítésű rendszert kapunk, amelyben az (1.3)-ban szereplő $\varepsilon_1$, illetve $\varepsilon_2$ szerepét $\varepsilon_1-\varepsilon$, illetve $\varepsilon_2+\varepsilon$ veszi át. Feltételezzük, hogy $0<\varepsilon<\varepsilon_1$. Ha ezeket az értékeket (1.5)-ben $\varepsilon_1$, illetve $\varepsilon_2$ helyébe beírjuk, azt tapasztaljuk, hogy *a fajok számarányával arányos lehalászás következtében a zsákmány átlagos mennyisége megnő, a ragadozó átlagos mennyisége pedig lecsökken.* Megfordítva, ha a lehalászást megszüntetjük, a zsákmány mennyisége lecsökken, a ragadozó mennyisége pedig megnő. Ez a modell tehát összhangban van az adriai halászok tapasztalataival.

A szóban forgó jelenségre egy másik jelentős és érdekes példa is ismeretes. 1868-ban véletlenül behurcolták Ausztráliából az Egyesült Államok nyugati vidékeire a pajzstetű egy fajtáját *(Icerya purchasi)*, amely tönkretette a citrus-félék termését. A kétségbeesett farmerek ezután behozatták ennek a pajzstetűnek ausztráliai természetes pusztítóját, egy katicabogár fajt *(Rodolia cardinalis)*. Ez hatásos módszernek bizonyult, a katicabogarak (ragadozó) jelentősen redukálták és bizony küszöbérték alatt tartották a pajzstetvek (zsákmány) mennyiségét. A DDT rovarirtószer feltalálása után elkezdték ezeket a citrusültetvényeken is alkalmazni. A hatás katasztrofális volt. Az előbbi modell segítségével felismert törvényszerűség szerint ugyanis a katicabogarak és a pajzstetvek egyidejű, rendszeres pusztítása nyomán a katicabogarak mennyisége lecsökkent, a pajzstetveké pedig megnőtt.

## 2. A Lotka—Volterra modell kritikája

Az előző pontban tárgyalt ragadozó–zsákmány modell több szempontból kifogásolható. Maga VOLTERRA is foglalkozott már annak módosításával, javításával.

Az első kifogás természetesen az, hogy amint erre már utaltunk, ragadozó hiányában, az $N_2=0$ esetben a modell (1.1)-re redukálódik, vagyis a malthusi exponenciális népesedési törvényre vezet. A modellnek ez a hibája viszonylag egyszerűen kiküszöbölhető. Már a múlt század közepén módosította VERHULST az egyetlen fajra vonatkozó (1.1) differenciálegyenletet úgy, hogy feltételezte: a természetes szaporodási ráta a faj egyedszámának növekedésével csökken, és ha a szám eléri azt a kritikus értéket, amennyit az ökológiai környezet tartósan fenntartani képes, akkor zérussá válik. Ezt, az állandó természetes szaporodási ráta feltételezésénél lényegesen realisztikusabb feltevést úgy érvényesíthetjük, hogy (1.1)-ben $\varepsilon_1>0$ helyébe $\varepsilon_1(1-N_1/K)$-t írunk, vagyis az

$$\dot{N}_1 = \varepsilon_1(1-N_1/K)N_1 \tag{2.1}$$

differenciálegyenletet tekintjük, ahol a $K>0$ állandó a *környezet fenntartó képessége.* Ez az ún. *logisztikai differenciálegyenlet,* amelynek tetszőleges $N_1(0)$ kezdeti értékhez tartozó megoldása

$$N_1(t) = \frac{KN_1(0)e^{\varepsilon_1 t}}{K+N_1(0)(e^{\varepsilon_1 t}-1)}$$

*a logisztikai függvény.* Ha $0<N_1(0)<K$, akkor a függvény szigorúan monoton növekedően $K$-hoz tart $t\to\infty$ esetén. Ha $N_1(0)>K$, akkor a függvény szigorúan monoton fogyólag tart $K$-hoz $t\to\infty$ esetén. Érdemes megjegyezni, hogy ha a kezdeti érték elég kicsi, pontosabban ha kisebb, mint a fenntartó képesség fele, $0<N_1(0)<K/2$, akkor az $N_1$ függvény eleinte konvex módon, vagyis növekvő deriválttal nő, majd egy inflexió után tart konkáv módon $K$-hoz. Ha $K/2\leqq N_1(0)<K$, akkor $N_1$ kezdettől fogva konkáv, vagyis csökkenő deriválttal tart $K$-hoz.

A második probléma, ami az (1.3) rendszerrel kapcsolatban felvethető, a következő. Az elfogadható, hogy a ragadozó-szám megváltozása azonnal, késleltetés nélkül hat a zsákmány természetes szaporodási rátájára, azonban a zsákmány-mennyiség megváltozásától reálisan azt várjuk, hogy csak bizonyos késleltetéssel hat a ragadozó természetes szaporodási rátájára, a késleltetés minimálisan pl. a ragadozó vemhességi ideje, vagy pl. a ragadozó ivarérett korba jutásának ideje. Más szóval olyan modell közelítené pontosabban a valóságos jelenséget, amelyben a ragadozó természetes szaporodási rátája a $t$ időpillanatban nem csupán a zsákmány $t$ időpontbeli mennyiségétől, hanem a zsákmánynak a $t$ időpontot megelőző időszakbeli, múltbeli mennyiségétől is függ. Ezt elérhetjük, ha (1.3) második egyenletében $N_1(t)$-t, a zsákmány pillanatnyi mennyiségét a zsákmány mennyiségének a $(-\infty, t)$ időintervallumra vonatkozó valamilyen súlyozott átlagával helyettesítjük. Legyen $G$: $[0, \infty)\mapsto R$ valamilyen integrálható súlyfüggvény, vagyis nem negatív és

$$\int_0^\infty G(s)\,ds = 1.$$

Ekkor $N_1$-nek $G$-vel súlyozott átlagértéke a $(-\infty, t)$ intervallumon:

$$N_3(t) = \int_{-\infty}^{t} N_1(\tau) G(t-\tau)\, d\tau. \tag{2.2}$$

Ezt az értéket kell (1.3) második egyenletében $N_1$ helyébe írni. $G$ megválasztásától függ, hogy a múlt milyen szakaszait milyen súllyal vesszük figyelembe.

Az (1.3) rendszerrel szemben támasztható, talán legsúlyosabb kifogás az, hogy az *nem strukturálisan stabilis*. Ezen azt értjük, hogy a rendszer jobb oldalának kis megváltoztatása, megzavarása, perturbációja az 1. ábrán látható pályák rendszerét minőségileg megváltoztathatja. Ez annak a következménye, hogy a rendszer egyensúlyi helyzete centrum, vagyis az ebben a pontban linearizált rendszer sajátértékeinek valós része zérus. Az (1.3) rendszer jobb oldalának tetszőleges kis perturbációja eredményeként a centrum körüli zárt pályákból álló fázissíkbeli kép összeomolhat és átadhatja helyét egy stabilis vagy instabilis fókusznak.

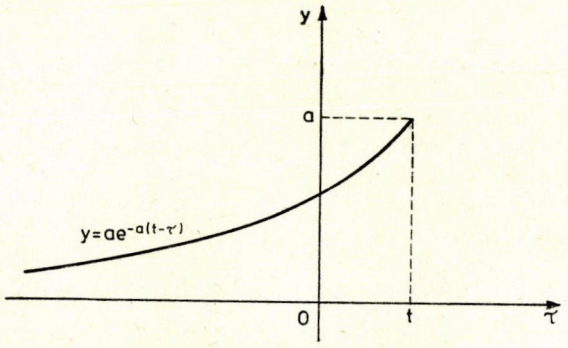
Az előbbiekben vázolt három vonatkozásban viszonylag könnyen javítható a modell. Az (1.3) rendszer első egyenletébe beírjuk a logisztikai növekedést biztosító tényezőt, a második egyenletbe pedig a múlt hatását figyelembe vevő (2.2) függvényt. Ha itt a $G$ sűrűségfüggvényt alkalmasan választjuk meg, akkor ezzel a rendszer egyben strukturálisan stabilissá is válik. Az (1.3) rendszer helyett tehát a következőt vizsgáljuk (lásd [2, 12, 13, 3, 4])

$$\begin{aligned} \dot{N}_1 &= \varepsilon_1(1-N_1/K)N_1 - \gamma_1 N_1 N_2 \\ \dot{N}_2 &= -\varepsilon_2 N_2 + \gamma_2 N_2 N_3, \end{aligned} \tag{2.3}$$

ahol $N_3$-at (2.2) szolgáltatja. A $G$ súlyfüggvény választásának egy lehetősége az, hogy ha a múlt késleltetett hatását a jelenre az időbeli távolsággal exponenciálisan lecsengő módon vesszük figyelembe, vagyis ha

$$G(s) = ae^{-as}, \quad a > 0. \tag{2.4}$$

Könnyű belátni, hogy $G$ kielégíti a súlyfüggvényekkel szemben támasztott követelményeket, és a múlt hatása annál erősebb, minél kisebb a pozitív $a$ állandó. Ez utób-




*3. ábra.*
Exponenciálisan lecsengő súlyfüggvény

bit például úgy is kifejezhetjük, hogy azt mondjuk, a késleltetés $1/a$-val arányos. A (2.2) függvény ekkor a következő alakot veszi fel:

$$N_3(t) = a \int_{-\infty}^{t} N_1(\tau) e^{-a(t-\tau)}\, d\tau.$$

Ha ezt (2.3)-ba helyettesítjük, látható, hogy $N_1$, $N_2$-re egy integro-differenciálegyenlet rendszert kell megoldanunk. Ebben az esetben azonban az integro-differenciálegyenlet vizsgálatának bonyolult feladata egy eggyel magasabb dimenziós, közönséges, autonóm differenciálegyenlet rendszer vizsgálatára redukálódik. Ugyanis, ha $N_3$ legutóbbi kifejezését differenciáljuk, azt kapjuk, hogy

$$\dot{N}_3(t) \equiv a\big(N_1(t) - N_3(t)\big).$$

Ezek szerint (2.3) ekvivalens az

$$\begin{aligned} \dot{N}_1 &= \varepsilon_1(1 - N_1/K)N_1 - \gamma_1 N_1 N_2 \\ \dot{N}_2 &= -\varepsilon_2 N_2 + \gamma_2 N_2 N_3 \\ \dot{N}_3 &= a(N_1 - N_3) \end{aligned} \tag{2.5}$$

differenciálegyenlet rendszerrel. A következőkben ennek a rendszernek a vizsgálatával fogunk foglalkozni.

Megjegyezzük, hogy a modell természetesen tovább javítható úgy, hogy jobban közelítse a valóságos viszonyokat. Így például figyelembe vehető a populációk kor szerinti, illetve nemek szerinti megoszlása, illetve az, hogy a populációk térbelileg nem egy pontszerű helyen vannak lokalizálva, hanem valamilyen területen térbelileg eloszlanak stb. (lásd pl. [15]). A modell ilyen jellegű finomításai azt egyben lényegesen bonyolultabbá teszik, így például közönséges differenciálegyenlet rendszer helyett parciális differenciálegyenlet rendszerre vezetnek. Egy másik lehetőség a jelenségeket befolyásoló sztochasztikus hatások figyelembevétele. Mindezekkel a lehetőségekkel itt nem foglalkozhatunk.

### 3. A módosított modell egyensúlyi helyzeteinek vizsgálata

A (2.5) differenciálegyenlet rendszer vizsgálata leegyszerűsödik, ha az $N_1 = Kn_1$, $N_2 = Kn_2$, $N_3 = Kn_3$ és a $t = s/\varepsilon_1$ transzformációval új „dimenziótlan" változókra térünk át. Az új változókban a differenciálegyenlet rendszer a következő alakot veszi fel

$$\begin{aligned} \frac{dn_1}{ds} &= n_1(1 - n_1) - n_1 n_2 \gamma_1 K/\varepsilon_1 \\ \frac{dn_2}{ds} &= -n_2\varepsilon_2/\varepsilon_1 + n_2 n_3 \gamma_2 K/\varepsilon_1 \\ \frac{dn_3}{ds} &= (n_1 - n_3)a/\varepsilon_1. \end{aligned} \tag{3.1}$$

A (3.1) rendszernek (ahol, ismételjük, $\varepsilon_i > 0$, $\gamma_i > 0$, $i = 1, 2$, $K > 0$, $a > 0$ állandók) három egyensúlyi helyzete van: (0, 0, 0), (1, 0, 1), ill.

$$(\bar{n}_1, \bar{n}_2, \bar{n}_3) = (\varepsilon_2/K\gamma_2, (1-\varepsilon_2/K\gamma_2)\varepsilon_1/\gamma_1 K, \varepsilon_2/K\gamma_2), \tag{3.2}$$

A (0, 0, 0) egyensúlyi helyzet instabilis és érdektelen. Az (1, 0, 1) egyensúlyi helyzet, amint erről az ebben a pontban linearizált (3.1) rendszer sajátértékeinek vizsgálatával könnyen meggyőződhetünk, aszimptotikusan stabilis, ha $\varepsilon_2/K\gamma_2 > 1$ és instabilis, ha $\varepsilon_2/K\gamma_2 < 1$. Ezek közül az első egyenlőtlenség szemléletes tartalma az, hogy a környezet fenntartó képessége és a zsákmánynak ragadozóvá történő feldolgozódási rátája kicsi a ragadozó mortalitásához viszonyítva, ami a ragadozó kihalását ($n_2 = 0$) vonja maga után. A (3.2) egyensúlyi helyzet akkor értelmes, ha az $n_1$, $n_2$, $n_3$ tér pozitív oktánsába esik. Ennek szükséges és elégséges feltétele nyilván az, hogy fennálljon az

$$\varepsilon_2/K\gamma_2 < 1 \tag{3.3}$$

egyenlőtlenség, ami egyben az (1, 0, 1) egyensúlyi helyzet instabilitásának feltétele.

A továbbiakban feltételezzük, hogy (3.3) fennáll. Ekkor tehát a (3.1) rendszernek legfeljebb egy stabilis egyensúlyi helyzete lehet, éspedig (3.2), amely a pozitív oktánsban helyezkedik el. A (3.2) egyensúlyi helyzet stabilitásának vizsgálata érdekében linearizáljuk a (3.1) rendszert ebben a pontban. A linearizált rendszer együttható mátrixa

$$\begin{bmatrix} -\varepsilon_2/K\gamma_2 & -\varepsilon_2\gamma_1/\varepsilon_1\gamma_2 & 0 \\ 0 & 0 & (1-\varepsilon_2/K\gamma_2)\gamma_2/\gamma_1 \\ a/\varepsilon_1 & 0 & -a/\varepsilon_1 \end{bmatrix},$$

a karakterisztikus egyenlet pedig

$$\lambda^3 + (\varepsilon_2/K\gamma_2 + a/\varepsilon_1)\lambda^2 + \lambda\varepsilon_2 a/\varepsilon_1 K\gamma_2 + (1-\varepsilon_2/K\gamma_2)a\varepsilon_2/\varepsilon_1^2 = 0. \tag{3.4}$$

A *Routh—Hurwitz kritérium* szerint ez akkor és csak akkor stabilis polinom (akkor és csak akkor negatívok gyökeinek valós részei), ha

$$(\varepsilon_2/K\gamma_2 + a/\varepsilon_1)\varepsilon_2 a/\varepsilon_1 K\gamma_2 > (1-\varepsilon_2/K\gamma_2)a\varepsilon_2/\varepsilon_1^2,$$

vagyis ha (3.3) mellett

$$a > K\gamma_2 - \varepsilon_2 - \varepsilon_1\varepsilon_2/K\gamma_2.$$

Vezessük be ezen egyenlőtlenség jobb oldalára az

$$a_0 = K\gamma_2 - \varepsilon_2 - \varepsilon_1\varepsilon_2/K\gamma_2$$

jelölést. Ha $a_0$ értéke negatív, vagy zérus, ami (3.3) figyelembe vételével azt jelenti, hogy $\varepsilon_1$, a zsákmány természetes szaporodási rátája elég nagy, akkor a (3.2) egyensúlyi helyzet minden pozitív $a$ érték mellett aszimptotikusan stabilis. Ez az eredmény összhangban van a szemlélettel: ha a környezet fenntartó képessége $K$, a zsákmánynak ragadozóvá való feldolgozódási rátája $\gamma_2$ és a zsákmány szaporodási rátája $\varepsilon_1$ elég nagy, továbbá a ragadozó mortalitása $\varepsilon_2$ kicsi, akkor a két faj tetszőlegesen nagy időkésleltetés (tetszőlegesen kicsiny $a$ érték) mellett is stabilisan együtt tud élni.

Érdekesebb a helyzet akkor, ha

$$a_0 = K\gamma_2 - \varepsilon_2 - \varepsilon_1\varepsilon_2/K\gamma_2 > 0. \tag{3.5}$$

Ez a feltétel implikálja (3.3)-at, az utóbbit tehát nem kell külön feltételezni. Ha $a>a_0$, vagyis a zsákmánynak ragadozóvá való feldolgozódása viszonylag kis késleltetéssel történik, akkor a (3.2) egyensúlyi helyzet aszimptotikusan stabilis. Ha azonban a késleltetés nagy ($a$ kicsi), pontosabban, ha $0<a<a_0$, akkor a (3.2) egyensúlyi helyzet instabilis. Más szavakkal, ha a késleltetést növeljük ($a$-t csökkentjük) és lefelé átlépjük az $a=a_0$ értéket, a korábban aszimptotikusan stabilis egyensúly destabilizálódik. Ez is példa annak a „szabálynak" az érvényesülésére, amely szerint *„a késleltetés destabilizál"*. Ezzel a „szabállyal" azonban csínján kell bánni. Számos példa mutatja ugyanis, hogy *a destabilizáló késleltetés további növelése ismét stabilizálhatja a rendszert.*

## 4. Az egyensúlyi helyzet stabilitásvesztése

A továbbiakban (3.5)-öt feltételezve (ami maga után vonja (3.3) teljesülését) azt vizsgáljuk, mi történik a rendszerrel, ha $a$ értékét $a_0$ alá csökkentjük, és így a rendszer egyetlen stabilis egyensúlyi helyzete elveszti stabilitását.

Vezessük be a $b=1/K\gamma_2$ jelölést. Ekkor a (3.5) feltétel az

$$1-\varepsilon_2 b-\varepsilon_1\varepsilon_2 b^2>0 \tag{4.1}$$

alakot ölti és

$$a_0=(1-\varepsilon_2 b-\varepsilon_1\varepsilon_2 b^2)/b.$$

A következményként teljesülő (3.3) feltétel most: $1-\varepsilon_2 b>0$. A (3.2) egyensúlyi helyzetben linearizált rendszer karakterisztikus polinomja az $a=a_0$ kritikus helyzetben:

$$\lambda^3+(1/b\varepsilon_1-\varepsilon_2/\varepsilon_1)\lambda^2+\lambda(1-\varepsilon_2 b-\varepsilon_1\varepsilon_2 b^2)\varepsilon_2/\varepsilon_1+$$
$$+(1-\varepsilon_2 b)(1-\varepsilon_2 b-\varepsilon_1\varepsilon_2 b^2)\varepsilon_2/b\varepsilon_1^2=$$
$$=\left(\lambda^2+(1-\varepsilon_2 b-\varepsilon_1\varepsilon_2 b^2)\varepsilon_2/\varepsilon_1\right)\left(\lambda+(1-\varepsilon_2 b)/b\varepsilon_1\right).$$

A sajátértékek: $\lambda_0(a_0)=-(1-\varepsilon_2 b)/b\varepsilon_1$, ami negatív, és $\lambda_{1,2}(a_0)=\pm i\omega$, ahol

$$\omega=\left((1-\varepsilon_2 b-\varepsilon_1\varepsilon_2 b^2)\varepsilon_2/\varepsilon_1\right)^{1/2}>0. \tag{4.2}$$

Mivel $\lambda_1(a_0)=i\omega$ egyszeres gyök, az implicit függvény tétel alkalmazásával, a (3.4) egyenletből elvileg kifejezhetjük $\lambda$-t mint az $a$ paraméter egyértelműen meghatározott, $a_0$ egy környezetében folytonosan differenciálható $\lambda_1(a)$ függvényét, amely az $a=a_0$ helyen az $i\omega$ értéket veszi fel. Szükségünk lesz e függvény $a$ szerinti deriváltjának valós részére az $a_0$ helyen, ami explicit módon előállítható:

$$\frac{d\,\mathrm{Re}\,\lambda_1(a_0)}{da}=\mathrm{Re}\frac{d\lambda_1(a_0)}{da}=$$

$$=-\frac{\varepsilon_2 b^2}{2}\,\frac{1-\varepsilon_2 b-\varepsilon_1\varepsilon_2 b^2}{\varepsilon_1\varepsilon_2 b^2(1-\varepsilon_2 b-\varepsilon_1\varepsilon_2 b^2)+(1-\varepsilon_2 b)^2}.$$

A (4.1) egyenlőtlenség következtében $d\,\mathrm{Re}\,\lambda_1(a_0)/da<0$, vagyis teljesülnek a *Hopf-féle bifurkációs tétel* feltételei (lásd pl. [7]). Bifurkációs paraméternek természetesen $a$-t tekintjük. Ahhoz, hogy a szokásos módon kimondott *Hopf-tétellel* összhangban

legyünk, a koordinátarendszer origóját eltoljuk a (3.2) egyensúlyi helyzetbe és bevezetjük például a $\mu = 1/a - 1/a_0$ bifurkációs paramétert. Ekkor $a$ csökkenésével, vagyis a késleltetés növekedésével $\mu$ nő, és $a = a_0$-nál lesz $\mu = 0$). Kissé leegyszerűsítve a Hopf-tétel állításait mondhatjuk, hogy a (3.2) egyensúlyi helyzetnek van olyan környezete és az $a_0$ érték bármely környezetében van olyan $a$ szám, hogy ha ez utóbbit írjuk a (3.1) rendszerbe, e rendszernek van a (3.2) pont adott környezetében levő zárt pályája (amely nem egyensúlyi helyzet). Más szóval bizonyos $a_0$-hoz közeli $a$ értékekre (3.1)-nek van a (3.2) egyensúlyi helyzethez közeli periodikus megoldása. Valójában létezik periodikus megoldásoknak egy egyparaméteres serege, amely a (3.2) pontra húzódik össze és amely az ettől a paramétertől folytonosan differenciálható módon függő $a$ értékekhez tartozik. E seregen kívül a (3.2) pont adott környezetében más periodikus megoldás nincs. Meg tudjuk adni a (3.2)-höz elég közeli periodikus megoldások periódusát is közelítőleg; ez a periódus közelítőleg $2\pi/\omega$-val egyenlő.

A *Hopf-féle bifurkációs tétel* azonban csak bizonyos kiegészítő feltevések mellett biztosítja azt, hogy a periodikus megoldások (a zárt pályák) $a_0$-nál *kisebb* $a$ értékekre jelentkezzenek, vagyis akkor, amikor az egyensúlyi helyzet már instabilissá vált, és maguk orbitálisan aszimptotikusan stabilisak legyenek. Márpedig ez, hogy ti. a bifurkáció szuperkritikus legyen, a stabilitásvesztésnek igen fontos és viszonylag kedvező esete. Ha ugyanis a késleltetés növelése ($a$ csökkentése) során átlépjük lefelé az $a_0$ értéket, és a bifurkáció *szuperkritikus,* akkor igaz ugyan, hogy a rendszernek nem lesz többé aszimptotikusan stabilis egyensúlyi helyzete, ezt azonban a rendszer jelzi azáltal, hogy stabilisan rezegni fog a korábbi egyensúlyi helyzet körül kis amplitudókkal. A késleltetés további növelése ekkor növeli a rezgések amplitudóját és végül teljes instabilitáshoz, illetve a rendszer kiszámíthatatlan viselkedéséhez vezethet. Ezt azonban, elvileg időben, előre észleljük és esetleg a késleltetés növelésének megállításával megakadályozhatjuk. Ha azonban a bifurkáció nem szuperkritikus, hanem pl. *szubkritikus*, vagyis a zárt pályák $a_0$-nál nagyobb $a$ értékek mellett jelentkeznek és instabilisak, akkor az $a_0$ küszöb lefelé történő átlépésekor az elveszett stabilis egyensúlyi helyzet szerepét semmi sem veszi át, és a rendszer viselkedése minden előrejelzés nélkül, azonnal kiszámíthatatlanná válik.

A *Hopf-bifurkáció* jellege, ti. az, hogy szuperkritikus vagy szubkritikus-e, egy mennyiség előjelén múlik, ennek a mennyiségnek a meghatározása azonban különösen kettőnél magasabb dimenziós rendszer esetén bonyolult feladat. A mi esetünkben a számítás elvégezhető és fennáll a következő

4.1. TÉTEL. *Ha* (4.1) *fennáll és a*

$$B = (1+2\varepsilon_1 b)\left[2\varepsilon_2 b(1-\varepsilon_2 b^2)+2\varepsilon_1\varepsilon_2 b^2\left(-(1-2\varepsilon_2 b)(1-\varepsilon_2 b-\varepsilon_1\varepsilon_2 b^2)-\right.\right.$$
$$-2\varepsilon_1\varepsilon_2 b^2(1-\varepsilon_2 b)+\varepsilon_1^2\varepsilon_2^2 b^4)\left.\right]+\left(\varepsilon_1\varepsilon_2^2 b^3(1-\varepsilon_2 b)/\omega^2+\varepsilon_1 b(1-\varepsilon_2 b)+\right.$$
$$\left.+\varepsilon_1^2\varepsilon_2 b^3\right)\left[(1-\varepsilon_2 b)(1-2\varepsilon_2 b)+2\varepsilon_1\varepsilon_2 b^2(1-\varepsilon_2 b-2\varepsilon_1\varepsilon_2 b^2)\right]$$

*kifejezés pozitív (negatív), akkor az $a_0$ értéknél bekövetkező Hopf-bifurkáció szuperkritikus (szubkritikus), vagyis van olyan $\delta > 0$, hogy minden $a \in (a_0-\delta, a_0)$-ra $(a \in (a_0, a_0+\delta)$-ra) $a$ (3.1) rendszernek van egyetlen, (3.2)-höz közeli, orbitálisan aszimptotikusan stabilis (instabilis) zárt pályája $2\pi/\omega$-hoz közeli periódussal.*

E tétel bizonyítása a [4] dolgozatban található. Ahhoz, hogy bebizonyítsuk, először meg kellett határozni a (3.2) egyensúlyi helyzethez és az $a = a_0$ értékhez

tartozó *központi sokaságot* és le kellett szűkítenünk a differenciálegyenlet rendszert erre a kétdimenziós invariáns sokaságra. Ezután egy *Ljapunov-függvényeken* alapuló módszert alkalmaztunk a kétdimenziós Hopf bifurkáció jellegének eldöntésére. A tételben szereplő $B$ előjele a kétdimenziós rendszer ún. *Poincaré-állandója* előjelének $-1$-szerese.

A 4.1. tétel egy „majdnem szükséges és elégséges feltételt" ad a bifurkáció jellegének eldöntésére, csupán a $B=0$ eset marad elintézetlen. $B$ azonban túlságosan bonyolult kifejezés, ezért egy elégséges feltételt is adunk a szuperkritikusságra.

4.2. TÉTEL. *Ha*

$$1-3\varepsilon_2 b-3\varepsilon_1\varepsilon_2 b^2 > 0 \tag{4.3}$$

*és*

$$1-\varepsilon_2 b-2\varepsilon_1^2\varepsilon_2 b^3 > 0, \tag{4.4}$$

*akkor* (4.1) *fennáll, vagyis* $a_0>0$ *és a 4.1. tételben szereplő* $B>0$, *tehát van olyan* $\delta>0$, *hogy minden* $a\in(a_0-\delta, a_0)$*-ra a* (3.1) *rendszernek van egyetlen,* (3.2)*-höz közeli, orbitálisan aszimptotikusan stabilis zárt pályája* $2\pi/\omega$*-hoz közeli periódussal.*

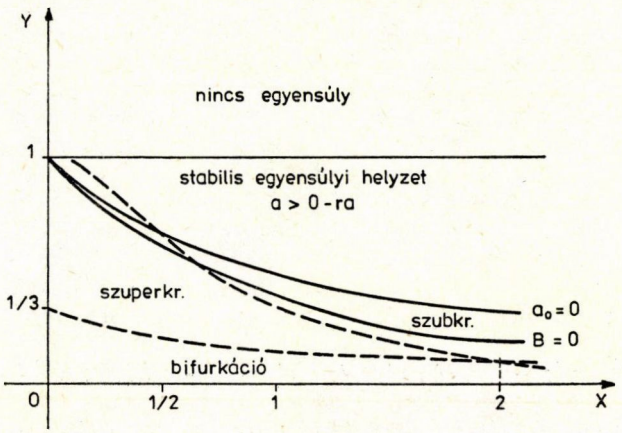
E tétel bizonyítására nézve lásd ugyancsak a [4] dolgozatot.

Megjegyezzük, hogy az eredeti $t$ időskálán *a bifurkálódó periodikus megoldások periódusa közelítőleg*

$$2\pi/\omega\varepsilon_1 = 2\pi\left(\varepsilon_1\varepsilon_2(1-\varepsilon_2/K\gamma_2-\varepsilon_1\varepsilon_2/K^2\gamma_2^2)\right)^{-1/2}.$$

Ezzel a közelítő periódussal oszcillálnak a létszámok a (3.2)-nek megfelelő

$$(N_1, N_2, N_3) = \left(\varepsilon_2/\gamma_2, (1-\varepsilon_2/K\gamma_2)\varepsilon_1/\gamma_1, \varepsilon_2/\gamma_2\right) \tag{4.5}$$




*4. ábra.*

Az $a_0=0$ $(y=(1+x)^{-1})$ görbe felett az egyensúlyi helyzet stabilis minden $a>0$-ra. E görbe és a $B(x, y)=0$ görbe között *szubkritikus Hopf bifurkáció* van $a=a_0>0$-nál. A $B(x, y)=0$ görbe alatt *szuperkritikus Hopf bifurkáció* van $a=a_0>0$-nál. Mindkét szaggatott görbe alatt áll fenn (4.3) és (4.4)

egyensúlyi helyzet körül, ha $a$ csak kevéssel kisebb $a_0$-nál (vagy ami ezzel ekvivalens, ha az amplitudó elég kicsi).

A 4. ábra az $x=\varepsilon_1 b$, $y=\varepsilon_2 b$ paramétersík pozitív kvadránsát ábrázolja. Könnyű látni, hogy $B$ is és a (4.3), (4.4) egyenlőtlenségek bal oldalai is csak $x$-en és $y$-on keresztül függenek a paraméterekből. Az $a_0=0$ görbe felett $a_0<0$, tehát ilyen paraméterértékek esetén a (3.2) egyensúlyi helyzet minden pozitív $a$-ra aszimptotikusan stabilis. $B(x, y)=0$ görbét számítógéppel határoztuk meg (a számítógépes kiértékelést Farkas Attila végezte, akinek ezért köszönetet mondok). A $B=0$ görbe fölött és az $a_0=0$ görbe alatt $B<0$, a *Hopf-bifurkáció* szubkritikus. A $B=0$ görbe alatt a *Hopf-bifurkáció* szuperkritikus. A szaggatott görbék alatti tartományok a (4.3), illetve a (4.4) egyenlőtlenségnek felelnek meg. A 4.2. tétel a tényleges szuperkritikus tartománynak csak a mindkét szaggatott görbe alá eső részét szolgáltatja.

Felhívjuk még a figyelmet arra, hogy a $\gamma_1$ ragadozási ráta sehol sem szerepel a feltételekben. Ez azért van így, mert a (2.3) rendszerben az $N_2$ változó helyett be lehetne vezetni a $\gamma_1 N_2$ változót anélkül, hogy a rendszer kvalitatív tulajdonságai megváltoznának. Ily módon a $\gamma_1$ paraméter „kitranszformálható".

## 5. A versengő kizárás elve

VOLTERRA idézett művében [17] nemcsak ragadozó—zsákmány modellekkel foglalkozik, hanem olyan ökológiai rendszereket is vizsgál, amelyekben két vagy több csoport (faj) verseng egy vagy több erőforrás (táplálék) megszerzéséért. A biológiában mintegy tapasztalati tényként már korábban megfogalmazódott a *„versengő kizárás elve" (competitive exclusion principle)*, amely szerint, ha két faj egyetlen táplálékforráson él és azért verseng, akkor a gyengébb paraméterekkel rendelkező faj kihal. A VOLTERRA által alkotott modell „hozza" ezt az elvet. Jelölje az $i$-edik faj mennyiségét (létszámát) $N_i$, természetes szaporodási rátáját akkor, amikor mindkét faj létszáma zérushoz közel van, $\varepsilon_i$ $(i=1, 2)$. Tételezzük fel, hogy a természetes szaporodási ráta a rendelkezésre álló táplálék mennyiségétől, ez pedig a két faj létszámától függ, és legyen $F(N_1, N_2)$ *az időegység alatt elfogyasztott táplálékmennyiség*, amikor az első, ill. a második faj $N_1$, ill. $N_2$ mennyiségben van jelen. Feltesszük, hogy ha $N_1=N_2=0$, akkor a táplálékmennyiség időegység alatt állandó értékekkel növekszik, és $F(N_1, N_2)$ valójában ebből a növekedésből vonódik le. A táplálékmennyiség időegység alatti növekedése (fogyasztó hiányában) azonban be van épülve az elvileg zérus $(N_1, N_2)$ érték melletti $\varepsilon_i$ természetes szaporodási rátába $(i=1, 2)$. Végül jelölje $\gamma_i$ azt az arányossági tényezőt, amely megmutatja, hogy a táplálékmennyiség időegység alatti egységnyi csökkenése mennyivel csökkenti az $i$-edik faj természetes szaporodási rátáját. E jelölésekkel a Volterra-féle modell a következő:

$$\dot{N}_1=(\varepsilon_1-\gamma_1 F(N_1, N_2))N_1, \quad \dot{N}_2=(\varepsilon_2-\gamma_2 F(N_1, N_2))N_2, \tag{5.1}$$

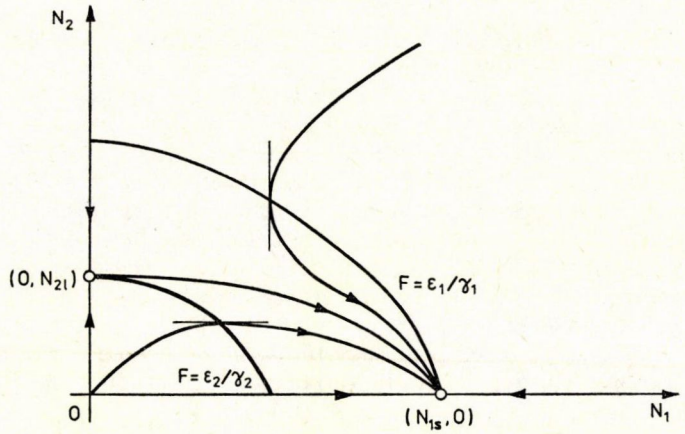
ahol $\varepsilon_i, \gamma_i>0$ állandók $(i=1, 2)$, az $F$ függvény pozitív, ha $N_1$ és $N_2$ nem negatívok és legalább az egyik pozitív, kellően sima, rögzített $N_1$ mellett $N_2$-ben, ill. $N_2$ mellett $N_1$-ben monoton növekedő, $F(0, 0)=0$, és ha $N_1^2+N_2^2\to\infty$, akkor $F(N_1, N_2)\to\infty$.

A továbbiakban feltételezzük, hogy

$$\varepsilon_1/\gamma_1 > \varepsilon_2/\gamma_2. \tag{5.2}$$

Ez a feltétel nem jelenti az általánosság megszorítását, mivel $\varepsilon_1, \gamma_1$, ill. $\varepsilon_2, \gamma_2$ szerepe szimmetrikus, és így csupán az $\varepsilon_1/\gamma_1 = \varepsilon_2/\gamma_2$ esetet zártuk ki. Ez utóbbi eset azonban egyrészt azt jelenti, hogy az adott helyzetben a két versengő faj lényegesen nem különbözik egymástól, másrészt strukturálisan nem stabilis, vagyis kis perturbációkra lényeges kvalitatív változással reagál. Világos az, hogy az $i$-edik faj fennmaradása szempontjából előnyös a nagy maximális természetes szaporodási ráta, $\varepsilon_i$. Ugyanakkor a $\gamma_i$ paraméter a faj táplálékszükségletére jellemző. Ha $\gamma_i$ kicsi, akkor a táplálékmennyiség csökkenése viszonylag csak kismértékű kedvezőtlen hatást gyakorol a faj természetes szaporodási rátájára. Tehát a faj fennmaradása szempontjából előnyös az, hogy $\gamma_i$ kicsi. Látható ezek alapján tehát az, hogy (5.2) fennállása esetén, intuitív, biológiai szempontból, egyértelműen az 1-es indexű faj életképesebb a 2-es indexűnél, hiszen az $\varepsilon_i/\gamma_i$ hányadost az életképesség mérőszámának tekinthetjük (az a jó, ha nagy, vagyis a számláló nagy és a nevező kicsi).

Az (5.1) autonóm rendszer viszonylag egyszerűen vizsgálható. Könnyű belátni, hogy a megoldások a $[0, \infty)$ időintervallumon korlátosak. Ehhez csak azt kell látni, hogy elég nagy $N_1^2+N_2^2$ esetén, az $N_1, N_2$ fázissík pozitív kvadránsában mindkét egyenlet jobb oldala negatív, vagyis $N_1$ és $N_2$ csökken. Lásd az 5. ábrát: Ha $(N_1, N_2)$ kívül van az $F(N_1, N_2)=\varepsilon_1/\gamma_1$ görbén, akkor $\dot{N}_1$ és $\dot{N}_2$ is negatív. Nyilvánvaló az is, hogy (5.2) fennállása következtében a rendszernek nincs egyensúlyi helyzete a fázissík pozitív kvadránsának belsejében és ennek következtében, a *Poincaré—Bendixson elmélet* alapján zárt pályája sincs. Pontosan három egyensúlyi helyzet van viszont a pozitív kvadráns határán, éspedig az origó, a $(0, N_{2l})$ pont, ahol $N_{2l}>0$ az $\varepsilon_2-\gamma_2 F(0, N_2)=0$ egyenlet megoldása, és az $(N_{1s}, 0)$ pont, ahol $N_{1s}>0$ az $\varepsilon_1-\gamma_1 F(N_1, 0)=0$ egyenlet megoldása. Egyszerű linearizálással megmutatható, hogy az origó instabilis csomópont, a $(0, N_{2l})$ egyensúlyi helyzet ugyancsak instabilis (labilis), ui. nyeregpont, az $(N_{1s}, 0)$ egyensúlyi helyzet viszont aszimptotikusan stabilis. A rendszer lehetséges $\omega$-határpontjainak vizsgálata alapján az is megmutatható, hogy az $(N_{1s}, 0)$ egyensúlyi helyzet *globális attraktor* (az $N_2$ tengelyen levők




*5. ábra.*

A versengő kizárás elve. $(N_{1s}, 0)$ stabilis csomópont, $(0, N_{2l})$ instabilis egyensúlyi helyzet (nyeregpont), $(0, 0)$ instabilis csomópont.
Az $F(N_1, N_2)=\varepsilon_i/\gamma_i$ görbe pontjaiban $\dot{N}_i=0$, $(i=1, 2)$

kivételével az összes első kvadránsbeli pálya ehhez a ponthoz tart, ha $t \to \infty$). Az (5.1) rendszer az (5.2) feltétellel tehát olyan, hogy bármilyen $N_1^0 > 0$, $N_2^0 > 0$ kezdeti feltételt adunk is meg, a megfelelő $(N_1(t, N_1^0, N_2^0), N_2(t, N_1^0, N_2^0))$ megoldás első koordinátája $N_{1s} > 0$-hoz, második koordinátája pedig zérushoz tart $t \to \infty$ esetén, vagyis az 1-es indexű faj fennmarad, a 2-es indexű pedig kihal.

A versengő kizárás elve érvényesülési körének vizsgálata matematikai modellek segítségével azóta is számos kutatót foglalkoztat. Az addig elért eredmények összefoglalását és továbbfejlesztését adják McGehee és Armstrong 1977-es [14] dolgozatukban. Először olyan $n$ számú ragadozóból és $k$ számú táplálékforrásból (korlátozó tényezőből) álló ökológiai rendszereket vizsgálnak, amelyek dinamikáját

$$\dot{N}_i = N_i u_i(z_1, \ldots, z_k), \quad (i = 1, 2, \ldots, n) \tag{5.3}$$

$$z_j = r_j(N_1, \ldots, N_n), \quad (j = 1, 2, \ldots, k)$$

típusú rendszerek határozzák meg. Itt $N_i$ az $i$-edik ragadozó, $z_j$ pedig a $j$-edik táplálékforrás mennyisége, $u_i, r_j \in C^\infty$. Könnyű látni, hogy ilyen a *Volterra-féle (5.1) modell* is, $n=2, k=1$. Bebizonyítják a következő tételeket. Ha $k<n$, akkor (5.3)-nak nem lehet „pont attraktora" (aszimptotikusan stabilis egyensúlyi helyzete) az $(N_1, \ldots, N_n)$ változók $R^n$ fázistere „pozitív $2^n$-ánsának" belsejében: $\{(N_1, \ldots, N_n) \in \in R^n: N_i \geqq 0, i=1, 2, \ldots, n\}$. Ha $n=2, k=1$, akkor (5.3)-nak semmiféle attraktora nem lehet a pozitív kvadráns belsejében, vagyis ebben az esetben (5.3)-ra teljesül a versengő kizárás elve. Ha $k<n$ és az $u_i$ függvények lineárisak, akkor sem lehet (5.3)-nak attraktora a pozitív kvadráns belsejében. McGehee és Armstrong foglalkoznak olyan ökológiai rendszerekkel is, amelyekben a *táplálékforrások önreprodukáló élőlények*. Ezeknek általános alakját a következőképpen veszik fel:

$$\dot{N}_i = N_i u_i(z_1, \ldots, z_k), \quad (i = 1, 2, \ldots, n) \tag{5.4}$$

$$\dot{z}_j = z_j s_j(N_1, \ldots, N_n, z_1, \ldots, z_k), \quad (j = 1, 2, \ldots, k),$$

ahol $N_i$ az $i$-edik ragadozó, $z_j$ pedig a $j$-edik zsákmány mennyisége, $u_i, s_j \in C^\infty$. Az (5.4) modellek annyiban speciálisak, hogy az $i$-edik ragadozó $u_i$ természetes szaporodási rátája közvetlenül csak a zsákmány mennyiségektől függ, a többi ragadozó mennyiségektől nem. Bebizonyítják, hogy ha $k<n$ és az $u_i$ függvények lineárisak, akkor (5.4)-nek nem lehet olyan attraktora, amelyre $N_i > 0$, $i=1, 2, \ldots, n$. Megmutatják továbbá, hogy az általános, nem lineáris esetben, ha $k \geqq n/2$, akkor lehetséges az $n$ számú ragadozó faj együttélése a $k$ számú táplálékforrás alapján.

Megjegyezzük, hogy az előbbiekben vázolt irányban 1977 óta is számos eredmény született. Jelenleg úgy tűnik, hogy az (5.3) típusú rendszerek esetében csupán az $n=2, k=1$ értékek mellett teljesül a versengő kizárás elve szükségképpen (más $n$ és $k$ értékek mellett lehet olyan rendszert konstruálni, amelyre nem teljesül). Az (5.4) típusú rendszerek esetében pedig úgy tűnik, hogy minden $n$ és $k$ érték mellett lehet olyan (nem lineáris szaporodási rátájú) modellt konstruálni, amelyre nem teljesül a versengő kizárás elve.

## 6. Versengés csökkenő mértékben szaporodó zsákmányért

Amint ezt már megjegyeztük, az előző pontban tárgyalt *Volterra-féle versengési modell* hallgatólagosan feltételezi, hogy a táplálék-(zsákmány) ellátás stacionárius, vagyis időegység alatt mindig ugyanolyan mennyiségű táplálék keletkezik, illetve érkezik be az ökológiai környezetbe. Ha az időegység alatt keletkező táplálékot $z$-vel jelöljük és $d_i>0$ az $i$-edik faj mortalitása táplálék hiányában, a $t$ időpontbeli természetes szaporodási ráta $\dot{N}_i/N_i=-d_i+\gamma_i(z-F(N_1, N_2))$, ahol $\gamma_i$ és $F$ jelentése ugyanaz, mint az 5. pontban. Feltéve, hogy

$$\varepsilon_i=-d_i+\gamma_i z>0, \tag{6.1}$$

kapjuk az (5.1) modellt.

Labóratóriumi környezetben megvalósítható, hogy időegység alatt mindig ugyanannyi táplálékot adagoljunk, valóságos körülmények között azonban természetesebb feltételezni azt, hogy az időegység alatt keletkező táplálék függ a meglevő táplálékmennyiségtől például úgy, hogy a táplálékmennyiség a logisztikai egyenlet szerint szaturálódik, telítődik. Az (5.1) modell egy másik fogyatékossága az, hogy amint ez a (6.1) formulából látszik, a $z$ táplálékmennyiség növelésével a kis $N_1, N_2$ létszámoknál érvényes $\varepsilon_i$ természetes szaporodási ráta minden határon túl nő. Azt várjuk azonban, és ezt a várakozást az erre vonatkozó vizsgálatok alátámasztják, hogy bármilyen bőségben álljon is rendelkezésre táplálék, a fogyasztók természetes szaporodási rátája egy bizonyos érték fölé nem növekedik. A ragadozónak ugyanis bizonyos időbe telik, amíg táplálkozás után a következő táplálékforrást megtalálja, időbe telik a táplálék elfogyasztása és feldolgozása is. Ezek a „holtidők" felülről behatárolják azt a táplálékmennyiséget, amit a ragadozó időegység alatt el tud fogyasztani.

E megfontolások alapján HSU, HUBBEL és WALTMAN [8, 9] az egyetlen táplálékul szolgáló és az ezért versengő két ragadozó fajból álló ökológiai rendszerre a következő modellt vezették be

$$\begin{aligned}\dot{S}&=\gamma S(1-S/K)-\frac{m_1N_1S}{a_1+S}-\frac{m_2N_2S}{a_2+S}\\ \dot{N}_1&=\frac{m_1N_1S}{a_1+S}-d_1N_1\\ \dot{N}_2&=\frac{m_2N_2S}{a_2+S}-d_2N_2,\end{aligned} \tag{6.2}$$

ahol $S(t)$ a zsákmány, $N_i(t)$ az $i$-edik ragadozó létszáma (mennyisége) a $t$ időpontban, $\gamma>0$ a *zsákmány természetes szaporodási rátája,* amikor létszáma közel van zérushoz és ragadozó nincs jelen, $K>0$ a *környezet fenntartó képessége* a zsákmányra nézve, $d_i>0$, illetve $m_i>0$ az $i$-edik ragadozó *mortalitása* (amikor nincs táplálék), illetve *maximális születési rátája* (akkor éri ezt el a ragadozó, amikor $S\to\infty$), $a_i>0$ pedig az $i$-edik ragadozó *„fél-telítődési állandója"*. A fél-telítődési állandó jelentése a következő: amikor $S$ eléri az $a_i$ értéket, akkor az $i$ indexű ragadozó születési rátája éppen a maximális fele, $m_i/2$ lesz.

A (6.2) modellben a táplálék nem burkoltan, időegység alatt állandó mennyiségben jelentkezik, hanem dinamikáját külön differenciálegyenlet határozza meg. Ha

ragadozó nincs jelen, akkor ez az $S$-re vonatkozó egyenlet a logisztikai differenciálegyenlet (v.ö. (2.1)). Ebben az esetben minden $S(0)>0$ kezdeti értékhez tartozó megoldás $t\to\infty$ esetén $K$-hoz, a környezet fenntartó képességéhez tart. Gyakorlatilag ez a táplálék lehetséges maximális mennyisége hosszú távon. Ha ragadozók is jelen vannak, $S$ időegység alatti növekedéséből levonódnak a ragadozók által időegység alatt elfogyasztott mennyiségek (az első egyenlet jobb oldalának második és harmadik tagja, amelyekből elhagytunk egy-egy állandó faktort anélkül, hogy ezzel a vizsgálat eredményeit befolyásolnánk). A ragadozók mortalitását állandónak vettük, születési rátájuk viszont a táplálékmennyiség függvénye; egyenlőnek vettük az egy ragadozó által időegység alatt elfogyasztott táplálékmennyiséggel, amit az ún. *Holling-féle függvénnyel,* $m_iS/(a_i+S)$-sel adtunk meg. Ez $S$-nek monoton növekedő, de korlátos függvénye. A (6.2) modell nagy előnye, hogy a benne szereplő paraméterek értéke kísérleti úton meghatározható anélkül, hogy a versengő ökológiai rendszert létrehoznánk és megnéznénk, mi történik; elegendő a zsákmányt egyetlen ragadozóval (először az egyikkel, azután a másikkal) összetelepíteni, így a modellben szereplő összes paraméter értéke becsülhető. Ez azt jelenti, hogy a modell nemcsak a rendszer a posteriori leírására alkalmas, hanem a priori következtetések levonására, előrejelzésre is.

A (6.2) rendszerben az $i$-edik ragadozó $(i=1,2)$ szaporodási rátája akkor, amikor a zsákmány mennyisége $S$, a megfelelő differenciálegyenlet jobb oldalán $N_i$ együtthatója:

$$m_iS/(a_i+S)-d_i. \tag{6.3}$$

Világos, hogy ez a szaporodási ráta akkor nagy, ha $m_i$ nagy, $d_i$ és $a_i$ pedig kicsi. A születési ráta monoton növekedőleg $m_i$-hez tart, ha $S\to\infty$. Nyilvánvaló, hogy az $i$-edik ragadozónak csak akkor lehet esélye a túlélésre, ha $m_i>d_i$. Bevezetjük a

$$\beta_i = m_i-d_i, \quad b_i = m_i/d_i, \quad (i=1,2)$$

paramétereket; $\beta_i$ nyilván az $i$-edik ragadozó *maximális természetes szaporodási rátája.* Az előbbiek szerint *annak szükséges feltétele, hogy az i-edik ragadozó ne haljon ki:*

$$\beta_i>0, \quad (b_i>1). \tag{6.4}$$

Az $a_i$ fél-telítődési állandó kicsinységének előnyös voltát a következőképpen tudjuk megvilágítani. Ha $a_i$ kicsi, akkor már kis zsákmánymennyiség esetén, ti. $S=a_i$-nél eléri a születési ráta a maximális felét, $m_i/2$-t. Az $i$-edik ragadozó túlélőképességének, *versenyképességének* jellemzésére bevezetjük még a

$$\lambda_i = \frac{a_id_i}{m_i-d_i} = \frac{a_i}{b_i-1}, \quad (i=1,2)$$

paramétert. Ez a $\lambda_i$ paraméter *az a zsákmánymennyiség, amelynél a (6.3) természetes* szaporodási *ráta zérus.* Ha $S<\lambda_i$, akkor (6.3) és így $\dot{N}_i$ is negatív, vagyis az $i$-edik ragadozó létszáma csökken, ha $S>\lambda_i$, akkor az $i$-edik ragadozó létszáma nő. Miután hosszú távon a zsákmánymennyiség maximuma $K$-nál nem lehet nagyobb, világos, hogy az $i$-edik ragadozó túléléséhez szükséges a $\lambda_i<K$ egyenlőtlenség teljesü-

lése. Valóban, könnyen bebizonyítható, hogy *az i-edik ragadozó túlélésének szükséges feltétele:*

$$0 < \lambda_i < K, \tag{6.5}$$

ami (6.4)-et is implikálja. Világos, hogy az $i$-edik ragadozó versenyképessége szempontjából az az előnyös, ha $\lambda_i$ kicsi, vagyis $a_i$ kicsi és $b_i$ nagy.

Könnyű belátni, hogy az $S, N_1, N_2$ háromdimenziós tér pozitív oktánsából induló *megoldások korlátosak és a pozitív oktánsban maradnak.* Ehhez csak azt kell látni, egyrészt, hogy ha $N_1$ vagy $N_2$ túl nagy, akkor (6.2) első egyenlete szerint $S$ csökken, ha pedig $S$ elég kicsi, akkor a második és harmadik egyenlet szerint $N_1$ és $N_2$ is csökken. Másrészt világos, hogy az $S=0$, ill. az $N_i=0$, $(i=1,2)$ koordinátasíkok invariánsak, vagyis ezeket nem metszheti át pálya.

HSU, HUBBEL és WALTMAN [9] a $\lambda_1<\lambda_2$ feltevés mellett végigvizsgálják az összes paraméterkonfigurációt és megállapítják, hogy az esetek nagy részében a 2-es indexű ragadozó kihal. Kivételt képez az az eset, amikor

$$0 < \lambda_1 < \lambda_2 < K,\ a_1 < a_2,\ b_1 < b_2 \quad \text{és} \quad K > (a_2b_1 - a_1b_2)/(b_2 - b_1).$$

Ezt az esetet számítógép segítségével vizsgálták [9]. A vizsgálat azt mutatta, hogy lehetséges a három faj együttélése a pozitív oktánsban elhelyezkedő zárt pályának megfelelő periodikus megoldás formájában. SMITH [16], illetve KEENER [10] bifurkációelmélet alkalmazásával, bizonyos feltételek mellett bizonyították, hogy a fenti esetben, kis $\lambda_2-\lambda_1$ és $a_2-a_1$ értékek esetén *léteznek stabilis periodikus megoldások a pozitív oktáns belsejében.*

Tételezzük fel, hogy a (6.2) modellben szereplő két ragadozó populáció $\lambda_1$ és $\lambda_2$ paraméterértéke megegyezik: $\lambda=\lambda_1=\lambda_2$. Ez ugyan nem tipikus eset, azonban vizsgálata a versengő kizárás elvének érvényesülésével kapcsolatban érdekes eredményekre vezet. Lásd [5, 6], (ezzel az esettel foglalkozik WILKEN [18] is). Ha $\lambda_1=\lambda_2$, akkor vagy $a_1=a_2$ és $b_1=b_2$, vagy pedig például $a_1>a_2$ (kedvezőtlen az 1-es indexűre nézve) és $b_1>b_2$ (kedvező az 1-es indexűre nézve). A populációdinamikával foglalkozó irodalomban az olyan fajt, amelynek magas az $a_1$ fél-telítődési állandója és a maximális természetes szaporodási rátája *„r-stratégiának"* is szokták nevezni. Jelen esetben magas $b_1=m_1/d_1$-ről beszélünk, ami nem feltétlenül jelenti azt, hogy a maximális természetes szaporodási ráta, $m_1-d_1$ nagy, de azt jelenti, hogy az $m_1$ születési ráta nagy a $d_1$ halálozási rátához viszonyítva. Az $r$-stratégiának sok táplálékra van szüksége ahhoz, hogy megfelelő ütemben szaporodjék, nehezen viseli a táplálékhiányt, viszont bőségben igen gyorsan szaporodik. Az olyan fajt, amelynek kicsi az $a_2$ fél-telítődési állandója és a maximális szaporodási rátája (jelen esetben a $b_2=m_2/d_2$ viszony) *„K-stratégiának"* szokták nevezni. A $K$-stratéga kevés táplálékon, mostoha körülmények közt is megél, szaporodik, de szaporodása táplálékbőség esetén is lassú.

## 7. Majdnem egyforma ragadozók versengése

Ebben a pontban a $\lambda=\lambda_1=\lambda_2$ esetnek azzal a speciális esetével foglalkozunk, amikor $a=a_1=a_2$ is fennáll. Azt vizsgáljuk, hogyan változik a rendszer dinamikája, ha a környezet $K>\lambda>0$ fenntartó képességét változtatjuk. A tételek bizonyításait lásd az [5] dolgozatban.

A vizsgált esetben nyilván $b=b_1=b_2$ is fennáll és legfeljebb a két faj maximális természetes szaporodási rátája különbözik egymástól. Tegyük fel például, hogy $\beta_2 \geqq \beta_1$. Mivel $b=m_1/d_1=m_2/d_2$, ezért $\varrho=d_2/d_1=m_2/m_1$. Nyilvánvaló, hogy $\varrho=\beta_2/\beta_1$ is fennáll, vagyis $\beta_2=\varrho\beta_1$, ahol legutóbbi feltevésünk szerint $\varrho \geqq 1$. E feltételek mellett könnyű látni, hogy a (6.2) rendszer a következő alakban írható,

$$\dot{S} = \gamma S(1-S/K) - \frac{(N_1+\varrho N_2)m_1 S}{a+S}$$

$$\dot{N}_1 = \beta_1 N_1 \frac{S-\lambda}{a+S} \tag{7.1}$$

$$N_2 = \varrho\beta_1 N_2 \frac{S-\lambda}{a+S},$$

ahol $\gamma, m_1, a, \beta_1 > 0$, $0 < \lambda < K$, $\varrho \geqq 1$ állandók. A 2-es indexű faj természetes szaporodási rátája nagyobb, mint az 1 indexűé, de ahányszor nagyobb a születési rátája, annyiszor nagyobb a halálozási rátája is.

Ennek a rendszernek egyensúlyi helyzetei az $S, N_1, N_2$ tér pozitív oktánsában $(0, 0, 0)$, $(K, 0, 0)$, illetve az

$$L = \left\{(S, N_1, N_2) \in R^3 : S = \lambda,\ N_1 \geqq 0,\ N_2 \geqq 0,\right.$$
$$\left. N_1 + \varrho N_2 = \frac{\gamma(a+\lambda)(K-\lambda)}{m_1 K}\right\}$$

egyenesszakasz pontjai. Könnyű látni, hogy az origó és a $(K, 0, 0)$ egyensúlyi helyzet instabilis. A következőkben az $L$ szakasz pontjainak, illetve magának az $L$ halmaznak a stabilitását vizsgáljuk.

Osszuk el a (7.1) rendszer harmadik egyenletét a másodikkal. Azt kapjuk, hogy a rendszer pályagörbéinek egyenletei kielégítik a $dN_2/dN_1 = \varrho N_2/N_1$ differenciálegyenletet. Innen könnyen adódik, hogy az $N_2/N_1^{\varrho}$ függvény a rendszer első integrálja, vagyis az

$$N_2 = cN_1^{\varrho}, \quad c \geqq 0 \tag{7.2}$$

parabolikus hengerek (tetszőleges nem negatív $c$ állandóval) a (7.1) rendszer invariáns felületei. Nyilvánvaló, hogy ezek a felületek egyrétűen kitöltik a pozitív oktánst, vagyis minden ponton át egy és csak egy felület halad ebből a felületseregből. Rögzítsük $c \geqq 0$ értékét és tekintsük a (7.1) rendszer leszűkítését a (7.2) invariáns felületre, amelyet az $S$ és $N_1$ változókkal paraméterezünk:

$$\dot{S} = \gamma S(1-S/K) - \frac{(N_1+\varrho c N_1^{\varrho})m_1 S}{a+S}$$

$$\dot{N}_1 = \beta_1 N_1 \frac{S-\lambda}{a+S}. \tag{7.3}$$

Az utóbbi rendszer egyensúlyi helyzetei: $(0, 0)$, $(K, 0)$, illetve az az egyetlen pont, amelyben az $L$ egyenesszakasz a (7.2) felületet metszi. Ennek a metszéspontnak a koordinátái $(\lambda, v_1(c, K), v_2(c, K))$, ahol

$$v_1 + \varrho v_2 = \gamma(a+\lambda)(K-\lambda)/m_1 K, \quad v_2 = cv_1^{\varrho},$$

vagyis $v_1(c, K)$ a

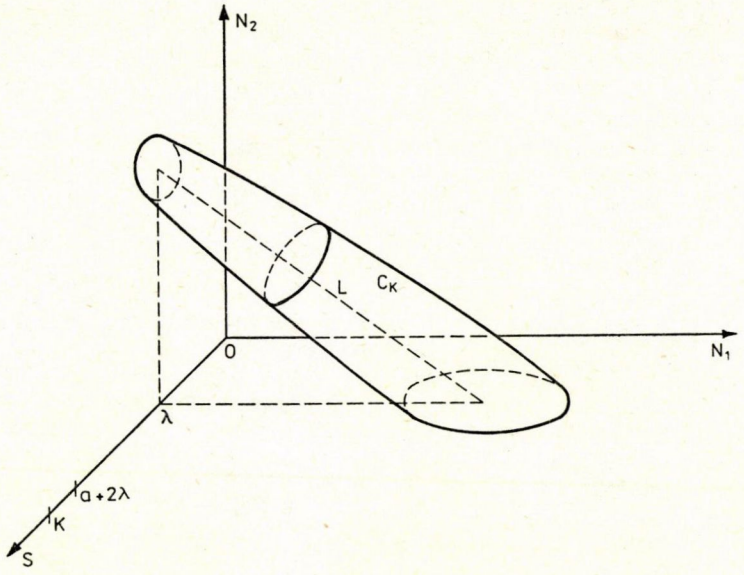
$$v_1 + \varrho c v_1^{\varrho} = \gamma(a+\lambda)(K-\lambda)/m_1 K$$

egyenlet egyetlen pozitív megoldása. Ha $K$-t változtatjuk, az $L$ egyenes önmagával párhuzamosan eltolódik és más és más $(\lambda, v_1(c, K), v_2(c, K))$ pontban metszi a (7.2) felületet. A $(\lambda, v_1(c, K))$ pont a (7.3) rendszer egyensúlyi helyzete. Erre vonatkozik a következő

7.1. TÉTEL. *Ha* $\lambda < K < a+2\lambda$, *akkor a (7.3) rendszer* $(\lambda, v_1(c, K))$ *egyensúlyi helyzete aszimptotikusan stabilis és attraktivitási tartománya az egész* $S>0$, $N_1>0$ *pozitív kvadráns; a* $K=a+2\lambda$ *értéknél a rendszer szuperkritikus Hopf-bifurkáción megy át, vagyis van olyan* $\delta>0$, *hogy minden az* $a+2\lambda < K < a+2\lambda+\delta$ *egyenlőtlenségnek eleget tevő K-ra a rendszernek egyetlen zárt pályája van az immár instabilis* $(\lambda, v_1(c, K))$ *pont egy környezetében, ez a zárt pálya körülveszi az egyensúlyi helyzetet és orbitálisan aszimptotikusan stabilis.*

Felhívjuk a figyelmet arra, hogy a kritikus paraméterérték, $K=a+2\lambda$, amelynél a bifurkáció történik, független a $c$ paraméter értékétől, vagyis a (7.2) invariáns felületsereg minden egyes egyedén ugyanazon $K=a+2\lambda$ értéknél bifurkálódik az egyensúlyi helyzet. A (7.2) sereg az $N_2=0$ koordinátasíkot tartalmazza, de az $N_1=0$ síkot nem. Könnyen belátható azonban, hogy az $N_1=0$ síkban ugyanolyan bifurkáció játszódik le és ugyanakkor, amilyen és amikor a (7.2) sereg egyedein.

7.2. KÖVETKEZMÉNY. *Az L egyenesszakasz pontjai a* (7.1) *rendszernek Ljapunov-értelemben stabilis egyensúlyi helyzetei, ha* $\lambda < K \leqq a+2\lambda$, *és instabilisak, ha* $K > a+2\lambda$.




*6. ábra.*
Az egyensúlyi helyzetek egyenesének bifurkálódása zárt pályákból álló topologikus hengerfelületté

A következő állításban egy $A$ halmaz „*környezetén*" az $A$ halmazt részhalmazként tartalmazó nyílt halmaznak és az $\{(S, N_1, N_2) \in R^3: S \geqq 0, N_1 \geqq 0, N_2 \geqq 0$ oktánsnak közös részét értjük.

7.3. KÖVETKEZMÉNY. *Ha* $\lambda < K < a+2\lambda$, *akkor az* $L$ *egyenesszakasz a (7.1) rendszer attraktora, vagyis* $L$*-nek van olyan „környezete", hogy minden e környezetből induló pálya* $L$*-hez tart, amikor* $t \to \infty$; *a* $K = a+2\lambda$ *értéknél az* $L$ *szakasz egy topologikus hengerfelületté bifurkálódik, pontosabban van olyan* $\delta > 0$, *hogy minden az* $a+2\lambda < K < a+2\lambda+\delta$ *egyenlőtlenségnek eleget tevő* $K$*-ra a (7.1) rendszernek van egy invariáns topologikus hengerfelülete* $C_K$, *amely zárt pályák egyesítése és a rendszernek attraktora, vagyis van olyan „környezete", hogy minden e környezetből induló pálya* $C_K$*-hoz tart, amikor* $t \to \infty$ (6. ábra).

Az eddigiekből könnyen látható, hogy $C_K$ egyszerű példa az *Aulbach-féle értelemben aszimptotikus fázissal rendelkező, zárt pályákból álló invariáns sokaságokra* (lásd [1]), amelyek olyanok, hogy minden egyes hozzájuk konvergáló pálya aszimptotikus fázissal valamely a sokaságon levő pályához konvergál.

A vizsgált esetben tehát viszonylag alacsony táplálékszintnél, $\lambda < K < a+2\lambda$, a rendszer az $L$ egyenes valamely pontjában stabilis egyensúlyban van. Ha innen kimozdítjuk, visszatér az $L$ egyenes valamely pontjába. Az $L$ szakasz minden pontja számításba jöhet beleértve az $N_1 = 0$, ill. az $N_2 = 0$ síkokban levő végpontokat is (az $L$ szakasz különböző pontjai különböző $N_1/N_2$ arányokat képviselnek, de ezek között nincs minőségi különbség). Ha a környezet $K$ fenntartó képességét növeljük és az túllépi az $a+2\lambda$ értéket a rendszer egyensúlyi helyzetei (egyszerre!) elvesztik stabilitásukat és a rendszer a $C_K$ felületen levő zárt pályák valamelyike mentén stabilisan rezegni kezd. Ezt a jelenséget az irodalomban a „*bőség paradoxonának*" nevezik: a táplálékbőség destabilizálja a rendszer egyensúlyi helyzetét és rezgésbe hozza a rendszert. A $C_K$ felületen levő pályák mindegyike felléphet a valóságos mozgás pályájaként. Ha a rendszert a $C_K$ felületből kimozdítjuk, oda visszatér. A *Hopf-tétel* közelítőleg megadja a bifurkálódó periodikus megoldások frekvenciáinak értékét. Egyszerű számítással adódik, hogy az $L$ egyenesszakasz $(\lambda, \nu_1, \nu_2)$ pontjából bifurkálódó periodikus megoldás $f(\nu_1)$ frekvenciája, mint a $\nu_1$ koordináta függvénye ($\nu_1$ már meghatározza $\nu_2$-t is) a következő

$$f(\nu_1) = (\lambda\beta_1)^{1/2}[\varrho\gamma(a+\lambda)^2/(a+2\lambda) - m_1\nu_1(\varrho-1)]^{1/2}/2\pi(a+\lambda).$$

Emlékeztetünk arra, hogy $\varrho \geqq 1$, ami azt jelenti, hogy a 2-es indexű ragadozó maximális szaporodási rátája nagyobb (egyenlő) az 1-es indexű ragadozóénál. Ezek szerint $\nu_1$ növekedésével ($\nu_2$ csökkenésével) az $f(\nu_1)$ frekvencia csökken. A környezet $K$ fenntartó képességének az $a+2\lambda$ érték fölé való emelkedésére a két ragadozó minőségileg hasonló módon reagál, létszámuk periodikusan változni fog az időben. A különbség csak az, hogy *azok az egyensúlyi helyzetek, amelyekben a nagy szaporodási rátájú faj dominál* ($\nu_2$ nagy, $\nu_1$ kicsi) *nagyobb frekvenciával, gyorsabban kezdenek rezegni, mint azok az egyensúlyi helyzetek, amelyekben a kis szaporodási rátájú faj dominál* ($\nu_2$ kicsi, $\nu_1$ nagy).

## 8. Zipzárbifurkáció

Ebben a pontban a $\lambda=\lambda_1=\lambda_2$ esettel foglalkozunk feltéve azt, hogy $a_1>a_2$, amiből $b_1>b_2$ következik. Most tehát az 1-es indexű ragadozó az $r$-stratéga, a 2-es indexű pedig a $K$-stratéga (lásd a 6. pont végét). Változtatjuk a környezet $K>\lambda>0$ fenntartó képességét és azt vizsgáljuk, hogyan alakul az $r$-stratéga és a $K$-stratéga versenye a változó körülmények között.

A (6.2) rendszer most a következő alakba írható át

$$\dot{S} = \gamma S(1-S/K) - \frac{m_1 N_1 S}{a_1+S} - \frac{m_2 N_2 S}{a_2+S} \tag{8.1}$$

$$\dot{N}_1 = \beta_1 N_1 \frac{S-\lambda}{a_1+S}$$

$$\dot{N}_2 = \beta_2 N_2 \frac{S-\lambda}{a_2+S}.$$

Ennek a rendszernek egyensúlyi helyzetei az $S, N_1, N_2$ tér pozitív oktánsában $(0, 0, 0)$, $(K, 0, 0)$, illetve az

$$L = \Big\{(S, N_1, N_2)\in R^3:\ S=\lambda,\ N_1 \geqq 0,\ N_2 \geqq 0,$$

$$\frac{m_1 N_1}{a_1+\lambda} + \frac{m_2 N_2}{a_2+\lambda} = \gamma(1-\lambda/K)\Big\}$$

egyenesszakasz pontjai.

Könnyű belátni linearizálással, hogy az első két egyensúlyi helyzet instabilis. Az $L$ egyenesszakasz pontjait a továbbiakban $(\lambda, v_1, v_2)$-vel fogjuk jelölni, vagyis $v_1$ és $v_2$ olyan számokat jelölnek, amelyekre

$$\frac{m_1 v_1}{a_1+\lambda} + \frac{m_2 v_2}{a_2+\lambda} = \frac{\gamma(K-\lambda)}{K}, \quad v_1 \geqq 0, \quad v_2 \geqq 0. \tag{8.2}$$

Az $L$ szakasz végpontjai a koordinátasíkokban:

$$P_2 = (\lambda, 0, v_2) = \big(\lambda, 0, \gamma(a_2+\lambda)(K-\lambda)/m_2 K\big),$$

$$P_1 = (\lambda, v_1, 0) = \big(\lambda, \gamma(a_1+\lambda)(K-\lambda)/m_1 K, 0\big).$$

Az $L$ szakaszon levő pontok stabilitásának vizsgálata érdekében linearizáljuk a (8.1) rendszert egy tetszőleges $(\lambda, v_1, v_2)$ pontban. A linearizált rendszer együttható mátrixa

$$\begin{bmatrix} -\frac{\gamma\lambda}{K} + \lambda\left(\frac{m_1 v_1}{(a_1+\lambda)^2} + \frac{m_2 v_2}{(a_2+\lambda)^2}\right) & -\frac{m_1\lambda}{a_1+\lambda} & -\frac{m_2\lambda}{a_2+\lambda} \\ \frac{\beta_1 v_1}{a_1+\lambda} & 0 & 0 \\ \frac{\beta_2 v_2}{a_2+\lambda} & 0 & 0 \end{bmatrix}.$$

E mátrix karakterisztikus polinomja:

$$D(\mu)=\mu\left[\mu^2+\mu\lambda\left(\frac{\gamma}{K}-\frac{m_1v_1}{(a_1+\lambda)^2}-\frac{m_2v_2}{(a_2+\lambda)^2}\right)+\right.$$

$$\left.+\lambda\left(\frac{\beta_1m_1v_1}{(a_1+\lambda)^2}+\frac{\beta_2m_2v_2}{(a_2+\lambda)^2}\right)\right].$$

A szögletes zárójelben álló másodfokú polinom akkor és csak akkor stabilis (gyökeinek valós részei akkor és csak akkor negatívok), ha

$$\frac{m_1v_1}{(a_1+\lambda)^2}+\frac{m_2v_2}{(a_2+\lambda)^2}<\frac{\gamma}{K}. \tag{8.3}$$

Ha $\lambda<K<a_2+2\lambda$, akkor

$$\frac{m_1v_1}{(a_1+\lambda)^2}+\frac{m_2v_2}{(a_2+\lambda)^2}\leqq\frac{1}{a_2+\lambda}\left(\frac{m_1v_1}{a_1+\lambda}+\frac{m_2v_2}{a_2+\lambda}\right)=$$

$$=\frac{\gamma}{a_2+\lambda}\left(1-\frac{\lambda}{K}\right)<\frac{\gamma}{a_2+\lambda}\left(1-\frac{\lambda}{a_2+2\lambda}\right)<\frac{\gamma}{K},$$

ha viszont $K>a_1+2\lambda$, akkor analóg becslés segítségével

$$\frac{m_1v_1}{(a_1+\lambda)^2}+\frac{m_2v_2}{(a_2+\lambda)^2}>\frac{\gamma}{K}$$

adódik. Ezek szerint, ha $\lambda<K<a_2+2\lambda$, akkor minden $(\lambda, v_1, v_2)$ egyensúlyi helyzethez egy zérus és két negatív valós részű sajátérték tartozik, ha viszont $K>a_1+2\lambda$, akkor minden $(\lambda, v_1, v_2)$ egyensúlyi helyzet instabilis. Az [5] dolgozat eredményeiből adódik a

8.1. TÉTEL. *Ha $\lambda<K<a_2+2\lambda$, vagyis a környezet fenntartó képessége alacsony, akkor az L egyenesszakasz minden pontja Ljapunov értelemben stabilis egyensúlyi helyzet, maga az L egyenesszakasz pedig a (8.1) rendszer attraktora (a 7.3. Következmény értelmében). Ha $K>a_1+2\lambda$, vagyis a környezet fenntartó képessége magas, akkor az L szakasz minden pontja (beleértve a $P_1$ pontot is) instabilis egyensúlyi helyzet.*

Az előző tétel nagyjából választ ad arra a kérdésre, hogy mi történik túl kicsi és túl nagy környezeti fenntartóképesség esetén. A válasz megfelel a várakozásnak. Ha a fenntartó képesség alacsony, akkor minden lehetséges $v_2/v_1$ arányban együttélhet a két ragadozó, beleértve azokat az eseteket is, amikor valamelyik hiányzik (vagyis a zérus, illetve a végtelen arányt is); a különböző arányoknak megfelelő egyensúlyi helyzetek egyformán stabilisak. Ha a rendszert kissé kimozdítjuk az $L$ egyenesen levő egyensúlyi helyzetéből, visszatér az $L$ egyenesbe, bár nem szükségképpen abba a pontba, ahonnan kimozdult. Ha a fenntartóképesség túl nagy, akkor nincs stabilis, egyensúlyi helyzettel jellemzett együttélés, ami egyrészt a $K$-stratéga hátrányba kerülésének, másrészt a bőség paradoxonának következménye.

Izgalmasnak tűnik a rendszer vizsgálata közbenső fenntartó képességek esetén, vagyis akkor, amikor $a_2+2\lambda\leqq K\leqq a_1+2\lambda$.

Könnyű látni, hogy ha a $(\lambda, v_1, v_2)$ pont az $L$ egyenesen a $P_2$ ponttól a $P_1$ pont felé mozog (vagyis $v_1$ növekszik a $[0, \gamma(a_1+\lambda)(K-\lambda)/m_1K]$ intervallumban és persze $v_2$ egyidejűleg csökken), akkor a (8.3) egyenlőtlenség bal oldala csökken. Ha $K \in [a_2+2\lambda, a_1+2\lambda]$, akkor van pontosan egy $(\lambda, v_1(K), v_2(K))$ pont az $L$ szakaszon, amelyben (8.3) egyenlőségjellel áll fenn. Ez a pont egyszerűen meghatározható a (8.2) és az

$$\frac{m_1 v_1}{(a_1+\lambda)^2} + \frac{m_2 v_2}{(a_2+\lambda)^2} = \frac{\gamma}{K}$$

egyenletekből álló egyenletrendszerből:

$$(8.4) \qquad (v_1(K), v_2(K)) = \left( \frac{\gamma(a_1+\lambda)^2(K-a_2-2\lambda)}{Km_1(a_1-a_2)}, \frac{\gamma(a_2+\lambda)^2(a_1+2\lambda-K)}{Km_2(a_1-a_2)} \right).$$

Az $L$ egyenesszakasznak a $(\lambda, v_1(K), v_2(K))$ ponttól balra levő pontjaihoz egy zérus és két pozitív valós részű sajátérték, az e ponttól jobbra levő pontokhoz egy zérus és két negatív valós részű sajátérték tartozik. Magához a $(\lambda, v_1(K), v_2(K))$ egyensúlyi helyzethez egy zérus és két konjugált tiszta képzetes sajátérték tartozik.

A következő állítások bizonyítása is az [5] dolgozatban található.

8.2. TÉTEL. *Minden $K$-ra, amelyre $a_2+2\lambda \leqq K \leqq a_1+2\lambda$, a $(\lambda, v_1(K), v_2(K))$ pont az $L$ egyenest két részre osztja; a (8.1) rendszer azon egyensúlyi helyzetei, amelyek a*

$$\{(\lambda, v_1, v_2) \in L: \ v_1 < v_1(K)\}$$

*szakaszt alkotják, instabilisak; az*

$$L_s = \{(\lambda, v_1, v_2) \in L: \ v_1 > v_1(K)\}$$

*szakaszt alkotó egyensúlyi helyzetek viszont Ljapunov értelemben stabilisak ($L_s$: „$L$ stabilis része", lásd a 7. ábrát).*

Eddig $K$ értékét rögzítettnek tekintettük. Legyen a továbbiakban $K$ változtatható paraméter, és növeljük értékét $a_2+2\lambda$-tól $a_1+2\lambda$-ig. Ekkor a $(\lambda, v_1(K), v_2(K))$ pont végighalad az $L$ egyenesszakaszon folytonosan annak bal oldali végpontjától a jobb oldali végpontig úgy, hogy a maga mögött hagyott pontok instabilissá válnak. Ezt a jelenséget *zipzárbifurkációnak* nevezzük. Megjegyezzük, hogy $K$ változtatásakor maga az $L$ egyenes is önmagával párhuzamosan elmozdul, ez azonban a jelenség lényegét nem befolyásolja. A $K=a_2+2\lambda$ értékhez tartozó $L$ szakasz bal oldali végpontja $P_2=(\lambda, v_1(a_2+2\lambda), v_2(a_2+2\lambda))=(\lambda, 0, \gamma(a_2+\lambda)^2/m_2(a_2+2\lambda))$, a $K=a_1+2\lambda$ értékhez tartozó $L$ szakasz jobb oldali végpontja

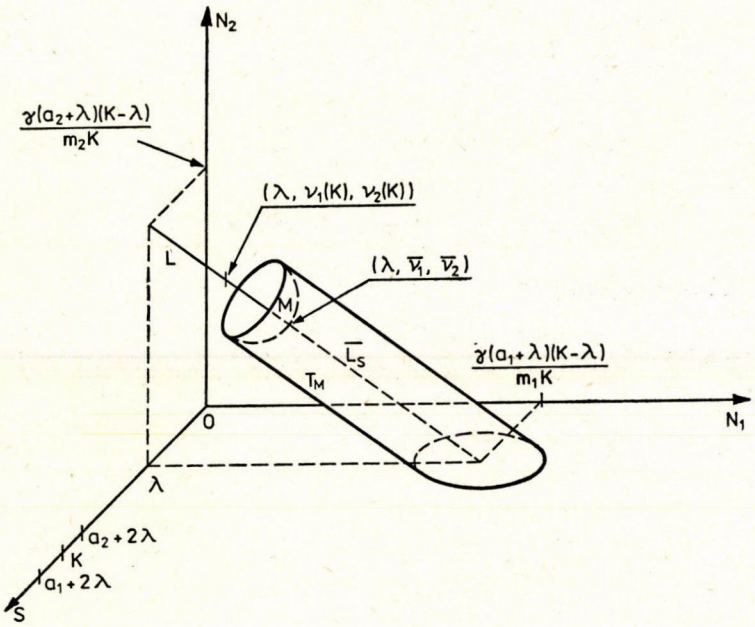
$$P_1 = (\lambda, v_1(a_1+2\lambda), v_2(a_1+2\lambda)) = (\lambda, \gamma(a_1+\lambda)^2/m_1(a_1+2\lambda), 0).$$

Legyen most (rögzített $K$ mellett) $(\lambda, \bar{v}_1, \bar{v}_2)$ egy tetszőleges pont $L$-en a $(\lambda, v_1(K), v_2(K))$ ponttól jobbra, vagyis legyen

$$(8.5) \qquad v_1(K) < \bar{v}_1 < \gamma(a_1+\lambda)(K-\lambda)/m_1K,$$

és tekintsük $L_s$-nek a $(\lambda, \bar{v}_1, \bar{v}_2)$ pont által meghatározott, következő, zárt, valódi részhalmazát:

$$\bar{L}_s = \{(\lambda, v_1, v_2) \in L: \ v_1 \leqq \bar{v}_1\}.$$



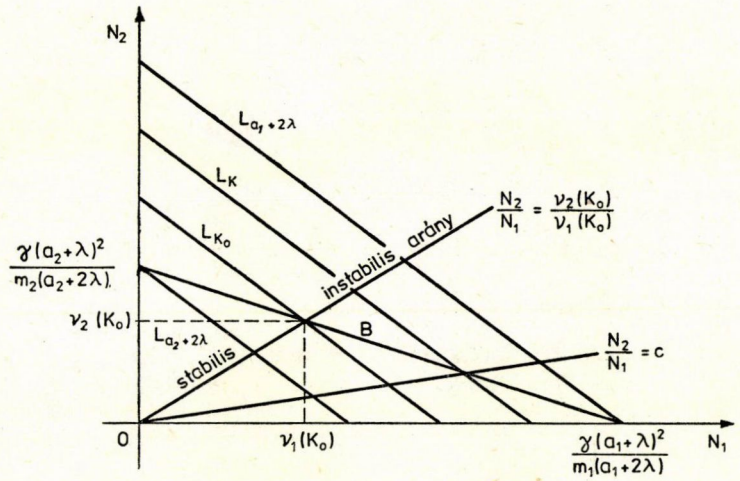

*7. ábra.*
Az $\bar{L}_s$ attraktor $T_M$ attraktivitási tartománya

Legyen $M$ egyszerű, sima felület, amely az $L$ egyenest a $(\lambda, \bar{\nu}_1, \bar{\nu}_2)$ pontban transzverzálisan metszi és a teret két nem összefüggő részre osztja úgy, hogy az egyik féltér tartalmazza az $\bar{L}_s$ szakaszt. Ez utóbbi félteret az $M$ *által meghatározott jobb oldali féltérnek* nevezzük. Az $\bar{L}_s$ halmaz egy környezetének, az $M$ által meghatározott jobb oldali féltérnek és a pozitív oktánsnak közös részét $\bar{L}_s$ *csőszerű* $T_M$ *környezetének* nevezzük (lásd 7. ábra).

8.3. TÉTEL. *Minden az* $a_2+2\lambda \leqq K < a_1+2\lambda$ *egyenlőtlenséget kielégítő* $K$*-ra és minden (*8.5*)-öt kielégítő* $\bar{\nu}_1$*-re az* $\bar{L}_s$ *szakasz a (*8.1*) rendszer attraktora a következő értelemben:* $\bar{L}_s$*-nek van olyan* $T_M$ *csőszerű környezete, hogy minden* $T_M$*-beli pontból induló megoldás pályája* $\bar{L}_s$*-hez tart, ha* $t$ *tart a végtelenhez.*

Az előbbi két tétel szemléletes jelentése a következő. Ha a táplálék mennyisége növekszik, pontosabban a környezet $K$ fenntartó képessége $a_2+2\lambda$-ról $a_1+2\lambda$-ra nő, a zipzárbifurkáció jelensége lép fel, ami azt jelenti, hogy *a három faj együttélését reprezentáló egyensúlyi helyzetek folyamatosan, sorban destabilizálódnak.* Ennek során először azok az egyensúlyi helyzetek destabilizálódnak, amelyek a 2-es indexű faj, a $K$-stratéga dominanciáját reprezentálják. Legtovább azok az egyensúlyi helyzetek maradnak stabilisak, amelyekben $N_2=\nu_2$ kicsi és $N_1=\nu_1$ nagy. Ha a táplálék mennyisége növekszik, a $K$-stratéga hátrányba kerül az $r$-stratégával szemben, nem indokolt tehát olyan egyensúlyi helyzetek fenntartása, amelyekben ő van többségben.

Az előbbieket még más nézőpontból is megvilágíthatjuk. Mint már említettük, az $L$ egyenes a $K$ paraméter változtatásával önmagával párhuzamosan mozog. Tüntessük fel a továbbiakban $L$ függését a $K$ paramétertől azáltal, hogy $L$ helyett $L_K$-t




*8. ábra.*

Arányok stabilitásvesztése $K$ növelése során. Az egyensúlyi helyzetek $L_K$ szakaszának eltolódása $K$ változtatása során. $B$: a $(\nu_1(K), \nu_2(K))$ bifurkációs pont pályája

írunk. Legyen $c$ tetszőleges nem negatív szám, vagy végtelen; azt mondjuk, hogy *a két ragadozó c aránya a fenntartó képesség K értékénél stabilis,* ha $(\lambda, \nu_1, \nu_2) \in L_K$, ahol $\nu_2/\nu_1 = c$, stabilis. Más szóval, minden egyes $c$ arányt, minden rögzített $K$ értéknél az $L_K$ szakasznak pontosan egy pontja reprezentálja; ha ez a pont stabilis, akkor $c$ stabilis arány $K$ szóban forgó értékénél. Rögzítsük most tetszőlegesen az $a_2+2\lambda \leqq \leqq K_0 \leqq a_1+2\lambda$ fenntartó képességet és tekintsük a (8.4) által meghatározott $(\lambda, \nu_1(K_0), \nu_2(K_0))$ bifurkációs pont által meghatározott $\nu_2(K_0)/\nu_1(K_0)$ arányt. Világos, hogy ez az arány stabilis $K < K_0$ esetén és instabilis $K > K_0$-ra. A 8. ábra mutatja $L_K$ eltolódását, amint $K$ változik $a_2+2\lambda$-tól $a_1+2\lambda$-ig és azt a $B$ egyenest, amelyen a $(\nu_1(K), \nu_2(K))$ pont mozog $K$ változtatása során. Az origóból kiinduló sugarak egy-egy aránynak felelnek meg. Ahogy $K$ nő, ugyanaz az arány egyre magasabb (létszám) szinten realizálódik, azonban egy bizonyos $K$ értéknél destabilizálódik, ti. annál a $K$ értéknél amelynek megfelelő $L_K$ egyenes átmegy az adott sugár és $B$ metszéspontján. Így például a $\nu_2(K_0)/\nu_1(K_0)$ arány éppen a $K_0$ értéknél destabilizálódik. Az ábrából világos, hogy minél kisebb egy arány (minél kisebb $\nu_2$ értéke $\nu_1$-hez képest), annál tovább marad stabilis $K$ növekedése során.

Az, hogy az $L_K$ egyenesszakaszok belső pontjaiban milyen a bifurkáció jellege, egyelőre nyitott kérdés. Ha $a_2+2\lambda < K_0 < a_1+2\lambda$, akkor a $(\lambda, \nu_1(K_0), \nu_2(K_0))$ pont bifurkációjára a *Hopf-tétel* nem alkalmazható, mivel e pontban $K=K_0$-nál három sajátérték valós része zérus. Az e ponton áthaladó központi sokaságnak nincs az $L_{K_0}$ egyenesre transzverzális, kétdimenziós, invariáns részsokasága.

Ezzel szemben az $L_K$ egyenesek végpontjainak bifurkációját jellemezni tudjuk. Ez különösen a $K_0 = a_1+2\lambda$ értéknél fontos, hiszen az eddigiekből csak azt tudjuk, hogy ha $K$ e fölé az érték fölé emelkedik, a (8.1) rendszernek nem marad egyetlen stabilis egyensúlyi helyzete sem a *zárt* pozitív oktánsban.

Az $S, N_i$ koordinátasikbeli $P_i(\lambda, \gamma(a_i+\lambda)(K-\lambda)/m_iK)$ pont (v.ö. a (8.2) formula utáni háromdimenziós felírásával) az erre a síkra leszűkített (8.1) rendszer, vagyis az

$$\dot{S} = \gamma S(1-S/K) - \frac{m_i N_i S}{a_i+S} \tag{8.6}$$

$$\dot{N}_i = \beta_i N_i \frac{S-\lambda}{a_i+S}$$

rendszer egyensúlyi helyzete ($i=1, 2$). Világos az előzőkből, hogy $\lambda<K<a_i+2\lambda$ esetén $P_i$ ennek a rendszernek aszimptotikusan stabilis, $K>a_i+2\lambda$ esetén pedig instabilis egyensúlyi helyzete. Érvényes a következő

8.4. TÉTEL. *A (8.6) rendszer $P_i$ egyensúlyi helyzete a $K=a_i+2\lambda$ kritikus értéknél szuperkritikus Hopf-bifurkáción megy át, vagyis van olyan $\delta_i>0$, hogy minden $K\in(a_i+2\lambda, a_i+2\lambda+\delta_i)$-ra a (8.6) rendszernek van lokálisan egyetlen zárt pályája, mely a $P_i$ pontot körülveszi és ez a zárt pálya orbitálisan aszimptotikusan stabilis* ($i=1, 2$).

Ez a tétel azt jelenti, hogy ha az $i$-edik ragadozó ($i=1, 2$) egyedül van a zsákmánnyal, $K>a_i+2\lambda$ esetén is stabilisan fennmarad, de a „bőség paradoxonának" megfelelően már nem állandó értéken, hanem a zsákmánnyal együtt időben periodikusan változó mennyiségben.

## IRODALOM


[1] AULBACH, B., "Behavior of solutions near manifolds of periodic solutions", *J. Diff. Equ.* **39** (1981) 345—377.

[2] CUSHING, J. M., *Integrodifferential Equations and Delay Models in Population Dynamics* (Springer, Berlin, 1978).

[3] FARKAS, M., "Stability of bifurcating orbits in a predator-prey model", in: Fourth International Conference on Mathematical Modelling, Zürich, 1983, (Pergamon Press, Oxford, New York, 1983), 925—927.

[4] FARKAS, M., "Stable oscillations in a predator-prey model with time lag", *J. Math. Anal. Appl.* **102** (1984), 175—188.

[5] FARKAS, M., "Zip bifurcation in a competition model", *Nonlinear Analysis TMA*, **8** (1984), 1295—1309.

[6] FARKAS, M., "Zip bifurcation arising in population dynamics", in: ICNO X. Varna, (Bulgarian Academy of Sci., Sofia, 1985), 150—155.

[7] HASSARD, B. D., KAZARINOFF, N. D. and WAN, Y. H., *Theory and Applications of Hopf Bifurcation*, (Cambridge Univ. Press, Cambridge, 1981).

[8] HSU, S. B., HUBBELL, S. P. and WALTMAN, P., "A contribution to the theory of competing predators", *Ecological Monographs* **48** (1978), 337—349.

[9] HSU, S. B., HUBBELL, S. P. and WALTMAN, P., "Competing predators", *SIAM J. Appl. Math.* **35** (1978) 617—625.

[10] KEENER, J. P., "Oscillatory coexistence in the chemostat: a codimension two unfolding", *SIAM J. Appl. Math.* **43** (1983) 1005—1018.

[11] LOTKA, A. J., *Elements of Mathematical Biology* (Dover, New York, 1965, eredetileg: Elements of Physical Biology, 1924).

[12] MACDONALD, N., "Time delay in prey-predator models 2. bifurcation theory", *Math. Biosci.* **33** (1977) 227—234.

[13] MACDONALD, N., *Time Lags in Biological Models* (Springer, Berlin, 1978).

[14] MCGEHEE, R. and ARMSTRONG, R. A., "Some mathematical problems concerning the ecological principle of competitive exclusion", *J. Diff. Equ.* **23** (1977) 30—52.

[15] Полуэктов, Р. А.: Динамическая теория буологических популаций (Наука, Москва, 1974).

[16] SMITH, H. L., "The interaction of steady state and Hopf bifurcations in a two predator-one-prey competition model", *SIAM J. Appl. Math.* **42** (1982) 27—43.
[17] VOLTERRA, V., *Leçons sur la théorie mathématique de la lutte pour la vie* (Gauthier-Villars, Paris, 1931).
[18] WILKEN, D. R., "Some remarks on a competing predators problem", *SIAM J. Appl. Math.* **42** (1982) 895—902.

Megjegyzés: Időközben sikerült a 4.2 tételt jelentősen élesíteni. Lásd: Farkas, A., Farkas, M., Kajtár, L., On Hopf bifurcation in a predator-prey model, in: Differential Equations: Qualitative Theory (Szeged, 1984), (North Holland, Amsterdam—New York, 1985), megjelenés alatt.

(Beérkezett: 1984. március 8.)

FARKAS MIKLÓS
BME MATEMATIKA TANSZÉKCSOPORT
1521 BUDAPEST, STOCZEK U. H ÉP. IV. E. 42.

## STABLE COEXISTENCE AND BIFURCATIONS IN POPULATION DYNAMICS

M. FARKAS

Defects of the classical *Lotka—Volterra model* for the interaction of a predator and a prey population are pointed out. Saturation and delay are introduced into the model. It is shown that as delay is increasing the equilibrium of the system gets unstable. Under some conditions imposed upon the parameters the system undergoes a *supercritical Hopf-bifurcation*, i.e. it begins to oscillate stably. The competition of two predator species for a single regenerating prey is also studied. Saturation is introduced at the prey, the predation rate and the birth rate of predators is assumed in *Holling's form*. We show that the increase of the carrying capacity leads to the bifurcation of stable equilibria. The phenomenon of zip bifurcation is discussed and it is shown that under some conditions this phenomenon occurs in the competition of an *r*- and a *K*-strategist.

*Alkalmazott Matematikai Lapok* **10** (1984) 231—256

# NEMLINEÁRIS DIFFERENCIÁLEGYENLETEK ATTRAKTORAI

KERTÉSZ VIKTOR

Budapest

A vizsgálat tárgya az

$$\dot{x} = g(t, x) + f(t, x); \quad x \in \mathbf{R}^n$$

perturbált egyenlet, ahol $f$-et tekintjük perturbációnak. Feltételezzük, hogy az

$$\dot{x} = g(t, x)$$

perturbálatlan egyenletnek van egy korlátos, aszimptotikusan stabilis $p(t)$ megoldása. A probléma abban áll, hogy a perturbált egyenletnek mikor van a $p(t)$ megoldás grafikonját tartalmazó attraktora, azaz aszimptotikusan stabilis invariáns halmaza, és ennek az attraktornak mi az attraktivitási tartománya.

## 1. Bevezetés

Tekintsük az

$$\dot{x} = g(t, x) \tag{1.1}$$

egyenletet, ahol $g \in C[\mathbf{R}^+ \times \Omega, \mathbf{R}^n]$, $\mathbf{R}^+ = [t_0, \infty]$, $\Omega$ $\mathbf{R}^n$-beli összefüggő, nyílt halmaz. Tételezzük fel, hogy (1.1)-nek létezik egy $\mathbf{R}^+$-on értelmezett korlátos $p(t)$ megoldása, amely minden $t \geqq t_0$-ra $\Omega$-ban marad. Jelölje $A_{pr}$ a $p(t)$ megoldás $r$-sugarú környezetét $(r > 0)$:

$$A_{pr} = \{(t, x) \in \mathbf{R}^+ \times \mathbf{R}^n : \; |x - p(t)| < r, \; t \in \mathbf{R}^+\},$$

és ennek lezártját: $\bar{A}_{pr}$. A jelölést célszerű az $r = 0$ és az $r = \infty$ esetre is kiterjeszteni:

$$A_{p0} = \{(t, x) \in \mathbf{R}^+ \times \mathbf{R}^n : \; x = p(t)\},$$

$$A_{p\infty} = \{\mathbf{R}^+ \times \mathbf{R}^n\}.$$

$A_{p0}$ nem más, mint $p(t)$ $\mathbf{R}^+ \times \mathbf{R}^n$-beli grafikonja. Nyilvánvalóan $A_{p0} = \bar{A}_{p0}$. Jelölésünkből következik, hogy $A_{00} = \mathbf{R}^+ \times \{0\}$. Tételezzük fel továbbá, hogy létezik $D > 0$ konstans, hogy

$$A_{pD} \subset \mathbf{R}^+ \times \Omega.$$

Végül feltesszük, hogy $p(t)$ aszimptotikusan stabilis, tehát $p(t)$ $\mathbf{R}^+ \times \mathbf{R}^n$-beli grafikonja (1.1) *attraktora.* A továbbiakban ezt a megoldást kitüntetett megoldásnak nevezzük. Kitüntetett megoldásként az aszimptotikusan stabilis *egyensúlyi helyzeteket, periodikus, kvázi-periodikus* vagy *majdnem-periodikus* megoldásokat [17], esetleg *különös attraktorokat* [14] célszerű kezelni.

Tekintsük most az

$$\dot{x} = g(t, x) + f(t, x) = h(t, x) \tag{1.2}$$

*perturbált* egyenletet. Milyen $f$ perturbációk esetén létezik az (1.2) egyenletnek $\bar{A}_{p\delta}$ attraktora, amelynek attraktivitási tartománya $A_{pd}$, ahol $0 \leqq \delta < d \leqq D$? Ebben a dolgozatban erre és az ezzel összefüggő kérdésekre igyekszünk választ adni.

[2, 3] periodikus kitüntetett megoldás és korlátos perturbáció esetében ad kérdésünkre választ, [4] pedig autonom perturbálatlan egyenletet vizsgál, amikor a kitüntetett megoldás egyensúlyi helyzet, a perturbáció ugyancsak korlátos. A FARKAS MIKLÓS által alkalmazott módszereknek a nem korlátos perturbációk esetében való alkalmazhatóságára mutat rá [8].

A felvetett problémának véleményünk szerint az elméleti jelentőségen túlmenően igen nagy gyakorlati jelentősége is van. Ugyanis, ha a perturbáció olyan, hogy az említett $\bar{A}_{p\delta}$ attraktor létezik, akkor azt mondhatjuk, hogy a $p(t)$ kitüntetett megoldás (1.2)-nek $\delta$-közelítéssel ugyancsak aszimptotikusan stabilis megoldása. Ha műszaki vagy egyéb tudományos problémákat oldunk meg, gyakran találjuk magunkat szemben a következő a) és b) szituációval.

a) Valamely valóságos folyamatot, jelenséget, berendezést stb. az $\dot{x} = h(t, x)$ differenciálegyenlet modellez matematikailag, természetszerűleg több-kevesebb hibával. A feladat abban áll, hogy megkeressük a folyamatnak, jelenségnek, berendezésnek stb. az egyenlet kitüntetett megoldása által leírható normál lefolyását, viselkedését, működését stb. Tételezzük fel, hogy a differenciálegyenlet „kezelhetetlenül" bonyolult. Ha sikerül az (1.2) egyenletnek megfelelően $h$-t úgy felbontani, hogy egyrészt (1.1) kitüntetett megoldását már meg tudjuk határozni, másrészt az $f$ perturbációnak tekintett tag olyan, hogy az $\bar{A}_{p\delta}$ létezik és $\delta > 0$ elég kicsiny, akkor — tekintettel a matematikai modellben eleve benne rejlő hibára — (1.1) megtalált kitüntetett megoldása éppen (1.2) keresett megoldásával azonosítható. Látnunk kell, hogy itt nemcsak számítástechnikai nyereséget könyvelhetünk el, hanem az egyszerűbb differenciálegyenlet a folyamat, jelenség, berendezés stb. lényegének jobb megértését, a lényegtelen tényezők felismerését is megkönnyíti (lásd pl. [6]).

b) Sok esetben nem matematikai megfontolások alapján, hanem a modellezett rendszer elemzése útján jutunk el ahhoz a felismeréshez, hogy zavarokkal kell számolni, amelyeket a matematikai modellben az $f$ perturbáció képvisel. A perturbáció lehet determinisztikus [1, 10, 11, 12] vagy sztochasztikus [15], és az is lehet, hoy csupán a normájára vonatkozó felső becslés az egyetlen rendelkezésre álló vagy felhasznált ismeret a perturbációra vonatkozólag [2, 3, 4, 8]. Fontos feladat annak megválaszolása, hogy milyen jellegű vagy milyen mértékű zavarok engedhetők még meg ahhoz, hogy a modellezett rendszer normális viselkedése ne változzon lényegesen meg. Ez utóbbi matematikailag ismét éppen azt jelenti, hogy létezik $\bar{A}_{p\delta}$ megfelelően kicsiny $\delta$-val.

Mi azzal az esettel fogunk foglalkozni, amikor a perturbációról becslések állnak rendelkezésre, illetve csak becsléseket veszünk figyelembe. E becsléseken túlmenően természesen azt is fel kell tételeznünk, hogy az (1.1) egyenlet és az (1.2) perturbált egyenlet vizsgált megoldásai egyáltalán *léteznek* $\mathbf{R}^+$-on. Az egyértelműségre nincs szükségünk.

Mint említettük különböző jellegű megoldások lehetnek kitüntetetten kezeltek. Ismeretes azonban, hogy az $x$ változó alkalmas transzformációjával elérhető, hogy a kitüntetett megoldás mindig a zérus egyensúlyi helyzet legyen. Legyen ugyanis az $y$ új változó:

$$y = x - p(t). \tag{1.3}$$

Ekkor (1.1), illetve (1.2) helyett az

$$\dot{y} = g(t, y+p(t)) - g(t, p(t)) = \tilde{g}(t, y), \tag{1.1'}$$

illetve

$$\dot{y} = \tilde{g}(t, y) + f(t, y+p(t)) = \tilde{g}(t, y) + \tilde{f}(t, y) \tag{1.2'}$$

írható, ahol

$$\tilde{g}(t, 0) \equiv 0.$$

Ezt a tényt messzemenően ki fogjuk használni és fő tételünket éppen az origó körüli attraktorokra fogjuk kimondani.

Befejezésképpen ismertetjük a dolgozat felépítését. A 2. szakasz az attraktivitással kapcsolatos definíciókat tartalmazza. A 3. szakasz az origó környezetében levő attraktorokra vonatkozó fő tételünket ismerteti. A 4. és 5. szakasz tételei, eredményei e fő tétel következményei. A dolgozatban több matematikai alkalmazási példát is mutatunk.

## 2. Az attraktivitással kapcsolatos definíciók

A legegyszerűbb eset, amikor az attraktor nem más, mint a kitüntetett megoldás grafikonja. Tárgyalásunkat ezért a megoldások stabilitására és attraktivitására vonatkozó definíciókkal kezdjük. Bármely megoldás — mint láttuk — az origóba transzformálható, ezért csak a zérus egyensúlyi helyzet stabilitását vizsgáljuk. Mivel a további szakaszokban kifejtett tételeink a megoldások normájára vonatkozó konkrét becslésekhez kapcsolódnak, ezért az ismertetett definíciókkal összefüggésben elsősorban olyan példákat mutatunk be, amelyeknél a megoldások normájára becslések állnak rendelkezésre.

Ezután halmazok stabilitását taglajuk. Megmutatjuk milyen halmazokat nevezhetünk attraktoroknak, és mi az attraktivitási tartomány.

$|\cdot|$-vel az euklideszi vektornormát fogjuk jelölni.

A 2.1.—2.14. definíciókat [17] alapján állítottuk össze. Jelölje $y(t, t_1, y_0)$ (1.1')-nek azt a megoldását, amelyre $y(t_1, t_1, y_0) = y_0$; egyébként jelölje $y(t)$ (1.1') valamely megoldását.

2.1. DEFINÍCIÓ. (1.1')-nek az $y(t) \equiv 0$ megoldása *stabilis*, ha bármely $\varepsilon > 0$-hoz, és $t_1 \geqq t_0$-hoz létezik $\delta > 0$, hogy $|y_0| < \delta$-ból $|y(t, t_1, y_0)| < \varepsilon$ következik minden $t \geqq t_1$-re.

2.1. PÉLDA. Tételezzük fel, hogy (1.1') minden olyan $y(t)$ megoldására, amelyre $y(t_1) < D$, igaz a

(i) $$|y(t)| \leqq c|y(t_1)| \exp \int_{t_1}^{t} (-\varrho(\tau))\, d\tau; \quad c > 0$$

becslés, ahol

(ii) $\varrho$ folytonos $\mathbf{R}^+$-on és $$\int_{t_1}^{\infty} \varrho(t)\, dt = \infty.$$

Ebben az esetben a zérus megoldás stabilis.

2.2. DEFINÍCIÓ. (1.1') zérus megoldása *egyenletesen stabilis*, ha a 2.1. definícióban szereplő $\delta$ független $t_1$-től.

Ha a fenti (i) és (ii) feltétel teljesül, abból még nem következik, hogy a nullmegoldás egyenletesen stabilis. Legyen ugyanis $\varrho$ a következő függvény:

2.2. PÉLDA.

$$\varrho(t) = n\varrho^*(t-n), \quad \text{ha} \quad t_0 \leqq n \leqq t < n+1,$$

ahol

$$\varrho^*(t) = \begin{cases} 16t, & \text{ha} \quad 0 \leqq t \leqq 1/4 \\ 8-16t, & \text{ha} \quad 1/4 \leqq t \leqq 1/2 \\ 4-8t, & \text{ha} \quad 1/2 \leqq t \leqq 3/4 \\ -8+8t, & \text{ha} \quad 3/4 \leqq t \leqq 1. \end{cases}$$

Ezzel (ii) nyilvánvalóan teljesül, továbbá:

$$\int_{n+1/2}^{n+1} \varrho(t)\,dt = -\frac{n}{2},$$

amelyből adódik, hogy $\delta$ valóban $t_1$-től is függ.

Ezzel szemben a zérus megoldás egyenletesen stabilis, ha (i) és (ii) mellett még létezik $t^* \geqq t_0$, hogy

(iii) $\varrho(t) \geqq 0$, ha $t \geqq t^*$.

2.3. DEFINÍCIÓ. (1.1') zérus megoldása *aszimptotikusan stabilis*, ha stabilis és létezik $t_1 \geqq t_0$ és $\delta_0 > 0$, hogy $|y_0| < \delta_0$-ból $\lim_{t\to\infty} y(t, t_1, y_0) = 0$ következik.

2.4. DEFINÍCIÓ. (1.1') zérus megoldása *kvázi-ekviaszimptotikusan stabilis*, ha bármely $\varepsilon > 0$-hoz és $t_1 \geqq t_0$-hoz létezik $\delta_0 > 0$ ($\delta_0$ független $\varepsilon$-tól) és $T > 0$, hogy $|y_0| < \delta_0$-ból és $t \geqq t_1 + T$-ből $|y(t, t_1, y_0)| < \varepsilon$ következik.

2.5. DEFINÍCIÓ. (1.1') zérus megoldása *ekviaszimptotikusan stabilis*, ha stabilis és kvázi-ekviaszimptotikusan stabilis.

A zérus egyensúlyi helyzet aszimptotikusan stabilis, kvázi-ekviaszimptotikusan stabilis és ekviaszimptotikusan stabilis, ha (i) és (ii) teljesül.

2.3. PÉLDA. Tekintsük az

(2.1) $$y\dot{y} + y^2 + ye^{-2t} = 0$$

differenciálegyenletet. Ennek az origó $(y(t) \equiv 0)$ egyensúlyi helyzete és bármely $y(t)$ megoldásra $\lim_{t\to\infty} y(t) = 0$. Ez a zérus megoldás azonban érdekes módon nem stabilis. Ugyanis a nullmegoldástól eltérő bármely megoldás ilyen alakban írható:

$$y(t) = ce^{-t} + e^{-2t}, \quad c \in \mathbf{R},$$

illetve formai átalakításokkal:

$$y(t, t_1, y_0) = (y_0 e^{t_1} - e^{-t_1})e^{-t} + e^{-2t}.$$

Ha $y_0 = 0$ (az origóban az unicitás nem teljesül!) akkor

$$y(t, t_1, 0) = -e^{-t-t_1} + e^{-2t}.$$

Tehát bármilyen kicsiny is $\delta$, létezik $\varepsilon$, hogy $|y_0|<\delta$-ból nem következik, hogy $|y(t, t_1, y_0)|<\varepsilon$, ha $t \geqq t_1$. Ebből következik, hogy egyenletünk zérus megoldása nem is aszimptotikusan stabilis és nem ekviaszimptotikusan stabilis. Nyilvánvalóan teljesül azonban a kvázi-ekviaszimptotikus stabilitás. (Megjegyezzük, hogy ezt az egyenletet (1.1) alakra átírva az így kapott $g$-re az (1.1)-nél tett feltevések nem teljesülnek.)

2.4. PÉLDA. Az origó az

$$\dot{y}+y+e^{-2t}=0$$

egyenlet összes megoldását vonzza olyan értelemben, hogy bármely $y(t)$ megoldásra éppúgy, mint (2.1)-nél: $\lim_{t\to\infty} y(t)=0$. Ennek ellenére a fenti definíciók közül egyetlen sem teljesül, hiszen itt $y\equiv 0$ nem megoldás.

2.6. DEFINÍCIÓ. (1.1′) zérus megoldása *kvázi-egyenletesen aszimptotikusan stabilis*, ha kvázi-ekviaszimptotikusan stabilis, továbbá $\delta_0$ és $T$ nem függ $t_1$-től.

2.7. DEFINÍCIÓ. (1.1′) zérus megoldása *egyenletesen aszimptotikusan stabilis*, ha egyenletesen stabilis és kvázi-egyenletesen aszimptotikusan stabilis.

(i) és (ii)-ből sem a kvázi-egyenletes aszimptotikus stabilitás, sem az egyenletes aszimptotikus stabilitás nem következik (de még (i), (ii) és (iii)-ből sem), hiszen nem szükségszerű, hogy $T$ független legyen $t_1$-től. Ezt az állítást a következő példával támasztjuk alá.

2.5. PÉLDA. Legyen $\varrho(t)=1/t$; $t_0=1$. $|y(t, t_1, y_0)|\leqq\varepsilon$, ha $|y_0|\leqq\delta_0$ és $t=t_1+T$, amelyből

$$\varepsilon = c\delta_0 \exp \int_{t_1}^{t_1+T} \left(-\frac{1}{t}\right) dt,$$

vagyis

$$T = \frac{t_1(c\delta_0-\varepsilon)}{\varepsilon}; \quad c\delta_0 > \varepsilon.$$

Tehát itt $T$ valóban függ $t_1$-től.

2.8. DEFINÍCIÓ. (1.1′) zérus megoldása *exponenciálisan aszimptotikusan stabilis*, ha létezik $\lambda>0$ szám, hogy bármely $\varepsilon>0$-hoz megadható $\delta>0$, amelyre $|y_0|<\delta$-ból $|y(t, t_1, y_0)|\leqq\varepsilon \exp\left(-\lambda(t-t_1)\right)$ következik minden $t\geqq t_1$-re.

Ez a definíció nem teljesül (i) és (ii)-vel, sem (i), (ii) és (iii)-vel, de teljesül, ha (i) és (ii) mellett

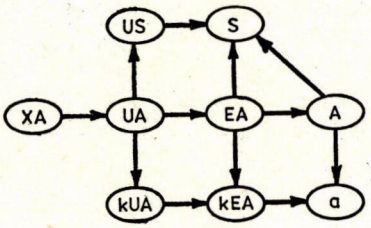
(iv) $\varrho(t)\geqq k>0$, ha $t\geqq t^*$ valamely $t^*\geqq t_0$-ra

vagy, ha ugyancsak (i) és (ii) mellett

(v) $\varrho(t)\geqq k+\omega(t)$,

ahol $k>0$ és $\omega(t)$ folytonos, periodikus függvény és $\int_0^\tau \omega(t)dt=0$, ahol $\tau$ a periódus.

Figyeljük meg, hogy ezzel szemben a (2.1) egyenletnél a zérus megoldás nem exponenciálisan aszimptotikusan stabilis.




*1. ábra*

$S$: stabilis
$A$: aszimptotikusan stabilis
$U$: egyenletesen-
$E$: ekvi-
$k$: kvázi-
$X$: exponenciálisan-
$EA = kEA + S$
$UA = kUA + US$
$US = S + t_1$-től függetlenség
$kUA = kEA + t_1$-től függetlenség
$a$: attraktor
$A = a + S$

A 2.1.—2.8. definíciók az 1. ábra szerint következnek egymásból. (Az attraktor definícióját l. később.)

Globális tulajdonságokról beszélhetünk, ha a fenti definíciók (1.1′) minden megoldására és nem csupán az origó egy környezetéből induló megoldásokra érvényesek. Ekkor azt is feltételezzük, hogy $\tilde{g}$ $\mathbf{R}^+ \times \mathbf{R}^n$-n folytonos.

2.9. DEFINÍCIÓ. (1.1′) zérus megoldása *globálisan aszimptotikusan stabilis*, ha stabilis és $\lim_{t \to \infty} y(t) = 0$ minden megoldásra.

2.10. DEFINÍCIÓ. (1.1′) zérus megoldása *globálisan kvázi-ekviaszimptotikusan stabilis*, ha bármely $\alpha > 0$-hoz, $\varepsilon > 0$-hoz és $t_1 \geqq t_0$-hoz létezik $T > 0$, hogy $|y_0| \leqq \alpha$-ból és $t \geqq t_1 + T$-ből $|y(t, t_1, y_0)| < \varepsilon$ következik.

2.11. DEFINÍCIÓ. (1.1′) zérus megoldása *globálisan ekviaszimptotikusan stabilis*, ha stabilis és globálisan kvázi-ekviaszimptotikusan stabilis.

A zérus megoldás globálisan aszimptotikusan stabilis, globálisan kvázi-ekviaszimptotikusan stabilis és globálisan ekviaszimptotikusan stabilis, ha (i) és (ii) teljesül $D = \infty$-re.

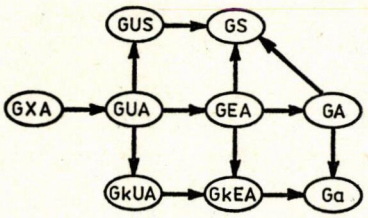
2.12. DEFINÍCIÓ. (1.1′) zérus megoldása *globálisan kvázi-egyenletesen aszimptotikusan stabilis*, ha globálisan kvázi-ekviaszimptotikusan stabilis, és $T$ független $t_1$-től.

2.13. DEFINÍCIÓ. (1.1′) zérus megoldása *globálisan egyenletesen aszimptotikusan stabilis*, ha globálisan kvázi-egyenletesen aszimptotikusan stabilis, egyenletesen stabilis és minden megoldás egyenletesen korlátos, vagyis minden megoldásra igaz, hogy bármely $\alpha > 0$-hoz létezik $\beta > 0$, amelyre $|y_0| < \alpha$-ból minden $t \geqq t_1$-re $|y(t, t_1, y_0)| < \beta$ következik.

2.14. DEFINÍCIÓ. (1.1′) zérus megoldása *globálisan exponenciálisan stabilis*, ha létezik $\lambda>0$, hogy bármely $\alpha>0$-hoz található $k>0$, amelyre $|y_0|\leqq\alpha$-ból minden $t\geqq t_1$-re $|y(t, t_1, y_0)|\leqq k|y_0|\exp(-\lambda(t-t_1))$ következik.

A zérus megoldás globálisan kvázi-egyenletesen aszimptotikusan stabilis, globálisan egyenletesen aszimptotikusan stabilis és globálisan exponenciálisan stabilis, ha (i) és (ii) mellett (iv) vagy (v) is teljesül és $D=\infty$.

A 2.9—2.14. definíciókra is változatlanul érvényes az 1. ábra, ha bevezetjük a globális stabilitás ($GS$) és a globális egyenletes stabilitás ($GUS$) fogalmát, és az ábra valamennyi fogalom rövidítése elé a „globális" jelzőt jelentő $G$ betűt tesszük (lásd 2. ábra). (1.1′) zérus megoldása *globálisan stabilis*, ha stabilis és *ekvi-korlátos*, vagyis minden $\alpha>0$-hoz és minden $t_1\geqq t_0$-hoz létezik $\beta$, hogy $|y_0|<\alpha$-ból minden $t\geqq t_1$-re $|y(t, t_1, y_0)|<\beta$ következik. Ha a megoldások unicitását feltételezzük, akkor a globális kvázi-ekviaszimptotikus stabilitásból és a stabilitásból a megoldásoknak a kezdeti értéktől való folytonos függése alapján egyszerűen következik az ekvi-korlátosság. Így a 2.11. definícióval ekvivalens az a definíció, ahol (az 1. és a 2. ábra jelöléseivel élve) $GEA=GkEA+S$ helyett $GEA=GkEA+GS$-t mondunk. A *globális egyenletes stabilitás*on a stabilitás és az *egyenletes korlátosság* (lásd a 2.13. definíció megfelelő részét) együttes teljesülését értjük.




*2. ábra*
$G$: globálisan-
$GS=S$+ekvi-korlátosság
$GUS=US$+egyenletes korlátosság
$GEA=GkEA+GS$
$GUA=GkUA+GUS$
$GkUA=GkEA+t_1$-től függetlenség
$GA=Ga+GS$

Ha valamelyik globális definíció teljesül, akkor ez maga után vonja a neki megfelelő nem globális definíció teljesülését, pl. $GkUA\to kUA$.

Ezekután ismertetjük az origó attraktivitására vonatkozó definíciókat [13] nyomán.

2.15. DEFINÍCIÓ. (1.1′) zérus megoldása *attraktor*, ha bármely $t_1\geqq t_0$-hoz létezik $\eta>0$, hogy minden $\varepsilon>0$-hoz és $|y_0|<\eta$-hoz található $T>0$, hogy $t\geqq t_1+T$ esetében $|y(t, t_1, y_0)|<\varepsilon$.

2.16. DEFINÍCIÓ. (1.1′) zérus megoldása *ekvi-attraktor*, ha attraktor és $T$ nem függ $y_0$-tól.

2.17. DEFINÍCIÓ. (1.1′) zérus megoldása *egyenletes attraktor*, ha attraktor és $\eta$ nem függ $t_1$-től és $T$ nem függ sem $t_1$-től, sem $y_0$-tól.

A zérus megoldás attraktor, ha teljesül (i) és (ii). A 2.1. egyenlet zérus megoldása attraktor, sőt ekvi-attraktor. Az ekvi-attraktivitás nyilvánvalóan ekvivalens a kvázi-ekviaszimptotikus stabilitással, az egyenletes attraktivitás pedig a kvázi-egyenletes aszimptotikus stabilitással. Az attraktivitás maga (2.15. definíció) azonban egyetlen említett definícióval sem ekvivalens. Ha a megoldások unicitása teljesül és bármely $y(t, t_1, y_0)$ megoldás $((t_1, y_0) \in \mathbf{R}^+ \times \Omega)$ folytatható $[t_0, t_1]$-ben, akkor az aszimptotikus stabilitásból az attraktivitás nyilván következik, és a stabilitás, valamint attraktivitás együtt ekvivalens az aszimptotikus stabilitással. A definícióknak ezeket az összefüggéseit az 1. ábrába is berajzoltuk.

Ha a zérus megoldás attraktor, akkor az $y(t, t_1, y_0)$ megoldásra igaz a $\lim_{t\to\infty} y(t, t_1, y_0)=0$ reláció, amennyiben $|y_0|<\eta$, ahol $\eta$ $t_1$-nek függvénye: $\eta=\eta(t_1)$. Ebben az értelemben azt mondhatjuk, hogy $K_{\eta(t_1)}$ (az origó $\eta(t_1)$ sugarú környezete) az origó, mint attraktor attraktivitási tartománya. Ezt a fogalmat általánosabban a soronkövetkező definíció határozza meg.

2.18. DEFINÍCIÓ. Az origónak az $A(t_1)$ halmaz $(t_1 \geqq t_0)$ *attraktivitási tartománya*, ahol

$$A(t_1) = \{y_0 \in \Omega: \lim_{t\to\infty} y(t, t_1, y_0) = 0\}.$$

Figyeljük meg, hogy ha az origónak bármely $t_1 \geqq t_0$-ra van nyílt és az origót is magába foglaló attraktivitási tartománya, akkor a zérus megoldás attraktor.

2.19. DEFINÍCIÓ. Az origónak az $A$ halmaz *egyenletes attraktivitási tartománya*, ha attraktivitási tartomány és független $t_1$-től.

2.20. DEFINÍCIÓ. Az origó *globális attraktor*, ha van $A$ egyenletes attraktivitási tartománya és $A=\Omega=\mathbf{R}^n$.

Ha az origó globális attraktor, akkor ebből még nem következik, hogy a zérus megoldás globálisan aszimptotikusan stabilis lenne, ugyanis az attraktivitás nem garantálja a stabilitást. A globális aszimptotikus stabilitásból a globális attraktivitás azonban következik (l. 2. ábra).

2.6. PÉLDA. Tekintsük [7] 2.2. következményét! E tételt a 2.1. példánk jelöléseivel a következőképpen tudjuk átfogalmazni. Az

$$\ddot{x}+p(t, x, \dot{x})\dot{x}+x=0; \quad t \geqq t_0$$

egyenlet

$$y(t) = \begin{bmatrix} x(t) \\ \dot{x}(t) \end{bmatrix}$$

vektor megoldásaira $|y(t_1)|<D$ esetében az (i) becslés érvényes, ha

$$\varrho(t) \leqq p(t, x, \dot{x}) \leqq \frac{1}{\varrho(t)}; \quad t \geqq t_1 \geqq t_0,$$

ahol $\varrho(t)$ (ii) szerinti, ezenkívül folytonosan deriválható, $\dot{\varrho}(t)\leqq 0$ és $0\leqq\varrho(t)\leqq 1$; $p(t, x, \dot{x})$ folytonos $A_{0D}$-ben.

A következők állapíthatók meg: A zérus megoldás egyenletesen stabilis, ekvi-aszimptotikusan stabilis és ekvi-attraktor, (attraktivitási tartománya tartalmazza $A_{0D}$-t), de általában nem egyenletesen aszimptotikusan stabilis. Ha $D=\infty$, akkor

globálisan ekviaszimptotikusan stabilis és globális attraktor, de általában nem globálisan egyenletesen aszimptotikusan stabilis.

Ha $\varrho(t)$-re a fenti feltételek mellett a $\lim_{t\to\infty} \varrho(t)>0$ is érvényes, akkor a zérus megoldás egyenletesen aszimptotikusan stabilis, exponenciálisan aszimptotikusan stabilis és egyenletes attraktor. Ha emellett még $D=\infty$, akkor globálisan egyenletesen aszimptotikusan stabilis, globálisan exponenciálisan aszimptotikusan stabilis.

Ha $p(t, x, \dot{x})$-ról a folytonosság mellett csupán annyit tudunk, hogy nem negatív (legalábbis $A_{0D}$-ben, valamely $D>0$-ra), akkor [7] 1.2. példája felhasználásával megállapíthatjuk, hogy a zérus megoldás egyenletesen stabilis, de általában nem aszimptotikusan stabilis és nem attraktor.

A következőkben [16] alapján halmazok stabilitásával foglalkozunk.

Legyen $M\subset\mathbf{R}^+\times\Omega$. $M(t)$-vel jelöljük $M$ és $\mathbf{R}^+\times\mathbf{R}^n$ $t$-hez tartozó hipersíkjának metszetét. Definícióink most erre az $M$ halmazra vonatkoznak.

2.21. DEFINÍCIÓ. $M$ (1.1)-nek *stabilis halmaza*, ha bármely $\varepsilon>0$, $\alpha>0$ és $t_1\geqq t_0$-hoz létezik $\delta>0$, hogy $d(y_0, M(t_1))<\delta$ és $|y_0|\leqq\alpha$-ból $d(y(t, t_1, y_0), M(t))<\varepsilon$ következik minden $t\geqq t_1$-re.

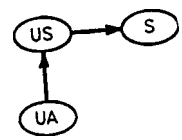
($d(P, R)$ a $P$ és $R$ halmazok távolságát jelöli.)

2.22. DEFINÍCIÓ. $M$ (1.1)-nek *(t-ben) egyenletesen stabilis halmaza*, ha stabilis halmaza és $\delta$ nem függ $t_1$-től.

A „$t$-ben" megjelölés itt és a továbbiakban azért szükséges, mert $\delta$ függhet még $\varepsilon$-tól és $\alpha$-tól, így az egyenletesség csak $t$-re vonatkozik.

2.23. DEFINÍCIÓ. $M$ (1.1)-nek *(t-ben) egyenletesen aszimptotikusan stabilis halmaza*, ha ($t$-ben) egyenletesen stabilis és bármely $\varepsilon>0$ és $\alpha>0$-hoz létezik $T>0$ és $\varepsilon$-tól független $\delta_0>0$, hogy $d(y_0, M(t_1))<\delta_0$ és $|y_0|\leqq\alpha$ esetén $d(y(t, t_1, y_0), M(t))<\varepsilon$ következik minden $t\geqq t_1+T$-re.

Ha $M=A_{00}$ (vagyis $M$ a zérus megoldás $\mathbf{R}^+\times\mathbf{R}^n$-beli grafikonja) és $g(t, 0)\equiv 0$, akkor a 2.21. definíció ekvivalens a 2.1. definícióval, a 2.22. definíció a 2.2.-vel, a 2.23. pedig a 2.7. definícióval ekvivalens. A 2.21.—2.23. definíciók közötti összefüggéseket a 3. ábrán szemléltetjük.




*3. ábra*

2.7. PÉLDA. Tételezzük fel, hogy (i), (ii) és (iv) vagy (v) teljesül, ha $0<\delta<y(t)<D\leqq\infty$. Ekkor $A_{0\delta}$ (vagy $\bar{A}_{0\delta}$) (1.1')-nek ($t$-ben) egyenletesen aszimptotikusan stabilis halmaza. Sőt, az $A_{0D}$-ból induló megoldások *véges* idő múlva elérik $\bar{A}_{0\delta}$-t.

2.24. DEFINÍCIÓ. $M$ (1.1)-nek *globálisan kvázi-aszimptotikusan stabilis halmaza*, ha minden megoldás tart $M$-hez, amint $t\to\infty$.

A 2.20. és 2.24. definíció ekvivalens, ha $M=A_{00}$ és $g(t, 0)\equiv 0$.

2.25. DEFINÍCIÓ. $M$ (1.1)-nek *globálisan kvázi-ekviaszimptotikusan stabilis halmaza,* ha bármely $\varepsilon>0$, $\eta>0$, $\alpha>0$ és $t_1\geqq t_0$-hoz létezik $T>0$, hogy $d(y_0, M(t_1))\leqq\eta$ és $|y_0|\leqq\alpha$-ból $d(y(t, t_1, y_0), M(t))<\varepsilon$ következik midnen $t\geqq t_1+T$-re.

Ez a definíció ekvivalens a 2.10. definícióval, ha $M=A_{00}$ és $g(t, 0)\equiv 0$.

2.26. DEFINÍCIÓ. $M$ (1.1)-nek *globálisan ekviaszimptotikusan stabilis halmaza,* ha stabilis halmaza, globálisan kvázi-ekviaszimptotikusan stabilis halmaza és végül (1.1) megoldásai ekvi-korlátosak $M$-re, vagyis bármely $\eta>0$, $\alpha>0$ és $t_1\geqq t_0$-hoz létezik $\beta>0$, hogy $d(y_0, M(t_1))\leqq\eta$ és $|y_0|\leqq\alpha$-ból $d(y(t, t_1, y_0), M(t))<\beta$ következik minden $t\geqq t_1$-re.

Ez a definíció $M=A_{00}$ és $g(t, 0)\equiv 0$ esetében ekvivalens a 2.11. definícióval, hiszen láttuk, hogy a kvázi-ekviaszimptotikus stabilitás és a stabilitás maga után vonja az ekvi-korlátosság teljesülését.

2.27. DEFINÍCIÓ. $M$ (1.1)-nek *globálisan (t-ben) kvázi-egyenletesen aszimptotikusan stabilis halmaza,* ha globálisan kvázi-ekviaszimptotikusan stabilis halmaza és $T$ nem függ $t_1$-től.

$M=A_{00}$ és $g(t, 0)\equiv 0$ esetében a 2.27. és a 2.12. definíciók ekvivalensek.

2.28. DEFINÍCIÓ. $M$ (1.1)-nek *globálisan (t-ben) egyenletesen aszimptotikusan stabilis halmaza,* ha ($t$-ben) egyenletesen stabilis halmaza, globálisan ($t$-ben) kvázi-egyenletesen aszimptotikusan stabilis halmaza, és végül (1.1) megoldásai $M$-re ($t$-ben) egyenletesen korlátosak, vagyis ekvi-korlátosak $M$-re és $\beta$ független $t_1$-től.

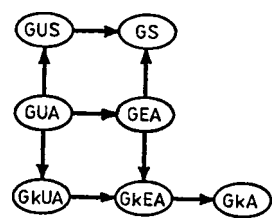
E definíció ekvivalens a 2.13. definícióval, ha $M=A_{00}$ és $g(t, 0)\equiv 0$.

Nézzük ismét a 2.7. példát. Ha csak (i) és (ii) teljesül $0<\delta<y(t)<\infty$ mellett, akkor $\bar{A}_{0\delta}$ globálisan kvázi-ekviaszimptotikusan stabilis halmaz, sőt globálisan ekviaszimptotikusan stabilis halmaz, azonban a 2.27. definíció (és így a 2.28. definíció) általában már nem teljesül. Valóban nem, például, ha $g(t)=1/t$ (lásd a 2.5. példát), mert ekkor $T$ nem független $t_1$-től. Ha (i) és (ii) mellett (iv)-t vagy (v)-t is feltételezzük, akkor a 2.27. és a 2.28. definíciók már teljesülnek, tehát ekkor $\bar{A}_{0\delta}$ globálisan ($t$-ben) egyenletesen aszimptotikusan stabilis halmaz.

A halmazok stabilitására vonatkozó fenti definíciók közötti összefüggéseket a 4. ábrán mutatjuk be.

A halmazok stabilitásával kapcsolatosan végezetül azt jegyezzük meg, hogy valamely $\Gamma$ $\mathbf{R}^n$-beli zárt trajektoriára vonatkozó orbitális stabilitási definíciók adhatók az $\mathbf{R}^+\times\Gamma$ halmazra vonatkozó fenti definíciókkal.

Következzen most az invariáns halmazokra vonatkozó definíció!




*4. ábra*
*GS*=*S*+ekvi-korlátosság *M*-re
*GUS*=*US*+egyenletes korlátosság *M*-re

2.29. DEFINÍCIÓ. $M$ (1.1)-nek *invariáns halmaza,* ha bármely $(t_1, y_0)\in M$-hez tartozó $y(t, t_1, y_0)$ megoldás értelmezett $t\geqq t_1$-re és $y(t, t_1, y_0)\in M$ minden $t\geqq t_1$-re.

Egyszerű belátni, hogy az $M$-re vonatkozó fenti stabilitási definíciók mindegyike ekvivalens az $\overline{M}$-re vonatkozó megfelelő stabilitási definícióval. Ebből nyilvánvalóan következik, hogy ha $M$ (1.1)-nek stabilis halmaza, akkor $\overline{M}$ (1.1)-nek invariáns halmaza. (Az invariáns halmaz nem feltétlenül zárt, de a stabilis halmaz csak akkor biztosan invariáns, ha zárt.)

Végezetül definíciókat adunk attraktorokra és attraktivitási tartományokra.

2.30. DEFINÍCIÓ. $\overline{M}$ (1.1)-nek *egyenletes attraktora,* ha ($t$-ben) egyenletesen aszimptotikusan stabilis (és ebből eredően invariáns) halmaza. $\overline{M}$ *attraktivitási tartománya* az

$$A = \{(\tau, s)\in \mathbf{R}^+\times\Omega\colon \lim_{t\to\infty} d(y(t, \tau, s), \overline{M}(t)) = 0\}$$

halmaz.

Legyen $\tilde{A}$ a következő halmaz:

$$\tilde{A} = \{(t, y)\in\mathbf{R}^+\times\Omega\colon d(y, M(t)) < \delta_0^*;\ |y|\leqq\alpha;\ \alpha > 0\},$$

ahol $\delta_0^*$ a 2.23. definícióban szereplő, $\alpha$-hoz tartozó $\delta_0$ pozitív számok supremuma. Nyilván $\tilde{A}\subset A$.

Ha $\overline{M} = A_{00}$ és e halmaz (1.1′)-nek attraktora a 2.30. definíció értelmében, akkor a zérus megoldás (1.1′)-nek attraktora a 2.15. definíció értelmében is (de fordítva már nem!), és ekkor a 2.30. definíció szerinti $A$ attraktivitási tartomány és a $t_1$-hez tartozó $\mathbf{R}^+\times\mathbf{R}^n$-beli hipersík metszete, valamint a 2.18. definíció szerinti $A(t_1)$ megegyezik.

2.8. PÉLDA. A 2.6. példában $p(t, x, \dot{x})$-ről a folytonosság mellett csak a nem negativitást tételezzük most fel. Ekkor (lásd [7] 1.2. példáját) $\bar{A}_{cd}$ bármely $d\geqq 0$-ra invariáns halmaz, de általában semmilyen $d$-re sem egyenletes attraktor. Arra hogy $\bar{A}_{0d}$ valóban ne legyen egyenletes attraktor elégséges feltétel, ha létezik $p_0(t)$ $\mathbf{R}^+$-on folytonos függvény, hogy

$$0\leqq p(t, x, \dot{x})\leqq p_0(t);\quad \int_{t_0}^{\infty} p_0(t)\,dt < \infty.$$

A 2.7. példában $\bar{A}_{0\delta}$ (1.1′)-nek egyenletes attraktora.

2.31. DEFINÍCIÓ. $\overline{M}$ (1.1)-nek *globális attraktora,* ha stabilis (és ebből eredően invariáns) halmaza, és ha globálisan kvázi-aszimptotikusan stabilis halmaza. Ekkor e globális attraktor *attraktivitási tartománya* $\mathbf{R}^+\times\mathbf{R}^n$.

Ha $\overline{M} = A_{00}$ és e halmaz globális attraktor a 2.31. definíció értelmében, akkor az origó globális attraktor a 2.20. definíció értelmében, de a fordított következtetés már nem igaz.

2.32. DEFINÍCIÓ. $\overline{M}$ (1.1)-nek *globális egyenletes attraktora,* ha globálisan egyenletesen aszimptotikusan stabilis halmaza. Ekkor attraktivitási tartománya $\mathbf{R}^+\times\mathbf{R}^n$.

A 2.7. példában $D = \infty$ esetében $\bar{A}_{0\delta}$ (1.1′)-nek globális egyenletes attraktora.

Annak érdekében, hogy a 2.31. definícióval analóg, nemglobális definíciót nyerjünk, célszerű bevezetni a kvázi-aszimptotikusan stabilis halmaz fogalmát.

2.33. DEFINÍCIÓ. $M$ (1.1)-nek *kvázi-aszimptotikusan stabilis halmaza*, ha bármely $t_1 \geqq t_0$-hoz és $\alpha > 0$-hoz létezik $\delta_0 > 0$, hogy minden $0 < \delta \leqq \delta_0$, $\varepsilon > 0$ és minden olyan $y_0 \in \Omega$ esetében, amelyre $d(y_0, M(t_1)) < \delta$ és $|y_0| \leqq \alpha$ létezik $T > 0$, hogy $t \geqq t_1 + T$-ből $d(y(t, t_1, y_0), M(t)) < \varepsilon$ következik.

Ha $M = A_{00}$, akkor ez a definíció a 2.15. definícióval ekvivalens.

2.34. DEFINÍCIÓ. $\bar{M}$ (1.1)-nek *attraktora*, ha stabilis halmaza és kvázi-aszimptotikusan stabilis halmaza. Ekkor $\bar{M}$ attraktivitási tartománya az

$$A = \{(\tau, s) \in \mathbf{R}^+ \times \Omega : \lim_{t \to \infty} d(y(t, \tau, s), \bar{M}(t)) = 0\}$$

halmaz.

Ha a 2.1. példában (i) és (ii) teljesül, ha $0 < \delta < y(t) < D \leqq \infty$, akkor $\bar{A}_{0\delta}$ (1.1')-nek attraktora, sőt az $A_{0D}$-ból induló megoldások véges időn belül elérik $\bar{A}_{0\delta}$-t. Ha $D = \infty$, akkor $\bar{A}_{0\delta}$ globális attraktor.

2.9. PÉLDA. [9]-ben a szerző kémiai reakciót leíró, nemlineáris harmadrendű autonom differenciálegyenlet rendszer pozitív megoldásainak aszimptotikus viselkedését vizsgálja. A vizsgálat során alapvető jelentőségű volt, hogy sikerült egy attraktort találni, amelynek attraktivitási tartománya tartalmazza az $\mathbf{R}^+ \times Q$ halmazt, ahol $Q$ az első oktáns, kivéve egy egydimenziós sokaságot. Ez az attraktor nem korlátos. Az $\mathbf{R}^+ \times Q$-ból kiinduló megoldások véges időn belül elérik az attraktort, az attraktor invariáns halmaz, így elegendő csupán az attraktoron belüli megoldásokkal foglalkozni. Ez a tény a további vizsgálatot megkönnyítette és így sikerült kimutatni, hogy az $\mathbf{R}^+ \times Q$-ból induló megoldások nem korlátosak.

2.10. PÉLDA. [2, 3, 4, 8]-ban szereplő attraktív halmazok egyenletes attraktorok.

## 3. Attraktorok az origó környezetében

[7] eredményeinek felhasználása céljából érdemes feltételeznünk, hogy az (1.1') egyenlet

$$\dot{x} = F(t, x)x \tag{3.1}$$

alakban adott. Az (1.2') perturbált rendszer ezzel:

$$\dot{x} = F(t, x)x + f(t, x). \tag{3.2}$$

Itt $F$ $\mathbf{R}^+ \times \Omega$-n értelmezett, folytonos $n \times n$-es mátrix függvény.

Az $\bar{A}_{0\delta}$ attraktor létezésének elégséges feltételét következő tételünkben adjuk meg.

3.1. TÉTEL. Tekintsük a (3.2) egyenletet. Legyen $F$ és $f$ értelmezve és folytonos $A_{0D}$-ben $0 < D \leqq \infty$-re. Tegyük fel a következőket:

(i) létezik $R(t)$ $n \times n$-es, $\mathbf{R}^+$-on folytonosan deriválható, reguláris mátrix függvény, és léteznek az $a, \delta$ nem negatív és $b, \lambda_1, \lambda_2, d$ pozitív konstansok, valamint $\varrho_0(t) > 0$ függvény, hogy

$$|R(t)| \leqq \frac{1}{\lambda_1}; \quad |R^{-1}(t)| \leqq \lambda_2; \quad t \geqq t_0,$$

$$a = \lambda_1 \delta; \quad b = \lambda_2 d \leqq \lambda_1 D,$$

és $a<|y|<b$, valamint $t\geqq t_0$ esetében

$$\mu\big(R^{-1}(t)F(t, R(t)y)R(t)-R^{-1}(t)\dot{R}(t)\big)\leqq -\varrho_0(t),$$

ahol $\mu(Q)$ a $Q$ mátrix szimmetrikus részének legnagyobb sajátértékét jelöli [5, 7], és ahol $|\cdot|$ az euklideszi vektornormával generált mátrixnormát jelöli (ekkor nyilván $\lambda_1\leqq\lambda_2$);

(ii) létezik $\varrho(t)>0$ folytonos függvény, hogy

$$-\varrho_0(t)+\frac{\lambda_2}{|y|}\big|f(t, R(t)y)\big|\leqq -\varrho(t);\quad t\geqq t_0;\quad a<|y|<b,$$

továbbá

$$\int_{t_0}^{\infty}\varrho(t)\,dt=\infty;$$

(iii) (3.2) $A_{0D}$-ből induló megoldásai értelmezettek minden $t\geqq t_0$-ra, akkor $\bar{A}_{0\delta}$ a (3.2) egyenlet attraktora, melynek attraktivitási tartománya tartalmazza $A_{0d}$-t. Ha $d=D=\infty$, akkor $\bar{A}_{0\delta}$ globális attraktor. Ha $\delta=0$, akkor (3.2)-nek $y\equiv 0$ ekvi-aszimptotikusan stabilis megoldása, ha emellett $d=D=\infty$, akkor ez a megoldás globálisan ekviaszimptotikusan stabilis. Ha $\varrho(t)\geqq k+\omega(t)$, ahol $\omega(t)$ folytonos, periodikus függvény, $k>0$, $\int_0^{\tau}\omega(t)dt=0$, ahol $\tau$ periódus, akkor valamennyi felsorolt esetben az egyenletesség is teljesül.

*Bizonyítás.* $x$-ről térjünk át az $y=R^{-1}(t)x$ változóra. Ekkor $y$-ra a következő differenciálegyenlet lesz érvényes:

$$\dot{y}=\big(R^{-1}(t)F(t, R(t)y)R(t)-R^{-1}(t)\dot{R}(t)\big)y+R^{-1}(t)f\big(t, R(t)y\big).$$

Vegyük figyelembe, hogy bármely olyan $y(t)$ megoldásra, amelyre $a<|y|<b$:

$$\frac{d}{dt}\ln|y(t)|=\frac{\frac{d}{dt}|y(t)|}{|y(t)|}=\frac{1}{2}\frac{\frac{d}{dt}|y(t)|^2}{|y(t)|^2}=\frac{1}{2}\frac{\frac{d}{dt}y^T(t)y(t)}{y^T(t)y(t)}=$$

$$=\frac{1}{2}\frac{\dot{y}^T(t)y(t)+y^T(t)\dot{y}(t)}{y^T(t)y(t)}\leqq\mu\big(R^{-1}(t)F\big(t, R(t)y(t)\big)R(t)-R^{-1}(t)\dot{R}(t)\big)+$$

$$+\frac{y^T(t)R^{-1}(t)f\big(t, R(t)y(t)\big)}{y^T(t)y(t)}\leqq-\varrho_0(t)+\frac{\lambda_2}{|y(t)|}f\big(t, R(t)y(t)\big)\leqq-\varrho(t),$$

ahol $y^T$ jelöli $y$ transzponáltját ($|y|^2=y^Ty$).

Tegyük fel, hogy $a<|y(t_1)|<b$; $t_1\geqq t_0$, akkor integrálás után:

$$|y(t)|\leqq|y(t_1)|\exp\int_{t_1}^{t}\big(-\varrho(\tau)\big)\,d\tau.$$

Ez az egyenlőtlenség mindaddig érvényes, amíg $a<|y(t)|<b$. Ha $|y|<a\neq 0$, akkor ez maga után vonja, hogy $|x|<a/\lambda_1=\delta$; ugyanígy $|x|<b/\lambda_2=d$-ből $|y|<b$ követ-

kezik. Ezzel beláthatjuk, hogy $\delta<|x(t)|<d$ esetében $a<|y|<b$ és így igaz a következő becslés:

$$|x(t)| \leqq \frac{\lambda_2}{\lambda_1} |x(t_1)| \exp \int_{t_1}^{t} (-\varrho(\tau))\, d\tau.$$

Figyelembe véve a 2. szakasz (i), (ii) és (v) feltételeinek következményeit, a tétel állításait igazoltuk!

Ha $\delta>0$, akkor az $A_{0d}$-ből induló megoldások véges időn belül elérik $\bar{A}_{0\delta}$-t. Ha $\delta=0$, akkor $\lim_{t\to\infty} x(t)=0$, amennyiben $x(t_1)\in A_{0d}$.

Tételünk alkalmazását néhány példán mutatjuk be.

3.1. PÉLDA. E példa keretében a 3.1. tétel egy következményét tárgyaljuk.

3.2. KÖVETKEZMÉNY. Tekintsük a (3.1) egyenletet. Más szóval: (3.2)-ben legyen $f(t, x)\equiv 0$. Teljesüljenek a 3.1. tétel feltételei a következő módosítással: $\delta=0$ és minden olyan $c$-hez, amelyre $0<c<b$, létezik $R_c(t)$, hogy

$$\mu\left(R_c^{-1}(t) F\left(t, R_c(t) y\right) R_c(t) - R_c^{-1}(t) \dot{R}_c(t)\right) \leqq -\varrho_c(t); \quad t \geqq t_0; \quad 0 < c < |y| < b,$$

valamely $\varrho_c(t)>0$ folytonos függvényre, amelyre

$$\int_{t_0}^{\infty} \varrho_c(t)\, dt = \infty.$$

Ekkor $A_{00}$ a (3.1) egyenlet attraktora.

*Bizonyítás.* Legyen $c$ tetszőlegesen kicsiny, de rögzített szám: $c=\lambda_1\varepsilon$. Ekkor a 3.1. tétel bizonyítását alkalmazva, ahol $\delta$ szerepét most $\varepsilon$ tölti be, kiderül, hogy minden olyan $x(t)$ megoldás, amelyre $|x(t_1)|<d$, véges időn belül eléri $\bar{A}_{0\varepsilon}$-t: $|x(t_1+T)|\leqq\varepsilon$!

Konkrét példaképpen vizsgáljuk az

$$\ddot{x}+h(x^2+\dot{x}^2)\dot{x}+x = 0; \quad x\in\mathbf{R}; \quad t \geqq t_0$$

másodrendű egyenletet, ahol $h\in C[I, \mathbf{R}]$, $I=[0, D]$. A 3.1. tételben szereplő $R(t)$ transzformáló mátrix [7] alapján, $c$ függvényében egyszerűen konstruálható az egyenlethez, így kimutatható a szóbanforgó tétel alapján, hogy

$$\lim_{t\to\infty} \left(x^2(t)+\dot{x}^2(t)\right) = 0$$

bármely olyan $x(t)$ megoldása, amelyre $x^2(t_1)+x^2(t_1)<D^2$. $R(t)$ konstruálásához vegyük figyelembe, hogy $x^2+\dot{x}^2\in[c, D]$ esetén $h(x^2+\dot{x}^2)$ korlátos: $0<k_c\leqq \leqq h(x^2+\dot{x}^2)\leqq K_c$. Ezzel elgég kis $0<\delta$-val a következő időtől független $\mathbf{R}$ mátrix megfelel a célnak:

$$R = \frac{1}{\sqrt{2+\delta}} \begin{bmatrix} -\frac{\delta}{2} & \sqrt{4-\delta^2} \\ 1 & 0 \end{bmatrix}.$$

3.2. PÉLDA. Megmutatjuk, hogy a 3.1. tétel nem korlátos becslésű perturbációk esetében is alkalmazható. Tekintsük az

$$\dot{x} = -x+f(t, x); \quad x\in\mathbf{R}; \quad t \geqq t_0$$

egyenletet, amelyről feltesszük, hogy az origó egy elég nagy környezetében bármely $x(t_1)=x_0$; $t_1 \geqq t_0$ kezdeti értékű megoldás létezik minden $t \geqq t_1$-re és

$$|f(t, x)| \leqq \varepsilon\left(\frac{1}{|x|}+x^2\right)$$

valamely $0<\varepsilon$ elég kicsiny konstanssal. Ekkor a 3.1. tételt alkalmazva egyszerűen belátható, hogy $\bar{A}_{0\delta}$ egyenletes attraktor, amelynek attraktivitási tartománya tartalmazza $A_{0d}$-t, ahol $\delta$, illetve $d$ az

$$\varepsilon\left(\frac{1}{x}+x^2\right)=x; \quad x>0$$

egyenlet kisebbik, illetve nagyobbik megoldása. A $0<\delta<d$ konstansok létezésének elégséges feltétele: $0<\varepsilon<\frac{1}{2}$.

3.3. PÉLDA. Legyen az

$$\dot{x} = B(t)x+f(t, x); \; x\in\mathbf{R}^n; \; t \geqq t_0$$

egyenletben $B(t)$ olyan folytonos, periodikus mátrix függvény, hogy a perturbálatlan lineáris rendszer karakterisztikus multiplikátorai abszolút értékben egynél kisebbek. Akkor egyrészt FLOQUET-nak a periodikus lineáris rendszerekre vonatkozó elméletéből, másrészt [5]-ből tudjuk, hogy létezik $R(t)$, 3.1. tétel szerinti transzformáló mátrix, amellyel

$$\mu\left(R^{-1}(t)B(t)R(t)-R^{-1}(t)\dot{R}(t)\right) = -\varrho_0$$

valamely $\varrho_0>0$ konstanssal. Az $f$-re vonatkozó egyéb feltételek teljesülése mellett egyenletünkre a 3.1. tétel alkalmazható.

3.4. PÉLDA. E példa keretében bebizonyítjuk a következő tételt.

3.3. TÉTEL. Tekintsük az

$$\ddot{x}+p(t, x, \dot{x})\dot{x}+x+f(t, x, \dot{x}) = 0; \quad x\in\mathbf{R}; \quad t \geqq t_0$$

egyenletet. Tételezzük fel, hogy $t \geqq t_1 \geqq t_0$-ra és mefelelően nagy $0<D$-re $x^2+\dot{x}^2<D^2$ esetén létezik $\delta$: $\mathbf{R}^+ \to \mathbf{R}$ függvény, amelyre

(i) $\delta(t) \leqq p(t, x, \dot{x}) \leqq \dfrac{1}{\delta(t)}$,

(ii) $0 < \delta(t) < 1$,

(iii) $\delta$ folytonosan deriválható,

(iv) $\dot{\delta}(t) \leqq 0$,

(v) $\displaystyle\int_{t_0}^{\infty} \delta(t)\,dt = \infty$,

(vi) $\dfrac{|f(t, x, \dot{x})|}{\delta(t)} \leqq \varepsilon(t)$; $\lim\limits_{t\to\infty} \varepsilon(t) = 0$; $\varepsilon$ folytonos.

E feltételekkel $A_{00}$ egyenletünknek attraktora.

*Bizonyítás.* Legyen $t \geqq t_1$ és írjuk át egyenletünket mátrixos alakra:

$$z_1 = x,$$

$$z_2 = \dot{x},$$

$$z = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix},$$

$$\dot{z} = \begin{bmatrix} 0 & 1 \\ -1 & -p(t, z) \end{bmatrix} z + \begin{bmatrix} 0 \\ -f(t, z) \end{bmatrix}.$$

[7] alapján alkalmazzuk az

$$R^{-1}(t) = \begin{bmatrix} 0 & \sqrt{2+\delta(t)} \\ \dfrac{2}{\sqrt{2-\delta(t)}} & \dfrac{\delta(t)}{\sqrt{2-\delta(t)}} \end{bmatrix}$$

transzformáló mátrixot, amelyre

$$|R(t)| = \frac{1}{\sqrt{2}} = \frac{1}{\lambda_1},$$

$$|R^{-1}(t)| = \sqrt{2}\sqrt{\frac{2+\delta(t)}{2-\delta(t)}} \leqq \sqrt{2}\sqrt{\frac{2+\delta(t_1)}{2-\delta(t_1)}} = \lambda_2,$$

$$y = R^{-1}(t)\, z,$$

$$\varrho_0(t) = \frac{\delta(t)}{3}.$$

Nyilvánvaló, hogy elég nagy $t_1$-hez létezik $0 \leqq a(t_1) < \dfrac{\lambda_1^2}{\lambda_2} D$, hogy

$$-\frac{\delta(t)}{3} + \frac{\lambda_2}{|y|}\,|f(t, R(t)y)| \leqq -\varepsilon_{t_1}(t) < 0; \quad a(t_1) < |y| < \lambda_1 D,$$

ahol

$$\int_{t_1}^{\infty} \varepsilon_{t_1}(t)\, dt = \infty.$$

Vagyis a 3.1. tétel szerint (az egyéb feltételek teljesülése mellett) $\bar{A}_{0\delta(t_1)}$ attraktor $(\delta(t_1) = a(t_1)/\lambda_1)$, és a megoldások véges időn belül elérik $\bar{A}_{0\delta(t_1)}$-t, és ezután benne is maradnak. Az is nyilvánvaló, hogy bármely $t^* \geqq t_1$-hez is létezik $a(t^*) < a(t_1)$, hogy $t \geqq t^*$ és $a(t^*) < |y| < \lambda_1 D$ esetében

$$-\frac{\delta(t)}{3} + \frac{\lambda_2}{|y|}\,|f(t, R(t)y)| \leqq -\varepsilon_{t^*}(t) < 0,$$

ahol

$$\int_{t^*}^{\infty} \varepsilon_{t^*}(t)\, dt = \infty.$$

Így az $\bar{A}_{0\delta(t_1)}$-beli megoldásokat nyomon követve a $t \geqq t^*$ időpontokban, világos, hogy $\bar{A}_{0\delta(t^*)}$ attraktor $(\delta(t^*)=a(t^*)/\lambda_1)$ és $\bar{A}_{0\delta(t_1)}$-ből a megoldások véges idő múlva elérik $\bar{A}_{0\delta(t^*)}$-t. Tekintettel arra, hogy $\lim_{t^*\to\infty} a(t^*)=0$, így $A_{00}$ is attraktor!

$A_{00}$ attraktivitási tartománya tartalmazza $A_{0\delta(t_1)}$-t.

### 4. Nemlineáris egyenletek közelítően periodikus megoldásai; rezonancia

A 3.1. tételt alkalmazzuk arra az esetre, amikor (1.1)-ben

$$g(t, x) = B(t)x+b(t), \tag{4.1}$$

ahol $B(t)$ folytonos, periodikus mátrix függvény és a homogén rendszer valamennyi karakterisztikus exponense negatív valós részű, $b(t)$ pedig folytonos, periodikus függvény. Ebben az esetben (4.1)-nek van pontosan egy, globálisan egyenletesen aszimptotikusan stabilis, majdnem periodikus $p(t)$ megoldása [17]. $p(t)$ nyilván korlátos. Tegyük fel, hogy valamely $K>0$ konstansra

$$|p(t)| \leqq K. \tag{4.2}$$

Alkalmazzuk az (1.3) transzformációt az (1.2) perturbált egyenletre. Ekkor

$$\dot{y} = B(t)y+f(t, y+p(t)), \tag{4.3}$$

amelyre a 3.1. tétel a 3.3. példában közölt módon alkalmazható.

Ha létezik $\bar{A}_{p\delta}$ attraktor elég kicsiny $\delta$-val, akkor — mint a bevezetőben említettük — azt mondhatjuk, hogy $p(t)$ $\delta$ közelítéssel (1.2)-nek is megoldása. Gyakorlati célokra a $\delta$ adat helyett jobban jellemzi a közelítés hibáját a $\delta/K$ viszonylagos érték. Ezért vezessük be az

$$\alpha = \frac{\delta}{K} \tag{4.4}$$

számot, mint az attraktor relatív átmérőjét. Hasonlóképpen $d$ helyett a

$$\beta = \frac{d}{K} \tag{4.5}$$

értékkel fogunk számolni.

Vizsgálódásaink esetei abban fognak különbözni egymástól, hogy $|f(t, x)|$-re milyen becslések adhatók.

Tételezzük fel például, hogy

$$|f(t, x)| \leqq \nu|x|^n; \quad \nu > 0 \tag{4.6}$$

valamely $n$ nem negatív egész számra. (A fenti $\alpha$ és $\beta$ betűknél ezt az $n$-t indexként fogjuk feltüntetni: $\alpha_n$, $\beta_n$).

Figyelembe véve (4.3)-at a 3.1. tétel (ii) feltétele így alakul:

$$-\varrho_0+\frac{\lambda_2}{|y|}\nu|R(t)y+p(t)|^n \leqq -\varrho(t) < 0.$$

$|R(t)y+p(t)|$-t felülről becsülve legyen

$$-\varrho_0+\frac{\lambda_2}{|y|}\nu\left(\frac{|y|}{\lambda_1}+K\right)^n \leqq -\varrho^* < 0 \tag{4.7}$$

valamely $\varrho^*>0$ konstanssal. Azt az $[a, b]$ intervallumot kellene meghatározni, amelyen belüli $|y|$-ra (4.7) teljesül. Ehelyett azt az $[a, b]$-t fogjuk keresni, amelyre (4.7) baloldali kifejezése nem pozitív. Ha létezik ilyen zérusnál hosszabb intervallum, akkor ennek minden belső rész-intervallumához már valóban teljesül (4.7) megfelelően kicsiny $0<\varrho^*$-ra. Ha $n=0$, akkor

$$\alpha_0 = \frac{\lambda_2}{\lambda_1}\frac{\nu}{K\varrho_0}, \tag{4.8}$$

$$\beta_0 = \infty. \tag{4.9}$$

Ha $n=1$, akkor

$$\alpha_1 = \frac{\frac{\lambda_2}{\lambda_1}\nu}{\varrho_0-\frac{\lambda_2}{\lambda_1}\nu}, \tag{4.10}$$

feltéve, hogy

$$\varrho_0 > \frac{\lambda_2}{\lambda_1}\nu, \tag{4.11}$$

továbbá:

$$\beta_1 = \infty. \tag{4.12}$$

Az $n=2$ esetre (4.7)-ből rendezés után

$$|y|^2+\left(2K\lambda_1-\frac{\varrho_0}{\nu}\frac{\lambda_1^2}{\lambda_2}\right)|y|+\lambda_1^2K \leqq 0,$$

amelyből $a\leqq b$ pontosan akkor létezik, ha

$$\frac{\varrho_0\lambda_1}{4\nu\lambda_2} \geqq K \tag{4.13}$$

és ekkor

$$a = \frac{\varrho_0\lambda_1^2}{2\nu\lambda_2}-K\lambda_1-\frac{1}{2}\sqrt{\frac{\varrho_0\lambda_1^2}{\nu\lambda_2}\left(\frac{\varrho_0\lambda_1^2}{\nu\lambda_2}-4K\lambda_1\right)},$$

$$b = \frac{\varrho_0\lambda_1^2}{2\nu\lambda_2}-K\lambda_1+\frac{1}{2}\sqrt{\frac{\varrho_0\lambda_1^2}{\nu\lambda_2}\left(\frac{\varrho_0\lambda_1^2}{\nu\lambda_2}-4K\lambda_1\right)}.$$

Ezenkívül az

$$\frac{a}{\lambda_1} \leqq \frac{b}{\lambda_2} \tag{4.14}$$

feltételnek is teljesülnie kell. Könnyen kimutatható, hogy a két feltétel ((4.13) és (4.14)) pontosan akkor teljesül, ha

$$\sqrt{\frac{\lambda_2}{\lambda_1}}\left(1+\sqrt{\frac{\lambda_2}{\lambda_1}}\right)^2 \leqq \frac{\varrho_0}{\nu K}. \tag{4.15}$$

Ekkor:

$$\alpha_2 = \frac{\varrho_0\lambda_1}{2\nu K\lambda_2} - 1 - \sqrt{\left[\frac{\varrho_0\lambda_1}{2\nu K\lambda_2}-1\right]^2-1}, \tag{4.16}$$

$$\beta_2 = \frac{\lambda_1}{\lambda_2}\left\{\frac{\varrho_0\lambda_1}{2\nu K\lambda_2} - 1 + \sqrt{\left[\frac{\varrho_0\lambda_1}{2\nu K\lambda_2}-1\right]^2-1}\right\}. \tag{4.17}$$

(4.15), (4.16) és (4.17) áttekinthetőbbé válik új változók bevezetésével. Ekkor összefoglalva az $n=2$ esetet, a következőket kapjuk. Legyen

$$u = \frac{\varrho_0\lambda_1}{2\nu K\lambda_2} - 1, \tag{4.18}$$

$$v = \sqrt{\frac{\lambda_1}{\lambda_2}}. \tag{4.19}$$

Ha

$$\frac{1}{v} + v \leqq 2u, \tag{4.20}$$

akkor

$$\alpha_2 = u - \sqrt{u^2-1}, \tag{4.21}$$

$$\alpha_2 < \beta_2 = v^2\left(u+\sqrt{u^2-1}\right) = \frac{v^2}{\alpha_2}. \tag{4.22}$$

Ha $u$ eléggé nagy, akkor $\alpha_2$ és $\beta_2$-re a következő közelítés adható:

$$\alpha_2 \approx \frac{1}{2u} \approx \frac{\nu K\lambda_2}{\varrho_0\lambda_1}, \tag{4.23}$$

$$\beta_2 \approx \frac{\varrho_0\lambda_1^2}{\nu K\lambda_2^2}. \tag{4.24}$$

Most tételezzük fel, hogy $n\geqq 2$. Használjuk fel azt az egyszerűen bizonyítható egyenlőtlenséget, hogy

$$(1+\varepsilon)^n < 1+2n\varepsilon,$$

ahol

$$\frac{1}{2n} > \varepsilon > 0; \quad n \geqq 1.$$

Ezzel

$$\left(K+\frac{|y|}{\lambda_1}\right)^n < K^n\left(1+2n\frac{|y|}{\lambda_1 K}\right),$$

ahol

$$|y| < \frac{\lambda_1 K}{2n}.$$

Ezt felhasználva ki tudjuk mutatni, hogy

$$0 < \alpha_n^* = \frac{\frac{\lambda_2}{\lambda_1} \nu K^{n-1}}{\varrho_0 - 2n \frac{\lambda_2}{\lambda_1} \nu K^{n-1}} < \frac{\lambda_1}{\lambda_2} \frac{1}{2n} = \beta_n^* \tag{4.25}$$

teljesülése esetén igaz, hogy

$$\alpha_n < \alpha_n^* < \beta_n^* < \beta_n. \tag{4.26}$$

Az elmondottak alkalmazását példákon mutatjuk be.

4.1. PÉLDA. Az

$$\dot{x} = -Ax + B \sin t; \quad x \in \mathbf{R}; \quad A > 0; \quad B > 0$$

egyenletnek a

$$p(t) = \frac{AB}{1+A^2} \sin t - \frac{B}{1+A^2} \cos t$$

függvény globálisan egyenletesen aszimptotikusan stabilis periodikus megoldása és

$$|p(t)| \leqq \frac{B}{\sqrt{1+A^2}} = K.$$

Tekintsük most az

$$\dot{x} = -Ax + B \sin t + f(t, x)$$

egyenletet, ahol

$$|f(t, x)| \leqq \nu |x|^n.$$

A 3.1. tételben szereplő egyéb állandók:

$$\varrho_0 = A,$$

$$\lambda_1 = \lambda_2 = 1.$$

Ezekkel

$$\alpha_0 = \nu \frac{\sqrt{1+A^2}}{BA}; \quad \beta_0 = \infty,$$

$$\alpha_1 = \frac{\nu}{A-\nu} \quad (\text{ha} \quad A > \nu); \quad \beta_1 = \infty.$$

Ha $\nu$ elegendően kicsi, akkor:

$$\alpha_2 \approx \nu \frac{B}{A\sqrt{1+A^2}}; \quad \beta_2 \approx \frac{1}{\nu} \frac{A\sqrt{1+A^2}}{B}.$$

4.2. PÉLDA. Vizsgáljuk most az

$$\ddot{x}+p\dot{x}+qx+f(t, x, \dot{x}) = b \sin \omega t; \quad x\in\mathbf{R}; \quad t \geqq t_0;$$
$$|f(t, x, \dot{x})| \leqq v(x^2+\dot{x}^2)^{n/2}$$

nemlineáris egyenletet, ahol $v, \omega, p, q$ és $b$ pozitív állandók, $n$ értéke 0,1 vagy 2. A perturbálatlan egyenlet ($f\equiv 0$) periodikus megoldása:

$$x_0(t) = b\frac{q-\omega^2}{(q-\omega^2)^2+\omega^2p^2}\sin\omega t - b\frac{\omega p}{(q-\omega^2)^2+\omega^2p^2}\cos\omega t =$$
$$= \frac{b}{\sqrt{(q-\omega^2)^2+\omega^2p^2}}\sin(\omega t+\varphi).$$

Kimutatható, hogy

$$\sqrt{x_0^2(t)+\dot{x}_0^2(t)} \leqq \frac{b}{\sqrt{(q-\omega^2)^2+\omega^2p^2}}\max(1;\ \omega) = K(\omega).$$

Differenciálegyenletünk perturbálatlan bal oldalának együttható mátrixa az egyenlet mátrixos átírásában:

$$A = \begin{bmatrix} 0 & 1 \\ -q & -p \end{bmatrix}.$$

Tekintsük a

$$V = \begin{bmatrix} \frac{1}{p}+\frac{p}{q}+\frac{q}{p} & \frac{1}{q} \\ \frac{1}{q} & \frac{1+q}{pq} \end{bmatrix}$$

mátrixot, amellyel

$$VA+A^TV = -2E,$$

ahol $E$ az egységmátrix. Bontsuk fel $V$-t

$$V = M^TD^2M$$

alakban, ahol $M$ ortogonális, $D^2$ diagonális, főátlójában $V$ pozitív sajátértékeivel ($\lambda_2^2 \geqq \lambda_1^2$). Legyen

$$R = M^TD^{-1},$$

akkor

$$R^{-1}AR+R^TA^TR^{T-1} = -2D^2,$$

amelyből:

$$\mu(R^{-1}AR) = -\varrho_0 = -\frac{1}{\lambda_2^2},$$

továbbá a részletszámításokat mellőzve:

$$|R^{-1}| = \lambda_2,$$
$$|R| = \frac{1}{\lambda_1},$$
$$0 < \lambda_1 = \sqrt{\frac{\gamma_1-\sqrt{\gamma_1\gamma_2}}{2pq}} < \lambda_2 = \sqrt{\frac{\gamma_1+\sqrt{\gamma_1\gamma_2}}{2pq}},$$

ahol

$$\gamma_1 = (1+q)^2+p^2; \quad \gamma_2 = (1-q)^2+p^2.$$

Az $\alpha$-ra és a $\beta$-ra vonatkozó számításokhoz minden adat a rendelkezésünkre áll. Egy konkrét számpélda kapcsán vizsgáljuk a rezonancia jelenségét. Induljunk ki a következő adatokból:

$$p = 10^{-4},$$

$$q = 1,$$

$$b = 1,$$

$$\nu = 10^{-7},$$

$$0 < \omega \quad (\omega \text{ paraméter}).$$

Ekkor

$$\lambda_1 \approx \lambda_2 \approx \sqrt{2} \cdot 10^2,$$

$$\varrho_0 \approx 5 \cdot 10^{-5},$$

$$K(\omega) = \frac{\max(1;\ \omega)}{\sqrt{(1-\omega^2)^2+\omega^2 10^{-8}}}; \quad K(1) = 10^4.$$

Ha $\omega$-t változtatjuk, akkor a perturbált egyenlet $\omega \approx 1$ esetén rezonanciába kerül, vagyis $K\omega$ függvényében az $\omega = 1$ értéknél közelítőleg maximumát veszi fel. (Valódi rezonancia akkor lenne, ha perturbálatlan egyenletnél $p=0$ lenne. Ekkor $\lim_{\omega \to 1} K(\omega) = \infty$. Ha differenciálegyenletünk, illetve ennek perturbálatlan változata valamely valóságos műszaki berendezés modellje, akkor elég nagy $K$-nál már a berendezés tönkremenetelével kell számolni. Ezért $p=10^{-4}$-nél az $\omega=1$ esetet is rezonanciának tekintjük.) Határozzuk meg $\alpha_n$ és $\beta_n$ értékét:

$$\alpha_0 \approx 2 \cdot 10^{-3}/K(\omega); \quad \beta_0 = \infty,$$

$$\alpha_1 \approx 2 \cdot 10^{-3}; \qquad \beta_1 = \infty,$$

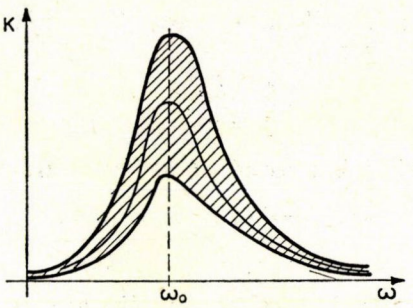
$$u \approx \frac{250}{K(\omega)} - 1; \qquad v \approx 1,$$

$$\alpha_2 = u - \sqrt{u^2-1}; \qquad \beta_2 \approx \frac{1}{\alpha_2}.$$

A következőket állapíthatjuk meg:

1. Ha $n=0$ (korlátos perturbáció), akkor a perturbálatlan egyenlet periodikus megoldása igen jó közelítéssel ($2 \cdot 10^{-5}\% \ldots 2 \cdot 10^{-3}\%$) a perturbált egyenletnek is megoldása. A közelítés a rezonancia felé haladva ($\omega \to 1$) egyre jobb.
2. Ha $n=1$, akkor ugyanaz igaz, mint 1. esetében, azzal a különbséggel, hogy a közelítés hibája független $\omega$-tól.
3. Ha $n=2$, akkor a közelítés hibája $\omega=1$-nél körülbelül 40%, a rezonanciától távolodva a közelítés egyre jobb. De még a 40%-os pontatlan közelítés is alkalmas arra, hogy megállapítsuk, a rezonancia jelensége a perturbált egyenletnél is fellép.

A rezonancia jelenségét az 5. ábrán szemléltettük. A perturbált rendszer egyenletes attraktorán belül haladó megoldások amplitudói a vonalkázott tartományba




*5. ábra*

esnek. A tartományon belül halad a perturbálatlan egyenlet $K(\omega)$ amplitudó-frekvencia görbéje. A rezonancia $\omega=\omega_0$-nál lép fel.

4.3. PÉLDA. Tételezzük fel, hogy (1.1)-ben nemcsak $g$, hanem $g'_x$ is folytonos, továbbá a $p(t)$ kitüntetett megoldás periódikus. Ekkor (1.2) így írható át:

$$\dot{x} = g'_x(t, p(t))x+[g(t, p(t))-g'_x(t, p(t))p(t)]+ \tag{4.27}$$

$$+[g(t, x)-g(t, p(t))-g'_x(t, p(t))(x-p(t))+f(t, x)].$$

A

$$B(t) = g'_x(t, p(t)),$$

$$b(t) = g(t, p(t))-g'_x(t, p(t))p(t),$$

$$\tilde{f}(t, x) = g(t, x)-g(t, p(t))-g'_x(t, p(t))(x-p(t))+f(t, x)$$

új jelölések bevezetésével elérjük, hogy (1.2) egyenletünk az e szakasz elején tárgyalt esetnek felel meg. Lényegében ezt az átírást alkalmazza [3].

## 5. Perturbált bifurkáció

A valóságos jelenségek matematikai modellezése során nem ritka és fontos esetnek számít, amikor a matematikai modellt képviselő differenciálegyenlet bifurkációs jelenséget mutat. [11] és [12] azt vizsgálja, hogy periodikus perturbáció miképpen befolyásolja a perturbálatlan egyenlet bifurkációs viselkedését. Az eddig megismert apparátusunk alkalmas arra, hogy megmutassuk, megfelelően kis perturbáció mellett a differenciálegyenlet jó közelítéssel megtartja a bifurkációs viselkedését. A módszert egy példa kapcsán világítjuk meg. A

$$\dot{z} = A(z, \mu)z+f(t, z, \mu); \quad z\in\mathbf{R}^2; \quad z=\begin{bmatrix} x \\ y \end{bmatrix}; \quad \mu\in\mathbf{R}; \quad t \geqq t_0,$$

$$A(z, \mu) = \begin{bmatrix} \mu-x^2-y^2 & 1 \\ -1 & \mu-x^2-y^2 \end{bmatrix}; \quad f = \begin{bmatrix} f_1 \\ f_2 \end{bmatrix}$$

differenciálegyenletnél a perturbálatlan esetben ($f \equiv 0$) a Hopf bifurkáció klasszikus esetét tanulmányozhatjuk. Az

$$x = r\cos\varphi$$

$$y = r\sin\varphi$$

változókra áttérve egyenletünk a következő alakot nyeri:

$$\dot{r} = r(\mu - r^2) + f_1\cos\varphi + f_2\sin\varphi$$

$$\dot{\varphi} = -1 + f_1\sin\varphi - f_2\cos\varphi$$

Ha $f_1 \equiv f_2 \equiv 0$, akkor $\mu \leqq 0$-nál az $r \equiv 0$ egyensúlyi helyzet aszimptotikusan stabilis és nincs (ezen kívül) periodikus megoldás. Ha $\mu > 0$, akkor az $r \equiv 0$ egyensúlyi helyzet labilis és (az ezen kívül létező) periodikus megoldás sereg $\Gamma$ trajektoriája az $r_0 = \sqrt{\mu}$ sugarú kör. Az $x$, $y$ síkban ez a trajektoria globálisan egyenletesen orbitálisan aszimptotikusan stabilis.

Legyen most $f$ $t$-ben és $z$-ben folytonos és korlátos:

$$|f| = \sqrt{f_1^2 + f_2^2} < \eta.$$

Határozzuk meg $\dot{r}$ előjelét. E célból a

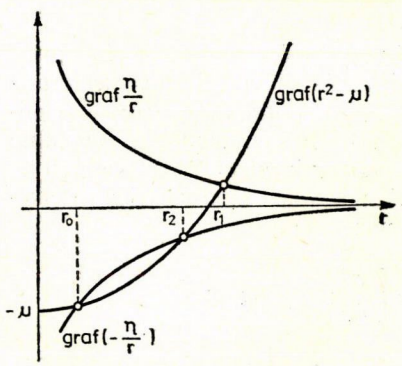
$$-\frac{\eta}{r} = r^2 - \mu,$$

illetve

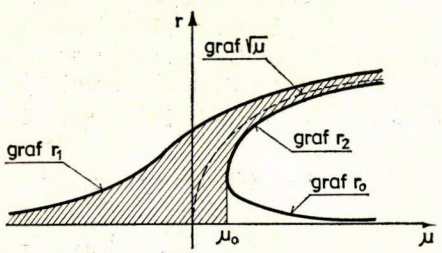
$$\frac{\eta}{r} = r^2 - \mu$$

egyenleteket kell megoldanunk (lásd a 6. ábrát). Ennek alapján felrajzolhatjuk a 7. ábrát. (5.1)-nek létezik $\mu$-től függően attraktora. $\bar{A}_{0r_1(\mu)}$ bármely $\mu$-re globális attraktor. Ha $\mu > \mu_0$, akkor $\bar{A}_{0r_1(\mu)} \setminus A_{0r_2(\mu)}$ is attraktor, de nem globális. Legyen $M(\mu)$ a következő halmaz:

$$M(\mu) = \begin{cases} \bar{A}_{0r_1(\mu)}, & \text{ha } \mu \leqq \mu_0, \\ \bar{A}_{0r_1(\mu)} \setminus A_{0r_2(\mu)}, & \text{ha } \mu > \mu_0. \end{cases}$$




6. ábra




7. ábra

A bevonalkázott tartomány a 7. ábrán az $M(\mu)$ halmaz $t$=állandó, $\varphi$=állandó metszete és a $\{\mu\colon \mu\in\mathbf{R}\}$ halmaz direkt szorzatát mutatja. $\mu>\mu_0$ esetén $M(\mu)$ attraktivitási tartománya $\mathbf{R}^2\setminus\bar{A}_{0r_0(\mu)}$. Az $M(\mu)$-ben haladó megoldások a közelítően periodikus megoldásoknak felelnek meg. Minél kisebb $\eta>0$, illetve minél nagyobb $\mu$, ezek a közelítőleg periodikus megoldások annál jobban közelíthetőek a perturbálatlan egyenlet megfelelő periodikus megoldásaival.

## IRODALOM


[1] FARKAS, M., „Autonom rendszerek periodikus perturbációiról", *Alkalmazott Matematikai Lapok* **1** (1975) 197—254.

[2] FARKAS, M., "The attractor of Duffing's equation under bounded perturbation". *Annali di Matematica pura ed applicata,* **128** (1980) 123—132.

[3] FARKAS, M., "Attractors of systems close to periodic ones", *Nonlinear Analysis TMA* **8** (1981) 845—851.

[4] FARKAS, M., "Attractors of systems close to autonomous ones", *Acta Scientarium Mathematicarum* **44** (1982) 329—334.

[5] GARAY, B. M. and KERTÉSZ, V., "Lozinsky's functional and the transformation of variables". *Zeitschrift für Analysis und ihre Anwendungen* Bd. **3** (1) (1984) 87—95.

[6] GÁTI, R., Turbófeltöltő rotor vizsgálata az ágyazás harmonikus rezgetése esetén. Műszaki doktori értekezés, Budapest, 1977.

[7] KERTÉSZ, V., „Pozitív definit kvadratikus Ljapunov függvények alkalmazása stabilitási vizsgálatokhoz". *Alkalmazott Matematikai Lapok* **9** (1983) 375—386.

[8] KERTÉSZ, V., "Notes on a theorem of M. Farkas", Megjelenés alatt.

[9] KERTÉSZ, V., "Global mathematical analysis of the explodator". *Nonlinear Analysis TMA* **8** (1984) 941—961.

[10] Малкин, И. Г., Некоторые задачи теории нелинейных коледаний (Госуд. Изд. Технико—Теоретической литературы, Москва, 1956).

[11] ROSENBLAT, S. and COHEN, D. S., "Periodically perturbed bifurcation-I. Simple bifurcation", *Studies in Applied Mathematics* **63** (1980) 1—23.

[12] ROSENBLAT, S. and COHEN, D. S., "Periodically perturbed bifurcation II. Hopf bifurcation", *Studies in Applied Mathematics* **64** (1981) 143—175.

[13] ROUCHE, N., HABETS, P. and LALOY, M., *Stability Theory by Liapunov's Direct Method* (Springer Vrlg., New York—Heidelberg—Berlin, 1977) (magyar kiadásban: ROUCHE, N., HABETS, P., LALOY, M., Stabilitáselmélet. A Ljapunov-féle direkt módszer, Műszaki Könyvkiadó, Budapest, 1984).

[14] *A Káosz*, szerkesztette *Szépfalusy Péter, Tél Tamás* (Akadémiai Kiadó, Budapest, 1982).

[15] TRAN VAN NHUNG, On stability of ordinary differential equations under random perturbations. Kandidátusi értekezés, Budapest, 1982.

[16] YOSHIZAWA, T., "Stability theory by Liapunov's second method", *The Mathematical Society of Japan,* 1966.

[17] YOSHIZAWA, T., *Stability Theory and the Existence of Periodic and Almost Periodic Solutions* (Springer Vrlg., New York—Heidelberg—Berlin, 1975).


(Beérkezett: 1984. június 22.)

KERTÉSZ VIKTOR
BME MATEMATIKA TANSZÉKCSOPORT
1521 BUDAPEST, STOCZEK U. H ÉP. IV. E. 44.

# ATTRACTORS OF NONLINEAR DIFFERENTIAL EQUATIONS

V. KERTÉSZ

The perturbed equation

$$\dot{x} = g(t, x) + f(t, x); \quad x \in \mathbf{R}^n$$

will be studied here, where $f$ is considered to be the perturbation. It is assumed that the equation

$$\dot{x} = g(t, x)$$

has a bounded, asymptotically stable solution $p(t)$. The problem is in which cases has the perturbed equation an attractor, that is an asymptotically stable invariant set consisting the graph of this mentioned solution $p(t)$ and what is the region of attractivity of this attractor.

*Alkalmazott Matematikai Lapok* **10** (1984) 257—272

# LINEÁRIS DIFFERENCIÁLEGYENLETEK NUMERIKUS MÓDSZEREINEK STABILITÁSA

GALÁNTAI AURÉL

Gödöllő

A dolgozatban kiterjesztjük az $A$-, $A(\alpha)$- és $A_0$-stabilitásokat inhomogén lineáris differenciálegyenletek esetére és vizsgáljuk e tulajdonság létezését. A kapott eredmények jellemzik az $A[\alpha]$-stabilis módszerek hatásmechanizmusát az ún. stiff típusú differenciálegyenleteken és választ adnak DAHLQUIST egy kérdésfeltevésére.

## 1. Bevezetés

A dolgozatban a konstans együtthatós, inhomogén

$$y' = Ay + g(t); \quad y(0) = y_0 \in \mathbf{C}^m \quad (A \in \mathbf{C}^{m\times m},\ g:\ R^+ \to \mathbf{C}^m) \tag{1.1}$$

alakú differenciálegyenletek numerikus megoldását vizsgáljuk a $[0, +\infty)$ intervallumon. Célunk az $A$-, $A(\alpha)$-, ill. $A_0$-stabilitások (továbbiakban közös néven $A[\alpha]$-stabilitás) kiterjesztése, pontosabban az $A[\alpha]$-stabilis módszerek viselkedésének jellemzése az (1.1) alakú *Cauchy-problémák* osztályán. A vizsgált tulajdonság a következőképpen fogalmazható meg.

Legyen $t_n = nh$ $(h>0,\ n=0, 1, \ldots)$ és

$$W(\alpha) = \begin{cases} \{z \mid z\in\mathbf{C},\ \mathrm{Re}\,(z) < 0\}, & \text{ha} \quad \alpha = \pi/2 \\ \{z \mid z\in\mathbf{C},\ \mathrm{Re}\,(z) < 0,\ |\arg(-z)| < \alpha\}, & \text{ha} \quad 0 < \alpha < \pi/2. \\ (-\infty, 0), & \text{ha} \quad \alpha = 0 \end{cases}$$

Jelölje továbbá $y_n$ az $y(t_n)$ pontos megoldásérték közelítését és tekintsük a

$$\sum_{i=0}^{k} a_i y_{n+i} = hG_f(t_n, y_n, y_{n+1}, \ldots, y_{n+k}, h) \quad (n = 0, 1, \ldots) \tag{1.2}$$

alakú numerikus módszereket, ahol $a_i \in R$, $G_f$: $R\times\mathbf{C}^{(k+1)m}\times R \to \mathbf{C}^m$ adottak ([20]).

1.1. DEFINÍCIÓ. ([10]). Egy numerikus módszert $A[\alpha, g]$-stabilisnak vagy $A[\alpha, g]$-elfogadhatónak nevezünk, ha minden rögzített $h>0$ lépéshossz és minden, a $\sigma(A)\subset W(\alpha)$ feltételt kielégítő együttható mátrix esetén az (1.1) *Cauchy-problémának* a módszer által kapott $y_n$ $(n=0, 1, 2, \ldots)$ közelítő megoldására fennáll az

$$y_n - y(t_n) \to 0 \quad (n \to +\infty) \tag{1.3}$$

feltétel.

Az $A[\alpha]$-stabilitás ekvivalens az $A[\alpha, 0]$-stabilitással. Az (1.3) feltételből, amely az $e_n := y_n - y(t_n)$ globális hiba végtelenben vett zérus konvergenciáját jelenti, követ-

keztetések vonhatók le a vizsgált módszerek hosszú számítási intervallumokon való viselkedéséről, önstabilizáló tulajdonságáról (STETTER [41]), valamint a közelítő megoldásnak az elméleti megoldás ún. sima komponenséhez való konvergenciájáról ([13], [20], [31], [41]). A sima megoldás alatt általában az inhomogén tagnak megfelelő megoldáskomponenst értik, azonban bizonyos esetekben ([11], [12], KREISS [30]) ettől különbözhet.

Az $A[\alpha]$-stabilitás ismert kiterjesztései főleg monotonitási ([7], [5], [35], [38]) vagy osztályra vonatkozó stabilitási, ill. globális hibakorlát feltételeket tartalmaznak ([29], [37], [36], [38], [45], [46—48], [17]). Kisparaméteres problémák sima megoldásához való konvergenciát vizsgálnak a [44], [41], [16], [47—48] munkák.

Vizsgálatainkat tehát az indokolja, hogy a kiterjesztések nem elsősorban az $A[\alpha]$-stabilitás hatását vizsgálják, hanem a módszerek hozzá kapcsolódó egyéb tulajdonságait, valamint az, hogy ez a hatásmechanizmus még a lineáris esetben sincs kielégítően jellemezve (DAHLQUIST [7]).

A globális hiba végtelenben vett konvergenciáját a lineáris *inhomogén Cauchy-problémák* esetén a [10] dolgozat, az

$$y'+g(y)=f(t); \quad y(0)=y_0\in R^m \tag{1.4}$$

alakú nemlineáris problémák és lineáris $k$-lépéses módszerek esetén NEVANLINNA [33], [34] dolgozata tanulmányozza. NEVANLINNA a $g$ függvényről felteszi, hogy folytonos, bijektív és monoton (azaz $\langle u-v,\ g(u)-g(v)\rangle\geqq 0,\ u, v\in R^m$). A jelen munkában a [10] dolgozat eredményeit terjesztjük ki a lineáris többlépéses multiderivatív, ill. a *Runge—Kutta-módszerek* osztályán. Az elért eredmények élesebbek, mint NEVANLINNA megfelelő eredményei. Ez részben a vizsgált problémaosztály linearitásából, részben pedig a vizsgálati módszerek különbözőségéből adódik.

## 2. Stabilitási tételek

Tekintsük mindazokat az (1.2) alakú közelítő módszereket, amelyeket a $\{t_n=nh|h>0,\ n=0, 1, \ldots\}\subset[0, +\infty)$ felosztáson az (1.1) problémára alkalmazva a

$$\sum_{i=0}^{k}\alpha_i(A, h)y_{n+i}=F(A, g, t_n, h) \quad (n=0, 1, \ldots;\ \alpha_i\in\mathbf{C}^{m\times m}) \tag{2.1}$$

differenciaegyenletet kapjuk, ahol az inhomogén tag kielégíti a

$$F(A, 0, t_n, h)\equiv 0 \quad (n\in\mathbf{N};\ F(A, g, \ldots)\colon R^2\to\mathbf{C}^m) \tag{2.2}$$

feltételt és $\det\alpha_k(A, h)\neq 0$. Az (1.1) típusú problémák és a (2.1) alakú diszkretizációs módszerek esetén a $t_{n+k}$ pontbeli képlethiba a következőképpen is felírható (lásd pl. [20]):

$$L(y(t_n);\ h)=\sum_{i=0}^{k}\alpha_i(A, h)y(t_{n+i})-F(A, g, t_n, h). \tag{2.3}$$

A (2.1) módszer $A[\alpha]$-stabilitása azzal ekvivalens, hogy minden $\sigma(A)\subset W(\alpha)$ és minden $h>0$ esetén a

$$\varrho(x, h) = \det\Big[\sum_{i=0}^{k} \alpha_i(A, h)x^i\Big] \tag{2.4}$$

polinom gyökei a nyílt $|z|<1$ egységkörbe esnek (lásd STETTER [40]).

Az $A[\alpha, g]$-stabilitás létezésére vonatkozóan igaz a

2.1. TÉTEL. Legyen a (2.1) módszer $A[\alpha]$-stabilis valamilyen $\alpha\in[0, \pi/2]$ esetén. Adott $g$ perturbációra a (2.1) módszer akkor és csak akkor $A[\alpha, g]$-stabilis, ha minden $h>0$ lépéshossz esetén teljesül, hogy

$$L(y(t_n); h) \to 0 \quad (n \to +\infty). \tag{2.5}$$

A tétel bizonyítása a következő lemmán ([10], [12]) és egy CHARTRES—STEPLEMAN-tól származó technikai fogáson alapul.

2.1. LEMMA. Tekintsük az

$$\sum_{i=0}^{k} A_i u_{n+i} = c_n \quad (n\in\mathbf{N},\ \det(A_k) \neq 0) \tag{2.6}$$

$k$-adrendű differencia egyenletet, ahol $A_i\in\mathbf{C}^{m\times m}$ konstans mátrix $(i=0, 1, ..., k)$ és $\{c_n\}_{n=0}^{\infty}\subset\mathbf{C}^m$. Ha a $p(x)=\det\big(\sum_{i=0}^{k} A_i x^i\big)$ polinom gyökei a $|z|<1$ nyílt egységkörben vannak és $c_n\to c$, akkor fennáll, hogy

$$u_n \to \Big(\sum_{i=0}^{k} A_i\Big)^{-1} c \quad (n \to +\infty). \tag{2.7}$$

*Bizonyítás.* Legyen $U_n=(u_{n+k-1}, ..., u_n)^T$ és $S_n=(A_k^{-1}c_n, 0, ..., 0)^T$. Írjuk át (2.6)-ot az

$$U_{n+1} = QU_n + S_n \quad (n = 0, 1, ...) \tag{2.8}$$

alakba, ahol $Q=(\delta_{i-1,j}E_m - A_k^{-1}A_{k-j}\delta_{i1})_{i,j=1}^{k}$ és $E_m\in\mathbf{C}^{m\times m}$ egységmátrix. Minthogy

$$\det(Q - E_{mk}x) = \det(-A_k^{-1}) \det\Big(\sum_{i=0}^{k} A_i x^i\Big),$$

a $Q$ mátrix sajátértékei a $|z|<1$ nyílt egységkörbe esnek. Ezért van olyan $Q\to D=P^{-1}QP$ hasonlósági transzformáció, amelyre $\|D\|_\infty<1$. A $Z_n=P^{-1}U_n$, $W_n=P^{-1}S_n$ jelölésekkel (2.8) átmegy a $Z_{n+1}=DZ_n+W_n$ alakba, ahol $W_n\to W_\infty := P^{-1}S_\infty$ és $S_\infty=(A_k^{-1}c, 0, ..., 0)^T$. Használjuk a $\|\cdot\|_\infty$ normát és bontsuk fel $Z_n$-et $Z_n=Z_n^*+Z_n^{**}$ alakban, ahol

$$Z_{n+1}^*=DZ_n^*+W_\infty,\ Z_{n+1}^{**}=DZ_n^{**}+(W_n-W_\infty) \quad (n=0, 1, ...). \tag{2.9}$$

A $\|D\|_\infty<1$ feltételből $Z_n^* \to (E_{mk}-D)^{-1}W_\infty$ következik. Feltéve, $\|W_n-W_\infty\| \leqq K$ $(n \geqq 0)$ és $\|W_n-W_\infty\|<\varepsilon$ $(n \geqq n_0=n_0(\varepsilon))$ kapjuk, hogy

$$\|Z_{n+1}^{**}\| \leqq \|D\|^{n+1}\|Z_0^{**}\| + \sum_{i=0}^{n} \|D\|^{n-i}\|W_i-W_\infty\| \leqq$$

$$\leqq \|D\|^{n+1}\|Z_0^{**}\| + K \sum_{i=0}^{n_0} \|D\|^{n-i} + \varepsilon \sum_{i=n_0+1}^{n} \|D\|^{n-i} \leqq$$

$$\leqq \|D\|^{n+1}\|Z_0^{**}\| + (K\|D\|^{n-n_0}+\varepsilon)/(1-\|D\|),$$

amiből $Z_n^{**} \to 0$ következik. Emiatt $Z_n \to (E_{mk}-D)^{-1}W_\infty$, illetve $U_n \to (E_{mk}-Q)^{-1}S_\infty$ teljesül, ahonnan viszont az

$$(E_{mk}-Q)\left[\left(\sum_{i=0}^{k} A_i\right)^{-1}c, \ldots, \left(\sum_{i=0}^{k} A_i\right)^{-1}c\right]^T = S_\infty$$

összefüggés miatt a bizonyítandó

$$U_n \to \left[\left(\sum_{i=0}^{k} A_i\right)^{-1}c, \ldots, \left(\sum_{i=0}^{k} A_i\right)^{-1}c\right]^T \quad (n \to +\infty)$$

állítás már következik.

*A 2.1. tétel bizonyítása.* A (2.1) összefüggésből a (2.3) egyenletet kivonva az $e_n=y_n-y(t_n)$ globális hibára a

$$\sum_{i=0}^{k} \alpha_i(A, h)e_{n+i} = -L(y(t_n); h) \quad (n = 0, 1, \ldots) \tag{2.10}$$

rekurziót kapjuk. A módszer $A[\alpha]$-stabilitása miatt a (2.10) rekurzió kielégíti a 2.1. lemma feltételeit. Tehát $L(y(t_n); h) \to 0$ esetén $e_n \to 0$ következik. Fordítva az állítás triviális.

A következőkben látni fogjuk, hogy a tétel által megkövetelt (2.5) tulajdonság adott módszerosztályok esetén elsősorban probléma (perturbáció) függő, míg az $A[\alpha]$-stabilitás módszerfüggő. Ez a tény egyúttal jellemzi is az $A[\alpha]$-stabilitás hatásmechanizmusát.

2.2. TÉTEL. Ha az $L(y(t_n); h)$ képlethiba operátor lineáris az $y$ függvényben és a módszer $A[\alpha]$-stabilis, akkor az $A[\alpha, g_1]$- és $A[\alpha, g_2]$-stabilitás fennállásából az $A[\alpha, \mu g_1+\nu g_2]$-stabilitás $(\mu, \nu \in \mathbf{C})$ fennállása következik.

*Bizonyítás.* Jelölje $y(t; y_0, g)$ az (1.1) probléma megoldását. Tegyük fel, hogy $\mu, \nu \neq 0$. Ekkor viszont fennáll, hogy

$$y(t; y_0, \mu g_1+\nu g_2) = \mu y\left(t; \frac{c_1}{\mu} y_0, g_1\right) + \nu y\left(t; \frac{c_2}{\nu} y_0, g_2\right) \quad (c_1+c_2=1)$$

és a képlethiba operátor linearitása miatt a (2.5) feltétel teljesül. Ha $\mu, \nu$ valamelyike zérus, akkor a fenti előállításból a neki megfelelő tagot elhagyva kapjuk meg az állítást. Ha mindkét paraméter zérus, akkor az állítás triviális.

A bizonyításra kerülő stabilitási tételek közös alapgondolata a képlethiba operátor *Peano-formában* történő felírása, azaz a

$$L(y(t_n);\, h) = \int_0^{kh} R_r(h, s)\, y^{(r)}(t_n+s)\, ds \quad (r\in I,\ I\subset \mathbf{N}) \tag{2.11}$$

előállítás, ahol $I\neq\emptyset$ és $\|R_r(h, s)\|\leqq K_r(A, h)$ $(0\leqq s\leqq kh,\ r\in I)$.

2.3. TÉTEL. Tegyük fel, hogy a (2.1) módszer $A[\alpha]$-stabilis valamilyen $\alpha\in[0, \pi/2]$ értékre és képlethibája előáll a (2.11) alakban. Ekkor a módszer $A[\alpha, g]$-stabilis minden olyan $g$ perturbációra, amelyre van olyan $r\in I$, hogy az (1.1) probléma megoldása kielégíti a

$$\int_0^{\omega} \|y^{(r)}(t+s)\|\, ds \to 0 \quad (t\to+\infty,\ \omega>0) \tag{2.12}$$

feltételt. Ha van olyan $r\in I$, hogy

$$\int_0^{kh} R_r(h, s)\, ds = 0 \quad (h>0), \tag{2.13}$$

akkor a módszer $A[\alpha, g]$-stabilis minden olyan $g$-re, amelyre

$$\exists d\in\mathbf{C}^m:\ y^{(r)}(t)\to d \quad (t\to+\infty). \tag{2.14}$$

*Bizonyítás.* A (2.11) előállítás miatt a képlethibára a

$$\|L(y(t_n);\, h)\| \leqq K_r(A, h) \int_0^{kh} \|y^{(r)}(t_n+s)\|\, ds$$

egyenlőtlenség teljesül, ahonnan a (2.12) feltétel esetén az állítás következik, ui. $L(y(t_n);\, h)\to 0$. Ha van $r\in I$ és $d\in\mathbf{C}^m$, hogy $y^{(r)}(t)\to d$, akkor (2.11) alapján

$$L(y(t_n);\, h) \to \int_0^{kh} R_r(h, s)\, dds \quad (t\to+\infty),$$

ahonnan (2.13) miatt az állítás második fele következik.

A (2.12) feltétel (2.14) következménye, ha $d=0$. A (2.14) feltétel akkor és csak akkor teljesül, ha a $z'=Az+g^{(r-1)}(t)$ differenciálegyenlet megoldásaira $z'(t)\to d$ $(t\to+\infty)$ fennáll. Ez pedig akkor és csak akkor teljesül, ha vagy $g^{(r-1)}(t)\to c$ $(t\to+\infty,\ c\in\mathbf{C}^m$ tetszőleges, $d=0)$, vagy $g^{(r)}(t)\to -Ad$ $(t\to+\infty)$, vagy $g^{(r-1)}(t)=\varphi'(t)-A\varphi(t)$, ahol $\varphi'(t)\to d$ $(t\to+\infty)$.

Az $y^{(r)}(t)\to d$ feltétel $r$-ben lokális jellegű, amennyiben minden $r\geqq 0$ és $d\in\mathbf{C}^m$ esetén létezik $y\colon R^+\to\mathbf{C}^m$, amelyre $y^{(r)}(t)\to d$ $(t\to+\infty)$, de tetszőleges $i\neq r$ esetén nem létezik olyan $c\in\mathbf{C}^m$, hogy $y^{(i)}(t)\to c$ $(t\to+\infty)$.

A [10], [34] dolgozatokban az $y'(t)\to 0$, illetve speciális esetben ([10]) az $y'(t)\to d$ feltétel szerepelnek.

Megjegyezzük, hogy $g^{(r)}(t)\to d$ esetén $y^{(r)}(t)\to -A^{-1}d$ és $y^{(r+1)}(t)\to 0$ $(t\to+\infty,\ r\geqq 0)$ egyidejűleg fennállnak. Általában pedig RUDIN [39] (124. o. 16. fel-

adat) általánosításaként kimondható, hogy $y^{(r+i)}(t)\to 0$ $(t\to+\infty,\ i=1,\ldots,q-1)$, ha $y^{(r)}(t)\to d$ és $y^{(r+q)}(t)$ korlátos $R^+$-on, vagy ha $y^{(r)}(t)$ korlátos $R^+$-on és $y^{(r+q)}(t)\to 0$ $(t\to+\infty,\ q\geqq 2)$. Ha $y^{(r)}(t)\to d$, akkor $y(t)$ legfeljebb $O(t^r)$ nagyságrendben nőhet, míg $y^{(r+i)}(t)$ tetszőleges nagyságrendű lehet, ha $i\geqq 1$ és $y^{(r+1)}(t)\to 0$.

A fentiekben vázoltakhoz hasonló eredmények mondhatók ki $y^{(r)}$ korlátosságára is, amelyeknek az $S(\alpha)$-stabilitás ([38], [46]) vizsgálatában lehet szerepe.

Tekintsük most a lineáris többlépéses multiderivatív módszereket, amelyek alakja az

$$y'=f(t,y);\quad y(0)=y_0\in\mathbf{C}^m \tag{2.15}$$

alakú differenciálegyenletek esetén

$$\sum_{i=0}^{k}\sum_{j=0}^{l_i} h^j a_{ij} y_{n+i}^{(j)}=0\quad (n=0,1,\ldots), \tag{2.16}$$

ahol $a_{ij}\in R$, $a_{k0}\neq 0$, $l_i\geqq 0$, $k>0$, $\sum_{i=0}^{k} a_{i0}=0$, $l=\max_i l_i>0$, $y_n^{(0)}=y_n$ és $y_n^{(j)}=f^{(j-1)}(t_n,y_n)$, ahol

$$f^{(j)}(t,y)=\frac{\partial f^{(j-1)}(t,y)}{\partial t}+\frac{\partial f^{(j-1)}(t,y)}{\partial y}f(t,y)\quad (j\geqq 1). \tag{2.17}$$

A (2.16) módszerosztály speciális esetként tartalmazza a *lineáris k-lépéses módszereket* ($l_i=1$, $i=0,1,\ldots,k$) és a *Taylor-sor módszereket* ($k=1$). A multiderivatív módszerek $A[\alpha]$-stabilitását ($l>1$) ENRIGHT [8], GENIN [15], HAIRER—WANNER [19], JELTSCH [21—26], KOBZA [28] és mások tanulmányozták.

Ha a (2.16) módszert az (1.1) problémára alkalmazzuk, akkor olyan (2.1) alakú rekurziót kapunk, amelyre a 2.1 tétel alkalmazható. Elégséges feltételt mond ki a

2.4. TÉTEL. Ha a (2.16) multiderivatív módszer $p$-ed rendű ($p>l$) és $A[\alpha]$-stabilis valamely $\alpha\in[0,\pi/2]$ esetén, akkor $A[\alpha,g]$-stabilis minden olyan $g\colon R^+\to\mathbf{C}^m$ perturbációra, amelyre az (1.1) egyenlet megoldása kielégíti a

$$\exists r\,\exists d_r\colon\ l\leqq r<p+1,\quad d_r\in\mathbf{C}^m,\quad y^{(r)}(t)\to d_r\quad (t\to+\infty), \tag{2.18}$$

vagy a

$$\exists r\colon\ l<r\leqq p+1,\quad \int_0^{\omega}\|y^{(r)}(t+s)\|\,ds\to 0\quad (t\to+\infty,\ \omega>0) \tag{2.19}$$

feltételek valamelyikét.

*Bizonyítás.* Az

$$y(t+h)=\sum_{i=0}^{k-1}\frac{y^{(i)}(t)}{i!}h^i+\frac{1}{(k-1)!}\int_0^h (h-s)^{k-1}y^{(k)}(t+s)\,ds \tag{2.20}$$

előállítás alapján könnyen belátható, hogy tetszőleges (2.15) alakú, elég sima megoldású probléma esetén a (2.16) módszer képlethibája előáll a (2.11) formában. Pontosabban $l<r\leqq p+1$ esetén

$$L(y(t_n);h)=\int_0^{kh}\Big(\sum_{i=0}^{k}\sum_{j=0}^{l_i}\frac{a_{ij}h^j}{(r-j-1)!}(ih-s)_+^{r-j-1}\Big)y^{(r)}(t_n+s)\,ds, \tag{2.21}$$

ahol

$$x_+^j = \begin{cases} x^j, & \text{ha } x \geqq 0 \\ 0, & \text{ha } x < 0. \end{cases}$$

Az $R_r(h, s)$ magfüggvények az $l<r<p+1$ indexekre kielégítik a (2.13) anullálási feltételt. Tehát a tétel az $y^{(l)}(t)\to d$ eset kivételével a 2.3. tétel következménye. Az $r=l$ esetben

$$L(y(t_n);\, h) = \int_0^{kh} \Big( \sum_{i=0}^{k} \sum_{j=0}^{\min\{l_i, l-1\}} \frac{a_{ij} h^j}{(l-j-1)!} (ih-s)_+^{l-j-1} \Big) y^{(l)}(t_n+s)\, ds + \\ + \sum_{i:\, l_i = l} a_{il} h^l y^{(l)}(t_{n+i}),$$

ahonnan $y^{(l)}(t)\to d$ esetén

$$L(y(t_n);\, h) \to \sum_{i=0}^{k} \sum_{j=0}^{l_i} \frac{a_{ij} i^{l-j}}{(l-j)!} dh^l \quad (n \to +\infty)$$

következik. Ez a határérték a módszer $p$-ed rendűsége ($p>l$) miatt zérussal egyenlő. Tehát az állítást igazoltuk.

A lineáris multiderivatív módszerek esetén a $K_r(A, h)$ korlát $A$-tól független. A lineáris $k$-lépéses módszerek ($l=1$) esetére pontos becslések találhatók JELTSCH és NEVANLINNA [27] munkájában.

Az *s-paraméteres Runge—Kutta módszereket* a (2.15) alakú problémák esetén az

$$y_{n+1} - y_n = h \sum_{i=1}^{s} c_i k_i \tag{2.22}$$

$$k_i = f\Big(t_n + a_i h,\, y_n + h \sum_{j=1}^{s} b_{ij} k_j\Big) \quad (i = 1, \ldots, s)$$

összefüggések definiálják, ahol $a_i, c_i, b_{ij} \in R$ $(i, j = 1, \ldots, s)$ és

$$\sum_{i=1}^{s} c_i = 1, \quad \sum_{j=1}^{s} b_{ij} = a_i \quad (i = 1, \ldots, s). \tag{2.23}$$

A *Runge—Kutta-módszerek* esetén a képlethiba a módszer nemlinearitása miatt nem mindig állítható elő a (2.11) alakban. Az előállíthatóság, amint azt látni fogjuk, nem a módszer rendjétől függ.

Legyen a továbbiakban $B=(b_{ij})_{i,j=1}^{s}$, $c=(c_1, \ldots, c_s)^T \in R^s$, $e=(1, \ldots, 1)^T \in R^s$ és $V=\operatorname{diag}(a_1, \ldots, a_s)$.

Tetszőleges $P \in \mathbf{C}^{k\times l}$, $Q \in \mathbf{C}^{r\times s}$ mátrixok *Kronecker szorzatát* a

$$P \otimes Q := (q_{ij} P)_{i,j=1}^{r,s} \tag{2.24}$$

definíció szerint használjuk.

A továbbiakban tegyük fel, hogy a (2.22) *Runge—Kutta-módszer* $p$-ed rendű ($p \geqq 1$), $A[\alpha]$-stabilis és alkalmazzuk ezt a módszert az (1.1) problémára. Ekkor az

$$y_{n+1} - \alpha(A, h) y_n = h(E_{ms} - hA \otimes B)^{-1} G_n \quad (n = 0, 1, \ldots) \tag{2.25}$$

rekurziót kapjuk, ahol $\alpha(A,h)=E_m+h(E_m\otimes c)^T(E_{ms}-hA\otimes B)^{-1}(A\otimes e)$, $G_n=(g(t_n+a_1h),\ldots,g(t_n+a_sh))^T$. Az $A[\alpha]$-stabilitás miatt $\alpha(A,h)$ sajátértékei a $|z|<1$ nyílt egységkörbe esnek (STETTER [40]) és így a 2.1. tétel alkalmazható.

Az (1.1) *Cauchy-probléma* esetén a *Runge—Kutta-módszer* képlethibája

$$L(y(t_n);\,h)=h(E_m\otimes c)^T(hA\otimes B-E_{ms})^{-1}[G_n+(A\otimes e)y(t_n)]+ \\ +[y(t_{n+1})-y(t_n)]. \tag{2.26}$$

Igaz a következő

2.2. LEMMA. Ha $y^{(k)}$ integrálható, akkor $k\geqq 2$ esetén (2.26) előáll a

$$L(y(t_n);\,h)=\sum_{i=2}^{k-1}\frac{h^i}{i!}\varphi_i(hA)y^{(i)}(t_n)+\int_0^h R_k(h,\tau)y^{(k)}(t_n+\tau)\,d\tau \tag{2.27}$$

alakban, ahol

$$\varphi_i(hA)=E_m+(E_m\otimes c)^T(hA\otimes B-E_{ms})^{-1}(iE_m\otimes V^{i-1}-hA\otimes V^i)(E_m\otimes e), \tag{2.28}$$

$$R_k(h,\tau)=\frac{1}{(k-1)!}(E_m\otimes c)^T(hA\otimes B-E_{ms})^{-1}[(k-1)hE_m\otimes D(h,\tau)^{k-2}- \\ -hA\otimes D(h,\tau)^{k-1}+(h-\tau)^{k-1}(hA\otimes V-E_{ms})](E_m\otimes e)$$

és $D(h,\tau)=\operatorname{diag}\left((ha_1-\tau)_+,\ldots,(ha_s-\tau)_+\right)$.

*Bizonyítás.* A (2.20) előállítás és a $g(t)=y'(t)-Ay(t)$ egyenlőség alapján fejtsük $h$ szerint sorba a (2.26) előállításban szereplő szögletes zárójelpárok közti mennyiségeket olymódon, hogy a *Peano-féle maradéktagban* mindig $y^{(k)}$ szerepeljen. Az így kapott előállítást az $y^{(i)}(t_n)$ $(i=1,\ldots,k)$ deriváltak szerint rendezve adódik a (2.27)—(2.28) összefüggés.

Megjegyezzük, hogy $\varphi_1(hA)\equiv 0$ és

$$\int_0^h R_i(h,\tau)\,d\tau=\frac{h^i}{i!}\varphi_i(hA)\quad(i=2,3,\ldots). \tag{2.29}$$

Szükségünk van a következő lemmára is.

2.3. LEMMA. Tetszőleges $2\leqq k\leqq p+1$ esetén a *Runge—Kutta módszer* képlethibájának akkor és csak akkor létezik

$$L(y(t_n);\,h)=\int_0^h R_k(h,\tau)y^{(k)}(t_n+\tau)\,d\tau \tag{2.30}$$

alakú előállítása az (1.1) problémák osztályán $(\sigma(A)\subset W(\alpha))$, ha

$$\varphi_i(hA)=0\quad(h>0,\ \sigma(A)\subset W(\alpha);\ i=1,\ldots,k-1). \tag{2.31}$$

*Bizonyítás.* Ha $\varphi_i(hA)=0$ $(i=1,\ldots,k-1)$ fennáll, akkor (2.27) miatt készen vagyunk. Fordítva, tegyük fel, hogy $k>2$, ugyanis $k=2$ esetén az állítás triviális $\varphi_1(hA)=0$ miatt. Tegyük fel, hogy létezik $h>0$ lépéshossz és $A$ mátrix $(\sigma(A)\subset W(\alpha))$, amelyre $\varphi_2(hA)\neq 0$. Ekkor van olyan $d\in\mathbf{C}^m$, hogy $\varphi_2(hA)d\neq 0$ és létezik olyan $\hat{g}(t)=\sum_{i=0}^{2}a_it^i\in\mathbf{C}^m$ polinomvektor és $y_0\in\mathbf{C}^m$ kezdetiérték, amelyre

$\hat{y}(t; y_0, \hat{g})$ szintén másodfokú polinomvektor és $\hat{y}^{(2)}(t) \equiv d$. Tehát a (2.27) előállítás szerint $L(\hat{y}(t_n); h) = \frac{h^2}{2!}\varphi_2(hA)d \neq 0$, míg a feltételezett (2.30) előállítás szerint $L(\hat{y}(t_n); h) = 0$, ami ellentmondás.

Tegyük fel, hogy a $\varphi_q(hA) = 0$ $(q = 2, \ldots, i-1;\ i \leqq k)$ feltételt már igazoltuk. Ha létezik $h > 0$ lépéshossz és $A$ mátrix $(\sigma(A) \subset W(\alpha))$, amelyre $\varphi_i(hA) \neq 0$, akkor létezik $d \in \mathbf{C}^m$, hogy $\varphi_i(hA)d \neq 0$. Továbbá van olyan $\tilde{g}(t) = \sum_{j=0}^{i} a_j t^j \in \mathbf{C}^m$ polinomvektor és $y_0 \in \mathbf{C}^m$ kezdetiérték, amelyre $\tilde{y}(t; y_0, \tilde{g})$ $i$-edfokú polinomvektor és $\tilde{y}^{(i)}(t) \equiv d$. Ekkor ismét (2.27) alapján $L(\tilde{y}(t_n); h) = \frac{h^i}{i!}\varphi_i(hA)d \neq 0$, míg (2.30) alapján $L(\tilde{y}(t_n); h) = 0$, ami ellentmondás.

A $k \leqq p+1$ feltétel azért teljesül, mert $k > p+1$ esetén a képlethiba rendje $O(h^{p+2})$ lenne, ami ellentmondás.

*Megjegyzés.* A $\varphi_1(hA) = 0$ miatt a (2.30) reprezentáció a $k = 2$ értékre mindig létezik.

A lemma állítása lényegében azt jelenti, hogy a (2.30) típusú hibaelőállítás akkor és csak akkor létezik, ha a legfeljebb $(k-1)$-edfokú polinom perturbációkkal rendelkező inhomogén differenciálegyenletek polinom megoldásaira $L(y(t_n); h) \equiv 0$ teljesül. Ez csak bizonyos értelemben analogonja annak a ténynek, hogy a lineáris többlépéses (multiderivatív) módszerek esetén a (2.30), ill. a (2.21) hibareprezentáció létezéséhez szükséges és elégséges a képlethiba eltűnése az $y(t) = t^i$ $(i = 0, 1, \ldots, k-1)$ függvényeken. Ez a feltétel a *Runge—Kutta-módszerek* esetén nem elégséges.

A lemma feltételei mellett, (2.29) alapján teljesül a

$$\int_0^h R_i(h, \tau)\, d\tau = 0 \quad (2 \leqq i \leqq k-1) \tag{2.32}$$

anullálási feltétel.

Tekintsük a BUTCHER ([4]) által bevezetett

$$B(\xi): \quad \sum_{i=1}^{s} c_i a_i^{k-1} = 1/k \quad (1 \leqq k \leqq \xi) \tag{2.33}$$

$$C(\xi): \quad \sum_{j=1}^{s} b_{ij} a_j^{k-1} = a_i^k/k \quad (1 \leqq i \leqq s,\ 1 \leqq k \leqq \xi)$$

feltételeket. A (2.23) feltétel ebben a megfogalmazásban a $B(1)$ és $C(1)$ feltételeknek felel meg.

Igaz a

2.4. LEMMA. Ha $B(q)$ és $C(q)$ $(q \geqq 1)$ egyidejűleg fennáll, akkor

$$\varphi_i(hA) = 0 \quad (i = 1, \ldots, q). \tag{2.34}$$

*Bizonyítás.* A (2.33) definíciók alapján kapjuk, hogy

$$\varphi_i(hA) = E_m + (E_m \otimes c)^T (hA \otimes B - E_{ms})^{-1}[iE_m \otimes V^{i-1} - ihA \otimes (BV^{i-1})](E_m \otimes e) =$$
$$= E_m + i(E_m \otimes c)^T (hA \otimes B - E_{ms})^{-1}[E_m \otimes V^{i-1} -$$
$$- h(A \otimes B)(E_m \otimes V^{i-1})](E_m \otimes e) = (1 - ic^T V^{i-1} e) E_m = 0.$$

BUTCHER [4] eredményei alapján könnyen látható, hogy mindazokra az $s$-paraméteres *Runge—Kutta módszerekre,* amelyek legalább $2s-2$ rendűek, a $C(s-2)$ és $B(s-2)$ szükségképpen teljesülnek (vö. FUCHS [9], BURRAGE [3]). BURRAGE [3] módszerosztályaira $C(s-1)$ és $B(s+t)$ $(t=0, 1, ..., s)$ teljesül. HAIRER igazolta ([18]), hogy $p\geqq 3$ és $c_i>0$ $(i=1, ..., s)$ esetén $C(k)$ teljesül a $k=1, 2, ..., [(p-1)/2]$ értékekre.

A *diagonálisan implicit Runge—Kutta módszerekre* (ALEXANDER [1]) $C(2)$ általában nem teljesül. Elemi számolásokkal igazolható, hogy ALEXANDER ([1]) $p=2, 3, 4$ rendű módszereire $\varphi_2(z)\not\equiv 0$ $(z=\lambda h,\ h>0,\ \lambda\in\mathbf{C})$ áll fenn. Ugyancsak könnyen kimutatható, hogy a $\mathrm{II}_A$ módszerosztály $s=2, p=3$ paraméterű tagjára $\varphi_3(z)\not\equiv 0$ és nem teljesül $C(3)$. A $\mathrm{III}_C$ osztály $s=2, p=2$ paraméterű tagjára $\varphi_2(z)\not\equiv 0$ és nem teljesül $C(2)$, az $s=3, p=4$ paraméterű tagjára pedig $\varphi_3(z)\not\equiv 0$ és nem teljesül $C(3)$. A $G$ osztály $s=2, p=4$ paraméterű tagjára $\varphi_3(z)\not\equiv 0$ és nem teljesül $C(3)$.

A fentiek alapján kézenfekvő az a sejtés, hogy a (2.30) előállítás akkor és csak akkor létezik, ha $B(k-1)$ és $C(k-1)$ teljesül.

Eredményeink összegzéseként a következő elégséges tétel mondható ki.

2.5. TÉTEL. Tegyük fel, hogy a (2.22) *Runge—Kutta módszer* $A[\alpha]$-stabilis és legyen $k^*\geqq 2$ az a maximális $k$, amelyre (2.30) fennáll. A módszer minden olyan $g$ esetén $A[\alpha, g]$-stabilis, amelyre az (1.1) egyenlet megoldása kielégíti a

$$\exists r \exists d_r\colon\ 1\leqq r<k^*,\quad d_r\in\mathbf{C}^m,\quad y^{(r)}(t)\to d_r\quad (t\to+\infty), \tag{2.35}$$

vagy a

$$\exists r\colon\ 2\leqq r\leqq k^*,\quad \int_0^{\omega}\|y^{(r)}(t+s)\|\,ds\to 0\quad (t\to+\infty,\ \omega>0) \tag{2.36}$$

feltételek valamelyikét.

*Bizonyítás.* A (2.30), (2.32) relációk és a 2.2. lemma alapján az állítás az $y'(t)\to d_1\in\mathbf{C}^m$ eset kivételével a 2.3. tétel következménye. Az $y'(t)\to d_1$ $(t\to+\infty)$ esetben (2.20) és (2.26) felhasználásával kapjuk, hogy

$$G_n+(A\otimes e)y(t_n)=\left\{y'(t_n+a_ih)-A\int_{t_n}^{t_n+a_ih}y'(\tau)\,d\tau\right\}_{i=1}^{s}\to$$

$$\to\{d_1-hAa_id_1\}_{i=1}^{s}=(E_{ms}-hA\otimes B)(E_m\otimes e)d_1\quad (n\to+\infty),$$

illetve

$$y(t_{n+1})-y(t_n)=\int_{t_n}^{t_{n+1}}y'(\tau)\,d\tau\to d_1h\quad (n\to+\infty).$$

Emiatt

$$L(y(t_n);h)\to -hd_1+hd_1=0\quad (n\to+\infty),$$

amiből az állítás már következik.

A szakasz stabilitási tételei a 2.1. tétel kivételével csak elégségesek. A 2.2. tételben megfogalmazott linearitási tulajdonság miatt az $A[\alpha, g]$-stabilitás fennáll az előző tételekben vizsgált perturbációk tetszőleges véges lineáris kombinációjára is. Ezért és a képlethiba operátor nem egyértelmű előállíthatósága, valamint szerkezete

miatt az összes $g$ megadása, amelyre az $A[\alpha, g]$-stabilitás fennáll, nem várható. Tehát az $A[\alpha]$-stabilitás hatásmechanizmusának pontosabb jellemzéséhez a $g$ perturbációk hatását instabilitási tételeken keresztül is vizsgálnunk kell. Ezt a következő szakaszban végezzük el.

Végül megjegyezzük, hogy a nemlineáris (1.4) differenciálegyenletek és a lineáris $k$-lépéses módszerek esetén NEVANLINNA globális hibabecslési tételéből ([33], [34]) a globális hiba végtelenben vett zéruskonvergenciája („$A[\alpha, g]$-stabilitása") következik, ha $L(y(t_n); h) \to 0$ $(n \to +\infty)$, valamint a $0 < \theta < 1$ feltétel teljesül. E tény bizonyítása a 2.1. lemma bizonyításában látottak szerint történhet.

## 3. Instabilitási tételek

A szakaszban az $A[\alpha, g]$-stabilitás nemlétezésére vonatkozóan a megoldás növekedésétől, illetve periodikus jellegű viselkedésétől függő tételeket igazolunk.

Az első típusú nemegzisztencia tételek közös alapja a képlethiba

$$L(y(t_n); h) = \sum_{i=q}^{j-1} \psi_i(A, h) y^{(i)}(t_n) + \int_0^{kh} R_j(h, \tau) y^{(j)}(t_n + \tau)\, d\tau \tag{3.1}$$

alakban történő előállíthatósága a (2.1) alakú módszerekre. Feltételezzük, hogy $\|R_j(h, s)\| \leqq K_j(A, h)$ $(0 \leqq s \leqq kh)$ és $j$ a $g$ függvénytől függő tetszőleges természetes szám.

A lineáris többlépéses multiderivatív módszerekre

$$\psi_i(A, h) = \frac{h^i}{i!} L(t_+^i; 1)|_{t=0} \quad (i \geqq l+1), \tag{3.2}$$

a *Runge—Kutta módszerekre* pedig a 2.2. lemma szerint

$$\psi_i(A, h) = \frac{h^i}{i!} \varphi_i(hA) \quad (i \geqq 2). \tag{3.3}$$

Legyen a (3.1) előállításban $q$ értéke

$$q = \min \{i | \psi_i(A, h) \not\equiv 0\}. \tag{3.4}$$

A lineáris multiderivatív módszerek esetén $q = p+1$, a *Runge—Kutta módszerek* esetén pedig sejtés, hogy $q = \max \{i | C(i) \text{ igaz}\} + 1$.

Igaz a következő

3.1. TÉTEL. Tegyük fel, hogy a (2.1) módszernek létezik a (3.1), (3.4) feltételeket kielégítő képlethiba reprezentációja. Ha a módszer $A[\alpha]$-stabilis, akkor nem igaz az, hogy minden a

$$\exists d \in \mathbf{C}^m: \; y^{(j)}(t) \to d \quad (t \to +\infty) \tag{3.5}$$

feltételt $(j \geqq q)$ kielégítő $g$ perturbáció esetén $A[\alpha, g]$-stabilis.

*Bizonyítás.* Legyen $h>0$, $A$ és $d\neq 0$ olyan, hogy $\int_0^{kh} R_q(h,\tau)dd\tau\neq 0$. Ekkor létezik olyan $\hat{g}$ $q$-adfokú polinomvektor és $y_0\in\mathbf{C}^m$, hogy $\hat{y}(t;\, y_0, \hat{g})$ szintén $q$-adfokú polinomvektor és $\hat{y}^{(q)}(t)\equiv d$. Ha feltételezzük, hogy erre a $\hat{g}$ perturbációra a módszer $A[\alpha, \hat{g}]$-stabilis, akkor a 2.1. tétel alapján $L(\hat{y}(t_n);\, h)\to 0$. Másrészt a (3.1) előállítás alapján

$$L(\hat{y}(t_n);\, h)=\int_0^{kh} R_q(h,\tau)\hat{y}^{(q)}(t_n+\tau)\,d\tau \to \int_0^{kh} R_q(h,\tau)\,dd\tau\neq 0,$$

ami ellentmondás. Tegyük most fel, hogy $j>q$. Tekintsünk ismét egy olyan $\hat{g}$ $j$-edfokú polinomvektort és $y_0\in\mathbf{C}^m$ kezdetiértéket, amelyre $\hat{y}(t;\, y_0, \hat{g})$ $j$-edfokú polinomvektor és $\hat{y}^{(j)}(t)\equiv d$ $(d\neq 0)$. Ha erre a $\hat{g}$ függvényre az $A[\alpha, \hat{g}]$-stabilitás fennállna, akkor $L(\hat{y}(t_n);\, h)\to 0$ kell, hogy teljesüljön. Másrészt a (3.1) előállítás alapján

$$L(\hat{y}(t_n);\, h)=\sum_{i=q}^{j-1}\psi_i(A,h)\hat{y}^{(i)}(t_n)+\int_0^{kh} R_j(h,\tau)\hat{y}^{(j)}(t_n+\tau)\,d\tau,$$

amelyre az

$$\hat{y}^{(q)}(t)=\frac{1}{(j-q)!}\,dt^{j-q}+O(t^{j-q-1})$$

reláció miatt a $\psi_q(A,h)d\neq 0$ feltételezése esetén

$$L(\hat{y}(t_n);\, h)=0(t_n^{j-q}),\quad \|L(\hat{y}(t_n);\, h)\|\to+\infty\quad (t_n\to+\infty)$$

teljesül. Ez ellentmondás, tehát tételünket igazoltuk.

A multiderivatív módszerek esetén (3.2) miatt elég $d\neq 0$-át feltételezni.

A „periodikus" eset vizsgálatához tegyük fel, hogy a vizsgált $\hat{g}$ perturbáció olyan, hogy az $y$ megoldás $q$-adik deriváltja integrálható, továbbá létezik $y_0\in\mathbf{C}^m$, hogy az $\hat{y}(t;\, y_0, \hat{g})$ partikuláris megoldás rendelkezik a következő tulajdonsággal:

(A) $\exists d\in\mathbf{C}^m\setminus\{0\},\quad x^*, T\in R^+,\quad 0\leqq x^*<T\colon\ \forall\varepsilon>0,\quad \exists\delta_0=\delta_0(\varepsilon)>0,$

$$\exists x_0=x_0(\varepsilon)>0\colon$$

$$\|\hat{y}^{(q)}(x+jT)-d\|<\varepsilon\quad (|x^*-x|<\delta_0,\ x^*+jT>x_0,\ j\in\mathbf{N}).$$

Ezenkívül tegyük fel, hogy létezik a

$$\lim_{h\to 0} h^{-r}\psi_q(A,h)\,d\neq 0\quad (r\geqq 0) \tag{3.6}$$

határérték és

$$\int_0^{kh} R_q(h,\tau)\,d\tau=\psi_q(A,h), \tag{3.7}$$

$$\|R_q(h,s)\|\leqq h^{r-1}K_q(A)\quad (0\leqq s\leqq kh,\ h\to 0).$$

Ekkor igaz a

3.2. TÉTEL. Ha a (2.1) módszer $A[\alpha]$-stabilis és kielégíti a (3.6), (3.7) feltételeket, akkor nem lehet $A[\alpha, \hat{g}]$-stabilis semmilyen az $(A)$ feltételt kielégítő $\hat{g}$ függvényre.

*Bizonyítás.* A tétel állításánál többet igazolunk, amennyiben megmutatjuk, hogy nincs olyan $h^*>0$ szám, amelyre minden $h\in(0, h^*)$ lépéshossz esetén teljesül $y_n-y(t_n)\to 0$ $(t_n\to+\infty)$.

Legyen $h=T/M$ és $M\in\mathbf{N}$ olyan nagy, hogy $2kh<\delta_0$ teljesüljön. Létezik olyan legkisebb $i, j$ index, hogy $x_0<t_i$, $x^*+jT\in[t_i, t_{i+1})$. Ekkor

$$\left\|L(\hat{y}(t_i);\, h)-\int_0^{kh} R_q(h,\tau)\,d\,d\tau\right\| \leqq \left\|\int_0^{kh} R_q(h,\tau)[\hat{y}^{(q)}(t_i+\tau)-d]\,d\tau\right\| \leqq$$

$$\leqq \varepsilon k h^r K_q(A),$$

ahonnan (3.6) miatt elég kis $h$ és $\varepsilon$ esetén

$$\left\|L(\hat{y}(t_i);\, h)\right\| \geqq \left\|\int_0^{kh} R_q(h,\tau)\,d\,d\tau\right\|-\varepsilon k h^r K_q(A) =$$

$$= h^r[h^{-r}\|\psi_q(A,h)\,d\|-\varepsilon k K_q(A)] \geqq \gamma(h) > 0$$

adódik. Az (A) feltevés miatt $\left\|L(\hat{y}(t_{i+kM});\, h)\right\|\geqq\gamma(h)>0$ $(k=0, 1, \ldots)$, ami ellentmond az $L(\hat{y}(t_n);\, h)\to 0$ feltételnek. Ezzel az állítást igazoltuk.

A tétel feltételeinek eleget tesz minden olyan $g$ perturbáció, amely periodikus, $q$-szor folytonosan differenciálható és a $q$-adik deriváltja nem zérus.

Vizsgáljuk most meg a (3.6)—(3.7) feltételek teljesülését a lineáris multiderivatív és a *Runge—Kutta módszerek* esetén.

A lineáris multiderivatív módszerekre (3.2) miatt a (3.6) feltétel minden $d\in\mathbf{C}^m$, $d\neq 0$ esetén teljesül $r=p+1$ értékkel. A (3.7) feltételek teljesülése szintén triviális. A *Runge—Kutta módszerek* esetén induljunk ki abból, hogy ha $\varphi_q(z)\not\equiv 0$ $(z\in W(\alpha))$, akkor előáll

$$\varphi_q(z) = \sum_{i=\rho}^{\infty} d_i z^i \quad (z\to 0)$$

alakban, ahol $d_\rho\neq 0$. Innen $\sigma(A)\subset W(\alpha)$ és $h\to 0$ esetén

$$\varphi_q(hA) = \sum_{i=\rho}^{\infty} d_i (hA)^i,$$

ahonnan a (3.3) feltétel és $A^{-1}$ létezése miatt a (3.7) feltétel, illetve tetszőleges $d\in\mathbf{C}^m\setminus\{0\}$ esetén a (3.6) feltétel következik.

BABUSKA—PRÁGER—VITÁSEK [2], valamint TYIHONOV—ARSZENYIN [42] alapján vizsgálhatjuk a következő problémát. Tegyük fel, hogy adott módszerre és adott $g$ perturbációra fennáll az $A[\alpha, g]$-stabilitás. Ha bevezetjük a $\|g\|_0=\sup_{t\geqq 0}\|g(t)\|$ normát, akkor kérdezhetjük, hogy az $A[\alpha, g]$-stabilitás fennállásából vajon következik-e az $A[\alpha, \hat{g}]$-stabilitás, ha $\|g-\hat{g}\|_0<\varepsilon$ teljesül, ahol $\varepsilon>0$ alkalmas szám. BABUSKA, PRÁGER és VITÁSEK erősen stabilis lineáris $k$-lépéses módszerekre és explicit egylépéses módszerekre igazolták, hogy a teljesen stabilis differenciálegyenletek kismértékű korlátos perturbációja esetén a hagyományos értelemben vett stabilitási korlát (erős stabilitás) az egész $R^+$-on fennáll. Hasonló állítás az $A[\alpha]$-stabilitás ($A[\alpha, 0]$-stabilitás) esetén nem igaz.

3.1. ÁLLÍTÁS. Minden a 3.2. tétel feltételeit kielégítő $A[\alpha]$-stabilis módszerre igaz az, hogy nincs olyan $\varepsilon>0$, hogy az $A[\alpha, g]$-stabilitásból következik az $A[\alpha, \hat{g}]$-stabilitás minden olyan $\hat{g}$ függvényre, amelyre teljesül $\|g-\hat{g}\|_0<\varepsilon$.

*Bizonyítás.* Legyen $\varepsilon>0$ tetszőleges és legyen $\hat{g}(t)=g(t)+\frac{\varepsilon}{2}e\sin(t)$. A linearitási tulajdonság (2.2. tétel) miatt az $A[\alpha, \hat{g}]$-stabilitás csak akkor teljesülhet, ha teljesül az $A\left[\alpha, \frac{\varepsilon}{2}e\sin(t)\right]$-stabilitás is. Ez pedig a 3.2. tétel miatt nem lehetséges.

A szakasz tételeinek több következménye is van.

A periodikus és majdnem periodikus perturbációk esetén nem érhető el önstabilizáló tulajdonság, csak a módszer „rendje" által nyújtott pontosság $(O(h^q))$. Ezt a viselkedést tapasztalati úton korábban már kimutatták és ilyen feladatok hatékony megoldására a (2.1) osztálytól eltérő jellegű módszereket ([14], [32]) dolgoztak ki.

A 3.1. állítás következménye, hogy az aszimptotikus stabilitás linearizálási elve nem teljesül az $A[\alpha]$-stabilis módszerekre a globális hiba vonatkozásában. Tehát az $y'=Ay$ alakú differenciálegyenletekre tapasztalt viselkedés nem várható el az általánosabb (1.1) problémák esetén sem. Ez azt is jelenti, hogy a „stiffségre" adott heurisztikus definíciók nem kielégítőek ([14], [11], [31], [32], [7]). A dolgozat eredményeiből jól látható, hogy a módszer viselkedése (önstabilizáló tulajdonsága, osztályra vonatkozó abszolút stabilitása, stb.) milyen mértékben és hogyan függ a módszertől, illetve a differenciálegyenlettől. Tehát a heurisztikus definíciók a módszerfüggetlenség feltételezése miatt ebből a szempontból sem kielégítőek.

Eredményeink élesebbek, mint az $S$-stabilitásra vonatkozó eredmények ([38], [46]). Az *S-stabilitás Verwer-féle gyengítése* ([46]) az $A[\alpha]$-stabilitás triviális következménye, ugyanis $y'$ korlátosságából azonnal következik $L(y(t_n); h)$, ill. $y_n-y(t_n)$ $(n=0, 1, \ldots)$ korlátossága.

Eredményeink negatív értelemben megválaszolják DAHLQUIST ([7]) egy problémáját, amennyiben sem az $A[\alpha]$-stabilitás, sem az $S$-stabilitás nem elégséges a lineáris stiff problémák és hatékony megoldásuk jellemzésére. Ezt KREISS ([29]) és VELDHUIZEN ([45]) eredményei is alátámasztják.

## IRODALOM


[1] ALEXANDER, R., "Diagonally implicit Runge—Kutta methods for stiff O. D. E.'s", *SIAM J. Number. Anal.* **14** (1977) 1006—1021.

[2] BABUSKA, I., PRÁGER, M. and VITÁSEK, E., *Numerical Processes in Differential Equations* (Wiley, London, 1966).

[3] BURRAGE, K., "A special family of Runge—Kutta methods for solving stiff differential equations", *BIT* **18** (1978) 22—41.

[4] BUTCHER, J. C., "Implicit Runge—Kutta processes", *Mathematics of Computation* **18** (1964) 50—64.

[5] BUTCHER, J. C., "A stability property of Runge—Kutta methods", *BIT* **15** (1975) 358—361.

[6] DAHLQUIST, G., "A special stability problem for linear multistep methods", *BIT* **3** (1963) 27—43.

[7] DAHLQUIST, G., "Recent work on stiff differential equations", Report TRITA-NA-7512, The Royal Institute of Technology, Stockholm, 1975.

[8] ENRIGHT, W. H., "Second derivative multistep methods for stiff ordinary differential equations", *SIAM J. Numer. Anal.* **11** (1974) 321—331.

[9] FUCHS, F., "*A*-stability of Runge—Kutta methods with single and multiple nodes", *Computing* **16** (1976) 39—48.

[10] GALÁNTAI, A., "New stability property concerning stiff methods", Colloquia Mathematicae Societatis János Bolyai, 22. Numerical Methods, Keszthely (Hungary) 1977, 203—212.

[11] GALÁNTAI, A., VICSEK, M., "On the concept of the stiffness", Conference „Numerische Behandlung von Anfangs- und 2-Punkt Randwertaufgaben bei gewoenlicher Differentialgleichungen", Berlin 2. und 3. Oktober, 1978. Sektion Mathematik, Humboldt Universität, Berlin, 1978, 96—109.

[12] GALÁNTAI, A., „Vizsgálatok a közönséges differenciálegyenletek közelítő módszereinek konvergencia és hibaanalízisének körében", kandidátusi értekezés, Budapest, 1978.

[13] GEAR, C. W., *Numerical Initial Value Problems in Ordinary Differential Equations* (Englewood Cliffs, Prentice-Hall, 1971).

[14] GEAR, C. W., "Numerical solution of ordinary differential equations: is there anything left to do?", SIAM REWIEW **23** (1981) 10—24.

[15] GENIN, Y., "An algebraic approach to $A$-stable linear multistep-multiderivative integration formulas", *BIT* **14** (1974) 382—406.

[16] GRIEPENTROG, E., „Numerische Integration steifer Differentialgleichungssysteme mit Einschrittverfahren", *Beiträge zur Num. Math.* **8** (1930) 59—74.

[17] GRIEPENTROG, E., „Zur numerischen Integration nichtlinearer Differentialgleichungen auf einem unendlichen Intervall", *Beiträge zur Num. Math.* **11** (1933) 21—31.

[18] HAIRER, E., "Highest possible order of algebraically stable diagonally implicit Runge—Kutta methods", *BIT* **20** (1930) 254—256.

[19] HAIRER, E. and WANNER, G., "Multistep-multistage-multiderivative methods for ordinary differential equations", *Computing* **11** (1973) 287—303.

[20] HALL, G. and WATT, J. M., *Modern Numerical Methods for Ordinary Differential Equations* (Clarendon Press, Oxford, 1976).

[21] JELTSCH, R., "Note on $A$-stability of multistep-multiderivative methods", *BIT* **16** (1976) 74—78.

[22] JELTSCH, R., "A necessary condition for $A$-stability of multistep-multiderivative methods", *Mathematics of Computation* **30** (1976) 739—746.

[23] JELTSCH, R., "Multistep methods using higher derivatives and damping at infinity", *Mathematics of Computation* **31** (1977) 124—138.

[24] JELTSCH, R. and KRATZ, L., "On the stability properties of Brown's multistep-multiderivative methods", *Numer. Math.* **30** (1978) 25—38.

[25] JELTSCH, R., "$A_0$-stability and stiff-stability of Brown's multistep-multiderivative methods", *Numer. Math.* **32** (1979) 167—181.

[26] JELTSCH, R. and NEVANLINNA, O., "Stability of explicit time discretizations for solving initial value problems", *Numer. Math.* **37** (1931) 61—91.

[27] JELTSCH, R. and NEVANLINNA, O., "Stability and accuracy of time discretizations for initial value problems", *Numer. Math.* **40** (1982) 245—296.

[28] KOBZA, J., "Stability of second derivative linear multistep formulas", *Acta Univ. Palackianae Olomucensis, Fac. Rer. Nat.* **T 53** (1977) 167—184.

[29] KREISS, H. O., "Difference methods for stiff ordinary differential equations", *SIAM J. Numer. Anal.* **15** (1978) 21—58.

[30] KREISS, H. O., "Problems with different time scales for ordinary differential equations", *SIAM J. Numer. Anal.* **16** (1979) 980—998.

[31] LAMBERT, J. D., *Computational Methods in Ordinary Differential Equations* (Wiley, London, 1973).

[32] MIRANKER, L. W., *The Computational Theory of Stiff Differential Equations*, Publication Mathematique d'Orsay, No. 219—7667, Université Paris, 1976.

[33] NEVANLINNA, O., "On the numerical integration of nonlinear initial value problems by linear multistep methods", *BIT* **17** (1977) 58—71.

[34] NEVANLINNA, O., "On the behaviour of global errors at infinity in the numerical integration of stable initial value problems", *Numer. Math.* **28** (1977) 445—454.

[35] NEVANLINNA, O. and LINIGER, W., "Contractive methods for stiff differential equations", Part I.: *BIT* **18** (1978) 457—474, Part II.: *BIT* **19** (1979) 53—72.

[36] NEVANLINNA, O. and ODEH, F., "Multiplier techniques for linear multistep methods", *Numer. Funct. Anal. and Optimiz.* **3** (1931) 377—423.

[37] ODEH, F. and LINIGER, W., "Nonlinear fixed-$h$ stability of linear multistep formulas", *Journ. of Math. Anal. and Appl.* **61** (1977) 691—712.

[38] PROTHERO, A. and ROBINSON, A., "On the stability and accuracy of one-step methods for solving stiff systems of ordinary differential equations", *Mathematics of Computation* **28** (1974), 145—162.

[39] RUDIN, W., *A matematikai analízis alapjai* (Műszaki Könyvkiadó, Budapest, 1978).

[40] STETTER, H. J., *Analysis of Discretization Methods for Ordinary Differential Equations* (Springer, Berlin, 1973).
[41] STETTER, H. J., "Towards a theory for discretizations of stiff differential systems", Proceedings of the Conference on Numerical Analysis, Springer Lecture Notes in Math., No. 506, Dundee, 1975, 190—201.
[42] TIKHONOV, A. N. and ARSENIN, V. Y., *Solutions of Ill-Posed Problems*, (Winston and Wiley, New York, 1977).
[43] TRIGIANTE, D., "Asymptotic stability and discretization on an infinite interval", *Computing* **18** (1977), 117—129.
[44] VELDHUIZEN, M. VAN, "Consistency and stability for one-step discretizations of stiff differential equations", Proceedings on Stiff Differential Systems (ed.: R. A. Willoughby), Plenum Press, New-York, 1974, 259—270.
[45] VELDHUIZEN, M. VAN, "*D*-stability", *SIAM J. Numer. Anal.* **18** (1981) 45—64.
[46] VERWER, J. G., "*S*-stability properties for generalized *Runge—Kutta methods*", *Numer. Math.* **27** (1977) 359—370.
[47] VERWER, J. G., "An analysis of Rosenbrock methods for non-linear stiff initial value problems", Report NW 90/80, Mathematisch Centrum, Amsterdam, 1980.
[48] VERWER, J. G., "Instructive experiments with some Runge—Kutta—Rosenbrock methods", Report NW 100/81, Mathematisch Centrum, Amsterdam, 1981.

(Beérkezett: 1984. február 3.)

GALÁNTAI AURÉL
AGRÁRTUDOMÁNYI EGYETEM
2103 GÖDÖLLŐ

## STABILITY OF NUMERICAL METHODS FOR LINEAR DIFFERENTIAL EQUATIONS

A. GALÁNTAI

In this paper we investigate the zero convergence of global errors at infinity in case of $A[\alpha]$-stable linear multistep-multiderivative and *Runge—Kutta methods* and linear inhomogeneous differential equations of the form (1.1). The concept of $A[\alpha, g]$-stability (see [10]) means that for fixed perturbation $g$ the global errors tend to zero at infinity for all stepsize $h>0$ and matrix $A$ satisfying $\sigma(A)\subset W(\alpha)$ (Definition 1.1).

It is proven that any $A[\alpha]$-stable discretization method of the form (2.1) is $A[\alpha, g]$-stable for a given $g$, iff the local errors $L(y(t_n); h)$ tend to zero at infinity. It is also shown that $A[\alpha, g]$-stability is linear in $g$, if the local error operator $L(y(t); h)$ is linear in $y$.

Using the representation (2.11) sufficient existence theorems are proven under the condition (2.12) or (2.14) and (2.13).

Nonexistence of $A[\alpha, g]$-stability is proven using the conditions (3.1), (3.4) or (3.1), (3.4), (A), (3.6) and (3.7).

The following consequences are only mentioned.

1. The linearization principle is not necessarily satisfied for the global errors of $A[\alpha]$-stable methods.
2. The method-independent heuristic stiff concepts based on the first approximation are not satisfactory.
3. $A[\alpha]$-stability and $S$-stability do not fully characterize the behaviour of numerical methods on linear differential equations (see [7]).

*Alkalmazott Matematikai Lapok* **10** (1984) 273—281

# EGY ÖSSZLÉPÉSES POLINOM-FAKTORIZÁCIÓS ELJÁRÁS TÖBBSZÖRÖS GYÖKÖKKEL IS RENDELKEZŐ POLINOMOKRA

VARGA GYULA

Budapest

A cikk egy összlépéses eljárást ad meg ismert multiplicitású többszörös gyökökkel is rendelkező polinomok első- és másodfokú tényezők hatványaira történő felbontására. Az eljárás a *Vieta-féle gyökfüggvényeknek* a *Newton—Girard képlettel* való transzformációján alapul. Konvergenciája másodrendű.

## 1. Bevezetés

Az összlépéses polinom-faktorizációs eljárások polinomok összes gyökének, ill. a hozzájuk tartozó gyöktényezőknek egyszerre történő kiszámítására szolgálnak. Használatuk feleslegessé teszi a szukcesszív faktorizációs eljárások alkalmazása során fellépő kerekítési hibák felhalmozódása miatt szükségessé váló utóiterációs eljárások alkalmazását.

Az eljárások alapötlete WEIERSTRASSTÓL való. Számítástechnikailag használható algoritmust elsőfokú tényezőkre [1], másodfokúakra [2] adott. Ezek az algoritmusok egyszeres valós vagy komplex gyöktényezők, ill. egyszeres másodfokú tényezők kiszámítására alkalmasak. Az eljárások általánosításai [3], [4], [5], [6] stb. magasabbrendű konvergenciát biztosítanak, vagy az első-, ill. másodfokú tényezők egyszeresség̈ére vonatkozó megszorításokon enyhítenek.

Az alábbiakban egy olyan összlépéses eljárást adunk meg, amely alkalmazható többszörös gyökökkel is rendelkező polinomok egyidejűleg történő tényezőkre bontására is, feltéve, hogy az összes gyök megfelelő közelítését és multiplicitását ismerjük. Az eljárást külön mutatjuk be első- és másodfokú tényezőkre bontásra, majd útmutatást adunk egy polinom általános, első- és másodfokú tényezőkre bontott alakjának előállítására. Az eljárást valós együtthatós polinomokra mutatjuk be, de ez nem jelent megszorítást. Végül példákon szemléltetjük az eljárás egyes változatainak alkalmazását.

## 2. Az eljárás leírása

Legyen

$$a(x) = \sum_{i=0}^{n} a_i x^i \quad (a_n = 1) \quad (n > 0) \tag{2.1}$$

valós együtthatós polinom. Legyen az együtthatók vektora

$$\hat{\mathbf{a}} = (a_{n-1}, a_{n-2}, \ldots, a_0)^T,$$

a gyökök vektora

$$\mathbf{z}^* = (z_1^*, \ldots, z_n^*)^T,$$

és ennek valamely közelítése

$$\mathbf{z} = (z_1, \ldots, z_n)^T.$$

Mint ismeretes, az $a(x)$ polinom gyökeinek vektora, $\mathbf{z}^*$, kielégíti a

$$\mathbf{V}(\mathbf{z}) - \hat{\mathbf{a}} = \mathbf{0} \tag{2.2}$$

$n$ egyenletből álló $n$ ismeretlenes nemlineáris egyenletrendszert, amelyben

$$\mathbf{V}(\mathbf{z}) = (V_1(\mathbf{z}), \ldots, V_n(\mathbf{z}))^T$$

a *Vieta-féle*

$$V_i(\mathbf{z}) \equiv (-1)^i \sum_{\substack{j_1=1 \\ j_1 \neq \ldots \neq j_i}}^{n} \ldots \sum_{j_i=1}^{n} z_{j_1} \ldots z_{j_i} \quad (i = 1, \ldots, n)$$

*gyökfüggvények vektora.*

A (2.2) egyenletrendszerrel ekvivalens, de egyszerűbb alakú és könnyebben kezelhető egyenletrendszert írhatunk fel

$$\mathbf{S}(\mathbf{z}) - \mathbf{s} = \mathbf{0} \tag{2.3}$$

alakban, ahol

$$\mathbf{S}(\mathbf{z}) = (S_1(\mathbf{z}), \ldots, S_n(\mathbf{z}))^T,$$

$$\mathbf{s} = (s_1, \ldots, s_n)^T,$$

és

$$S_i(\mathbf{z}) \equiv \sum_{k=1}^{n} z_k^i \quad (i = 1, \ldots, n)$$

a gyökök $i$-edik hatványainak összege, az $\mathbf{s}$ vektor komponenseit pedig a *Newton—Girard-féle*

$$s_i = -\sum_{k=1}^{i} a_{n-k} s_{i-k} \quad s_0 = s_0(i) = i \quad (i = 1, \ldots, n) \tag{2.4}$$

rekurziós képlet segítségével kaphatjuk meg a polinom együtthatóiból.

Tekintsük ezután kitűzött feladatunk különböző eseteit.

I. Legyen $a(x)$ olyan $n$-edfokú polinom, amely többszörös gyökökkel is rendelkezik. Legyen az összes különböző gyökök vektora

$$\mathbf{x}^* = (x_1^*, \ldots, x_m^*)^T \quad (m \leqq n),$$

ennek valamely közelítése

$$\mathbf{x} = (x_1, \ldots, x_m)^T,$$

és legyenek az ismert multiplicitások rendre $n_1, \ldots, n_m$, akkor nyilvánvalóan

$$n = \sum_{i=1}^{m} n_i.$$

Az $a(x)$ polinomot a gyökök multiplicitásának ismeretében

$$a(x) = \prod_{k=1}^{m} (x - x_k^*)^{n_k} \tag{2.5}$$

alakban akarjuk az elsőfokú tényezők egyidejű kiszámításával felbontani.

A (2.3) egyenletrendszer, amelynek csak az első $m$ egyenletére van szükségünk $\mathbf{x}^*$ kiszámítására, most az alábbi alakú:

$$f_i(\mathbf{x}) \equiv \sum_{k=1}^{m} n_k x_k^i - s_i = 0 \quad (i = 1, \ldots, m). \tag{2.6}$$

Az egyenletrendszert a *Newton—Raphson módszerrel* oldjuk meg. Írjuk fel az egyenletrendszer *Jacobi mátrixát:*

$$\mathbf{U} = [u_{ik}]; \quad u_{ik} = \frac{\partial f_i(\mathbf{x})}{\partial x_k} = i n_k x_k^{i-1} \quad (i = 1, \ldots, m), \quad (k = 1, \ldots, m). \tag{2.7}$$

A Jacobi mátrix az alábbi szorzatalakban is felírható:

$$\mathbf{U} = \begin{bmatrix} 1 & & \\ & \ddots & \\ & & m \end{bmatrix} \begin{bmatrix} \cdots & 1 & \cdots \\ & x_k & \\ & \vdots & \\ \cdots & x_k^{m-1} & \cdots \end{bmatrix} \begin{bmatrix} n_1 & & \\ & \ddots & \\ & & n_m \end{bmatrix}. \tag{2.8}$$

A tényezők közül a két szélső nemszinguláris diagonális mátrix, a közbülső pedig *Vandermonde mátrix*, amelyről tudjuk, hogy egymástól különböző $x_k$ generáló elemek esetén (ez megfelelő kezdő közelítésekkel biztosítható) nemszinguláris. Tehát $\mathbf{U}$ nemszinguláris, sőt inverze egyszerűen, képlettel is felírható:

$$\mathbf{U}^{-1} = [w_{ki}]; \quad w_{ki} = \frac{b_{k,i-1}}{i n_k b_k(x_k)} \quad (k = 1, \ldots, m), \quad (i = 1, \ldots, m), \tag{2.9}$$

ahol

$$b_k(x) = \prod_{\substack{j=1 \\ j \neq k}}^{m} (x - x_j) \quad (k = 1, \ldots, m).$$

Az iterációhoz a korrekciós tagot a

$$\Delta \mathbf{x} = -\mathbf{U}^{-1}\mathbf{f}(\mathbf{x})$$

képlettel adhatjuk meg. Ezt részletesen kiírva kapjuk:

$$\Delta x_k = -\frac{\sum_{i=1}^{m} b_{k,i-1} f_i(\mathbf{x})/i}{n_k b_k(x_k)} \quad (k = 1, \ldots, m). \tag{2.10}$$

A korrekciós vektor valamennyi komponensének kiszámítása után az iterációt

$$x_i^{(l+1)} = x_i^{(l)} + \Delta x_i^{(l)} \quad (i = 1, \ldots, m), \quad (l = 0, 1, \ldots) \tag{2.11}$$

alakban írhatjuk fel.

Minthogy a $\varphi(\mathbf{x})$ iterációs függvény $k$-adik komponensére, a

$$\varphi_k(\mathbf{x}) = x_k - \frac{\sum_{i=1}^{m} b_{k,i-1} f_i(\mathbf{x})/i}{n_k b_k(x_k)} \quad (k = 1, \ldots, m) \tag{2.12}$$

függvényre fennállnak a

$$\varphi_k(\mathbf{x}^*) = x_k^*, \quad \left.\frac{\partial \varphi_k(\mathbf{x})}{\partial x_j}\right|_{\mathbf{x}=\mathbf{x}^*} = 0 \quad (k = 1, \ldots, m), \quad (j = 1, \ldots, m)$$

egyenlőségek, azért az eljárás konvergenciája másodrendű.

A képletben szereplő $b_k(x)$ polinom együtthatóit úgy kapjuk meg, hogy először kiszámítjuk a

$$b(x) = \prod_{j=1}^{m} (x - x_j) \tag{2.13}$$

közelítő redukált polinomot a

$$b_j = \delta_{j,0}$$

$$b_j = b_{j-1} - x_l b_j$$

$$(l = 1, \ldots, m; \quad j = l, \ldots, 0) \tag{2.14}$$

rekurzióval, (itt és a továbbiakban $\delta$ a *Kronecker szimbólumot* jelöli) majd a $b(x)$ polinomot az $x - x_k$ tényezővel elosztjuk a

$$b_{m,k} = 0$$

$$b_{i,k} = b_{i+1} + b_{i+1,k} x_k$$

$$(i = m-1, \ldots, 0; \quad k = 1, \ldots, m) \tag{2.15}$$

rekurzió segítségével.

Az iteráció sikeres befejezése után rendelkezésünkre állnak a polinom gyökei és a gyököket egyszeres multiplicitással tartalmazó redukált polinom együtthatói az iterációs eljárásban megkívánt pontosságra.

Az I. változat alkalmazása csupa valós gyökökkel rendelkező polinomokra célszerű.

II. Legyen most

$$a(x) = \sum_{i=0}^{2n} a_i x^i \quad (a_{2n} = 1), \quad (n > 0) \tag{2.16}$$

páros fokszámú valós együtthatós polinom. Legyen

$$\hat{\mathbf{a}} = (a_{2n-1}, a_{2n-2}, \ldots, a_0)^T$$

az együtthatók vektora. A polinomot

$$a(x) = \prod_{i=1}^{m} (x^2 + r_{2i}^* x + r_{2i-1}^*)^{n_i} \tag{2.17}$$

alakban akarjuk felbontani az összes különböző másodfokú tényező együtthatóinak egyidejű kiszámításával. Legyen

$$\mathbf{r}^* = (r_1^*, \ldots, r_{2m}^*)^T$$

a másodfokú tényezők együtthatóinak vektora, és legyenek $n_1, \ldots, n_m$ az egyes másod-

fokú tényezők ismert multiplicitásai, amelyekre fennáll az

$$n = \sum_{i=1}^{m} n_i$$

egyenlőség. Legyen továbbá

$$\mathbf{r} = (r_1, \ldots, r_{2m})^T$$

$\mathbf{r}^*$ valamely közelítése, és

$$\mathbf{x}^* = (x_1^*, \ldots, x_{2m}^*)^T$$

a polinom gyökeinek vektora az alábbi értelmben: Jelöljük a (2.17) $i$-edik másodfokú tényezőjének diszkriminánsát $d_i$-vel. Akkor

$$x_{2i-1}^* = \begin{cases} (-r_{2i}^* + \sqrt{d_i})/2, & \text{ha } d_i \geqq 0 \\ -r_{2i}^*/2, & \text{ha } d_i < 0 \end{cases} \tag{2.18}$$

$$x_{2i}^* = \begin{cases} (-r_{2i}^* - \sqrt{d_i})/2, & \text{ha } d_i \geqq 0 \\ \sqrt{|d_i|}/2, & \text{ha } d_i < 0. \end{cases}$$

Legyen végül

$$\mathbf{x} = (x_1, \ldots, x_{2m})^T$$

$\mathbf{x}^*$ valamely közelítésének a vektora a fentiekkel megegyező értelemben.

A (2.3) egyenletrendszert az I. változatéhoz hasonlóan felírva kapjuk a következőt:

$$f_i(\mathbf{x}) \equiv \sum_{k=1}^{m} n_k p_{ik} - s_i = 0 \quad (i = 1, \ldots, 2m), \tag{2.19}$$

ahol $s_i$ az I. változatéval megegyezően a (2.4) rekurziós képletből számítható ki, a $p_{i,k}$ mennyiségek pedig a $k$-adik közelítő másodfokú tényező gyökeinek $i$-edik hatványösszegei. Ezek komplex gyökök esetén is kiszámíthatók valós úton az alábbi rekurzió segítségével:

$$\begin{aligned} p_{0,k} &= 2 \\ p_{1,k} &= -r_{2k} \\ p_{i,k} &= -p_{i-1,k} r_{2k} - p_{i-2,k} r_{2k-1} \quad (i = 2, \ldots, 2m). \end{aligned} \tag{2.20}$$

A *Newton—Raphson módszer* alkalmazásához szükséges *Jacobi mátrix* inverze az I. változatéhoz hasonlóan írható fel. Először kiszámítjuk a

$$g(x) = \sum_{i=1}^{m} (x^2 + r_{2i}x + r_{2i-1}) = \sum_{i=0}^{2m} g_i x^i \tag{2.21}$$

közelítő redukált polinom együtthatóit az alábbi rekurzió segítségével:

$$\begin{aligned} g_j &= \delta_{j,0} \\ g_j &= g_{j-2} + g_{j-i} r_{2l} + g_j r_{2l-1} \\ (l &= 1, \ldots, m;\ j = 2l, \ldots, 0). \end{aligned} \tag{2.22}$$

A továbbiakban a közelítő redukált polinomot rendre elosztjuk a $k$-adik közelítő másodfokú tényezővel. Így kapjuk a $g_k(x)$ polinomokat. Az osztáshoz az alábbi rekurziót használjuk:

$$g_{k,2m-1} = g_{k,2m} = 0$$

$$g_{k,i} = g_{i+2} - r_{2k}\, g_{k,i+1} - r_{2k-1}\, g_{k,i+2} \tag{2.23}$$

$$(i = 2m-2, \ldots, 0)$$

$$(k = 1, \ldots, m).$$

A (2.18)-cal értelmezett $x_{2k-1}$ és $x_{2k}$ mennyiségekre vonatkozó korrekciós tagok felírásánál $d_k$ előjelétől függően megkülönböztetést kell tennünk.

1. $d_k > d > 0$ esetén

$$\Delta x_{2k-j} = (-1)^{j+1} \frac{\sum_{i=1}^{2m} (g_{k,i-2} - x_{2k-1+j}\, g_{k,i-1}) f_i(\mathbf{x})/i}{n_k (x_{2k} - x_{2k-1})\, g_k(x_{2k-j})} \quad (j = 0, 1), \tag{2.24}$$

2. $d_k < -d < 0$ esetén

$$[\Delta x_{2k-1}, \Delta x_{2k}] = -\frac{\sum_{i=1}^{2m} (g_{k,i-2} - [x_{2k-1}, x_{2k}] g_{k,i-1}) f_i(\mathbf{x})/i}{n_k [0, \sqrt{|d_k|}]\, g_k([x_{2k-1}, x_{2k}])}. \tag{2.25}$$

(A fenti összegezésekhez kiegészítésül kell a $g_{k,-1} = 0$, $g_{k,2m} = 0$ definíció.)

A (2.25) képletben a szögletes zárójelben valamely komplex szám valós és képzetes része áll. Látható, hogy itt komplex műveletek is szerepelnek, ti. a valós együtthatós $g_k(x)$ polinomba történő komplex behelyettesítés, továbbá egy szorzatösszeg kiszámítása.

Az előző változatéhoz hasonlóan az iterációs függvénynek és első parciális deriváltjainak a vizsgálatával itt is belátható az eljárás másodrendű konvergenciája, ha az egyes másodfokú tényezőknek nincs közös zérushelyük, továbbá, ha minden $k$-ra $(k = 1, \ldots; m)$ $|d_k| > d > 0$. A $d_k \to 0$ esetet, vagyis a teljes négyzet alakú másodfokú tényező esetét a III. változattal intézhetjük el. A II. változat ebben az esetben nem biztosít másodrendű konvergenciát.

III. Az előző két változat kombinációjaként ismert multiplicitású valós gyökök és konjugált komplex gyökpárok megfelelő kezdő közelítéseinek birtokában az $n$-edfokú valós együtthatós

$$a(x) = \sum_{i=0}^{n} a_i x^i \quad (n > 0)$$

polinom legáltalánosabb

$$a(x) = \prod_{k=1}^{\mu} (x - q_k^*)^{\alpha_k} \prod_{k=1}^{\nu} (x^2 + r_{2k}^* x + r_{2k-1}^*)^{\beta_k} \tag{2.26}$$

$$n = \sum_{k=1}^{\mu} \alpha_k + 2 \sum_{k=1}^{\nu} \beta_k$$

alakú felbontását valósíthatjuk meg valamennyi első-, ill. másodfokú tényezőjének

egyidejű kiszámításával. Ha az egyes első- és másodfokú tényezőknek közös gyökeik nincsenek, akkor az eljárás, kiindulva valamely alkalmas $q_1, \ldots, q_\mu$ és $r_1, \ldots, r_{2\nu}$ közelítő értékekből, másodrendben konvergál az összes tényező megfelelő $q_k^*$ $(k=1, \ldots, \mu)$, ill. $r_k^*$ $(k=1, \ldots, 2\nu)$ együtthatóihoz.

## 3. Megjegyzések

1. Az I. és II. változat egyes részlépéseinek végrehajtásához mellékelt rekurziós eljárások mutatják, hogy 1—1 gyökre vonatkozóan a műveletigény iterációs lépésenként $O(m)$.

2. A II. változatban szereplő komplex műveleteket valós számpárokra vonatkozó aritmetikával is elvégezhetjük.

3. A III. változatban a másodfokú tényezők kizárólag csak konjugált komplex gyökpárokhoz tartozhatnak.

4. A *Newton—Girard képlet* alkalmazása után a tényezőkre bontandó polinom együtthatóira már nincs szükségünk, azok az eljárás végrehajtása során kiszámított redukált polinom együtthatóival felülíródnak.

5. Ha a tényezőkre bontandó polinom valamennyi különböző gyökének megfelelő közelítéseit ismerjük, a gyökök multiplicitásának meghatározása egyszerűen elvégezhető komplex függvénytani segédeszközökkel. Ismeretes, hogy az $a(z)$ polinomra és a komplex sík egy zárt egyszerű $G$ görbéjére, amely $a(z)$ egyik zérushelyén sem halad át, érvényes az

$$\frac{1}{2\pi\sqrt{-1}} \oint_G \frac{a'(z)}{a(z)}\, dz = N$$

egyenlőség, ahol $N$ az $a(z)$ $G$ belsejében levő gyökeinek száma, tehát a polinom logaritmikus deriváltjának a $G$ mentén vett integrálja segítségével a $G$-n belüli gyökök száma multiplicitásukkal véve kiszámítható. Mivel az eredmény valós pozitív egész szám, elegendő az egyes közelítő gyökök körül a numerikus integrálást $0<\varepsilon<1/2$ pontosságra olyan $G$ görbe mentén elvégezni, amelynek belsejében más gyök nem helyezkedik el.

## 4. Teszteredmények

A fenti eljárás mindhárom változatára a próbafuttatások az MTA CDC 3300 gépén történtek. Az egyes változatokra USASI FORTRAN nyelven írt szubrutinokat az alábbi tesztfeladatokra próbáltuk ki:

*I. változat:*

$$a(x) = (x-2)^4(x+2)^4(x-1)^3(x+4).$$

A megoldandó nemlineáris egyenletrendszer:

$$\begin{aligned}
4x_1+4x_2+3x_3+x_4+1 &= 0\\
4x_1^2+4x_2^2+3x_3^2+x_4^2-51 &= 0\\
4x_1^3+4x_2^3+3x_3^3+x_4^3+61 &= 0\\
4x_1^4+4x_2^4+3x_3^4+x_4^4-387 &= 0.
\end{aligned}$$

A kezdő közelítések:

$$1{,}8 \quad -2{,}1 \quad 0{,}85 \quad -3{,}92.$$

Az iterációs eljárás végrehajtása során kapott közelítések:

| | | | |
|---|---|---|---|
| 2,035 998 22 | −1,994 853 35 | 0,949 395 05 | −4,012 764 62 |
| 2,001 401 59 | −2,000 013 47 | 9,998 170 46 | −4,000 063 86 |
| 2,000 002 17 | −2,000 000 02 | 0,999 997 13 | −4,000 000 01 |
| 2,000 000 00 | −2,000 000 00 | 1,000 000 00 | −4,000 000 00. |

*II. változat:*

$$a(x) = (x^2+x-1)^3(x^2-x+2)^2.$$

A megoldandó nemlineáris egyenletrendszer:

$$3p_{11}+2p_{12}+1 = 0$$
$$3p_{21}+2p_{22}+3 = 0$$
$$3p_{31}+2p_{32}+22 = 0$$
$$3p_{41}+2p_{42}-23 = 0.$$

A kezdő közelítések a másodfokú tényezők együtthatóihoz:

$$-0{,}9 \quad 1{,}1 \quad 1{,}9 \quad -0{,}8,$$

a gyökök kezdő közelítései ezekből adódnak egy-egy másodfokú egyenlet megoldásaként.

Az iterációs eljárás végrehajtása során kapott gyökközelítések:

| | | | |
|---|---|---|---|
| 0,546 585 61 | −1,646 585 61 | 0,400 000 00 | 1,319 090 60 |
| 0,613 809 14 | −1,617 062 01 | 0,502 439 65 | 1,315 924 45 |
| 0,618 097 64 | −1,618 042 24 | 0,499 958 45 | 1,322 900 40 |
| 0,618 033 99 | −1,618 033 99 | 0,499 999 99 | 1,322 875 65 |
| 0,618 033 99 | −1,618 033 99 | 0,500 000 00 | 1,322 875 66. |

A másodfokú tényezők együtthatóinak végértékei:

$$-1{,}000\,000\,00 \quad 1{,}000\,000\,00 \quad 2{,}000\,000\,00 \quad -1{,}000\,000\,00.$$

*III. változat:*

$$a(x) = (x^2-x+1)^4(x+1)^3(x-2).$$

A megoldandó nemlineáris egyenletrendszer:

$$4p_{11}+3x_3+x_4-3 = 0$$
$$4p_{21}+3x_3^2+x_4^2-3 = 0$$
$$4p_{31}+3x_3^3+x_4^3+3 = 0$$
$$4p_{41}+3x_4^3+x_4^4-15 = 0.$$

A kezdő közelítések a másodfokú tényezők együtthatóihoz és az elsőfokú tényezőkhöz tartozó gyökökhöz:

$$0{,}8 \quad -1{,}2 \quad -1{,}3 \quad 2{,}2.$$

Az iterációs eljárás végrehajtása során kapott gyökközelítések:

| | | | |
|---|---|---|---|
| 0,600 000 00 | 0,663 324 96 | −1,300 000 00 | 2,200 000 00 |
| 0,501 513 71 | 0,946 263 21 | −1,045 398 20 | 2,124 084 92 |
| 0,499 625 51 | 0,873 089 07 | −1,055 538 14 | 2,013 585 27 |
| 0,499 989 50 | 0,866 091 77 | −1,000 027 43 | 2, 000 166 30 |
| 0,499 999 99 | 0,866 025 41 | −1,000 000 00 | 2,000 000 02 |
| 0,500 000 00 | 0,866 025 40 | −1,000 000 00 | 2,000 000 00. |

A másodfokú tényező együtthatóinak végértékei:

$$1{,}000\,000\,00 \quad -1{,}000\,000\,00.$$

Mindhárom változatban eredményül kapjuk még a redukált polinom együtthatóit is.

## IRODALOM


[1] KERNER, I. O., „Ein Gesamtschrittverfahren zur Berechnung der Nullstellen von Polynomen", *Numer. Math.* **8** (1966) 290—294.

[2] FILIPPI, S., „Ein verallgemeinertes Bairstow-Verfahren zur gleichzeitigen Ermittlung aller Nullstellen eines Polynoms", *Beiträge Numer. Math.* **4** (1975) 83—89.

[3] EHRLICH, L. W., "A modified Newton method for polynomials", *CACM* **10** (1967).

[4] MAESS, G., „Simultane Polynomaufspaltung in Quadratfaktoren", *Rostock Math. Kolloq.* **18** (1981) 89—96.

[5] VARGA, GY., „Párhuzamos algoritmus polinomok másodfokú tényezőkre bontására", *Alk. Mat. Lapok*, **10** (1984) 177—183.

[6] VARGA, GY., "On a generalization of the *Newton—Kerner procedure* for simultaneous calculation of all zeros of polynomials", MTA SZTAKI Working Paper, Budapest, 1983.


(Beérkezett: 1983. június 7.)

VARGA GYULA
MTA SZÁMÍTÁSTECHNIKAI ÉS AUTOMATIZÁLÁSI KUTATÓ INTÉZET
1502 BUDAPEST, KENDE U. 13—17.

## A TOTAL-STEP PROCEDURE FOR FACTORIZATION OF POLYNOMIALS WITH MULTIPLE ZEROS OF KNOWN MULTIPLICITY

GY. VARGA

The paper gives a total-step procedure for factorization of polynomials into the powers of linear and/or quadratic factors. The polynomial in question may have multiple zeros of known multiplicity too. The procedure is based on the transformation of the *Vieta root functions* by means of the *Newton—Girard formula*. Its convergence is quadratic.

*Alkalmazott Matematikai Lapok* **10** (1984) 283—288

# A GEOMETRIAI PROGRAMOZÁS ÉS AZ $l_p$ PROGRAMOZÁS GYENGE DUALITÁSI TÉTELÉNEK EGY ÚJ BIZONYÍTÁSA

TERLAKY TAMÁS

Budapest

Cikkünkben a geometriai és az $l_p$ programozás gyenge dualitás tételére adunk új, rövid bizonyítást. Bizonyításunk a megelőző bizonyításokkal ellentétben nem igényli külön segédtételek bizonyítását a homogén esetre, mint ahogy az [1] és [3]-ban található, hanem közvetlenül alkalmazható az inhomogén feladatra, így lényegesen egyszerűbb.

## 1. Bevezetés

A geometriai programozási feladatpár megfogalmazása és tulajdonságainak részletes vizsgálata [1]-ben, az $l_p$ programozási feladatpár megfogalmazása és tulajdonságainak részletes vizsgálata [3]-ban található. Az érthetőség céljából az alábbiakban definiáljuk az $l_p$ programozási és a geometriai programozási feladatpárt, valamint a bizonyításhoz szükséges fogalmakat.

*a) A geometriai programozási feladatpár*

Legyen $\mathbf{A}=(\mathbf{a}_i)=(\alpha_{ij})$ egy $n\times m$-es mátrix, $f\in R^m$, $\mathbf{c}=(\gamma_i)$ egy $n$ dimenziós vektor. Legyen továbbá $I_1, \ldots, I_r$ egy diszjunkt felbontása az $\{1, \ldots, n\}$ indexhalmaznak, azaz $I_k \cap I_j = \emptyset$, ha $k \neq j$ és $\bigcup_{k=1}^{r} I_k = \{1, \ldots, n\}$.

*A geometriai programozás primál feladata* ($\mathscr{P}1$)*:*

Határozzuk meg azt az $\mathbf{y}=(\eta_j)\in R^m$ vektort, melyre

$$\mathbf{fy} \text{ maximális}$$

feltéve, hogy

$$\sum_{i\in I_k} e^{\mathbf{a}_i \mathbf{y}-\gamma_i} \leqq 1, \quad k = 1, \ldots, r.$$

*A geometriai programozás duál feladata* ($\mathscr{D}1$)*:*

Határozzuk meg azt az $\mathbf{x}=(\xi_i)\in R^n$ vektort, melyre

$$\mathbf{xc} + \sum_{k=1}^{r} \log \frac{\prod_{i\in I_k} \xi_i^{\xi_i}}{\left(\sum_{i\in I_k} \xi_i\right)^{\sum_{i\in I_k} \xi_i}}$$

minimális, feltéve, hogy

$$\mathbf{xA} = \mathbf{f}$$

$$\mathbf{x} \geqq \mathbf{0}.$$

1.1. DEFINÍCIÓ. A geometriai programozás duál feladata *Slater reguláris,* ha a duál feladatnak létezik olyan megengedett $\mathbf{x}\in R^n$ megoldása, melyre $\mathbf{x}>\mathbf{0}$.

*b) Az $l_p$ programozási feladatpár*

Legyen $\mathbf{A}, \mathbf{f}, \mathbf{c}, \mathbf{y}, \mathbf{x}, I_1, \ldots, I_r$ ugyanaz, mint az előbb. Legyen továbbá $\mathbf{B}=(\mathbf{b}_k)=(\beta_{kj})$ egy $r\times m$-es mátrix és $\mathbf{d}=(\delta_k)\in R^r$ egy $r$ dimenziós vektor és $p_1, \ldots, p_n>1$ valós számok, valamint $\dfrac{1}{q_i}=1-\dfrac{1}{p_i}$ $i=1, \ldots, n$.

*Az $l_p$ programozás primál feladata ($\mathscr{P}2$):*

Határozzuk meg azt az $\mathbf{y}\in R^m$ vektort, melyre

$$\mathbf{fy} \text{ maximális,}$$

feltéve, hogy

$$G_k(\mathbf{y}) = \sum_{i\in I_k} \frac{1}{p_i} |\mathbf{a}_i\mathbf{y}-\gamma_i|^{p_i}+\mathbf{b}_k\mathbf{y}-\delta_k \leqq 0, \quad k=1, \ldots, r.$$

*Az $l_p$ programozás duál feladata ($\mathscr{D}2$):*

Határozzuk meg azt az $(\mathbf{x}, \mathbf{z})=(\xi_1, \ldots, \xi_n, \zeta_1, \ldots, \zeta_r)\in R^{n+r}$ vektort, melyre

$$\mathbf{xc}+\mathbf{zd}+\sum_{\substack{k=1\\ \zeta_k>0}}^{r} \zeta_k \sum_{i\in I_k} \frac{1}{q_i} \left|\frac{\xi_i}{\zeta_k}\right|^{q_i} .$$

minimális, feltéve, hogy

$$\mathbf{xA}+\mathbf{zB}=\mathbf{f}$$

$$\mathbf{z}\geqq\mathbf{0}$$

$$\zeta_k=0 \Rightarrow \xi_i=0, \quad i\in I_k, \quad k=1, \ldots, r.$$

1.2. DEFINÍCIÓ. Az $l_p$ programozás duál feladata *Slater reguláris,* ha van olyan $(\mathbf{x}, \mathbf{z})\in R^{n+r}$ megengedett megoldása a duál feladatnak, melyre $\mathbf{z}>\mathbf{0}$.

Tudjuk, hogy a ($\mathscr{D}1$) és ($\mathscr{D}2$) feladatok célfüggvényei pozitív homogén, szubadditív és így konvex függvények. Igaz továbbá, hogy a ($\mathscr{P}1$), illetve a ($\mathscr{P}2$) feladatok tetszőleges megengedett megoldásához tartozó célfüggvény érték kisebb vagy egyenlő mint a ($\mathscr{D}1$), illetve a ($\mathscr{D}2$) feladatok tetszőleges megengedett megoldásához tartozó célfüggvény érték. Ezen állítások bizonyításai szintén [1], illetve [3]-ban találhatók.

*Megjegyzés.* KLAFSZKY EMIL [1] dolgozatában „kanonikus" elnevezéssel szerepelt a *Slater reguláris geometriai programozási duál feladat.* Mivel a két fogalom egybeesik, így indokoltnak tartjuk az általában használatos „*Slater reguláris*" elnevezés használatát ebben az esetben is.

A bizonyítás során felhasználjuk a *Farkas tétel* konvex függvényekre vonatkozó változatát, így most ezt is megfogalmazzuk. Bizonyítása [2]-ben található.

Legyenek a $g_1, \ldots, g_r, f$ konvex függvények egy $\mathscr{C}\subset R^m$ konvex halmazon.

1.3. DEFINÍCIÓ. A $g_1, \ldots, g_r$ függvényrendszert *Slater regulárisnak* nevezzük, ha van olyan $\mathbf{x}^0$ relatív belső pontja $\mathscr{C}$-nek, hogy

$g_j(\mathbf{x}^0)\leqq 0$, ha $g_j$ lineáris függvény és

$g_j(\mathbf{x}^0)<0$, ha $g_j$ nemlineáris függvény.

1.3. TÉTEL. Tegyük fel, hogy a $g_1, \ldots, g_r$ függvények kielégítik a *Slater regularitási feltételt*. Ekkor, ha a

$$g_k(\mathbf{x}) \leqq 0, \quad k = 1, \ldots, r$$

$$f(\mathbf{x}) < 0$$

$$\mathbf{x} \in \mathscr{C} \subset R^m$$

rendszernek nincs megoldása, akkor van olyan $\mathbf{y}=(\eta_1, \ldots, \eta_r)$ nemnegatív vektor, hogy

$$f(\mathbf{x}) + \sum_{k=1}^{r} \eta_k g_k(\mathbf{x}) \geqq 0 \quad \text{minden } \mathbf{x} \in \mathscr{C} \text{ esetén.}$$

A tétel erejét mutatja, hogy megfordítása is igaz. A fordított állítás bizonyítása triviális.

Most rátérünk a geometriai, majd az $l_p$ programozás gyenge dualitási tételének bizonyítására.

## 2. A geometriai programozás gyenge dualitási tételének bizonyítása

Az egyszerűség kedvéért jelöljük $\{\mathscr{D}1\}$-gyel a $(\mathscr{D}1)$ feladat megengedett megoldásainak halmazát és legyen

$$v = \inf_{\mathbf{x} \in \{\mathscr{D}1\}} \left\{ \mathbf{xc} + \sum_{k=1}^{r} \log \frac{\prod_{i \in I_k} \xi_i^{\xi_i}}{\left(\sum_{i \in I_k} \xi_i\right)^{\sum_{i \in I_k} \xi_i}} \right\}.$$

2.1. TÉTEL. Ha a $(\mathscr{D}1)$ feladat *Slater reguláris* és célfüggvénye alulról korlátos, akkor a $(\mathscr{P}1)$ feladatnak létezik $\mathbf{y}^* \in R^m$ optimális megoldása, melyre

$$\mathbf{fy}^* = v.$$

*Bizonyítás.* A $v$ érték definíciója miatt az

$$\mathbf{xA} = \mathbf{f}$$

$$\mathbf{xc} + \sum_{k=1}^{r} \log \frac{\prod_{i \in I_k} \xi_i^{\xi_i}}{\left(\sum_{i \in I_k} \xi_i\right)^{\sum_{i \in I_k} \xi_i}} - v < 0$$

feladatnak nincs megoldása a $\mathscr{C}=R_+^n$ konvex halmazon.

Mivel a $(\mathscr{D}1)$ feladat *Slater reguláris*, tehát alkalmazhatjuk a *Farkas tételt*, így létezik olyan $\mathbf{y}^*=(\eta_1^*, \ldots, \eta_m^*)$ vektor, hogy

$$\mathbf{xc} + \sum_{k=1}^{r} \log \frac{\prod_{i \in I_k} \xi_i^{\xi_i}}{\left(\sum_{i \in I_k} \xi_i\right)^{\sum_{i \in I_k} \xi_i}} - v + \mathbf{fy}^* - \mathbf{xAy}^* \geqq 0 \tag{2.1}$$

minden $\mathbf{x} \in \mathscr{C}$ esetén.

Mivel $\mathbf{x}\in\mathscr{C}$ esetén $\vartheta\mathbf{x}\in\mathscr{C}$ minden $\vartheta\geqq 0$ mellett, valamint $\mathbf{x}$-nek pozitív homogén függvénye a (2.1) baloldal $\mathbf{x}$-et tartalmazó minden tagja (ugyanis a ($\mathscr{D}1$) feladat célfüggvénye pozitív homogén), így

$$\mathbf{xc}+\sum_{k=1}^{r}\log\frac{\prod_{i\in I_k}\xi_i^{\xi_i}}{\left(\sum_{i\in I_k}\xi_i\right)^{\sum_{i\in I_k}\xi_i}}-\mathbf{xAy}^*\geqq 0 \tag{2.2}$$

minden $\mathbf{x}\in\mathscr{C}$ esetén.

Mivel $\mathbf{x}\in\mathscr{C}$ tetszőleges, így választhatjuk tetszőleges rögzített $k$ index mellett a

$$\xi_i=\begin{cases}e^{\mathbf{a}_i\mathbf{y}^*-\gamma_i}, & \text{ha}\quad i\in I_k\\ 0, & \text{ha}\quad i\notin I_k\end{cases}.$$

értéket. Ezt (2.2)-be helyettesítve kapjuk, hogy:

$$\sum_{i\in I_k}e^{\mathbf{a}_i\mathbf{y}^*-\gamma_i}\gamma_i+\log\frac{\prod_{i\in I_k}(e^{\mathbf{a}_i\mathbf{y}^*-\gamma_i})^{e^{\mathbf{a}_i\mathbf{y}^*-\gamma_i}}}{\left(\sum_{i\in I_k}e^{\mathbf{a}_i\mathbf{y}^*-\gamma_i}\right)^{\sum_{i\in I_k}e^{\mathbf{a}_i\mathbf{y}^*-\gamma_i}}}-\sum_{i\in I_k}e^{\mathbf{a}_i\mathbf{y}^*-\gamma_i}\mathbf{a}_i\mathbf{y}^*\geqq 0.$$

Átalakítva:

$$-\sum_{i\in I_k}(\mathbf{a}_i\mathbf{y}^*-\gamma_i)e^{\mathbf{a}_i\mathbf{y}^*-\gamma_i}+\sum_{i\in I_k}e^{\mathbf{a}_i\mathbf{y}^*-\gamma_i}(\mathbf{a}_i\mathbf{y}^*-\gamma_i)-$$

$$-\sum_{i\in I_k}e^{\mathbf{a}_i\mathbf{y}^*-\gamma_i}\log\sum_{i\in I_k}e^{\mathbf{a}_i\mathbf{y}^*-\gamma_i}\geqq 0.$$

Mivel $\omega\log\omega\leqq 0$-ból következik, hogy $\omega\leqq 1$, így

$$\sum_{i\in I_k}e^{\mathbf{a}_i\mathbf{y}^*-\gamma_i}\leqq 1.$$

Ezt az eljárást tetszőleges $k$ indexre elvégezhetjük ($k=1,\ldots,r$), így $\mathbf{y}^*$ megengedett megoldása a ($\mathscr{P}1$) feladatnak.

A (2.1) egyenlőtlenség minden $\mathbf{x}\in\mathscr{C}$ esetén fennáll, így ha az $\mathbf{x}=\mathbf{0}$ vektort választjuk, akkor az $\mathbf{fy}^*\geqq v$ egyenlőtlenséghez jutunk, amiből következik, hogy $\mathbf{fy}^*=v$, mivel $\mathbf{fy}^*\leqq v$ igaz minden megengedett megoldására ($\mathscr{P}1$)-nek.

Tehát $\mathbf{y}^*$ optimális megoldása a ($\mathscr{P}1$) feladatnak, így tételünket bebizonyítottuk.

## 3. Az $l_p$ programozás gyenge dualitás tételének bizonyítása

Az egyszerűség kedvéért jelöljük $\{\mathscr{D}2\}$-vel a ($\mathscr{D}2$) feladat megengedett megoldásainak halmazát, és legyen

$$v=\inf_{(\mathbf{x},\mathbf{z})\in\{\mathscr{D}2\}}\left\{\mathbf{xc}+\mathbf{zd}+\sum_{\substack{k=1\\ \zeta_k>0}}^{r}\zeta_k\sum_{i\in I_k}\frac{1}{q_i}\left|\frac{\xi_i}{\zeta_k}\right|^{q_i}\right\}.$$

3.1. TÉTEL. Ha a ($\mathscr{D}2$) feladat *Slater reguláris* és célfüggvénye alulról korlátos, akkor a ($\mathscr{P}2$) feladatnak van olyan $\mathbf{y}^*\in R^m$ optimális megoldása, melyre

$$\mathbf{fy}^*=v.$$

*Bizonyítás.* A $v$ érték definíciója miatt az

$$\mathbf{xA}+\mathbf{zB}=\mathbf{f}$$

$$\mathbf{xc}+\mathbf{zd}+\sum_{\substack{k=1\\ \zeta_k>0}}^{r}\zeta_k\sum_{i\in I_k}\frac{1}{q_i}\left|\frac{\xi_i}{\zeta_k}\right|^{q_i}-v<0$$

feladatnak nincs megengedett megoldása a $\mathscr{C}=\{(x,z)\in R^{n+r}|\mathbf{z}\geqq\mathbf{0},\ \zeta_k=0\Rightarrow\xi_i=0,\ i\in I_k,\ k=1,\ldots,r\}$ konvex halmazon. Mivel a ($\mathscr{D}2$) feladat *Slater reguláris*, a *Farkas tétel* szerint létezik olyan $\mathbf{y}^*=(\eta_1^*,\ldots,\eta_m^*)$ vektor, hogy

$$\mathbf{xc}+\mathbf{zd}+\sum_{\substack{k=1\\ \zeta_k>0}}^{r}\zeta_k\sum_{i\in I_k}\frac{1}{q_i}\left|\frac{\xi_i}{\zeta_k}\right|^{q_i}-v+\mathbf{fy}^*-\mathbf{xAy}^*-\mathbf{zBy}^*\geqq 0 \tag{3.1}$$

minden $(\mathbf{x},\mathbf{z})\in\mathscr{C}$ esetén.

Mivel $(\mathbf{x},\mathbf{z})\in\mathscr{C}$ esetén $\vartheta(\mathbf{x},\mathbf{z})\in\mathscr{C}$ minden $\vartheta\geqq 0$ mellett, valamint az $\mathbf{x}$-et és $\mathbf{z}$-t tartalmazó tagok (3.1)-ben pozitív homogén függvényei $(\mathbf{x},\mathbf{z})$-nek (mivel a ($\mathscr{D}2$) feladat célfüggvénye pozitív homogén függvény), így

$$\mathbf{xc}+\mathbf{zd}+\sum_{\substack{k=1\\ \zeta_k>0}}^{r}\zeta_k\sum_{i\in I_k}\frac{1}{q_i}\left|\frac{\xi_i}{\zeta_k}\right|^{q_i}-\mathbf{xAy}^*-\mathbf{zBy}^*\geqq 0 \tag{3.2}$$

minden $(\mathbf{x},\mathbf{z})\in\mathscr{C}$ esetén.

Mivel $(\mathbf{x},\mathbf{z})\in\mathscr{C}$ tetszőleges, így választhatjuk rögzített $k$ index mellett az alábbi értékeket:

$$\zeta_j=\begin{cases}1, & \text{ha } j=k\\ 0, & \text{ha } j\neq k,\end{cases}\qquad \xi_i=\begin{cases}0 & \text{ha } i\notin I_k,\\ (\mathbf{a}_i\mathbf{y}^*-\gamma_i)^{p_i-1} & \text{ha } i\in I_k \text{ és } \mathbf{a}_i\mathbf{y}^*\geqq\gamma_i,\\ -(\gamma_i-\mathbf{a}_i\mathbf{y}^*)^{p_i-1} & \text{ha } i\in I_k \text{ és } \mathbf{a}_i\mathbf{y}^*<\gamma_i,\end{cases}$$

Legyen továbbá $I_{k_1}=\{i\in I_k|\mathbf{a}_i\mathbf{y}^*\geqq\gamma_i\}$ és $I_{k_2}=\{i\in I_k|\mathbf{a}_i\mathbf{y}^*<\gamma_i\}$. Helyettesítsük a fenti értékeket (3.2)-be, így:

$$\sum_{i\in I_{k_1}}(\mathbf{a}_i\mathbf{y}^*-\gamma_i)^{p_i-1}\gamma_i-\sum_{i\in I_{k_2}}(\gamma_i-\mathbf{a}_i\mathbf{y}^*)^{p_i-1}\gamma_i+\delta_k-\mathbf{b}_k\mathbf{y}^*+$$

$$+\sum_{i\in I_k}\frac{1}{q_i}|\mathbf{a}_i\mathbf{y}^*-\gamma_i|^{q_i(p_i-1)}-\sum_{i\in I_{k_1}}\mathbf{a}_i\mathbf{y}^*(\mathbf{a}_i\mathbf{y}^*-\gamma_i)^{p_i-1}+$$

$$+\sum_{i\in I_{k_2}}\mathbf{a}_i\mathbf{y}^*(\gamma_i-\mathbf{a}_i\mathbf{y}^*)^{p_i-1}\geqq 0.$$

Átalakítva:

$$-\sum_{i\in I_{k_1}}(\mathbf{a}_i\mathbf{y}^*-\gamma_i)^{p_i}-\sum_{i\in I_{k_2}}(\gamma_i-\mathbf{a}_i\mathbf{y}^*)^{p_i}+\sum_{i\in I_k}\frac{1}{q_i}|\mathbf{a}_i\mathbf{y}^*-\gamma_i|^{p_i}-$$

$$-\mathbf{b}_k\mathbf{y}^*+\delta_k\geqq 0.$$

Mivel $(\mathbf{a}_i\mathbf{y}^*-\gamma_i)=|\mathbf{a}_i\mathbf{y}^*-\gamma_i|$, ha $i\in I_{k_1}$, $(\gamma_i-\mathbf{a}_i\mathbf{y}^*)=|\mathbf{a}_i\mathbf{y}^*-\gamma_i|$, ha $i\in I_{k_2}$ és $\frac{1}{p_i}=1-\frac{1}{q_i}$, így:

$$\sum_{i\in I_k}\frac{1}{p_i}|\mathbf{a}_i\mathbf{y}^*-\gamma_i|^{p_i}+\mathbf{b}_k\mathbf{y}^*-\delta_k\leqq 0.$$

Mivel ezt az eljárást minden $k=1, \ldots, r$ indexre elvégezhetjük, így $\mathbf{y}^*$ megengedett megoldása a ($\mathscr{P}2$) feladatnak.

A (3.1) egyenlőtlenség minden $(\mathbf{x}, \mathbf{z}) \in \mathscr{C}$ esetén fennáll, így ha az azonosan nulla vektort választjuk, akkor az $\mathbf{f}\mathbf{y}^* \geqq v$ egyenlőtlenséghez jutunk. Ebből az $\mathbf{f}\mathbf{y}^* = v$ egyenlőség következik, mivel $\mathbf{f}\mathbf{y}^* \leqq v$ igaz minden megengedett megoldására a ($\mathscr{P}2$) feladatnak.

Tehát $\mathbf{y}^*$ optimális megoldása a ($\mathscr{P}2$) feladatnak, így tételünket bebizonyítottuk.

## IRODALOM

[1] KLAFSZKY, E., „Geometriai programozás és néhány alkalmazása", MTA SZTAKI Tanulmányok, 8/1973.

[2] STOER, J. and WITZGALL, CH., *Convexity and Optimization in Finite Dimensions, I.* (Springer Verlag, 1970).

[3] TERLAKY, T., „Az $l_p$ programozásról", *Alkalmazott Matematikai Lapok* **6** (1980) 27—63.

(Beérkezett: 1983. december 21.)

TERLAKY TAMÁS
ELTE TTK OPERÁCIÓKUTATÁSI TANSZÉK
1088 BUDAPEST, MÚZEUM KRT. 6—8.

## A NEW PROOF FOR THE WEAK DUALITY THEOREM OF GEOMETRICAL AND $l_p$ PROGRAMMING

T. TERLAKY

This paper gives a new proof for the weak duality theorem of geometrical and $l_p$ programming. We prove directly the inhomogen case, so our proof is much more simple than the proofs in [1] and [3].

*Alkalmazott Matematikai Lapok* 10 (1984) 289—296

# A „CRISS-CROSS MÓDSZER" LINEÁRIS PROGRAMOZÁSI FELADATOK MEGOLDÁSÁRA ÉS VÉGESSÉGÉNEK BIZONYÍTÁSA

TERLAKY TAMÁS

Budapest

Cikkünk egy új, lineáris programozási feladatok megoldására szolgáló „*criss-cross módszer*" leírását tartalmazza. Egy általában sem primál sem duál megengedett megoldásból indulva véges számú iteráción keresztül optimális megoldáshoz jutunk, ha létezik a feladatnak optimális megoldása. Ellenkező esetben kimutatjuk, hogy nincs primál megengedett, illetve duál megengedett megoldás. Az eljárás végességét bizonyítjuk.

Az eljárás új abban a vonatkozásban is, hogy primál, illetve duál megengedett megoldás esetén sem egyezik meg a primál, illetve duál szimplex módszerrel.

## 1. Bevezetés

A „*criss-cross módszer*" létrehozása során felhasználtuk BLAND [2] és ZIONTS [5] módszereinek alapötletét, és így új, véges eljárást adunk LP feladatok megoldására. Módszerünk ZIONTS [5] módszeréhez hasonlóan nem követel meg sem primál megengedett, sem duál megengedett megoldást indulómegoldásként. Primál, illetve duál iterációkon keresztül jutunk el az optimális megoldáshoz. Így nincs szükségünk a közönséges szimplex módszer alkalmazása során fellépő első fázis feladatára, s így csak egyetlen feladatot kell megoldanunk. Eljárásunk így valószínűleg gyorsabb is lesz.

Eljárásunk abban a vonatkozásban is különbözik ZIONTS [5] módszerétől és a szimplex módszertől, hogy primál, illetve duál megengedett megoldás esetén sem egyezik meg a primál, illetve duál szimplex módszerrel. Lehetséges az is, hogy primál, illetve duál megengedett megoldás után ismét sem primál sem duál megengedett megoldást kapunk. Kimutatjuk, hogy véges számú iteráció után optimális megoldást nyerünk, illetve azt, hogy nem létezik primál, illetve duál megengedett megoldás.

A fentiekből már látható, hogy módszerünk lényegesen eltér ZIONTS [5] módszerétől. Eljárásunk végességének bizonyításában felhasználtuk BLAND [2] ötletét a szimplex módszer ciklizálásának elkerülésére.

## 2. A szimplex tábla alaptulajdonságai

Röviden közöljük a cikkünkben használt jelöléseket és a szimplex tábla alaptulajdonságait, melyeket felhasználunk bizonyításaink során.

A mátrixokat nagy, a vektorokat kis latin betűkkel jelöljük, a skalárokat és a vektorok koordinátáit a megfelelő görög betűkkel. $I_B$-vel jelöljük a **B** bázisnak megfelelő indexhalmazt.

Vizsgáljuk az alábbi lineáris programozási feladatot:

$$\min \mathbf{cx} \tag{2.1}$$

$$\mathbf{Ax} = \mathbf{b}$$

$$\mathbf{x} \geqq \mathbf{0}.$$

Ahol $\mathbf{A}$: $m \times n$ tetszőleges mátrix, $\mathbf{c}=(\gamma_1, \ldots, \gamma_n)$, $\mathbf{x}=(\xi_1, \ldots, \xi_n)$, $\mathbf{b}=(\beta_1, \ldots, \beta_m)$ tetszőleges vektorok. Egy adott $\mathbf{B}$ bázisnak megfelelően írjuk fel a szimplex táblát.

| | | $a_1 \ldots a_j \ldots a_n$ | |
|---|---|---|---|
| | | $\vdots$ | $\vdots$ |
| $i \in I_B$ | $a_i$ | $\tau_{i1} \ldots \tau_{ij} \ldots \tau_{in}$ | $\xi_i$ |
| | | $\vdots$ | $\vdots$ |
| | | $\ldots \zeta_j - \gamma_j \ldots$ | $\zeta_0$ |

*1. ábra*

Legyen $\mathbf{t}_i=(\tau_{i1}, \ldots, \tau_{in})$, $i \in I_B$ esetén és $\mathbf{d}^{(j)}=(\delta_1^{(j)}, \ldots, \delta_n^{(j)})$ $j \notin I_B$,

ahol
$$\delta_i^{(j)} = \begin{cases} \tau_{ij}, & i \in I_B \\ -1, & \text{ha } i = j \\ 0, & \text{máskor.} \end{cases}$$

Ismertek, illetve könnyen ellenőrizhetők az alábbi észrevételek.

*2.1. Megjegyzés.* Ha $\mathbf{x} \geqq \mathbf{0}$ és $\mathbf{z}-\mathbf{c} \leqq \mathbf{0}$, akkor a $\mathbf{B}$ bázis optimális.

*2.2. Megjegyzés.* Az $\mathbf{a}_i$ vektor nem primál megengedett, ha $\xi_i < 0$. Ekkor $i \in I_B$, mivel $i \notin I_B$ esetén $\xi_i = 0$.

*2.3. Megjegyzés.* Az $\mathbf{a}_j$ vektor nem duál megengedett, ha $\zeta_j - \gamma_j > 0$. Ekkor $j \notin I_B$, mivel $i \in I_B$ esetén $\zeta_i - \gamma_i = 0$.

*2.4. Megjegyzés.* Minden $i \in I_B$ esetén $\mathbf{t}_i \in \operatorname{Lin} \mathbf{A}$, mivel $\mathbf{t}_i = \mathbf{e}_k \mathbf{B}^{-1}\mathbf{A}$, ha az $\mathbf{a}_i$ vektor a $k$-adik elem a $\mathbf{B}$ bázisban.

*2.5. Megjegyzés.* Minden $j \notin I_B$ esetén $\mathbf{d}^{(j)} \perp \operatorname{Lin} \mathbf{A}$, mivel $\mathbf{A}\mathbf{d}^{(j)} = \mathbf{B}\mathbf{B}^{-1}\mathbf{a}_j - \mathbf{a}_j = 0$.

*2.6. Megjegyzés.* Mivel $z = \mathbf{c}_B \mathbf{B}^{-1}\mathbf{A}$, így $\mathbf{z} \in \operatorname{Lin} \mathbf{A}$, ahol $\mathbf{c}_B$ a $\mathbf{c}$ vektor $\mathbf{B}$ bázisnak megfelelő koordinátáit tartalmazza.

*2.7. Megjegyzés.* Tetszőleges két $\mathbf{x}''$ és $\mathbf{x}'$ megoldásra (melyekre $\mathbf{A}\mathbf{x}'' = \mathbf{b}$, $\mathbf{A}\mathbf{x}' = \mathbf{b}$) $\mathbf{x}'' - \mathbf{x}' \perp \operatorname{Lin} \mathbf{A}$, mivel $\mathbf{A}(\mathbf{x}'' - \mathbf{x}') = \mathbf{b} - \mathbf{b} = 0$.

Most két lemmát bizonyítunk, melyek biztosítani fogják eljárásunk végrehajthatóságát.

2.8. LEMMA. Ha $\mathbf{B}'$ bázis és $\xi_i' < 0$, $\tau_{ij}' \geqq 0$, $j \notin I_{B'}$, esetén, akkor nem létezik primál megengedett megoldása a (2.1) feladatnak.

*Bizonyítás.* Indirekt tegyük fel, hogy létezik $\mathbf{B}''$ primál megengedett bázis, és a neki megfelelő $\mathbf{x}''$ primál megengedett megoldás. Mivel $\mathbf{x}'_B = \mathbf{B}'^{-1}\mathbf{b} = \mathbf{B}'^{-1}\mathbf{A}\mathbf{x}''$, így

$$\xi'_i = \sum_{j=1}^{n} \tau'_{ij}\xi''_j = \xi''_i + \sum_{j \notin I_{B'}} \tau'_{ij}\xi''_j.$$

Feltevésünk szerint $\tau'_{ij} \geqq 0$, $j \notin I_{B'}$ és $\xi''_j \geqq 0$, $j=1, \ldots, n$, így $0 > \xi'_i \geqq \xi''_i$, ami ellentmondás, tehát állításunkat bebizonyítottuk.

2.9. LEMMA. Ha $\mathbf{B}'$ bázis és $\zeta'_j - \gamma_j > 0$, $\tau'_{ij} \leqq 0$, $i \in I_{B'}$ esetén, akkor nem létezik duál megengedett megoldása a (2.1) feladatnak.

*Bizonyítás.* Tegyük fel indirekt, hogy létezik $\mathbf{B}''$ duál megengedett bázis, azaz $\mathbf{z}'' - \mathbf{c} \leqq \mathbf{0}$. Így $\mathbf{c}_{B''}\mathbf{B}''^{-1}\mathbf{B}' \leqq \mathbf{c}_{B'}$. Mivel feltételeink miatt $\mathbf{B}'^{-1}\mathbf{a}_j \leqq \mathbf{0}$, így

$$\zeta''_j - \gamma_j = \mathbf{c}_{B''}\mathbf{B}''^{-1}\mathbf{a}_j - \gamma_j = (\mathbf{c}_{B''}\mathbf{B}''^{-1}\mathbf{B}')(\mathbf{B}'^{-1}\mathbf{a}_j) - \gamma_j \geqq$$
$$\geqq \mathbf{c}_{B'}\mathbf{B}'^{-1}\mathbf{a}_j - \gamma_j = \zeta'_j - \gamma_j > 0.$$

Ellentmondásra jutottunk, lemmánkat beláttuk.

Most rátérünk módszerünk ismertetésére és végességének bizonyítására.

## 3. A „criss-cross módszer"

A „*criss-cross módszer*" a szimplex tábla tömörített változatát használja. (Az egységmátrixnak megfelelő részt elhagyjuk, erre a számítási eljárás során nincs szükség.) A szimplex tábla oszlopai a bázison kívüli változóknak, sorai pedig a bázisváltozóknak felelnek meg. A bázisból kimenő és a bázisba bejövő vektor meghatározása után a szokásos módon hajtjuk végre a bázistranszformációt. A pivotálási szabály a következő.

Rögzítsük az eljárás idejére a változók egy tetszőleges sorrendjét.

*Pivotálási szabály (P):*

— Adott $\mathbf{B}$ bázis esetén keressük ki azon változókat, melyekre $\xi_i < 0$ ($\mathbf{a}_i$ nem primál megengedett) vagy $\zeta_j - \gamma_j > 0$ ($\mathbf{a}_j$ nem duál megengedett). Legyen $k$ a nem megengedett változók indexei közül a minimális.

— Ha $\xi_k < 0$, akkor $\mathbf{a}_k$ távozik a bázisból és az az $\mathbf{a}_j$ vektor kerül be a bázisba, melyre $\tau_{kj} < 0$ és $j = \min \{s | \tau_{ks} < 0, s \notin I_B\}$.

— Ha $\zeta_k - \gamma_k > 0$, akkor $\mathbf{a}_k$ a bázisba bevonandó vektor, és az az $\mathbf{a}_i$ vektor távozik a bázisból, melyre $\tau_{ik} > 0$ és $i = \min \{r | \tau_{rk} > 0, r \in I_B\}$.

A fenti kiválasztási szabály alkalmazásával hajtsuk végre rendre a bázistranszformációkat. A bázisba bejövő és a bázisból kimenő vektorok egyértelműen meghatározottak, mivel $\xi_i < 0$ és $\zeta_i - \gamma_i > 0$ egyidejűleg nem fordulhat elő, ugyanis $i \in I_B$ esetén $\zeta_i - \gamma_i = 0$ és $i \notin I_B$ esetén $\xi_i = 0$.

Eljárásunk az alábbi három eset valamelyikénél ér véget:

I. Legyen $\mathbf{a}_k$ a bázisból távozó vektor és nem tudunk a (P) szabály alkalmazásával a bázisba bevonandó vektort választani. Ekkor $\xi_k < 0$ és $\tau_{kj} \geqq 0$, $j \notin I_B$. Ami a 2.8. lemma miatt azt jelenti, hogy nincs primál megengedett megoldása a feladatnak.

II. Legyen $\mathbf{a}_k$ a bázisba bevonandó vektor és nem tudunk a (P) szabály alkalmazásával a bázisból távozó vektort választani. Ekkor $\zeta_k - \gamma_k > 0$ és $\tau_{ik} \leqq 0$, $i \in I_B$. Ami a 2.9. lemma miatt azt jelenti, hogy nincs duál megengedett megoldása a feladatnak.

III. A (P) szabály alkalmazásával sem a bázisba bevonandó, sem onnan távozó vektort nem tudunk kijelölni. Ekkor $\xi_i \geqq 0$, $i \in I_B$ és $\zeta_j - \gamma_j \leqq 0$, $j \notin I_B$, azaz optimális megoldáshoz jutottunk.

Mielőtt módszerünk végességét bizonyítanánk, eljárásunkat egy egyszerű példán illusztráljuk.

*Példa*

Az alábbi, ZIONTS [5] által bemutatott, és az általa adott *criss-cross módszerrel* is megoldott feladatot oldjuk meg.

$$\min (-3\xi_1 + 4\xi_2)$$

$$\xi_1 + 2\xi_2 \geqq 2$$

$$3\xi_1 + \xi_2 \geqq 4 \quad \xi_1 \geqq 0, \quad \xi_2 \geqq 0$$

$$\xi_1 - \xi_2 \leqq 1$$

$$\xi_1 + \xi_2 \leqq 3$$

*Induló megoldás*

| | | $\mathbf{a}_1$ | $\mathbf{a}_2$ |
|---|---|---|---|
| $\mathbf{a}_3$ | $-2$ | $-1$ | $-2$ |
| $\mathbf{a}_4$ | $-4$ | $-3$ | $-1$ |
| $\mathbf{a}_5$ | $1$ | $1$ | $-1$ |
| $\mathbf{a}_6$ | $3$ | $1$ | $1$ |
| | $0$ | $3$ | $-4$ |

*Magyarázat*

Nem megengedett az $\mathbf{a}_1, \mathbf{a}_3, \mathbf{a}_4$ változó. Primál iterációval kezdünk, mivel $\zeta_1 - \gamma_1 = 3$, azaz az $\mathbf{a}_1$ vektor nem duál megengedett. Az $\mathbf{a}_5$ vektor távozik a bázisból.

*Az első iteráció után*

| | | $\mathbf{a}_5$ | $\mathbf{a}_2$ |
|---|---|---|---|
| $\mathbf{a}_3$ | $-1$ | $1$ | $-3$ |
| $\mathbf{a}_4$ | $-1$ | $3$ | $-4$ |
| $\mathbf{a}_1$ | $1$ | $1$ | $-1$ |
| $\mathbf{a}_6$ | $2$ | $-1$ | $2$ |
| | $-3$ | $-3$ | $-1$ |

Nem megengedett az $\mathbf{a}_3, \mathbf{a}_4$ változó. Duál iteráció következik, mivel $\zeta_3 < 0$, azaz az $\mathbf{a}_3$ vektor nem primál megengedett. Az $\mathbf{a}_2$ vektor jön be a bázisba.

*A második iteráció után*

| | | $\mathbf{a}_5$ | $\mathbf{a}_3$ |
|---|---|---|---|
| $\mathbf{a}_2$ | $\frac{1}{3}$ | $-\frac{1}{3}$ | $-\frac{1}{3}$ |
| $\mathbf{a}_4$ | $\frac{1}{3}$ | $\frac{5}{3}$ | $-\frac{4}{3}$ |
| $\mathbf{a}_1$ | $\frac{4}{3}$ | $\frac{2}{3}$ | $-\frac{1}{3}$ |
| $\mathbf{a}_6$ | $\frac{4}{3}$ | $-\frac{1}{3}$ | $\frac{2}{3}$ |
| | $-\frac{8}{3}$ | $-\frac{10}{3}$ | $-\frac{1}{3}$ |

Optimális megoldás!

A feladat megoldásához eggyel kevesebb iteráció kellett, mint ZIONTS [5] módszerének alkalmazásával. A kétfázisú szimplex módszer alkalmazása esetén csak az első fázis végrehajtásához legalább két iteráció kellene.

## 4. A „criss-cross módszer" végességének bizonyítása

Az eljárás végességének bizonyításához azt kell csak belátni, hogy nem ciklizálhat, mivel véges sok különböző bázis van egy lineáris programozási feladat esetében. Így ha eljárásunk nem ciklizálhat, akkor I., II. vagy III. esetnél véges lépésben véget ér.

4.1. TÉTEL. A (P) pivotálási szabály alkalmazásával a „*criss-cross módszer*" nem ciklizálhat.

*Bizonyítás.* Tegyük fel indirekt, hogy eljárásunk ciklizál. Legyen $I^*$ azon indexek halmaza, melyek a ciklus során ki-, illetve bekerülnek a bázisba. Megjegyezzük, ha $i \notin I^*$, akkor $\mathbf{a}_i$ a ciklus során vagy végig a bázisban volt, vagy végig a bázison kívül volt.

Legyen $q = \max\{i | i \in I^*\}$. Vizsgáljuk azt az esetet, amikor $\mathbf{a}_q$ bekerül, illetve, amikor $\mathbf{a}_q$ kikerül a bázisból a ciklus során. Legyen $\mathbf{B}'$ az előbbi és $\mathbf{B}''$ az utóbbi bázis. Legyen továbbá $\mathbf{a}_r$ a bázisból távozó vektor, amikor $\mathbf{a}_q$ a bázisba bejön, illetve $\mathbf{a}_s$ a bázisba bejövő vektor, amikor $\mathbf{a}_q$ kikerül a bázisból.

Az alábbi négy eset lehetséges:

*a)* $\mathbf{a}_q$ primál iterációnál kerül be és $\mathbf{a}_q$ primál iterációnál kerül ki a bázisból,
*b)* $\mathbf{a}_q$ primál iterációnál kerül be és $\mathbf{a}_q$ duál iterációnál kerül ki a bázisból,
*c)* $\mathbf{a}_q$ duál iterációnál kerül be és $\mathbf{a}_q$ primál iterációnál kerül ki a bázisból,
*d)* $\mathbf{a}_q$ duál iterációnál kerül be és $\mathbf{a}_q$ duál iterációnál kerül ki a bázisból.

Vizsgáljuk rendre a fenti eseteket, mind a négy esetben ellentmondásra fogunk jutni, így egyik eset sem lehetséges.

*a)* Primál iterációnál, $\mathbf{B}'$ bázis esetén, $\mathbf{a}_q$ bejön a bázisba és $\mathbf{a}_r$ távozik, valamint primál iterációnál, $\mathbf{B}''$ bázis esetén, $\mathbf{a}_q$ távozik a bázisból és $\mathbf{a}_s$ kerül be a bázisba.

Tudjuk, hogy $\mathbf{z}' \in \operatorname{Lin} \mathbf{A}$ és $\mathbf{d}^{(s)'} \perp \operatorname{Lin} \mathbf{A}$ (2.6. és 2.5. megjegyzés). Így

$$0 = \mathbf{z}'\mathbf{d}^{(s)'} = \sum_{k \in I_{B''}} \zeta'_k \tau''_{ks} - \zeta'_s = \sum_{k \in I_{B''}} (\zeta'_k - \gamma_k)\tau''_{ks} - (\zeta'_s - \gamma_s) + \sum_{k \in I_{B''}} \gamma_k \tau''_{ks} - \gamma_s =$$

$$= \sum_{k \in I_{B''} \setminus I_{B'}} (\zeta'_k - \gamma_k)\tau''_{ks} - (\zeta'_s - \gamma_s) + (\zeta''_s - \gamma_s) > \sum_{k \in I_{B''} \setminus I_{B'}} (\zeta'_k - \gamma_k)\tau''_{ks}.$$

Ahol felhasználtuk, hogy $\zeta'_k - \gamma_k = 0$, $k \in I_{B'}$, valamint, hogy $\zeta'_s - \gamma_s \leqq 0$, mivel $s \in I^*$, $s < q$ és (P) szabály szerint $\mathbf{a}_q$ volt a legalacsonyabb indexű nem megengedett vektor $\mathbf{B}'$ esetén. Továbbá $\zeta''_s - \gamma_s > 0$ mivel $\mathbf{a}_s$ a bázisba bejövő vektor $\mathbf{B}''$ bázisnál.

Tudjuk, hogy $\zeta'_q - \gamma_q > 0$ és $\tau''_{qs} > 0$, mivel $\mathbf{a}_q$ a bázisba bekerülő vektor $\mathbf{B}'$ és primál iteráció mellett, valamint $\tau''_{qs}$ a pivot elem szintén primál iterációnál. Így

$$0 = \mathbf{z}'\mathbf{d}^{(s)'} > \sum_{\substack{k \in I_{B''} \setminus I_{B'} \\ k \neq q}} (\zeta'_k - \gamma_k)\tau''_{ks} + (\zeta'_q - \gamma_q)\tau''_{qs} >$$

$$> \sum_{\substack{k \in I_{B''} \setminus I_{B'} \\ k \neq q}} (\zeta'_k - \gamma_k)\tau''_{ks} \geqq 0.$$

Az utolsó egyenlőtlenség igaz, mivel $I_{B''} \setminus I_{B'} \subset I^*$, így a (P) szabály miatt $\zeta'_k - \gamma_k \leqq 0$, $\tau''_{ks} \leqq 0$, $k \in I^*$, $k \neq q$. Ellentmondásra jutottunk, tehát ez az eset nem lehetséges.

*b)* Primál iterációnál, $\mathbf{B}'$ bázis esetén, $\mathbf{a}_q$ bejön a bázisba és $\mathbf{a}_r$ távozik, valamint duál iterációnál, $\mathbf{B}''$ bázis esetén, $\mathbf{a}_q$ távozik a bázisból és $\mathbf{a}_s$ kerül be a bázisba.

Tudjuk, hogy $\mathbf{z}', \mathbf{z}'' \in \operatorname{Lin} \mathbf{A}$ (2.6. megjegyzés) és így $(\mathbf{z}'' - \mathbf{z}') \in \operatorname{lin} \mathbf{A}$ is fennáll, valamint $(\mathbf{x}'' - \mathbf{x}') \perp \operatorname{Lin} \mathbf{A}$ (2.7. megjegyzés). Így

$$0 = (\mathbf{z}' - \mathbf{z}'')(\mathbf{x}'' - \mathbf{x}') = \sum_{k=1}^{n} (\zeta'_k - \zeta''_k)(\xi''_k - \xi'_k) =$$

$$= \sum_{k=1}^{n} [(\zeta'_k - \gamma_k) - (\zeta''_k - \gamma_k)](\xi''_k - \xi'_k) =$$

$$= \sum_{k=1}^{n} (\zeta'_k - \gamma_k)(\xi''_k - \xi'_k) - \sum_{k=1}^{n} (\zeta''_k - \gamma_k)(\xi''_k - \xi'_k) =$$

$$= \sum_{k \in I_{B''} \setminus I_{B'}} (\zeta'_k - \gamma_k)\xi''_k + \sum_{k \in I_{B'} \setminus I_{B''}} (\zeta''_k - \gamma_k)\xi'_k =$$

$$= \sum_{\substack{k \in I_{B''} \setminus I_{B'} \\ k \neq q}} (\zeta'_k - \gamma_k)\xi''_k + (\zeta'_q - \gamma_q)\xi''_q + \sum_{k \in I_{B'} \setminus I_{B''}} (\zeta''_k - \gamma_k)\xi'_k.$$

Ahol felhasználtuk, hogy $\zeta_k - \gamma_k = 0$, $k \in I_B$ esetén és $\xi_k = 0$, $k \notin I_B$ esetén. Tudjuk továbbá, hogy $\zeta'_q - \gamma_q > 0$, $\xi''_q < 0$ a (P) szabály miatt. Mivel $I_{B''} \setminus I_{B'} \setminus \{q\} \subset I^* \setminus \{q\}$ és $I_{B'} \setminus I_{B''} \subset I^* \setminus \{q\}$, így a (P) szabály miatt $\zeta'_k - \gamma_k \leqq 0$, $\xi''_k \geqq 0$, $\zeta''_k - \gamma_k \leqq 0$, $\xi'_k \geqq 0$ $k \in I^* \setminus \{q\}$, azaz $0 = (\mathbf{z}' - \mathbf{z}'')\ (\mathbf{x}'' - \mathbf{x}') < 0$, ami ellentmondás, tehát ez az eset sem lehetséges.

*c)* Duál iterációnál, $\mathbf{B}'$ bázis esetén, $\mathbf{a}_q$ bejön a bázisba és $\mathbf{a}_r$ távozik, valamint primál iterációnál, $\mathbf{B}''$ bázis esetén, $\mathbf{a}_q$ távozik a bázisból és $\mathbf{a}_s$ kerül a bázisba.

Mivel az első esetben $\mathbf{a}_r$ a bázisból távozó vektor, így tudjuk, hogy $\mathbf{t}'_r \in \operatorname{Lin} \mathbf{A}$ és $\mathbf{d}^{(s)''} \perp \operatorname{Lin} \mathbf{A}$ (2.4. és 2.5. megjegyzés). Így

$$0 = \mathbf{t}'_r \mathbf{d}^{(s)''} = \sum_{k=1}^{n} \tau'_{rk} \delta_k^{(s)''} = \sum_{k \in I_{B''} \setminus I_{B'}} \tau'_{rk} \tau''_{ks} + \tau''_{rs} - \tau'_{rs} =$$

$$= \sum_{\substack{k \in I_{B''} \setminus I_{B'} \\ k \neq q}} \tau'_{rk} \tau''_{ks} + \tau'_{rq} \tau''_{qs} + \tau''_{rs} - \tau'_{rs}.$$

Ahol felhasználtuk, hogy $\tau'_{rk} = 0$, ha $k \in I_{B'} \setminus \{r\}$ és $\tau'_{rr} = 1$, valamint $\delta_k^{(s)''} = 0$, ha $k \notin I_{B''} \cup \{s\}$ és $\delta_s^{(s)''} = -1$.

Tudjuk továbbá, hogy $I_{B''} \setminus I_{B'} \setminus \{q\} \subset I^* \setminus \{q\}$, valamint a (P) szabály miatt $\tau'_{rk} \geqq 0$, $\tau''_{ks} \leqq 0$, $k \in I^* \setminus \{q\}$. Így $0 = \mathbf{t}'_r \mathbf{d}^{(s)''} \leqq \tau'_{rq} \tau''_{qs} + \tau''_{rs} - \tau'_{rs} < 0$.

Ez utóbbi egyenlőtlenség azért igaz, mivel (P) szabály miatt $\tau'_{rq} < 0$ és $\tau''_{qs} > 0$, valamint $s < q$, $r < q$, $s, r \in I^*$ és (P) szabály miatt $\tau''_{rs} \leqq 0$, $\tau'_{rs} \geqq 0$. Ellentmondásra jutottunk, így ez az eset sem lehetséges.

*d)* Duál iterációnál, $\mathbf{B}'$ bázis esetén, $\mathbf{a}_q$ bejön a bázisba és $\mathbf{a}_r$ távozik, valamint duál iterációnál, $\mathbf{B}''$ bázis esetén, $\mathbf{a}_q$ távozik a bázisból és $\mathbf{a}_s$ kerül be a bázisba.

Tudjuk, hogy $\mathbf{t}'_r \in \operatorname{Lin} \mathbf{A}$ és $(\mathbf{x}'' - \mathbf{x}') \perp \operatorname{Lin} \mathbf{A}$ (2.4. és 2.7. megjegyzés). Így

$$0 = \mathbf{t}'_r (\mathbf{x}'' - \mathbf{x}') = \sum_{k=1}^{n} \tau'_{rk} (\xi''_k - \xi'_k) = \sum_{k \in I_{B''} \setminus I_{B'}} \tau'_{rk} \xi''_k + (\xi''_r - \xi'_r).$$

Ahol felhasználtuk, hogy $\tau'_{rk} = 0$, ha $k \in I_{B'} \setminus \{r\}$, $\tau'_{rr} = 1$, $\xi'_i = 0$, $i \notin I_{B'}$ és $\xi''_i = 0$, ha $i \notin I_{B''}$.

Tudjuk (P) szabály miatt, hogy $\xi'_r < 0$ és $r < q$, $r \in I^*$ miatt $\xi''_r \geqq 0$, így

$$0 = \mathbf{t}'_r (\mathbf{x}'' - \mathbf{x}') > \sum_{\substack{k \in I_{B''} \setminus I_{B'} \\ k \neq q}} \tau'_{rk} \xi''_k + \tau'_{rq} \xi''_q > \sum_{\substack{k \in I_{B''} \setminus I_{B'} \\ k \neq q}} \tau'_{rk} \xi''_k.$$

Ahol felhasználtuk, hogy $\tau'_{rq} < 0$, $\xi''_q < 0$ a (P) szabály miatt. Mivel $I_{B''} \setminus I_{B'} \setminus \{q\} \subset I^* \setminus \{q\}$, így (P) szabály miatt $\tau'_{rk} \geqq 0$, $\xi''_k \geqq 0$, $k \in I^* \setminus \{q\}$, azaz

$$0 = \mathbf{t}'_r (\mathbf{x}'' - \mathbf{x}') > 0.$$

Ellentmondásra jutottunk, így ez az eset sem lehetséges.

Mivel a négy lehetséges közül egyik szituáció sem fordulhat elő, így bebizonyítottuk, hogy a (P) szabály alkalmazásával ciklizálás nem fordulhat elő, azaz eljárásunk véges számú iteráció után végetér.

## 5. Megjegyzések a módszer alkalmazásához

Ha eljárásunk I. vagy II. esetnél ér véget, akkor azt tudjuk, hogy nem létezik primál, illetve duál megengedett megoldás. Ekkor vizsgálhatjuk azt a kérdést, hogy az I., illetve II. esetben létezik-e duál, illetve primál megengedett megoldás. Ennek a kérdésnek az eldöntésére egyszerű eljárást tudunk adni. Az első esetben legyen $\mathbf{b} = \mathbf{0}$ ($\mathbf{x} = \mathbf{0}$), a II. esetben legyen $\mathbf{c} = \mathbf{0}$ ($\mathbf{z} - \mathbf{c} = \mathbf{0}$) a továbbiakban, és így alkalmazzuk a „*criss-cross módszert*". Ennek eredményeként, mivel eljárásunk véges, az I. esetben vagy azt kapjuk, hogy II. is fennáll, azaz duál megengedett megoldás sem létezik,

vagy duálmegengedett megoldásnál ér véget eljárásunk. A II. esetben vagy primál megengedett megoldásnál ér véget eljárásunk, vagy I. eset is előáll, azaz primál megengedett megoldás sem létezik.

Megjegyezzük, hogy $\mathbf{b}=\mathbf{0}$ $(\mathbf{x}=\mathbf{0})$, illetve $\mathbf{c}=\mathbf{0}$ $(\mathbf{z}-\mathbf{c}=\mathbf{0})$ esetén a *„criss-cross módszer"* a primál, illetve duál szimplex módszerre redukálódik, ha a szimplex módszer esetében a BLAND [2] által megfogalmazott pivotálási szabályt alkalmazzuk. Amennyiben sem $b$, sem $c$ nem egyezik meg a zérus vektorral, akkor a *„criss-cross módszer"* még akkor sem egyezik meg a primál, illetve duál szimplex módszerrel, ha primál, illetve duál megengedett megoldást kaptunk. A *„criss-cross módszer"* alkalmazása esetén előállhat az az eset is, hogy primál megengedett megoldás után ismét sem primál sem duál megengedett megoldást kapunk, illetve ugyanaz a helyzet állhat elő duál megengedett megoldás esetén is.

Egyenlőtlenséges feltételek esetén a slack változók jó induló megoldást biztosítanak módszerünkhöz, ami általában se nem primál, se nem duál megengedett. Amennyiben egyenlőséges feltételek is szerepelnek, akkor annyi pivot művelet segítségével, ahány egyenlőséges feltételünk van, megkaphatjuk az induló bázist, illetve annak inverzét.

## IRODALOM

[1] BALINSKI, M. L. and TUCKER, A. W., "Duality theory of linear programs: A constructive approach with applications", *SIAM Rewiew* **11** (1969) 347—377.
[2] BLAND, R. G., "A new pivoting rule for the simplex method", *Mathematics of Operations Research* **2** (1977) 102—108.
[3] DANTZIG, G. B., *Linear Programming and Extensions* (Princeton University Press, Princeton, 1963).
[4] PRÉKOPA, A., *Lineáris programozás* (Budapest, 1968).
[5] ZIONTS, S., "The criss-cross method for solving linear programming problems", *Management Science* **15** (1969) 426—445.
[6] ZIONTS, S., "Some empirical tests of the criss-cross method", *Management Science* **19** (1972). 406—410.

(Beérkezett: 1984. február 13.)

TERLAKY TAMÁS
EÖTVÖS LORÁND TUDOMÁNYEGYETEM OPERÁCIÓKUTATÁSI TANSZÉK
1088 BUDAPEST, MÚZEUM KRT. 6—8.

## A FINITE "CRISS-CROSS METHOD", FOR SOLVING LINEAR PROGRAMMING PROBLEMS

T. TERLAKY

Our paper treats of a new" *criss-cross method*" for solving linear programming problems. Starting from a neither primal nor dual feasible solution, we reach an optimal solution in finite number of steps if it exists. If there is no optimal solution, then we show that there is not primal feasible or dual feasible solution. We prove the finiteness of this procedure.

Our procedure is not the same as the primal or dual simplex method if we have a primal or dual feasible solution, so we have constructed a quite new procedure for solving linear programming problems.

*Alkalmazott Matematikai Lapok* **10** (1984) 297—304

# ARMA FOLYAMATOK EGZAKT SŰRŰSÉGFÜGGVÉNYE

HUHN EDIT

Szeged

A dolgozat valamely ARMA folyamatra vonatkozó statisztikai minta elemei együttes sűrűségfüggvényének meghatározásával foglalkozik. Abból a tényből indulunk ki, hogy minden ARMA folyamat valamely részlegesen megfigyelhető *több dimenziós elemi Gauss-folyamat* egyik komponense. Módszerünk lényege, hogy a nem-megfigyelhető komponenseket a *Kálmán-szűrővel* becsüljük, és a szűrőegyenleteket megoldva, a kérdéses sűrűségfüggvény meghatározható.

## 1. Bevezetés

Legyen $\zeta(n)$, $E\{\zeta(n)\}=0$ $(n=0, \pm 1, \ldots)$ *valós értékű stacionárius Gauss-folyamat,* amelynek spektrális sűrűségfüggvénye

$$f(\lambda)=\frac{1}{2\pi}\left|\frac{\sum_{j=0}^{q}\beta_j e^{i(q-j)\lambda}}{\sum_{j=0}^{p}\alpha_j e^{i(p-j)\lambda}}\right|^2,$$

ahol $\alpha_0=1$ és a $P(z)=\sum_{j=0}^{p}\alpha_j z^{p-j}=0$ egyenlet gyökei az egységkörön belül helyezkednek el. Ekkor $\zeta(n)$ megoldása egy

$$\zeta(n)+\alpha_1\zeta(n-1)+\ldots+\alpha_p\zeta(n-p)=\beta_0\varepsilon(n)+\beta_1\varepsilon(n-1)+\ldots+\beta_q\varepsilon(n-q)$$

differenciaegyenletnek [3], ahol $\varepsilon(n)$ $(n=0, \pm 1, \ldots)$, $E\{\varepsilon(n)\}=0$, $E\{\varepsilon(n^2)\}=1$, független normális eloszlású valószínűségi változók sorozata. Az ilyen folyamatot ARMA $(p, q)$ folyamatnak nevezzük.

Legyen $\Sigma_N$ a $\zeta(0), \ldots, \zeta(N-1)$ valószínűségi változók kovariancia-mátrixa. Ekkor e változók együttese sűrűségfüggvénye

$$f(x_0, \ldots, x_{N-1})=\{(2\pi)^N|\Sigma_N|\}^{-1/2}\exp\{-1/2\mathbf{x}^T\Sigma_N^{-1}\mathbf{x}\},$$

$\mathbf{x}=(x_0, \ldots, x_{N-1})^T$. Az alapvető problémát a $\Sigma_N$ inverzének, illetve determinánsának meghatározása jelenti. $\Sigma_N^{-1}$-et autoregressziós folyamatok esetén $(q=0)$ ARATÓ [9], SIDDIQUI [6], illetve PARZEN [4] határozták meg. Tiszta mozgó átlag folyamatokra $(p=0)$ vonatkozóan a kérdést a [9, 8, 5] dolgozatok tárgyalják. ARMA (1, 1) folyamatok $\Sigma_N$ kovariancia-mátrixának inverzét TIAO és ALI [7] dolgozatukban adják

meg. DZSAPARIDZE [10] a *stacionárius Gauss-folyamatok* olyan osztályára vonatkozóan vizsgálja a problémát, amelyik az ARMA folyamatokat is tartalmazza.

A dolgozatban a feladatot az előzőektől különböző módszerrel, a *Kálmán-szűrők* alkalmazásával oldjuk meg.

## 2. Az ARMA folyamatok és az elemi Gauss-folyamatok kapcsolata

2.1. DEFINÍCIÓ. A $\zeta(n)$ folyamatot *elemi Gauss-folyamatnak* nevezzük, ha *stacionárius Gauss—Markov folyamat.* Ismeretes a következő két tétel [2, 1]:

2.1. TÉTEL. A $\zeta(n)$ folyamat akkor és csakis akkor elemi *Gauss-folyamat*, ha megoldása egy

$$\zeta(n+1) = \mathbf{Q}\zeta(n)+\varepsilon(n+1), \quad n = 0, \pm 1, \ldots \tag{2.1}$$

sztochasztikus differenciálegyenletnek, ahol $\varepsilon(n)$ normális eloszlású fehér zaj, amelyre $E\{\varepsilon(n)\varepsilon(n)^T\}=\mathbf{B}_\varepsilon \neq 0$ és $\mathbf{Q}$ olyan nem-szinguláris mátrix, amelynek minden sajátértéke az egységkörön belül helyezkedik el, és $\mathbf{B}(0)=\operatorname{cov}\{\zeta(n), \zeta(n)\}$ a

$$\mathbf{B}(0) = \mathbf{Q}\mathbf{B}(0)\mathbf{Q}^T+\mathbf{B}_\varepsilon \tag{2.2}$$

mátrixegyenlet megoldása.

2.2. TÉTEL. Minden $\zeta(n)$, $E\{\zeta(n)\}=0$ *Gauss*-ARMA *folyamathoz* megadható olyan $\zeta(n)=(\zeta^1(n), \ldots, \zeta^k(n))^T$ *elemi Gauss-folyamat*, hogy $\zeta^1(n)=\zeta(n)$. A megfelelő $\zeta(n)$ folyamat ARMA $(0, q)$, illetve ARMA $(p, q)$ esetben a következő módon adható meg.

Ha $\xi(n)$ mozgó átlag folyamat, akkor a $\zeta(n)=(\zeta^1(n), \ldots, \zeta^{q+1}(n))^T$ folyamatot a

$$\begin{cases} \zeta^1(n) = \zeta(n) \\ \zeta^2(n) = \varepsilon(n) \\ \zeta^j(n) = \zeta^{j-1}(n-1), \quad 3 \leqq j \leqq q+1, \quad n = 0, \pm 1, \ldots \end{cases} \tag{2.3}$$

egyenlőségekkel definiálva egyszerű számolásokkal adódik, hogy $\zeta(n)$ kielégíti a (2.1) egyenletet, ha

$$\mathbf{Q} = \begin{bmatrix} 0 & \beta_1 & \ldots & \beta_q \\ 0 & 0 & \ldots & 0 \\ 0 & 1 & 0\ldots & 0 \\ \vdots & & & \\ 0 & \ldots & 0\,1\,0 & \end{bmatrix}_{(q+1)\times(q+1)}, \quad \varepsilon(n) = \begin{bmatrix} \beta_0\varepsilon(n) \\ \varepsilon(n) \\ 0 \\ \vdots \\ 0 \end{bmatrix}_{(q+1)\times 1}.$$

ARMA $(p, q)$ folyamat esetén pedig $\zeta(n)$ legyen a következő:

$$\begin{cases} \zeta^1(n) = \zeta(n) \\ \zeta^j(n) = \zeta^{j-1}(n+1), \quad j = 2, \ldots, p \\ \zeta^{p+1}(n) = \varepsilon(n+p-1) \\ \zeta^j(n) = \zeta^{j-1}(n-1), \quad j = p+2, \ldots, p+q. \end{cases} \tag{2.4}$$

Könnyen látható, hogy a (2.4)-gyel definiált $\zeta(n)$ megoldása (2.1)-nek és

$$\mathbf{Q} = \begin{bmatrix} 0 & 1 & 0.. & 0 & & 0 & \dots & 0 \\ \vdots & & & & & & & \\ 0 & & 0.. & 0 & 1 & 0 & \dots & 0 \\ -\alpha_p & & \dots & -\alpha_2 & -\alpha_1 & \beta_1 & \dots & \beta_q \\ 0 & & \dots & & 0 & 0 & \dots & 0 \\ 0 & & \dots & & 0 & 1 & 0.. & 0 \\ \vdots & & & & & & & \\ 0 & & \dots & & 0 & 0 & ..1 & 0 \end{bmatrix}_{(p+q)\times(p+q)},$$

$$\boldsymbol{\varepsilon}(n) = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \beta_0\varepsilon(n+p-1) \\ \varepsilon(n+p-1) \\ 0 \\ \vdots \\ 0 \end{bmatrix}_{(p+q)\times 1}.$$

## 3. A Kálmán-szűrő általánosítása

Legyen $(\mathfrak{I}(n), \xi(n))$ $(n=0, \pm 1, \dots)$, $\mathfrak{I}(n)=(\vartheta_1(n), \dots, \vartheta_k(n))^T$, $\xi(n)=(\xi_1(n), \dots, \xi_l(n))^T$ olyan részlegesen megfigyelhető folyamat ($\xi(n)$ a megfigyelhető rész), amelyre

$$(3.1)\quad \begin{cases} \mathfrak{I}(n+1) = \mathbf{a}_0(n,\xi)+a_1(n,\xi)\mathfrak{I}(n)+b_1(n,\xi)\boldsymbol{\varepsilon}_1(n+1)+b_2(n,\xi)\boldsymbol{\varepsilon}_2(n+1) \\ \xi(n+1) = \mathbf{A}_0(n,\xi)+A_1(n,\xi)\mathfrak{I}(n)+B_1(n,\xi)\boldsymbol{\varepsilon}_1(n+1)+B_2(n,\xi)\boldsymbol{\varepsilon}_2(n+1) \end{cases}$$

differenciálegyenlet-rendszer megoldása, ahol

$$\boldsymbol{\varepsilon}_1(n) = (\varepsilon_{11}(n), \dots, \varepsilon_{1k}(n))^T, \quad \boldsymbol{\varepsilon}_2(n) = (\varepsilon_{21}(n), \dots, \varepsilon_{2l}(n))^T$$

független komponensű, független *Gauss-vektorok* és

$$E\{\varepsilon_{1i}(n)\} = E\{\varepsilon_{2j}(n)\} = 0, \quad E\{|\varepsilon_{1i}(n)|^2\} = E\{|\varepsilon_{2j}(n)|^2\} = 1,$$
$$i = 1, \dots, k, \quad j = 1, \dots, l.$$

A $(\mathfrak{I}_0, \xi_0)$ kezdeti vektorról feltesszük, hogy $(\boldsymbol{\varepsilon}_1(n), \boldsymbol{\varepsilon}_2(n))$-től független $(n=1, 2, \dots)$.

Legyen $\mathbf{m}_n = E\{\mathfrak{I}(n)|F_n^\xi\}$ a feltételes várhatóérték-vektor $(F_n^\xi = \sigma\{\xi(s), s=0, \dots, n\})$ és

$$\gamma_n = E\{(\mathfrak{I}(n)-\mathbf{m}_n)(\mathfrak{I}(n)-\mathbf{m}_n)^T|F_n^\xi\}.$$

A (3.1) egyenletben szereplő $a_i, A_i, b_i, B_i$ együtthatókra tett bizonyos feltevések mellett érvényes a következő tétel [11]:

3.1. TÉTEL. Ha a $(\vartheta(n), \xi(n))$ *Gauss-folyamat* eleget tesz a (3.1) egyenletrendszernek, akkor az $\mathbf{m}_n$ vektor, illetve a $\gamma_n$ mátrix a következő rekurzív egyenletekből határozható meg (az egyszerűség kedvéért az együtthatók argumentumait nem írjuk ki):

$$\mathbf{m}_n = [\mathbf{a}_0 + a_1 \mathbf{m}_{n-1}] + \tag{3.2}$$

$$+[b \circ B + a_1 \gamma_{n-1} A_1^T][B \circ B + A_1 \gamma_{n-1} A_1^T]^+ [\xi(n) - \mathbf{A}_0 - A_1 \mathbf{m}_{n-1}],$$

$$\gamma_n = [a_1 \gamma_{n-1} a_1^T + b \circ b] - \tag{3.3}$$

$$-[b \circ B + a_1 \gamma_{n-1} A_1^T][B \circ B + A_1 \gamma_{n-1} A_1^T]^+ [b \circ B + a_1 \gamma_{n-1} A_1^T]^T,$$

és $\mathbf{m}_0 = E\{\vartheta_0 | \xi_0\}$, $\gamma_0 = E\{(\vartheta_0 - \mathbf{m}_0)(\vartheta_0 - \mathbf{m}_0)^T | \xi_0\}$. A fenti tételben

$$b \circ b = b_1 b_1^T + b_2 b_2^T$$

$$b \circ B = b_1 B_1^T + b_2 B_2^T$$

$$B \circ B = B_1 B_1^T + B_2 B_2^T,$$

és $[B \circ B + A_1 \gamma_{n-1} A_1^T]^+$ az illető mátrix pszeudoinverzét jelöli. LIPCER és SIRJAJEV [11] könyvükben megmutatják továbbá, hogy a (3.2) egyenlet jobb oldalán szereplő $\xi(n) - \mathbf{A}_0 - A_1 \mathbf{m}_{n-1}$ tényező egy valószínűséggel fehér zaj folyamat, amelynek kovariancia-mátrixa $B \circ B + A_1 \gamma_{n-1} A_1^T$.

## 4. A Kálmán-szűrő alkalmazása ARMA folyamatok sűrűségfüggvényének meghatározására

A 2. szakaszban a (2.3), illetve a (2.4) egyenletekkel definiált $\zeta(n)$ folyamatok olyanok, hogy az első esetben $\zeta^1(n)$, a másodikban $\zeta^1(n), \ldots, \zeta^p(n)$ megfigyelhetők. Ha $\xi(n)$-nel, illetve $\vartheta(n)$-nel jelöljük a megfigyelhető, illetve a nem-megfigyelhető komponensekből álló vektorokat, akkor látható, hogy mindkét esetben a megfelelő (2.1) egyenlet (3.1) alakba írható, és az $a_1, A_1, b_j, B_j$ $(j = 1, 2)$ együtthatók konstansok, az $\mathbf{a}_0(n, \xi(n))$, $\mathbf{A}_0(n, \xi(n))$ pedig $\xi(n)$-nek lineáris függvényei. Ekkor könnyen belátható, hogy $\gamma_n = E\{(\vartheta(n) - \mathbf{m}_n)(\vartheta(n) - \mathbf{m}_n)^T | F_n^\xi\} = E\{(\vartheta(n) - \mathbf{m}_n)(\vartheta(n) - \mathbf{m}_n)^T\}$, és a 3. szakaszban foglaltak alapján $\mathbf{m}_n$, $\gamma_n$ a (3.2), illetve a (3.3) egyenleteknek az

$$\mathbf{m}_0 = E\{\vartheta_0 | \xi_0\} = D_{\vartheta\xi} D_{\xi\xi}^+ \xi_0, \tag{4.1}$$

$$\gamma_0 = E\{(\vartheta_0 - \mathbf{m}_0)(\vartheta_0 - \mathbf{m}_0)^T\} = D_{\vartheta\vartheta} - D_{\vartheta\xi} D_{\xi\xi}^+ D_{\vartheta\xi}^T \tag{4.2}$$

kezdeti értékekhez tartozó megoldásai, ahol

$$D_{\vartheta\vartheta} = \operatorname{cov}\{\vartheta_0, \vartheta_0\}, \quad D_{\vartheta\xi} = \operatorname{cov}\{\vartheta_0, \xi_0\}, \quad D_{\xi\xi} = \operatorname{cov}\{\xi_0, \xi_0\}.$$

A

$$\mathbf{B}(0) = \operatorname{cov}\{\xi(0), \xi(0)\} = \begin{bmatrix} D_{\xi\xi} & D_{\vartheta\xi} \\ D_{\vartheta\xi}^T & D_{\vartheta\vartheta} \end{bmatrix}$$

mátrix pedig (2.2)-ből határozható meg.

Az előző szakasz végén mondottak szerint

$$\xi(n+1) = \mathbf{A}_0(n, \xi) + A_1 \mathbf{m}_n + [B \circ B + A_1 \gamma_n A_1^T]^{1/2} \tilde{\varepsilon}(n+1) \tag{4.3}$$

egy valószínűséggel teljesül, ahol az $\tilde{\varepsilon}(n)$-ek független komponensű *Gauss-vektorok* és $E\{\tilde{\varepsilon}(n)\}=0$, $E\{\tilde{\varepsilon}(n)\tilde{\varepsilon}(k)^T\}=\delta_{nk}\mathbf{I}$. Mivel $\mathbf{m}_n$ a $\xi(0), \ldots, \xi(n)$ vektorok lineáris függvénye, (4.3) a következő módon írható:

$$\xi(n+1) = \sum_{i=0}^{n} M_i^{(n)}\xi(i)+\mathbf{E}_n^{1/2}\tilde{\varepsilon}(n+1) \quad (n = 0, 1, \ldots), \tag{4.4}$$

ahol $\mathbf{E}_n=B\circ B+A_1\gamma_n A_1^T$.

(4.4) alapján a $\xi(0), \ldots, \xi(N-1)$ vektorváltozók együttes sűrűségfüggvénye

$$f(\mathbf{x}_0, \ldots, \mathbf{x}_{N-1}) = \left\{(2\pi)^{Nl}|D_{\xi\xi}| \prod_{i=1}^{N-1}|E_{i-1}|\right\}^{-1/2}\times$$

$$\times\exp\left\{-\frac{1}{2}\sum_{i=1}^{N-1}\Big(\sum_{s=0}^{i-1} M_s^{(i-1)}\mathbf{x}_s-\mathbf{x}_i\Big)^T E_{i-1}^{-1}\Big(\sum_{s=0}^{i-1} M_s^{(i-1)}\mathbf{x}_s-\mathbf{x}_i\Big)+\mathbf{x}_0^T D_{\xi\xi}^{-1}\mathbf{x}_0\right\}, \tag{4.5}$$

$\mathbf{E}_i=B\circ B+A_1\gamma_i A_1^T$, és $l$ a megfigyelhető komponensek száma. (4.5)-ből a $\zeta(0), \ldots, \zeta(N-1)$ együttes sűrűségfüggvénye meghatározható, mivel $\zeta(i)$ a $\xi(i)$ vektor első komponense $(i=0, \ldots, N-1)$.

## 5. Példák

1. *Elsőrendű mozgóátlag folyamat.* Ekkor $\zeta(n)$ a

$$\zeta(n) = \beta_0\varepsilon(n)+\beta_1\varepsilon(n-1)$$

egyenlet megoldása. A (2.1) egyenlet, amelynek a $\zeta(n)=(\zeta^1(n), \zeta^2(n))^T$ a megoldása, ebben az esetben a következő:

$$\begin{cases}\zeta^1(n+1) = \beta_1\zeta^2(n)+\beta_0\varepsilon(n+1)\\ \zeta^2(n+1) = \varepsilon(n+1).\end{cases} \tag{5.1}$$

Nyilván $\zeta^1(n)=\zeta(n)$ a megfigyelhető komponens (az előző szakasz jelöléseivel $\xi(n)=\zeta^1(n)$ és $\vartheta(n)=\zeta^2(n)$). (5.1)-et összevetve (3.1)-gyel kapjuk, hogy $\mathbf{a}_0(n, \xi)=a_1(n, \xi)=\mathbf{A}_0(n, \xi)=0$, $A_1=\beta_1$, $b_1=1$, $B_1=\beta_0$, $b_2=B_2=0$ és $\gamma_n$ valamint $m_n$ a

$$\beta_1^2\gamma_{n+1}\gamma_n+\beta_0^2\gamma_{n+1}-\beta_1^2\gamma_n = 0$$

$$m_{n+1} = \frac{\beta_0}{\beta_0^2+\beta_1^2\gamma_n}\zeta(n+1)-\frac{\beta_0\beta_1}{\beta_0^2+\beta_1^2\gamma_n}m_n \tag{5.2}$$

egyenleteknek a

$$\gamma_0 = \frac{\beta_1^2}{\beta_1^2+\beta_0^2}, \quad m_0 = \frac{\beta_0^2}{\beta_1^2+\beta_0^2}\zeta(0)$$

kezdeti feltételekhez tartozó megoldása, azaz

$$\gamma_n = \gamma_0 \frac{\left(\frac{\beta_1}{\beta_0}\right)^{2n}}{1+\gamma_0 \sum_{j=1}^{n}\left(\frac{\beta_1}{\beta_0}\right)^{2j}},$$

$$m_n = \left\{\prod_{j=0}^{n-1} \frac{-\beta_0\beta_1}{\beta_0^2+\beta_1^2\gamma_j}\right\} m_0 + \sum_{s=1}^{n} \frac{\beta_0}{\beta_0^2+\beta_1^2\gamma_{s-1}} \left\{\prod_{j=s}^{n-1} \frac{-\beta_0\beta_1}{\beta_0^2+\beta_1^2\gamma_j}\right\} \zeta(s).$$

(4.3) szerint

(5.3) $$\zeta(n+1) = \beta_1 m_n + [\beta_0^2+\beta_1^2\gamma_n]^{1/2}\tilde{\varepsilon}(n+1).$$

(5.3)-at véve $n=0, 1, \ldots, N-1$-re adódik, hogy

(5.4)
$$\begin{cases} \zeta(1) = M_0^{(0)}\zeta(0) + [\beta_0^2+\beta_1^2\gamma_0]^{1/2}\tilde{\varepsilon}(1) \\ -\sum_{s=1}^{n-1} M_s^{(n-1)}\zeta(s) + \zeta(n) = M_0^{(n-1)}\zeta(0) + [\beta_0^2+\beta_1^2\gamma_{n-1}]^{1/2}\tilde{\varepsilon}(n), \quad n = 2, \ldots, N-1, \end{cases}$$

ahol

$$M_0^{(0)} = \frac{\left(\frac{\beta_1}{\beta_0}\right)^2}{1+\left(\frac{\beta_1}{\beta_0}\right)^2},$$

$$M_0^{(n-1)} = (-1)^{n-1} \frac{\left(\frac{\beta_1}{\beta_0}\right)^n}{\left[1+\left(\frac{\beta_1}{\beta_0}\right)^2\right] \prod_{j=0}^{n-2}\left[1+\left(\frac{\beta_1}{\beta_0}\right)^2\gamma_j\right]}, \quad n = 2, \ldots, N-1$$

$$M_s^{(n-1)} = (-1)^{n-s-1} \frac{\left(\frac{\beta_1}{\beta_0}\right)^{n-s}}{\prod_{j=s-1}^{n-2}\left[1+\left(\frac{\beta_1}{\beta_0}\right)^2\gamma_j\right]}, \quad n = 2, \ldots, N-1, \quad s = 1, \ldots, n-1.$$

(5.4) alapján $\zeta(0), \ldots, \zeta(N-1)$ együttes sűrűségfüggvénye

$$f(x_0, \ldots, x_{N-1}) = \left\{(2\pi)^N D_{\xi\xi} \prod_{j=0}^{N-2} (\beta_0^2+\beta_1^2\gamma_j)\right\}^{-1/2} \times$$

$$\times \exp\left\{-\frac{1}{2}\left[\frac{x_0^2}{D_{\xi\xi}} + \sum_{j=0}^{N-1} \frac{\left(x_j - \sum_{s=0}^{j-1} M_s^{(j-1)} x_s\right)^2}{\beta_0^2+\beta_1^2\gamma_{j-1}}\right]\right\},$$

és $D_{\xi\xi}=\beta_0^2+\beta_1^2$.

2. ARMA $(p, 1)$ *folyamat.* Most $\zeta(n)$ a

$$\zeta(n)+\alpha_1\zeta(n-1)+\ldots+\alpha_p\zeta(n-p) = \beta_0\varepsilon(n)+\beta,\varepsilon(n-1)$$

egyenlet megoldása. A (2.1) egyenlet megfelelője

$$(5.5)\qquad \begin{cases} \zeta^j(n+1) = \zeta^{j+1}(n), \quad j = 1, \ldots, p-1, \\ \zeta^p(n+1) = -\sum_{j=1}^{p} \alpha_j \zeta^{p+1-j}(n) + \beta_1 \zeta^{p+1}(n) + \beta_0 \varepsilon(n+p), \\ \zeta^{p+1}(n+1) = \varepsilon(n+p). \end{cases}$$

Nyilván $\zeta^1(n), \ldots, \zeta^p(n)$ a megfigyelhető komponensek és $\zeta^1(n) = \zeta(n)$. A (3.2), (3.3) egyenletek megoldásai ebben az esetben

$$\gamma_n = \gamma_0 \frac{\left(\frac{\beta_1}{\beta_0}\right)^{2n}}{1 + \gamma_0 \sum_{j=1}^{n} \left(\frac{\beta_1}{\beta_0}\right)^{2j}},$$

$$m_{n+1} = \frac{(-\beta_0\beta_i)^{n+1}}{\prod_{j=0}^{n} (\beta_0^2 + \beta_1^2 \gamma_j)} m_0 + \sum_{s=0}^{n+p-1} \left\{ \sum_{j=\max\{1, p-s\}}^{\min\{p, n+p-s\}} \alpha_j \frac{\beta_0(-\beta_0\beta_1)^{n+p-j-s}}{\prod_{l=s+j-p}^{n} (\beta_0^2 + \beta_1^2 \gamma_l)} \right\} \zeta(s) +$$

$$+ \sum_{s=p}^{n+p} \frac{\beta_0(-\beta_0\beta_1)^{n-s+p}}{\prod_{j=s-p}^{n} (\beta_0^2 + \beta_1^2 \gamma_j)} \zeta(s) = \sum_{s=0}^{n+p} K_s^{(n+1)} \zeta(s),$$

figyelembe véve, hogy (4.1) szerint az $m_0$ kezdeti érték $\sum_{s=0}^{p-1} K_s^{(0)} \zeta(s)$ alakú.

$\zeta(n) = (\zeta^1(n), \ldots, \zeta^p(n))^T$ utolsó komponensére most a következő előállítás adódik:

$$(5.6)\qquad \zeta^p(n+1) = -\sum_{j=1}^{p} \alpha_j \zeta^{p+1-j}(n) + \beta_1 m_n + (\beta_0^2 + \beta_1^2 \gamma_n)^{1/2} \tilde{\varepsilon}_p(n+1).$$

(5.6)-ot véve $n = 0, 1, \ldots, N-p-1$-re kapjuk, hogy

$$(5.7)\qquad \begin{cases} \zeta(p) = \sum_{s=0}^{p-1} M_s^{(p-1)} \zeta(s) + (\beta_0^2 + \beta_1^2 \gamma_0)^{1/2} \tilde{\varepsilon}_p(1), \\ \sum_{s=p}^{n-1} M_s^{(n-1)} \zeta(s) + \zeta(n) = \sum_{s=0}^{p-1} M_s^{(n-1)} \zeta(s) + (\beta_0^2 + \beta_1^2 \gamma_{n-p})^{1/2} \tilde{\varepsilon}_p(n+1-p) \\ \qquad\qquad n = p+1, \ldots, N-1, \end{cases}$$

$$M_s^{(p-1)} = \beta_1 K_s^{(0)} - \alpha_{p-s}, \quad 0 \leqq s \leqq p-1$$

$$M_s^{(n-1)} = \begin{cases} \left.\begin{array}{ll} \beta_1 K_s^{(n-p)}, & 0 \leqq s \leqq n-p-1 \\ \beta_1 K_s^{(n-p)} - \alpha_{n-s}, & n-p \leqq s \leqq p-1 \\ \alpha_{n-s}, & p \leqq s \leqq n-1 \end{array}\right\} p+1 \leqq n \leqq 2p-1 \\ \left.\begin{array}{ll} \beta_1 K_s^{(n-p)}, & 0 \leqq s \leqq p-1 \\ -\beta_1 K_s^{(n-p)}, & p \leqq s \leqq n-p-1 \\ \alpha_{n-s} - \beta_1 K_s^{(n-p)}, & n-p \leqq s \leqq p-1 \end{array}\right\} 2p \leqq n \leqq N-1 \end{cases}$$

(5.7)-ből a keresett sűrűségfüggvény már felírható:

$$f(x_0, \ldots, x_{N-1}) = \left\{(2\pi)^N \prod_{j=0}^{N-p-1} (\beta_0^2+\beta_1^2\gamma_j)\,|\right\} D_{\xi\xi}|^{-1/2} \times$$

$$\times \exp\left\{-\frac{1}{2}\sum_{j=p}^{N-1} \frac{\left[\sum_{s=p}^{j-1} M_s^{(j-1)}x_s + x_j - \sum_{s=0}^{p-1} M_s^{(j-1)}x_s\right]^2}{\beta_0^2+\beta_1^2\gamma_{j-p}} + \mathbf{x}_0^T\, \mathbf{D}_{\xi\xi}^{-1}\mathbf{x}_0\right\},$$

$$\mathbf{x}_0 = (x_0, \ldots, x_{p-1})^T.$$

Végezetül meg kívánom köszönni ARATÓ MÁTYÁSnak, hogy a kérdésre figyelmemet felhívta és munkám elkészítése során értékes tanácsaival támogatott.

## IRODALOM

[1] ARATÓ, M., BENCZÚR, A., KRÁMLI, A. and PERGEL, J., "Statistical problems of the elementary Gaussian processes" (MTA SZTAKI Tanulmányok, 1975/41).
[2] ARATÓ, M., *Linear Stochastic Systems with Constant Coefficients* (Lecture Notes in Control and Information Sciences 45., Springer Berlin—Heidelberg—New York, 1982).
[3] DOOB, J. L., *Stochastic Processes* (New York John Wiley & Sons, London—Chapman Hall, 1953).
[4] PARZEN, E., "An approach to time series analysis", *Ann. Math. Statist.* **32** (1961) 951—989.
[5] SHAMAN, P., "On the inverse of the covariance matrix of a first order moving average", *Biometrika* **56** (1969) 595—600.
[6] SIDDIQUI, M. M., "On the inversion of the sample covariance matrix in a stationary autoregressive process", *Ann. Math. Statist.* **29** (1958) 585—588.
[7] TIAO, G. C. and ALI, M. M., "Analysis of correlated random effects: linear models with two random components" *Biometrika* **58** (1971) 37—51.
[8] UPPULURI, V. R. R. and CARPENTER, J. A., "The inverse of a matrix occouring in first order moving average models", *Sankhyā, Ser. A.* **31** (1961) 79—82.
[9] ARATÓ, M., "Sufficient statistics for stationary Gaussian processes", *Theory of Probability and Applic.* **6** (1961) 216—218 (in Russian).
[10] Джапаридзке, К. О., Оценка параметров и проверка гипотез в спектральном анаиизе с тационарных временных рядов (Изд-во Тбилиского Университета, Тбилиси, 1981).
[11] Липцер, Р. Ш., Ширяев, А. Н., Статистика случайных процессов (Наука, Москва, 1974).

(Beérkezett: 1984. május 11.)

HUHN EDIT
SZEGEDI ORVOSTUDOMÁNYI EGYETEM SZÁMÍTÁSTECHNIKAI KÖZPONT
6720 SZEGED, PÉCSI U. 4/A.

## ON THE EXACT LIKELIHOOD FUNCTION OF ARMA RPOCESSES

E. HUHN

In the paper the generalized *Kálmán filter* is proposed to get the exact likelihood function of a stationary autoregressive moving average process.

*Alkalmazott Matematikai Lapok* **10** (1984) 305—321

# LINEÁRIS EGYÜTTHATÓJÚ DIFFÚZIÓS FOLYAMATOK PARAMÉTEREINEK BECSLÉSE

KONCZ KÁROLY

A dolgozatban az (1.1) sztochasztikus vektoregyenletet vizsgáljuk. Bebizonyítjuk a 4.1. tételt, amely egyszerű előállítást ad az elégséges statisztikák eloszlásának a *Laplace transzformáltjára*. Ez NOVIKOV [3] egy dimenzióra elért eredményének az általánosítása. A stacionárius esetben bebizonyítjuk az 5.2. tételt, mellyel általánosítjuk ARATÓ és BENCZÚR eredményét (lásd [1]).

## 1. Bevezetés

Ebben a cikkben egy módszert fogunk bemutatni, amelynek segítségével megadhatjuk többdimenziós lineáris együtthatójú diffúziós folyamatok paramétereire vonatkozó elégséges statisztikák és becslések eloszlását, illetve *Laplace transzformáltját*.

A felhasznált eljárással kapcsolatban utalunk NOVIKOV [3] 1972-ben megjelent dolgozatára, amelyben módszert és pontos formulát közöl egydimenziós, ill. kétdimenziós „szimmetrikus" (komplex) folyamat esetén a maximum-likelihood becslések eloszlásáról és *Laplace transzformáltjáról*. A pontos eloszlások meghatározása speciális esetekben szerepel az [1] és [7] könyvekben.

A többdimenziós esetben csak asszimptotikus közelítő formulák voltak ismertek, az erre vonatkozó irodalom megtalálható az [1], [2], [4], [5], [7] könyvekben.

## 2. Az elégséges statisztikák

Vizsgáljuk a következő sztochasztikus vektoregyenletet

$$d\zeta(t) = \mathbf{A}\zeta(t)\,dt + B_{\mathbf{W}}^{1/2}\,d\mathbf{W}(t), \quad 0 \leqq t \leqq T, \tag{2.1}$$

ahol $\zeta(t)=(\zeta_1(t), \ldots, \zeta_n(t))^*$ $n$ dimenziós sztochasztikus folyamat,
$\mathbf{W}(t)=(W_1(t), \ldots, W_n(t))^*$ $n$ dimenziós standard *Wiener folyamat*,
$\mathbf{B}_{\mathbf{W}}^{1/2} \neq \mathbf{0}$ pozitív szemidefinit $n\times n$-es konstans mátrix, melynek rangja $r \leqq n$,
$\mathbf{A}$ $n\times n$-es konstans mátrix, melynek sajátértékei a negatív (komplex) félsíkban vannak, és
$F_t^{\mathbf{W}}=\sigma\{\mathbf{W}(s),\ 0 \leqq s \leqq t\}$ a $\mathbf{W}(s)$ folyamat által generált $\sigma$-algebra $(s \leqq t)$.

Legyen $\zeta(0)$ független az $F_T^{\mathbf{W}}$-től, $n$ dimenziós (esetleg elfajuló) *Gauss eloszlású* valószínűségi változó, amelyre $E\zeta(0)=\mathbf{0}$ teljesül.

A $\mathbf{B}_{\mathbf{W}}^{1/2}$ mátrixot ismerjük, a $\zeta(t)$ folyamatot megfigyeljük a $[0, T]$-n, ebből becsüljük az $\mathbf{A}$ mátrix elemeit. Ehhez a következő tétel nyújt segítséget (l. [1] 2.3. § 2. tétel (2.3.25) formula).

2.1. TÉTEL. A $C^r[0, T]$-n generált $P_\zeta$ és $P_w$ mértékek ekvivalensek, és

$$(2.2)\quad \frac{dP_\zeta}{dP_W}(T, \zeta) = f(\zeta(0)) \exp\left\{\int_0^T \zeta^*(t)\mathbf{A}^*\mathbf{B}^+ d\zeta(t) - \frac{1}{2}\int_0^T \zeta^*(t)\mathbf{A}^*\mathbf{B}^+\mathbf{A}\zeta(t)dt\right\},$$

ahol $\mathbf{B} = \mathbf{B}_W^{1/2}\mathbf{B}_W^{1/2*}$ és $\mathbf{B}^+$ a $\mathbf{B}$ mátrix ún. általánosított inverze, valamint $f$ a $\zeta(0)$ sűrűségfüggvénye (a *Lebesgue mérték* szerint).

*Megjegyzés.* (2.2)-ből látható, hogy az $\mathbf{A}$-ra vonatkozó feltételes $(\zeta(0)=x)$ elégséges statisztika a

$$(2.3)\quad \left\{\int_0^T \zeta_i(t)\zeta_j(t)dt,\ \int_0^T \zeta_k(t)d\zeta_l(t) \quad i, j, k, l = 1, 2, \ldots, n\right\}.$$

Az $\mathbf{A}$-ra vonatkozó elégséges statisztika pedig a

(2.4)

$$\left\{\int_0^T \zeta_i(t)\zeta_j(t)dt,\ \int_0^T \zeta_k(t)d\zeta_l(t),\ \zeta_m(0)\zeta_s(0),\ \zeta_p(T)\zeta_q(T) \quad i, j, k, l, m, s, p, q = 1 \ldots n\right\}$$

kifejezés lesz (pl. stacionárius indításnál).

Így az $\mathbf{A}$-ra vonatkozó becslésekhez elegendő a (2.3), ill. (2.4)-ben szereplő valószínűségi változók együttes eloszlását megadni.

Legyen most

$$(2.5)\quad d\boldsymbol{\eta}(t) = a\boldsymbol{\eta}(t)dt + \mathbf{B}_W^{1/2}\, d\mathbf{W}(t), \quad 0 \leqq t \leqq T,$$

és tegyük fel, hogy $\boldsymbol{\eta}(0)=\zeta(0)$.

A továbbiakban gyakran fogjuk alkalmazni a 2.1. tétel egy másik formáját (l. [7], (7.138) formula).

2.2. TÉTEL. Ha az $\mathbf{A}, \mathbf{a}, \mathbf{B}_W^{1/2}$ mátrixokra teljesül, hogy létezik olyan $\mathbf{F}$ mátrix, amelyre

$$(2.6)\quad \mathbf{B}_W^{1/2}\mathbf{F} = \mathbf{A} - \mathbf{a}$$

igaz, akkor a $P_\zeta$ és a $P_\eta$ mértékek ekvivalensek, a $C^r[0, T]$ téren és

(2.7)

$$\frac{dP_\zeta}{dP_\eta}(T, \boldsymbol{\eta}) = \exp\left\{\int_0^T \boldsymbol{\eta}^*(t)(\mathbf{A}-\mathbf{a})^*\mathbf{B}^+ d\boldsymbol{\eta}(t) - \frac{1}{2}\int_0^T \boldsymbol{\eta}^*(t)(\mathbf{A}^*\mathbf{B}^+\mathbf{A} - \mathbf{a}^*\mathbf{B}^+\mathbf{a})\boldsymbol{\eta}(t)dt\right\} =$$

$$= \exp\left\{\int_0^T \boldsymbol{\eta}^*(t)(\mathbf{A}-\mathbf{a})^*\mathbf{B}^+\mathbf{B}_w^{1/2} d\mathbf{W}(t) - \frac{1}{2}\int_0^T \boldsymbol{\eta}^*(t)(\mathbf{A}-\mathbf{a})^*\mathbf{B}^+(\mathbf{A}-\mathbf{a})\boldsymbol{\eta}(t)dt\right\}.$$

Gyakran fogjuk használni a

$$\int_0^T \mathbf{x}\, d\mathbf{x}, \quad \text{ill.} \quad \int_0^T \mathbf{x}\, dt$$

jelöléseket. Ezek nem közönséges integrálokat jelölnek, hanem a $(C^r[0, T], B^r[0, T])$ mérhető tér egy-egy mérhető függvényét, amelyet az $\int_0^T \zeta(t)d\zeta(t)$ és $\int_0^T \eta(t)d\eta(t)$, ill. az $\int_0^T \eta(t)dt$ és $\int_0^T \zeta(t)dt$ generálnak. Mivel a $P_\zeta$ és $P_\eta$ mértékek ekvivalensek, a generált funkcionálok jól értelmezettek (l. [7] 4.10. lemma).

Legyen most

$$\psi_T(\mathbf{A}, \mathbf{C}) = E\exp\left\{\int_0^T \xi^*(t)\mathbf{C}\xi(t)dt\right\} = \tag{2.8}$$

$$= E\exp\left\{\sum_{i,j=1}^{n} c_{ij}\int_0^T \zeta_i(t)\zeta_j(t)\,dt\right\} = E_{\mathbf{A}}\exp\left\{\int_0^T \mathbf{x}^*\mathbf{C}\mathbf{x}\,dt\right\}$$

az $\left\{\int_0^T \zeta_i(t)\zeta_j(t)dt \quad i,j=1,\ldots,n\right\}$ valószínűségi változók együttes eloszlásának *Laplace transzformáltja* és $\mathbf{C}$ negatív szemidefinit szimmetrikus mátrix.

A $\psi_T(\mathbf{A}, \mathbf{C})$ függvény meghatározza a (2.3)-ban szereplő elégséges statisztika együttes eloszlásának *Laplace transzformáltját* is, igaz ugyanis a következő:

2.1. ÁLLÍTÁS.
$$E\exp\left\{\int_0^T \zeta^*(t)\mathbf{C}\zeta(t)dt + \int_0^T \zeta^*(t)\,Gd\zeta(t)\right\} =$$

$$= \psi_T\left(\mathbf{A}+\mathbf{BG}^*,\ \mathbf{C}+\frac{1}{2}(\mathbf{GA}+\mathbf{A}^*\mathbf{G}^*+\mathbf{GBG}^*)\right).$$

*Bizonyítás.* A 2.2 tétel alapján

$$E\exp\left\{\int_0^T \zeta^*(t)\mathbf{C}\zeta(t)dt + \int_0^T \zeta^*(t)\mathbf{G}\,d\zeta(t)\right\} =$$

$$= E_{\mathbf{A}}\exp\left\{\int_0^T \mathbf{x}^*\mathbf{C}\mathbf{x}\,dt + \int_0^T \mathbf{x}^*\mathbf{G}\,d\mathbf{x}\right\} =$$

$$= E_{\mathbf{A}+\mathbf{BG}^*}\exp\left\{\int_0^T \mathbf{x}^*\left[\mathbf{C}+\frac{1}{2}(\mathbf{GA}+\mathbf{A}^*\mathbf{G}^*+\mathbf{GBG}^*)\right]\mathbf{x}\,dt\right\} =$$

$$= \psi_T\left(\mathbf{A}+\mathbf{BG}^*,\ \mathbf{C}+\frac{1}{2}(\mathbf{GA}+\mathbf{A}^*\mathbf{G}^*+\mathbf{GBG}^*)\right),$$

ahol az $E_{\mathbf{A}}$ a $P_\zeta$ szerint vett várható értéket jelenti a $C^r[0, T]$ téren.

Tehát a $\psi_T(\mathbf{A}, \mathbf{C})$ függvény segítségével megadhatók az $\mathbf{A}$-ra vonatkozó (feltételes) becslések jellemzői.

### 3. A maximum-likelihood becslések momentumai

Tekintsük **A**-nak a $\zeta(0)=0$ melletti feltételes maximum-likelihood becslését, erre az

$$\hat{\mathbf{A}}(\zeta)=\Big(\int_0^T \zeta(t)d\zeta^*(t)\Big)^*\Big(\int_0^T \zeta(t)\zeta^*(t)dt\Big)^{-1} \tag{3.1}$$

kifejezés adódik. (l. [1] (4.6.1) formula). Innen az $\hat{\mathbf{A}}(\zeta)$ momentumai explicite kifejezhetők a $\psi_T(\mathbf{A},\mathbf{C})$ függvény segítségével.

3.1. ÁLLÍTÁS. Legyen $\alpha=\mathbf{B}^+[\hat{\mathbf{A}}(\zeta)-\mathbf{A}]$. Igazak a következő formulák:

$$E(\alpha)=\frac{d}{d\mathbf{A}}E\Big(\int_0^T \zeta(t)\zeta^*(t)dt\Big)^{-1}, \tag{3.2}$$

$$E(\alpha_{im}\cdot\alpha_{pn})=\sum_{j,q}\frac{\partial^2}{\partial A_{ij}\partial A_{pq}}E[K_{jm}(\zeta)K_{qn}(\zeta)]+B^{ip}E[K_{mn}(\zeta)], \tag{3.3}$$

$$E(\alpha\alpha^*)=\frac{d}{d\mathbf{A}}E\Big(\int_0^T \zeta(t)\zeta^*(t)dt\Big)^{-2}\frac{d}{d\mathbf{A}^*}+\mathbf{B}^+\operatorname{sp}E\Big(\int_0^T \zeta(t)\zeta^*(t)dt\Big)^{-1}, \tag{3.4}$$

$$E\Big(\int_0^T \zeta(t)\zeta^*(t)dt\Big)^{-k}=\frac{1}{(k-1)!}\int_0^\infty s^{k-1}E\exp\Big\{-s\int_0^T \zeta(t)\zeta^*(t)dt\Big\}ds,\quad k\in N, \tag{3.5}$$

ahol a $\frac{d}{d\mathbf{A}}$ mátrix $(i,j)$-edik eleme $\frac{\partial}{\partial A_{ij}}$, $B^{ip}$ a $\mathbf{B}^+$ mátrix megfelelő eleme, $K_{mn}(\zeta)$ pedig a $\mathbf{K}(\zeta)=\Big(\int_0^T \zeta(t)\zeta^*(t)dt\Big)^{-1}$ mátrix eleme.

*Bizonyítás.* Jelöljük (átmenetileg)

$$\mathrm{Exp}=\exp\Big\{\int_0^T \mathbf{x}^*\mathbf{A}^*\mathbf{B}^+d\mathbf{x}-\frac{1}{2}\int_0^T \mathbf{x}^*\mathbf{A}^*\mathbf{B}^+\mathbf{A}\mathbf{x}\,dt\Big\}.$$

A (3.2) és (3.3) formulák a következő azonosságok alapján igazolhatók

$$\frac{\partial}{\partial A_{ij}}E_{(\mathbf{I}-\mathbf{BB}^+)\mathbf{A}}[\mathrm{Exp}]=E_{(\mathbf{I}-\mathbf{BB}^+)\mathbf{A}}\Big[\sum_{k=1}^n B^{ik}\Big(\int_0^T x_j\,dx_k-\sum_{s=1}^n\int_0^T x_jx_s\,dt\Big)\cdot\mathrm{Exp}\Big], \tag{3.6}$$

$$\frac{\partial^2}{\partial A_{pq}\partial A_{ij}}E_{(\mathbf{I}-\mathbf{BB}^+)\mathbf{A}}[\mathrm{Exp}]=E_{(\mathbf{I}-\mathbf{BB}^+)\mathbf{A}}\Big\{\Big[\sum_{k,l=1}^n B^{ik}B^{pl}\Big(\int_0^T x_j\,dx_k-\sum_{s=1}^n A_{ks}\int_0^T x_jx_s\,dt\Big)\times$$

$$\times\Big(\int_0^T x_q\,dx_l-\sum_{r=1}^n A_{lr}\int_0^T x_qx_r\,dt\Big)-B^{ip}\int_0^T x_jx_q\,dt\Big]\cdot\mathrm{Exp}\Big\}. \tag{3.7}$$

A (3.2) formula bizonyítása a 2.2 tételen alapszik:

$$E(\alpha) = E_{\mathbf{A}}\Big[\mathbf{B}^{+}\Big(\int_0^T d\mathbf{x}\cdot\mathbf{x}^{*} - \mathbf{A}\int_0^T \mathbf{x}\mathbf{x}^{*}dt\Big)K(\mathbf{x})\Big] =$$

$$= E_{(\mathbf{I}-\mathbf{B}\mathbf{B}^{+})\mathbf{A}}\Big[\mathbf{B}^{+}\Big(\int_0^T d\mathbf{x}\cdot\mathbf{x}^{*} - \mathbf{A}\int_0^T \mathbf{x}\mathbf{x}^{*}dt\Big)K(\mathbf{x})\cdot\mathrm{Exp}\Big].$$

Az $(i,j)$-edik elemet kiírva a (3.6) összefüggést felhasználva kapjuk

$$E(\alpha_{ij}) = E_{(\mathbf{I}-\mathbf{B}\mathbf{B}^{+})\mathbf{A}}\Big[\sum_{k,r,l} B^{ik}\Big(\int_0^T x_r dx_k - \sum_s A_{ks}\int_0^T x_s x_r dt\Big)\cdot\mathrm{Exp}\cdot K_{rj}(\mathbf{x})\Big] =$$

$$= \sum_r \frac{\partial}{\partial A_{ir}} E_{(\mathbf{I}-\mathbf{B}\mathbf{B}^{+})\mathbf{A}}[K_{rj}(\mathbf{x})\cdot\mathrm{Exp}] = \sum_r \frac{\partial}{\partial A_{ir}} E_{\mathbf{A}}[K_{rj}(\mathbf{x})] =$$

$$= \sum_r \frac{\partial}{\partial A_{ir}} EK_{rj}(\zeta).$$

Amiből következik (3.2).

A (3.3) formula ugyanígy bizonyítható. A 2.2. tétel alapján

$$E(\alpha_{im}\cdot\alpha_{pn}) = E\Big\{\sum_{k,j} B^{ik}\Big[\int_0^T \zeta_j(t)d\zeta_k(t) - \sum_s A_{ks}\int_0^T \zeta_s(t)\zeta_j(t)dt\Big]\cdot K_{jm}(\zeta)\times$$

$$\times\sum_{l,q} B^{pl}\Big[\int_0^T \zeta_q(t)d\zeta_l(t) - \sum_r A_{lr}\int_0^T \zeta_r(t)\zeta_q(t)dt\Big]\cdot K_{qn}(\zeta)\Big\} =$$

$$= E_{(\mathbf{I}-\mathbf{B}\mathbf{B}^{+})\mathbf{A}}\Big\{\sum_{k,j} B^{ik}\Big[\int_0^T x_j dx_k - \sum_s A_{ks}\int_0^T x_s x_j dt\Big]\cdot K_{jm}(\mathbf{x})\times$$

$$\times\sum_{l,q} B^{pl}\Big[\int_0^T x_q dx_l - \sum_r A_{lr}\int_0^T x_r x_q dt\Big]\cdot K_{qn}(\mathbf{x})\cdot\mathrm{Exp}\Big\} =$$

$$= E_{(\mathbf{I}-\mathbf{B}\mathbf{B}^{+})\mathbf{A}}\Big\{\sum_{k,j,l,q} B^{ik}B^{pl}\Big[\int_0^T x_j dx_k - \sum_s A_{ks}\int_0^T x_s x_j dt\Big]\times$$

$$\times\Big[\int_0^T x_q dx_l - \sum_r A_{lr}\int_0^T x_r x_q dt\Big]\cdot\mathrm{Exp}\cdot K_{jm}(\mathbf{x})\cdot K_{qn}(\mathbf{x})\Big\}.$$

A kapott eredményt összehasonlítva (3.7)-tel:

$$E(\alpha_{im}\cdot\alpha_{pn}) = E_{(\mathrm{I}-\mathbf{BB}^+)\mathbf{A}}\left\{\sum_{j,q} K_{jm}(\mathbf{x})K_{qn}(\mathbf{x})\cdot\left[\frac{\partial^2}{\partial A_{ij}\partial A_{pq}}\mathrm{Exp}+B^{ip}\int_0^T x_j x_q dt\cdot\mathrm{Exp}\right]\right\} =$$

$$= \sum_{j,q}\frac{\partial^2}{\partial A_{ij}\partial A_{pq}} E_{(\mathrm{I}-\mathbf{BB}^+)\mathbf{A}}\{K_{jm}(\mathbf{x})K_{qn}(\mathbf{x})\cdot\mathrm{Exp}\}+$$

$$+\sum_{j,q} B^{ip}E_{(\mathrm{I}-\mathbf{BB}^+)\mathbf{A}}\left\{K_{jm}(\mathbf{x})K_{qn}(\mathbf{x})\cdot\int_0^T x_j x_q dt\cdot\mathrm{Exp}\right\} =$$

$$= \sum_{j,q}\frac{\partial^2}{\partial A_{ij}\partial A_{pq}} E\{K_{jm}(\zeta)K_{qn}(\zeta)\}+\sum_{j,q} B^{ip}E\left\{K_{jm}(\zeta)K_{qn}(\zeta)\cdot\int_0^T \zeta_j(t)\zeta_q(t)dt\right\}.$$

Így $K(\zeta)$ definíciója alapján éppen (3.3)-at kapjuk.
A (3.4) formula a (3.3)-ból azonnal adódik.
A (3.5) formula a *Fubini-tétel* következménye.
A 3.1. állításban az **A**-ra vonatkozó feltételes maximum-likelihood becslés várható értékének és szórásának meghatározásához szükség van még a (3.5) jobboldalán az integrandusban szereplő $E\exp\left\{-s\int_0^T \zeta(t)\zeta^*(t)dt\right\}$ mátrix $\psi_T(\mathbf{A},\mathbf{C})$-vel való kifejezésére.

3.2. ÁLLÍTÁS. $E\exp\left\{-s\int_0^T \zeta(t)\zeta^*(t)dt\right\}=\sum_{k=0}^{\infty}\frac{(-s)^k}{k!}\left(\frac{d\psi}{d\mathbf{C}}\Big|_{\mathbf{C}=0}\right)^k$, ahol a $\frac{d\psi}{d\mathbf{C}}\Big|_{\mathbf{C}=0}$ mátrix $(i,j)$-edik eleme $\frac{\partial\psi_T(\mathbf{A},\mathbf{C})}{\partial c_{ij}}\Big|_{\mathbf{C}=0}$.

*Bizonyítás.* A $\psi_T(\mathbf{A},\mathbf{C})$ függvény $c_{ij}$ szerint vett parciális deriváltja

$$\frac{\partial\psi_T(\mathbf{A},\mathbf{C})}{\partial c_{ij}} = \frac{\partial}{\partial c_{ij}}E\exp\left\{\sum_{k,l} c_{kl}\int_0^T \zeta_k(t)\zeta_l(t)\,dt\right\} =$$

$$= E\left[\int_0^T \zeta_i(t)\zeta_j(t)\,dt\cdot\exp\left\{\sum_{k,l} c_{kl}\int_0^T \zeta_k(t)\zeta_l(t)\,dt\right\}\right].$$

Így teljesül a

$$\frac{\partial\psi_T(\mathbf{A},\mathbf{C})}{\partial c_{ij}}\Big|_{\mathbf{C}=0} = E\int_0^T \zeta_i(t)\zeta_j(t)\,dt$$

egyenlőség. Ebből következik, hogy

$$E\left(\int_0^T \zeta(t)\zeta^*(t)dt\right)^k = \left(\frac{d\psi}{d\mathbf{c}}\Big|_{\mathbf{C}=0}\right)^k, \quad \forall k\in N.$$

Innét sorfejtéssel kapjuk az állítást.

## 4. Az eloszlásfüggvény Laplace-transzformáltjának meghatározása

A 2.1. és 3.1. állításokból következik, hogy az **A** (feltételes) becsléseihez elegendő meghatározni a $\psi_T(\mathbf{A}, \mathbf{C})$ függvényt.

4.1. TÉTEL. Tegyük fel, hogy teljesülnek (2.1) és (2.8) feltételei. Ekkor

$$\psi_T(\mathbf{A}, \mathbf{C}) = e^{-T \operatorname{sp} \mathbf{BD}} \cdot \det [\mathbf{I}_{2n} - 2\overline{\mathbf{D}}\overline{\mathbf{\Gamma}}(T)]^{-1/2}, \tag{4.1}$$

a **D** szimmetrikus és $\tilde{\mathbf{a}}$ $n \times n$ konstans (valós) mátrixokra teljesülnek az

$$\mathbf{DA} + \mathbf{A}^*\mathbf{D} - 2\mathbf{DBD} = \mathbf{C}, \tag{4.2}$$

(az ún. *algebrai mátrix Riccati egyenlet*)

$$2\mathbf{BD} = \mathbf{A} - \tilde{\mathbf{a}}, \tag{4.3}$$

összefüggések, továbbá $\mathbf{\Gamma}(t)$ szimmetrikus $n \times n$-es mátrix a

$$\dot{\mathbf{\Gamma}}(t) = \tilde{\mathbf{a}}\mathbf{\Gamma}(t) + \mathbf{\Gamma}(t)\tilde{\mathbf{a}}^* + \mathbf{B}, \quad \mathbf{\Gamma}(0) = E(\boldsymbol{\eta}(0)\boldsymbol{\eta}^*(0)) = E(\boldsymbol{\zeta}(0)\boldsymbol{\zeta}^*(0)), \tag{4.4}$$

lineáris differenciál egyenletrendszer megoldása, azaz

$$\mathbf{\Gamma}(t) = e^{\tilde{\mathbf{a}}t}\mathbf{\Gamma}(0)e^{\tilde{\mathbf{a}}^*t} + \int_0^t e^{\tilde{\mathbf{a}}s}\mathbf{B}e^{\tilde{\mathbf{a}}^*s}\,ds,$$

$\overline{\mathbf{D}}$ és $\overline{\mathbf{\Gamma}}(t)$ $2n \times 2n$-es hipermátrixokat jelölnek, melyekre

$$\overline{\mathbf{D}} = \begin{pmatrix} -\mathbf{D} & \mathbf{0} \\ \mathbf{0} & \mathbf{D} \end{pmatrix}, \quad \overline{\mathbf{\Gamma}}(t) = \begin{pmatrix} \mathbf{\Gamma}(0) & \mathbf{\Gamma}(0) \cdot e^{\tilde{\mathbf{a}}^*t} \\ e^{\tilde{\mathbf{a}}t}\mathbf{\Gamma}(0) & \mathbf{\Gamma}(t) \end{pmatrix}, \tag{4.5}$$

és $I_{2n}$ a $2n \times 2n$-es egységmátrixot jelöli.

*Bizonyítás.* A 2.2. tételt alkalmazva a (2.1) és (2.5) összefüggést kielégítő $\boldsymbol{\zeta}(t)$ és $\boldsymbol{\eta}(t)$ folyamatokra kapjuk, hogy

$$\psi_T(\mathbf{A}, \mathbf{C}) = E \exp\left\{\int_0^T \boldsymbol{\zeta}^*(t)\mathbf{C}\boldsymbol{\zeta}(t)\,dt\right\} = E_{\mathbf{A}} \exp\left\{\int_0^T \mathbf{x}^*\mathbf{C}\mathbf{x}\,dt\right\} =$$

$$= E_{\mathbf{a}} \exp\left\{\int_0^T \mathbf{x}^*\mathbf{C}\mathbf{x}\,dt - \frac{1}{2}\int_0^T \mathbf{x}^*(\mathbf{A}-\mathbf{a})^*\mathbf{B}^+(\mathbf{A}-\mathbf{a})\mathbf{x}\,dt + \int_0^T \mathbf{x}^*(\mathbf{A}-\mathbf{a})^*\mathbf{B}^+\mathbf{B}_{\mathbf{W}}^{1/2}\,d\mathbf{W}(t)\right\},$$

vagyis

$$\psi_T(\mathbf{A}, \mathbf{C}) = \tag{4.6}$$

$$= E \exp\left\{\int_0^T \boldsymbol{\eta}^*(t)\left[\mathbf{C} - \frac{1}{2}(\mathbf{A}-\mathbf{a})^*\mathbf{B}^+(\mathbf{A}-\mathbf{a})\right]\boldsymbol{\eta}(t)\,dt + \int_0^T \boldsymbol{\eta}^*(t)(\mathbf{A}-\mathbf{a})^*\mathbf{B}^+\mathbf{B}_{\mathbf{W}}^{1/2}\,d\mathbf{W}(t)\right\},$$

ha az **a** mátrixra teljesül a 2.2. tétel feltétele.

Legyen most

$$v(t) = \boldsymbol{\eta}^*(t)\mathbf{D}\boldsymbol{\eta}(t),$$

ahol $\mathbf{D}$ $n\times n$-es konstans szimmetrikus mátrix. Ekkor az *Ito formula* alapján:

$$(4.7)\quad v(T)-v(0)-T\,\mathrm{sp}\,\mathbf{BD}=\int_0^T \boldsymbol{\eta}^*(t)(\mathbf{Da}+\mathbf{a}^*\mathbf{D})\boldsymbol{\eta}(t)dt+2\int_0^T \boldsymbol{\eta}^*(t)\mathbf{DB}_W^{1/2}d\mathbf{W}(t).$$

Úgy választjuk meg az $\mathbf{a}$ és $\mathbf{D}$ mátrixokat, hogy (4.6) jobboldalán a kitevőben szereplő kifejezés teljes (sztochasztikus) differenciál legyen, és teljesüljön (2.6). A megoldáspárt (amennyiben létezik) $(\tilde{\mathbf{a}}, \mathbf{D})$-vel jelölve.

$$(4.8)\quad \mathbf{C}-\frac{1}{2}(\mathbf{A}-\tilde{\mathbf{a}})^*\mathbf{B}^+(\mathbf{A}-\tilde{\mathbf{a}})=\mathbf{D}\tilde{\mathbf{a}}+\tilde{\mathbf{a}}^*\mathbf{D},$$

$$(4.9)\quad (\mathbf{A}-\tilde{\mathbf{a}})^*\mathbf{B}^+\mathbf{B}_W^{1/2}=2\mathbf{DB}_W^{1/2},$$

$$(4.10)\quad \mathbf{B}_W^{1/2}\mathbf{F}=\mathbf{A}-\tilde{\mathbf{a}},$$

valamilyen $\mathbf{F}$ konstans mátrixra.

Megmutatható, hogy a (4.8), (4.9), (4.10) egyenletrendszer ekvivalens a (4.2), (4.3)-mal. Ugyanis (4.10) mindkét oldalát adjungálva, és $(\mathbf{A}-\mathbf{a})^*$-ot (4.9)-be helyettesítve

$$\mathbf{F}^*\mathbf{B}_W^{1/2*}\mathbf{B}_W^{1/2*+}\mathbf{B}_W^{1/2+}\mathbf{B}_W^{1/2}=2\mathbf{DB}_W^{1/2}.$$

Mindkét oldalt $\mathbf{B}_W^{1/2*}$-gal szorozva, és az „általánosított" invert azt a tulajdonságát használva, mely szerint $\mathbf{B}_W^{1/2+}\mathbf{B}_W^{1/2}\mathbf{B}_W^{1/2*}=\mathbf{B}_W^{1/2*}$

$$\mathbf{F}^*\mathbf{B}_W^{1/2*}\mathbf{B}_W^{1/2*+}\mathbf{B}_W^{1/2*}=\mathbf{F}^*\mathbf{B}_W^{1/2*}=(\mathbf{A}-\tilde{\mathbf{a}})^*=2\mathbf{DB},$$

vagyis (4.3)-at kaptuk. Ebből $\tilde{\mathbf{a}}$-ot kifejezve, és (4.8)-ba helyettesítve a (4.2) egyenlőséghez jutunk. Megfordítva, ha egy $(\tilde{\mathbf{a}}, \mathbf{D})$ pár kielégíti (4.2) és (4.3)-at, akkor könnyen ellenőrizhető, hogy (4.8), (4.9) és (4.10) is teljesül.

Kiválasztva egy olyan $(\tilde{\mathbf{a}}, \mathbf{D})$ párt, amely kielégíti (4.2) és (4.3)-at:

$$\psi_T(\mathbf{A}, \mathbf{C})=E\exp\{v(T)-v(0)-T\,\mathrm{sp}\,\mathbf{BD}\},$$

vagyis $v(t)$ definíciója alapján

$$(4.11)\quad \psi_T(\mathbf{A}, \mathbf{C})=e^{-T\,\mathrm{sp}\,\mathbf{BD}}E\exp\{\boldsymbol{\eta}^*(T)\mathbf{D}\boldsymbol{\eta}(T)-\boldsymbol{\eta}^*(0)\mathbf{D}\boldsymbol{\eta}(0)\}.$$

Mivel $\boldsymbol{\eta}(t)$ a (2.5) egyenlet (erős) megoldása, így $\boldsymbol{\eta}(t)$ Gauss folyamat, vagyis a $\boldsymbol{\Gamma}(t)=E(\boldsymbol{\eta}(t)\boldsymbol{\eta}^*(t))$ kovarianciamátrix a (4.4)-nek megoldása (l. [7] 15.1. tétel). Az

$$\begin{pmatrix}\boldsymbol{\eta}(0)\\ \boldsymbol{\eta}(T)\end{pmatrix}$$

valószínűségi változók együttes kovarianciamátrixa éppen a (4.5)-ben definiált $\bar{\boldsymbol{\Gamma}}(T)$ mátrix. *Gauss eloszlású* valószínűségi változók együttes sűrűségfüggvénye meghatározható a kovarianciamátrix-szal. Így a (4.11)-ben szereplő várható értékre, amely normális eloszlású vektor változók kvadratikus alakja, pl. [7] (11.48) formula alapján a (4.1) képlet adódik.

*4.1. Megjegyzés.* Ha $\boldsymbol{\zeta}(0)=\boldsymbol{\eta}(0)=0$ akkor (4.1) a

$$\psi_T(\mathbf{A}, \mathbf{C})=e^{-T\,\mathrm{sp}\,\mathbf{BD}}\det[\mathbf{I}_n-2\mathbf{D}\boldsymbol{\Gamma}(T)]^{-1/2}$$

egyszerű alakot nyeri, ahol a $\mathbf{D}$ és $\mathbf{\Gamma}(t)$ a (4.2), (4.3) és (4.4) egyenletekkel adhatók meg, és $\mathbf{I}_n$ az $n\times n$-es egységmátrix.

*4.2. Megjegyzés.* A 4.1. tételhez hasonló módon meghatározhatjuk a

$$\varphi_T(\mathbf{A}, \mathbf{C}_1, \mathbf{C}_2, \mathbf{C}_3) = E\exp\left\{\int_0^T \zeta^*(t)\mathbf{C}_1\zeta(t)\,dt+\zeta^*(0)\mathbf{C}_2\zeta(0)+\zeta^*(T)\mathbf{C}_3\zeta(T)\right\}$$

függvényt is, amely a $\left\{\int_0^T \zeta_i(t)\zeta_j(t)dt,\ \zeta_k(0)\zeta_l(0),\ \zeta_m(T)\zeta_p(T),\ i,j,k,l,m,p=\right.$ $\left.=1,\ldots,n\right\}$ valószínűségi változók együttes eloszlásának *Laplace-transzformáltja,* ahol $\mathbf{C}_1, \mathbf{C}_2, \mathbf{C}_3$ mind negatív szemidefinit szimmetrikus mátrixok.

Erre a

$$\varphi_T(\mathbf{A}, \mathbf{C}_1, \mathbf{C}_2, \mathbf{C}_3) = e^{-T\,\mathrm{sp}\,\mathbf{BD}}\det\left[\mathbf{I}_{2n}-2\overline{\mathbf{D}}_1\overline{\mathbf{\Gamma}}(T)\right]^{-1/2}$$

adódik, ahol $\overline{\mathbf{\Gamma}}(T)$ és $\mathbf{D}$ a (4.2), (4.3), (4.4), (4.5)-ben definiált mátrixok, $\overline{\mathbf{D}}_1$ pedig olyan $2n\times 2n$-es hipermátrix, amely

$$\overline{\mathbf{D}}_1 = \begin{pmatrix} \mathbf{C}_2-\mathbf{D} & 0 \\ 0 & \mathbf{C}_3+\mathbf{D} \end{pmatrix}$$

alakú.

Ezt összevetve a 2.1. állítással, megkapjuk a (2.4)-ben szereplő nem feltételes elégséges statisztika együttes eloszlásának *Laplace transzformáltját* is.

A (4.2) és (4.3) egyenletek megoldhatóságáról a következőket mondhatjuk.

4.1. ÁLLÍTÁS. A 4.1. tétel feltételei mellett a (4.2) egyenletnek mindig van valós szimmetrikus pozitív szemidefinit $\mathbf{D}$ megoldása. A (4.3)-ból kapott $\tilde{\mathbf{a}}$ mátrix sajátértékei is a negatív félsíkba esnek.

*Bizonyítás.* A (4.2) egyenlet egy ún. *algebrai mátrix Riccati egyenlet* $\mathbf{D}$-re, amelyről ismert, hogy ha $\mathbf{C}\geqq\mathbf{0}\leqq\mathbf{B}$ és $\mathbf{A}$ saját értékei a negatív komplex félsíkba esnek, akkor van $\mathbf{D}\geqq\mathbf{0}$ megoldása (l. 5.1. tétel). Az állítás második fele közvetlenül adódik a pozitív szemidefinit mátrixok tulajdonságaiból.

*4.3. Megjegyzés.* A 4.1. állítás úgy is megfogalmazható, hogy ha a (2.1) egyenlet stabil, akkor a (4.2), (4.3)-mal nyert (2.5) egyenlet ($\mathbf{a}=\tilde{\mathbf{a}}$ helyettesítéssel) is az.

## 5. Stacionárius folyamatok

Az alkalmazások szempontjából igen fontos stacionárius esettel külön foglalkozunk.

Legyen tehát $\zeta(t)$ stacionárius megoldása (2.1)-nek, ilyen a feltételek mellett mindig pontosan egy létezik. Az eddig felsorolt tételek és állítások természetesen ebben az esetben is működnek. Ugyanakkor a 2.1. állításban és a 2.2., 4.1. tételekben a „transzformált" $\eta(t)$ folyamat az $\eta(0)=\zeta(0)$ feltétel miatt már nem lesz stacionárius. Ez a „szépséghiba" a konkrét számolásokat erősen elbonyolítja. Természetes igény az is, hogy a stacionárius folyamatok köréből ne vezessenek ki a transzformációk. Ez úgy érhető el, hogy a 2.2. tétel helyett olyan „transzformációs képletet" használunk, amelyben megszabadulunk a problémás feltételtől (l. [7] 17.6. lemmát).

5.1. TÉTEL. Legyen (2.1)-ben $\zeta(0)$ eloszlása $F_A$, a (2.5)-ben szereplő $\eta(0)$ eloszlása $F_a$. Ha az $\mathbf{A}$, $\mathbf{B}_W^{1/2}$, $\mathbf{a}$ mátrixokra teljesül a (2.6) feltétel, és az $(\mathbf{R}^n, B^n)$ mérhető téren az $F_A$ és $F_a$ eloszlások ekvivalensek, akkor a $P_\zeta$ és a $P_\eta$ mértékek ekvivalensek a $C^r[0, T]$-n és

$$(5.1)\qquad \frac{dP_\zeta}{dP_\eta}(T, \eta) =$$

$$= \frac{dF_A}{dF_a}(\eta(0))\exp\left\{\int_0^T \eta^*(t)(\mathbf{A}-\mathbf{a})^*\mathbf{B}^+ d\eta(t) - \frac{1}{2}\int_0^T \eta^*(t)(\mathbf{A}^*\mathbf{B}^+\mathbf{A}-\mathbf{a}^*\mathbf{B}^+\mathbf{a})\eta(t)\,dt\right\} =$$

$$= \frac{dF_A}{dF_a}(\eta(0))\times$$

$$\times\exp\left\{-\frac{1}{2}\int_0^T \eta^*(t)(\mathbf{A}-\mathbf{a})^*\mathbf{B}^+(\mathbf{A}-\mathbf{a})\eta(t)\,dt + \int_0^T \eta^*(t)(\mathbf{A}-\mathbf{a})^*\mathbf{B}^+\mathbf{B}_W^{1/2}\,d\mathbf{W}(t)\right\}.$$

*5.1. Megjegyzés.* Az $\eta(t)$ folyamatról a továbbiakban feltesszük, hogy szintén stacionárius. Ilyen választással az $F_A$ és $F_a$ eloszlások ekvivalensek lesznek.

Legyen

$$(5.2)\quad \varphi_T(\mathbf{A}, \mathbf{C}_1, \mathbf{C}_2, \mathbf{C}_3) = E\exp\left\{\int_0^T \zeta^*(t)\mathbf{C}_1\zeta(t)\,dt + \zeta^*(0)\mathbf{C}_2\zeta(0) + \zeta^*(T)\mathbf{C}_3\zeta(T)\right\}$$

és $\mathbf{C}_1, \mathbf{C}_2, \mathbf{C}_3$ negatív szemidefinit szimmetrikus mátrixok. A $\varphi_T(\mathbf{A}, \mathbf{C}_1, \mathbf{C}_2, \mathbf{C}_3)$ függvény már meghatározza a (2.4)-beli elégséges statisztika együttes eloszlásának *Laplace transzformáltját* is, igaz ugyanis a következő

5.1. ÁLLÍTÁS. Tegyük fel, hogy $\mathbf{G}$ olyan mátrix, amelyre az $\mathbf{A}+\mathbf{BG}^*$ saját értékei a negatív félsíkba esnek. Ekkor

$$(5.3)\quad E\exp\left\{\int_0^T \zeta^*(t)\mathbf{C}_1\zeta(t)\,dt + \zeta^*(0)\mathbf{C}_2\zeta(0) + \zeta^*(T)\mathbf{C}_3\zeta(T) + \int_0^T \zeta^*(t)\mathbf{G}\,d\zeta(t)\right\} =$$

$$= \det(\mathbf{U}^{-1}\mathbf{V})^{1/2}\varphi_T\left(\mathbf{A}+\mathbf{BG}^*, \mathbf{C}_1 + \frac{1}{2}(\mathbf{GA}+\mathbf{A}^*\mathbf{G}^*+\mathbf{GBG}^*), \mathbf{C}_2 + \frac{1}{2}\mathbf{V}^{-1} - \frac{1}{2}\mathbf{U}^{-1}, \mathbf{C}_3\right),$$

ahol

$$(5.4)\qquad \mathbf{U} = \int_0^\infty e^{\mathbf{A}t}\mathbf{B}e^{\mathbf{A}^*t}dt \quad \text{és} \quad \mathbf{V} = \int_0^\infty e^{(\mathbf{A}+\mathbf{BG}^*)t}\mathbf{B}e^{(\mathbf{A}+\mathbf{BG}^*)^*t}\,dt.$$

*Bizonyítás.* Az 5.1. tétel alapján:

$$E\exp\left\{\int_0^T \zeta^*(t)\mathbf{C}_1\zeta(t)\,dt + \zeta^*(0)\mathbf{C}_2\zeta(0) + \zeta^*(T)\mathbf{C}_3\zeta(T) + \int_0^T \zeta^*(t)\mathbf{G}\,d\zeta(t)\right\} =$$

$$= E_A\exp\left\{\int_0^T \mathbf{x}^*\mathbf{C}_1\mathbf{x}\,dt + \mathbf{x}^*(0)\mathbf{C}_2\mathbf{x}(0) + \mathbf{x}^*(T)\mathbf{C}_3\mathbf{x}(T) + \int_0^T \mathbf{x}^*\mathbf{G}\,d\mathbf{x}\right\} =$$

$$= E_{A+BG^*}\left[\frac{dF_A}{dF_{A+BG^*}}(\mathbf{x}(0))\exp\left\{\int_0^T \mathbf{x}^*\left[\mathbf{C}_1 + \frac{1}{2}(\mathbf{GA}+\mathbf{A}^*\mathbf{G}^*+\mathbf{GBG}^*\right]\mathbf{x}\,dt + \right.\right.$$

$$\left.\left. + \mathbf{x}^*(0)\mathbf{C}_2\mathbf{x}(0) + \mathbf{x}^*(T)\mathbf{C}_3\mathbf{x}(T)\right\}\right].$$

A $\zeta(t)$ folyamat szórásmátrixra az (5.4)-ben definiált $\mathbf{U}$, az $\mathbf{A}+\mathbf{BG}^*$-hoz tartozó transzformált stacionárius folyamat szórásmátrixa $\mathbf{V}$.

Ezekkel kifejezve

$$\frac{dF_{\mathbf{A}}}{dF_{\mathbf{A}+\mathbf{BG}^*}}(\mathbf{y}) = \frac{f_{\mathbf{A}}(\mathbf{y})}{f_{\mathbf{A}+\mathbf{BG}^*}(\mathbf{y})} = |\mathbf{U}^{-1}\mathbf{V}|^{1/2}\exp\left\{-\frac{1}{2}(\mathbf{V}^{-1}-\mathbf{U}^{-1})\mathbf{y}\right\},$$

ahol $f_{\mathbf{A}}(\mathbf{y})$ a $\zeta(0)$ sűrűségfüggvénye. Ezt felhasználva az előzőből (5.3)-at kapjuk.

Most megfogalmazzuk a 4.1. tételt és annak a 4.2. megjegyzésben szereplő általánosítását a stacionárius esetre.

5.2. TÉTEL. Teljesüljenek az 5.1. tétel és az 5.1. megjegyzés feltételei. Ekkor

$$\varphi_T(\mathbf{A}, \mathbf{C}_1, \mathbf{C}_2, \mathbf{C}_3) = e^{-T\,\mathrm{sp}\,\mathbf{BD}}|\mathbf{U}^{-1}\mathbf{V}|^{1/2}|\mathbf{I}_{2n}-2\overline{\mathbf{D}}\overline{\mathbf{\Gamma}}(T)|^{-1/2}, \tag{5.5}$$

a $\mathbf{D}$ szimmetrikus és $\tilde{\mathbf{a}}$ $n\times n$-es konstans (valós) mátrixokra teljesülnek az

$$\mathbf{DA}+\mathbf{A}^*\mathbf{D}-2\mathbf{DBD} = \mathbf{C}_1, \tag{5.6}$$

$$2\mathbf{BD} = \mathbf{A}-\tilde{\mathbf{a}}, \tag{5.7}$$

összefüggések, valamint $U$ és $V$ az

$$\mathbf{AU}+\mathbf{UA}^*+\mathbf{B} = \mathbf{0} \quad \text{és az} \quad \tilde{\mathbf{a}}\mathbf{V}+\mathbf{V}\tilde{\mathbf{a}}^*+\mathbf{B} = \mathbf{0}$$

egyenletek szimmetrikus megoldásai, azaz

$$\mathbf{U} = \int_0^\infty e^{\mathbf{A}s}\mathbf{B}e^{\mathbf{A}^*s}ds \quad \text{és} \quad \mathbf{V} = \int_0^\infty e^{\tilde{\mathbf{a}}s}\mathbf{B}e^{\tilde{\mathbf{a}}^*s}ds, \tag{5.8}$$

továbbá

$$\overline{\mathbf{\Gamma}}(t) = \begin{pmatrix} \mathbf{V} & \mathbf{V}e^{\tilde{\mathbf{a}}^*t} \\ e^{\tilde{\mathbf{a}}t}\mathbf{V} & \mathbf{V} \end{pmatrix} \quad \text{és} \quad \overline{\mathbf{D}} = \begin{pmatrix} \mathbf{C}_2-\mathbf{D}+\frac{1}{2}\mathbf{V}^{-1}\frac{1}{2}-\mathbf{U}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_3+\mathbf{D} \end{pmatrix}. \tag{5.9}$$

*5.2. Megjegyzés.* Ha csak $\int_0^T \zeta^*(t)\mathbf{C}\zeta(t)dt$ *Laplace transzformáltját* kívánjuk meghatározni (vesd össze (4.1) képlettel)

$$\psi_T(\mathbf{A}, \mathbf{C}) = e^{-T\,\mathrm{sp}\,\mathbf{BD}}|\mathbf{U}^{-1}\mathbf{V}|^{1/2}|\mathbf{I}_{2n}-2\overline{\mathbf{D}}\overline{\mathbf{\Gamma}}(\mathbf{T})|^{-1/2},$$

ahol $\mathbf{D}, \overline{\mathbf{D}}, \overline{\mathbf{\Gamma}}(T)$ a fentiek a $\mathbf{C}_1=\mathbf{C}$, $\mathbf{C}_2=\mathbf{C}_3=0$ helyettesítéssel.

A felsorolt állítások és tételek jól mutatják az ún. „*Randon—Nikodym transzformáció*"-k hatékonyságát, amely 1 dimenziós esetben szerepel a már említett [3] cikkben.

Az előbbiekben bemutatottak közvetlenül általánosíthatók arra az esetre, amikor (2.1)-ben az $\mathbf{A}$ mátrix $t$-nek függvénye.

Ekkor a 2.1. tétel helyett egy általánosabbat kell használni, amely megtalálható [7]-ben (l. 6. § 7.4.). A 4.1. tételben pedig (4.2) helyett egy *mátrix Riccati differenciálegyenletet* kapunk, amelyre vonatkozó eredmények összefoglalása, és irodalomjegyzék [4]-ben van.

Az 5.2. tétel lehetőséget ad arra, hogy megszabaduljunk az eredeti $\zeta$ folyamat indítására vonatkozó feltételtől.

A 4.1., illetve az 5.1. tételben ugyanis csak az $\eta(0)$ normalizálását használtuk ki, $\zeta(0)$-ról elég csak annyit feltenni, hogy abszolút folytonos eloszlású.

Az irányításelmélet *Riccati egyenletekre* vonatkozó kvalitatív eredményei jól felhasználhatók a becslések viselkedésének jellemzésére.

## 6. Példák

A következőkben az eddig felsorolt tételeket alkalmazzuk konkrét példákra, főként autoregressziós esetben.

Tekintsük tehát az $n$-edrendű autoregressziós folyamatot, azaz legyen (2.1)-ben

$$\text{(6.1)} \qquad \mathbf{A} = \begin{pmatrix} 0 & 1 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & & \\ 0 & 0 & \dots & 1 \\ -A_n & -A_{n-1} & \dots & -A_1 \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} 0 \dots 0 \\ 0 \dots 0 \\ \vdots \\ 0 \dots 1 \end{pmatrix}.$$

Ekkor létezik pontosan egy stacionárius folyamat, amely kielégíti (2.1)-et, ennek $E\zeta(0)\zeta^*(0)=\mathbf{U}$ kovarianciamátrixára a

$$\text{(6.2)} \qquad \mathbf{AU}+\mathbf{UA}^*+\mathbf{B}=\mathbf{0}$$

egyenlet teljesül. Ebben az esetben elég meghatározni a

$$\text{(6.3)} \qquad \varphi_T(\mathbf{A}, \mathbf{C}_1, \mathbf{C}_2, \mathbf{C}_3) = E\exp\left\{\int_0^T \zeta^*(t)\mathbf{C}_1\zeta(t)dt+\zeta^*(0)\mathbf{C}_2\zeta(0)+\zeta^*(T)\mathbf{C}_3\zeta(T)\right\}$$

függvényt, ahol $\mathbf{C}_1=\operatorname{diag}(-c_1, -c_2, \dots, -c_n)$, $c_i \geqq 0$ és $\mathbf{C}_2, \mathbf{C}_3$ negatív szemidefinit szimmetrikus mátrixok (l. [6] 9. formula).

A $\varphi$ függvényt az 5.2. tétel alapján határozhatjuk meg. Az (5.6) alapegyenlet megoldásához az

$$\mathbf{M} = \begin{pmatrix} \mathbf{A} & -2\mathbf{B} \\ \mathbf{C}_1 & -\mathbf{A}^* \end{pmatrix}$$

2×2-es hipermátrix vizsgálatára van szükség. Az $\mathbf{M}$ mátrix jól kiválasztott $n$ darab sajátértékéhez tartozó $\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_n$ általánosított sajátvektorai segítségével határozható meg a $\mathbf{D}$ mátrix. Legyenek ugyanis $\mathbf{x}_i, \mathbf{y}_i$ $(i=1, \dots, n)$ $n$ dimenziós (oszlop) vektorok, amelyekre

$$\mathbf{z}_i = \begin{bmatrix} \mathbf{x}_i \\ \hline \mathbf{y}_i \end{bmatrix}, \quad i = 1, 2, \dots, n.$$

Az (5.6) egyenlet valós, szimmetrikus, pozitív definit megoldása a

$$\mathbf{D} = [\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n][\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n]^{-1},$$

(l. [4] 1. tétel és 4. tétel). (5.7)-ből az $\tilde{\mathbf{a}}$ mátrix közvetlenül kifejezhető a most kapott $\mathbf{D}$-vel. Megmutatható, hogy a kapott $\tilde{\mathbf{a}}$ mátrix is (6.1) alakú, tehát a transzformált folyamat is autoregressziós. Ennek $\mathbf{V}$ szórásmátrixa $\tilde{\mathbf{a}}$-ból (5.8) alapján meghatároz-

ható. Ebből, az (5.9)-ben szereplő hipermátrixok determinánsát kifejezve kapjuk a

$$\varphi_T(\mathbf{A}, \mathbf{C}_1, \mathbf{C}_2, \mathbf{C}_3) = e^{-T\,\mathrm{sp}(\tilde{\mathbf{a}}+\mathbf{BD})} \cdot \{|\mathbf{C}_3+\mathbf{D}|\,|\mathbf{I}_n+2\mathbf{UD}-2\mathbf{UC}_2|\times \tag{6.4}$$

$$\times|e^{-\tilde{\mathbf{a}}T}[(\mathbf{C}_3+\mathbf{D})^{-1}-2\mathbf{V}]e^{-\tilde{\mathbf{a}}^*T}+2\mathbf{V}-2(2\mathbf{D}+\mathbf{U}^{-1}-2\mathbf{C}_2)^{-1}|\}^{-1/2}$$

formulát.

A (4.1)-ben szereplő $\psi$ függvényre pedig a

$$\psi_T(\mathbf{A}, \mathbf{C}) = e^{-T\,\mathrm{sp}(\tilde{\mathbf{a}}+\mathbf{BD})}\{|\mathbf{D}|\,|\mathbf{I}_n+2\boldsymbol{\Gamma}(0)\mathbf{D}|\times \tag{6.5}$$

$$\times|e^{-\tilde{\mathbf{a}}T}(\mathbf{D}^{-1}-2\mathbf{V})e^{-\tilde{\mathbf{a}}^*T}+2\mathbf{V}-2(2\mathbf{D}+\boldsymbol{\Gamma}(0)^{-1})^{-1}|\}^{-1/2}$$

képletet kapjuk, ahol $\boldsymbol{\Gamma}(0)$ a $\zeta$ folyamat kezdeti kovarianciamátrixa, vagyis $\boldsymbol{\Gamma}(0)=E\zeta(0)\zeta^*(0)$.

Amennyiben $\zeta$ stacionárius, úgy $\boldsymbol{\Gamma}(0)=\mathbf{U}$ lesz, és a (6.4) és (6.5) formulák $\mathbf{C}_1=\mathbf{C}$, $\mathbf{C}_2=\mathbf{C}_3=0$ helyettesítéssel megegyeznek.

Tekintsük a másodrendű autoregressziós folyamatot zero indítással, amely a

$$d\zeta(t) = \begin{pmatrix} 0 & 1 \\ -A_2 & -A_1 \end{pmatrix}\zeta(t)\,dt + \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} d\mathbf{W}(t), \quad A_1, A_2 > 0,$$

egyenlettel, és $\zeta(0)=O$ feltétellel van megadva.

A 2.1. tételt alkalmazva:

$$\frac{dP_\zeta}{dP_{\mathbf{W}}}(T, \zeta) = \exp\left\{\int_0^T [-A_2\zeta_1(t)-A_1\zeta_2(t)]\,d\zeta_2(t) - \frac{1}{2}\int_0^T [A_1\zeta_2(t)+A_2\zeta_1(t)]^2\,dt\right\}.$$

Meghatározzuk a

$$\psi_T(\mathbf{A}, \mathbf{C}) = E\exp\left\{\int_0^T \zeta^*(t)\mathbf{C}\zeta(t)\,dt\right\} = E\exp\left\{\int_0^T -c_1\zeta_1^2(t)-c_2\zeta_2^2(t)\,dt\right\} \tag{6.6}$$

függvényt, ahol $c_1, c_2 \geqq 0$. A 4.1 tételt használjuk. A (4.2) és (4.3) egyenletek megoldása ekkor:

$$\tilde{\mathbf{a}} = \begin{pmatrix} 0 & 1 \\ -a_2 & -a_1 \end{pmatrix}, \quad \mathbf{D} = \begin{pmatrix} d_1 & d_2 \\ d_2 & d_3 \end{pmatrix}, \tag{6.7}$$

ahol

$$a_2 = \sqrt{A_2^2+2C_1}, \quad a_1 = \sqrt{A_1^2+2c_2-2A_2+2a_2}, \tag{6.8}$$

$$d_1 = \frac{1}{2}(a_1a_2-A_1A_2), \quad d_2 = \frac{1}{2}(a_2-A_2), \quad d_3 = \frac{1}{2}(a_1-A_1). \tag{6.9}$$

Legyenek $\lambda_1$ és $\lambda_2$ az a mátrix különböző sajátértékei, azaz a $\lambda^2+a_1\lambda+a_2=0$ egyenlet gyökei. A (4.4) egyenlet megoldását kiszámolva

$$\boldsymbol{\Gamma}(t) = \begin{pmatrix} \Gamma_1(t) & \Gamma_2(t) \\ \Gamma_2(t) & \Gamma_3(t) \end{pmatrix},$$

ahol

$$\Gamma_1(t) = \frac{1}{a_1^2-4a_2}\left[\frac{1}{2\lambda_1}e^{2\lambda_1 t}+\frac{1}{2\lambda_2}e^{2\lambda_2 t}+\frac{2}{a_1}e^{-a_1 t}\right]+\frac{1}{2a_1a_2},$$

$$\Gamma_2(t) = \frac{1}{a_1^2-4a_2}\left[\frac{1}{2}\cdot e^{2\lambda_1 t}+\frac{1}{2}\cdot e^{2\lambda_2 t}-e^{-a_1 t}\right],$$

$$\Gamma_3(t) = \frac{1}{a_1^2-4a_2}\left[\frac{\lambda_1}{2}e^{2\lambda_1 t}+\frac{\lambda_2}{2}e^{2\lambda_2 t}+\frac{2a_2}{a_1}e^{-a_1 t}\right]+\frac{1}{2a_1}\Big].$$

Ebből (4.1) alapján

$$\begin{aligned}(6.10)\quad \psi_T(\mathbf{A},\mathbf{C}) = e^{1/2A_1T}2a_1\sqrt{a_2}\cdot\{&e^{-a_1T}[(a_1a_2-A_1A_2)(a_1-A_1)-(a_2-A_2)^2]+\\ &+e^{a_1T}[(a_1a_2+A_1A_2)(a_1+A_1)-(a_2-A_2)^2]+\\ &+\frac{8a_2}{a_1^2-4a_2}[A_2(-a_1^2+A_1^2)-(a_2-A_2)^2]+\\ &+e^{(\lambda_1-\lambda_2)T}\frac{a_1^2}{a_1^2-4a_2}[(a_1a_2-A_1A_2)(a_1+A_1)+(A_2-a_2)(3a_2+A_2)+2\lambda_1A_1(a_2-A_2)]+\\ &+e^{(\lambda_2-\lambda_1)T}\frac{a_1^2}{a_1^2-4a_2}[(a_1a_2-A_1A_2)(a_1+A_1)+(A_2-a_2)(3a_2+A_2)+2\lambda_2A_1(a_2-A_2)]\}^{-1/2}\end{aligned}$$

speciálisan, ha $A_2=0$ és $C_1=0$ akkor

$$\psi_T(A,C) = E\exp\left\{-c_2\int_0^T\zeta_2^2(t)dt\right\} = \left\{\frac{2a_1e^{TA_1}}{(a_1-A_1)e^{-a_1T}+(a_1+A_1)e^{a_1T}}\right\}^{1/2}$$

adódik, vagyis az elsőrendű AR folyamatra vonatkozó [3] 15. formulát kapjuk (l. még [7] 17.3. §).

A továbbiakban a stacionárius indítású másodrendű autoregressziós folyamatot vizsgáljuk. A *Randon—Nikodym derivált*

$$\begin{aligned}&\frac{dP_\zeta}{dP_W}(\zeta,T) =\\ &= f(\zeta(0))\exp\left\{\int_0^T[-A_2\zeta_1(t)-A_1\zeta_2(t)]d\zeta_2(t)-\frac{1}{2}\int_0^T[A_1\zeta_2(t)+A_2\zeta_1(t)]^2dt\right\} =\\ &= \frac{A_1\sqrt{A_2}}{\pi}\exp\left\{\frac{A_2^2}{2}\int_0^T\zeta_1^2(t)\,dt+\frac{2A_2-A_1^2}{2}\int_0^T\zeta_2^2(t)dt+\frac{A_1T}{2}-\right.\\ &\left.-\frac{A_1}{2}[\zeta_2^2(T)+\zeta_2^2(0)]-A_2[\zeta_1(T)\zeta_2(T)-\zeta_1(0)\zeta_2(0)]-\frac{A_1A_2}{2}[\zeta_1^2(T)+\zeta_1^2(0)]\right\}\end{aligned}$$

alakú.

Az 5.2. tételt alkalmazva meghatározzuk a (6.6)-beli $\psi$ függvényt. Az (5.6) és (5.7) egyenletek megoldása a (6.7), (6.8) és (6.9)-ben definiált $\tilde{\mathbf{a}}$ és $\mathbf{D}$ mátrixok lesz-

nek. (5.8)-ból

$$\mathbf{U} = \frac{1}{2A_1A_2}\begin{pmatrix}1 & 0\\ 0 & A_2\end{pmatrix}, \quad \mathbf{V} = \frac{1}{2a_1a_2}\begin{pmatrix}1 & 0\\ 0 & a_2\end{pmatrix}.$$

A (6.4) formulát alkalmazva a $\mathbf{C}_2=\mathbf{C}_3=\mathbf{0}$ helyettesítéssel, a $\psi$ függvényre a

$$\psi_T(\mathbf{A},\mathbf{C}) = E\exp\left\{-c_1\int_0^T \zeta_1^2(t)\,dt - c_2\int_0^T \zeta_2^2(t)\,dt\right\} = \tag{6.11}$$

$$= e^{-1/2A_1T}4a_1A_1\sqrt{a_2A_2}\cdot\{e^{-a_1T}[(a_1a_2-A_1A_2)(a_1-A_1)-(a_2-A_2)^2]^2+$$

$$+e^{a_1T}[(a_1a_2+A_1A_2)(a_1+A_1)-(a_2-A_2)^2]^2+$$

$$+\frac{8a_2}{a_1^2-4a_2}[A_2(-a_1^2+A_1^2)-(a_2-A_2)^2]^2-$$

$$-e^{(\lambda_1-\lambda_2)T}\frac{a_1^2}{a_1^2-4a_2}[(a_1a_2-A_1A_2)(a_1+A_1)+(A_2-a_2)(3a_2+A_2)+2\lambda_1A_1(a_2-A_2)]^2-$$

$$-e^{(\lambda_2-\lambda_1)T}\frac{a_1^2}{a_1^2-4a_2}[(a_1a_2-A_1A_2)(a_1+A_1)+(A_2-a_2)(3a_2+A_2)+2\lambda_2A_1(a_2-A_2)]^2\}^{-1/2}$$

összefüggés adódik.

Legyen most a $\zeta(t)$ kétdimenziós folyamat az

$$\mathbf{A} = \begin{pmatrix}A_1 & A_2\\ A_3 & A_4\end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}$$

mátrixokkal és a $\zeta(0)=\mathbf{0}$ feltétellel megadva. A sajátértékei a negatív félsíkba esnek.

Meghatározzuk a

$$\psi_T(\mathbf{A},\mathbf{C}) = E\exp\left\{\int_0^T \zeta^*(t)\mathbf{C}\zeta(t)\,dt\right\}$$

függvényt, ahol

$$\mathbf{C} = \begin{pmatrix}c_1 & c_2\\ c_2 & c_4\end{pmatrix}$$

negatív szemidefinit szimmetrikus mátrix. A 4.1. tételt használjuk.

Legyen

$$\mathbf{P} = \begin{pmatrix}p_1 & p_2\\ p_2 & p_4\end{pmatrix} = \mathbf{A}^*\mathbf{A}-2\mathbf{C},$$

valamint

$$f = A_2-A_3, \quad h^2 = p_1p_4-p_2^2, \quad g^2 = p_1+p_4+2h-f^2.$$

Ezekkel felírhatók a (4.2) és (4.3) egyenletek megoldása

$$a_1 = \frac{g(p_1+h)+fp_2}{f^2+g^2}, \qquad a_2 = \frac{f(p_4+h)+gp_2}{f^2+g^2}, \tag{6.12}$$

$$a_3 = \frac{-f(p_1+h)+gp_2}{f^2+g^2}, \quad a_4 = \frac{g(p_4+h)-fp_2}{f^2+g^2},$$

$$d_i = \frac{1}{2}(A_i-a_i), \quad i = 1, 2, 4.$$

A feltételekből adódik, hogy $h^2$ és $g^2$ nemnegatívok, azaz az $a_i$ és $d_i$ valósak lesznek ($i=1, 2, 3, 4$).

A (4.4) egyenletet megoldva, a (4.1) és (6.5) összefüggések alkalmazásával a következő eredményt kapjuk:

$$
\begin{aligned}
(6.13)\qquad \psi_T(\mathbf{A},\mathbf{C}) = e^{-1/2T(A_1+A_4)}\cdot\Big\{ & e^{(a_1+a_4)T}\frac{DK}{k(a_1+a_4)^2}+ \\
+\frac{e^{(\lambda_1-\lambda_2)T}}{k(\lambda_1-\lambda_2)^2}\big[ & k[(d_4-d_1)(a_4-a_1)-2d_2(a_2+a_3)]-D[(a_2+a_3)^2+(a_4-a_1)^2]+ \\
& +\lambda_2[-d_1(a_2^2+a_4^2-k)+2d_2(a_1a_2+a_3a_4)-d_4(a_1^2+a_3^2-k)]\big]+ \\
+\frac{e^{(\lambda_2-\lambda_1)T}}{k(\lambda_1-\lambda_2)^2}\big[ & k[(d_4-d_1)(a_4-a_1)-2d_2(a_2+a_3)]-D[(a_2+a_3)^2+(a_4-a_1)^2]+ \\
& +\lambda_1[-d_1(a_2^2+a_4^2-k)+2d_2(a_1a_2+a_3a_4)-d_4(a_1^2+a_3^2-k)]\big]+ \\
+\frac{e^{-(a_1+a_4)T}}{k(a_1+a_4)^2}\big[ & DK+(a_1+a_4)^2k+(a_1+a_4)[(a_1^2+a_3^2+k)d_4+(a_2^2+a_4^2+k)d_1- \\
& -2(a_1a_2+a_3a_4)d_2]\big]+\frac{4(a_3-a_2)}{(\lambda_1-\lambda_2)^2(a_1+a_4)^2}(a_1+a_4)[-a_2d_1+(a_4-a_1)d_2+a_3d_4] \\
& -2D(a_2-a_3)]\Big\}^{-1/2},
\end{aligned}
$$

ahol $k=\det(\tilde{\mathbf{a}})=a_1a_4-a_2a_3$, $K=a_1^2+a_2^2+a_3^2+a_4^2+2k$, $D=\det(\mathbf{D})=d_1d_4-d_2^2$, valamint $\lambda_1$ és $\lambda_2$ az $\tilde{\mathbf{a}}$ mátrix különböző sajátértékei. Speciálisan, ha $\zeta(t)$ „komplex" folyamat, $\zeta(0)=O$ indítással, azaz

$$
\mathbf{A}=\begin{pmatrix}-\lambda & -\omega\\ \omega & -\lambda\end{pmatrix},\quad \mathbf{B}=\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix},\quad \mathbf{C}=\begin{pmatrix}-c & 0\\ 0 & -c\end{pmatrix},\quad \lambda,\omega,c\geqq 0,
$$

akkor (6.12)-ből

$$
\begin{aligned}
a_1 &= -\sqrt{\lambda^2+2c}, & a_2 &= -\omega,\\
a_3 &= \omega, & a_4 &= -\sqrt{\lambda^2+2c}.
\end{aligned}
$$

Észrevehetjük, hogy a $\zeta\to\eta$ transzformáció során a „periódus" változatlan marad, a „hullámhossz" viszont növekszik. Ezeket az értékeket (6.13)-ba helyettesítve:

$$
\begin{aligned}
(6.14)\qquad \psi_T(\mathbf{A},\mathbf{C}) &= E\exp\Big\{-c\int_0^T \zeta_1^2(t)+\zeta_2^2(t)\,dt\Big\} = \\
&= \frac{2e^{\lambda T}\sqrt{\lambda^2+2c}}{(\lambda+\sqrt{\lambda^2+2c})e^{\sqrt{\lambda^2+2c}\cdot T}-(\lambda-\sqrt{\lambda^2+2c})e^{-\sqrt{\lambda^2+2c}\cdot T}}
\end{aligned}
$$

adódik.

A (6.14) formulát összevetve az [1] (4.2.16.) képletével, azt tapasztaljuk, hogy a „konstans szorzók" megegyeznek, csak stacionárius esetben éppen a négyzetük szerepel. Ugyanez figyelhető meg a (6.10) és (6.11) formuláknál, az autoregressziós folyamatok esetén is.

Ezúton szeretném megköszönni ARATÓ MÁTYÁS professzornak, hogy felhívta a figyelmemet a problémára, és a dolgozat megírása során adott hasznos tanácsait.

## IRODALOM

[1] ARATÓ, M., *Linear Stochastic Systems with Constant Coefficients* (Lecture Notes in Control and Information Sciences, Springer Verlag, 1982).
[2] BASAWA, I. W. and PRAKASA-RAO, R., *Statistical Inference for Stochastic Processes* (Academic Press, London, 1980).
[3] NOVIKOV, A. A., "On the estimation of parameters of diffusion processes" (in Russian), *Studia Scientiarum Mathematicae Hungaricae* 7 (1972), 201—209.
[4] KUCERA, V., "A review of the matrix Riccati equation", *Kybernetika* 9 (1973) 43—61.
[5] KUTOJANC, J. A., "Parameter estimation of random processes" (in Russian), Ak. Nauk, Jerevan, 1980.
[6] Арато, М., Точные формулы для плотностей мер элементарных гауссовских процессов *Studia Scientiarum Mathematicae Hungaricae* 5 (1970) 17—27.
[7] Липцер, Р. и Ширяев, А. Н. Статистика случайных процессов (Наука, Москва, 1974).

(Beérkezett: 1984. március 5.)

(Átdolgozva beérkezett: 1984. április 2.)

KONCZ KÁROLY
SZÁMALK
1115 Budapest, Szakasits Á. u. 68.

## ON THE ESTIMATION OF PARAMETERS OF A DIFFUSIONAL TYPE PROCESS WITH CONSTANT DRIFT

K. KONCZ

We consider the multidimensional diffusional type stochastic process given by (2.1). In this paper we prove Theorem 4.1, that in the nonstationary case, which means $\xi(0)=0$, the *Laplace transform* (2.8) of the distribution of sufficient statistics has the simple form (4.1). This is the generalization of a result of NOVIKOV [3] for the one dimensional case. In the stationary case (5.2) we prove foula (5.5) (Theorem 5.2), generalizing a result of ARATÓ—BENCZÚR (see in [1]).

*Alkalmazott Matematikai Lapok* **10** (1984) 323—352

# TAPASZTALATI FÜGGVÉNYEK SIMÍTÁSA $l_p$ PROGRAMOZÁSSAL

TERLAKY TAMÁS

Budapest

Ekvidisztáns pontokban adott közelítő függvényértékek simítására három modellt mutatunk be. Az első modell megegyezik a jól ismert *Whittaker-féle simítási modellel*, a második és harmadik modell új.

Mindhárom modell megoldására az $l_p$ programozás egyensúlyi feltételein alapuló interációs eljárást adunk, így a *Whittaker simítási modell* megoldására is egy új megoldási módszert adunk

## 1. Bevezetés

Mérési eredmények feldolgozása során számos problémával találkozunk, mivel ezek a megfigyelési értékek hibával terheltek. Így amennyiben ezekből a hibás adatokból határozzuk meg a pontokban a $k$-adik differenciákat, melyekkel a $k$-adik differenciálhányadost közelítjük, akkor azt tapasztaljuk, hogy az így számított értékek használhatatlanok a halmozott hibák miatt. Ezért függvényértékeinket simítani kell.

Az irodalomban számos simítási eljárás ismert. Ezek közül talán WHITTAKER [8] módszere a legelterjedtebb. WHITTAKER az általa több, mint hatvan évvel ezelőtt konstruált simítási modellre megoldási eljárást is adott. WHITTAKER módszerét NYÍRI [1] 1980-ban általánosította. NYÍRI dolgozata hívta fel figyelmünket a problémára. Dolgozatunkban új iteratív megoldási módszert adunk WHITTAKER simítási modelljére, valamint két új simítási modellt konstruálunk, melyek megoldására egyszerű iteratív eljárást közlünk. A modellek összehasonlítására mintafeladatokkal nyert számítástechnikai tapasztalatokat is közlünk.

A szükséges $l_p$ programozási ismeretek TERLAKY [7] dolgozatában találhatók, de a dolgozatunkban felhasznált eredményeket a *Függelékben* közöljük.

## 2. A simítási modellek konstruálása

Legyenek $\gamma_1, \ldots, \gamma_n$ egy $f$ függvény ismeretlen $y_1, \ldots, y_n$ értékeit ekvidisztáns pontokban közelítő mérési, megfigyelési értékek. Legyenek továbbá $\varepsilon_1, \ldots, \varepsilon_{n-k}$ az $f$ függvény ismeretlen $\Delta^k y_1, \ldots, \Delta^k y_{n-k}$ $k$-adik differenciáit közelítő mérési, megfigyelési értékek $\left(\Delta^k y_i = \sum_{j=0}^{k} (-1)^{k-j} \binom{k}{j} y_{i+j}, \quad i=1, \ldots, n-k\right)$. Természetesen a megfigyelési eredmények véletlen hibával terheltek, így a $\gamma_1, \ldots, \gamma_n$ értékekből számított $k$-adik differenciák nagymértékben eltérhetnek a valamilyen más módon mért $\varepsilon_1, \ldots, \varepsilon_{n-k}$ megfigyelt $k$-adik differenciáktól. Ezért az $y_1, \ldots, y_n$ függvényértékeket

úgy kell megválasztanunk, hogy ezek a megfigyelt $\gamma_1, \ldots, \gamma_n$ értékektől se térjenek el nagyon, de a megfigyelt $\varepsilon_1, \ldots, \varepsilon_{n-k}$ $k$-adik differenciák is jó közelítései legyenek a valódi $\Delta^k y_i$ $i=1, \ldots, n$ értékeknek, azaz a $\Delta^k y_i$ $k$-adik differenciák „jól simuljanak" a megfigyelési értékekhez.

Valószínűségszámítási megfontolások alapján három simítási modellt konstruálunk. A modellek konstruálásához az alábbi két feltevéssel élünk.

Tegyük fel, hogy a $\gamma_1, \ldots, \gamma_n$ értékek mérési hibái egymástól független, azonos, zérus várható értékű és $\sigma/\lambda$ szórású normális eloszlású valószínűségi változók.

Tegyük fel, hogy az $\varepsilon_1, \ldots, \varepsilon_{n-k}$ értékek mérési hibái egymástól és a $\gamma_i$ értékek mérési hibáitól is független, azonos, zérus várható értékű és $\sigma$ szórású normális eloszlású valószínűségi változók.

A fenti feltételeket figyelembe véve feladatunk azon $y_1, \ldots, y_n$ értékek meghatározása, melyek a fenti megfigyelési értékek mellett a legvalószínűbb értékei az ismeretlen $f$ függvénynek.

*I. Modell.* A maximum likelihood elv alapján feladatunk azon $y_1, \ldots, y_n$ értékek meghatározása, melyek maximalizálják a $\gamma_1, \ldots, \gamma_n, \varepsilon_1, \ldots, \varepsilon_{n-k}$ értékek megfigyelésének valószínűségét, azaz az alábbi $L$ likelihood függvényt:

$$L(\gamma_1, \ldots, \gamma_n, \varepsilon_1, \ldots \varepsilon_{n-k}, y_1, \ldots, y_n) = \prod_{i=1}^{n} \frac{\lambda}{\sqrt{2\pi}\,\sigma} e^{-\frac{\lambda^2}{2\sigma^2}(y_i-\gamma_i)^2} \times$$

$$\times \prod_{i=1}^{n-k} \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2\sigma^2}(\Delta^k y_i-\varepsilon_i)^2}.$$

Az $L$ függvény maximalizálása ekvivalens.

$$\lambda^2 \sum_{i=1}^{n} |y_i-\gamma_i|^2 + \sum_{i=1}^{n-k} |\Delta^k y_i-\varepsilon_i|^2$$

minimumának meghatározásával.

*II. Modell.* Tartsuk meg a $\Delta^k y_i$ $k$-adik differenciákra vonatkozó feltételezéseinket, azaz, hogy az $\varepsilon_1, \ldots, \varepsilon_{n-k}$ értékek mérési hibái egymástól független, azonos eloszlású valószínűségi változók, zérus várható értékkel, $\sigma$ szórással. A $\gamma_1, \ldots, \gamma_n$ mérési eredményekről azonban csak azt tételezzük fel, hogy jó közelítései az $y_1, \ldots, y_n$ értékeknek, vagyis hogy a pontonkénti mérési hibák négyzetösszege nem haladja meg az adott $\delta^2$ értéket, azaz az $(y_1, \ldots, y_n)$ és $(\gamma_1, \ldots, \gamma_n)$ pontok euklideszi távolsága nem haladja meg $\delta$-t.

A maximum likelihood elv alapján feladatunk azon $y_1, \ldots, y_n$ értékek meghatározása, melyek kielégítik a $\gamma_1, \ldots, \gamma_n$ értékek hibáira ismert feltételt, és egyidejűleg maximalizálják az $\varepsilon_1, \ldots, \varepsilon_{n-k}$ $k$-adik differenciák megfigyelésének valószínűségét, azaz feladatunk maximalizálni a

$$\prod_{i=1}^{n-k} \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2\sigma^2}(\Delta^k y_i-\varepsilon_i)^2}$$

függvényt, feltételezve, hogy

$$\sum_{i=1}^{n} |y_i-\gamma_i|^2 \leqq \delta^2.$$

Ez a feladat pedig ekvivalens a következő feladattal:

$$\min \sum_{i=1}^{n-k} |\Delta^k y_i - \varepsilon_i|^2$$

$$\sum_{i=1}^{n} |y_i - \gamma_i|^2 \leqq \delta^2.$$

*III. Modell.* Tételezzük fel ugyanazokat a $\gamma_1, \ldots, \gamma_n$ mérési értékek hibáira, mint az első modell esetén, azaz legyenek egymástól független, azonos, zérus várható értékű normális eloszlású valószínűségi változók. Az ismeretlen $\Delta^k y_i$ $k$-adik differenciákról viszont azt tételezzük fel, hogy közel vannak a megfigyelt $\varepsilon_1, \ldots, \varepsilon_{n-k}$ értékekhez, azaz a pontonkénti hibák négyzetösszege nem haladja meg az adott $\delta^2$ értéket, vagyis a $(\Delta^k y_1, \ldots, \Delta^k y_{n-k})$ és $(\varepsilon_1, \ldots, \varepsilon_{n-k})$ pontok euklideszi távolsága ne legyen több $\delta$-nál $R^{n-k}$-ban.

A maximum likelihood elv alapján feladatunk azon $y_1, \ldots, y_n$ értékek meghatározása, melyek kielégítik a $k$-adik differenciákra tett feltételünket, és egyidejűleg maximalizálják a $\gamma_1, \ldots, \gamma_n$ értékek megfigyelésének valószínűségét, azaz feladatunk maximalizálni a

$$\prod_{i=1}^{n} \frac{\lambda}{\sqrt{2\pi}\,\sigma} e^{-\frac{\lambda^2}{2\sigma^2}(y_i - \gamma_i)^2}$$

függvényt, feltéve, hogy

$$\sum_{i=1}^{n-k} |\Delta^k y_i - \varepsilon_i|^2 \leqq \delta^2.$$

Ez a feladat pedig ekvivalens az alábbi feladattal:

$$\min \sum_{i=1}^{n} |y_i - \gamma_i|^2$$

$$\sum_{i=1}^{n-k} |\Delta^k y_i - \varepsilon_i|^2 \leqq \delta^2.$$

A további fejezetekben a fent megfogalmazott három modell megoldására adunk módszert, majd számítási tapasztalatainkat közöljük.

## 3. A Whittaker-féle simítás

### *3.1. Dualitás.*

Mint az előző fejezetben láttuk, első modellünk szerint feladatunk a

$$\lambda^2 \sum_{i=1}^{n} |y_i - \gamma_i|^2 + \sum_{i=1}^{n-k} |\Delta^k y_i - \varepsilon_i|^2 \tag{3.1}$$

függvényt minimalizáló $y_1, \ldots, y_n$ értékek meghatározása. A $\lambda^2 > 0$ simítási paramétert a mért értékek vizsgálata alapján kell megválasztani, esetleg más $\lambda$ értékekkel újra számolni.

A WHITTAKER [8] és NYÍRI [1] által adott eljárások mindegyike differencia egyenletrendszer megoldásán alapszik. Most egy, ezektől lényegesen eltérő eljárást közlünk a (3.1) függvényt minimalizáló $y_1, \ldots, y_n$ értékek meghatározására.

Feladatunk tehát meghatározni

$$\min\left\{\lambda^2 \sum_{i=1}^{n} |y_i-\gamma_i|^2+\sum_{i=1}^{n-k} |\Delta^k y_i-\varepsilon_i|^2\right\}.$$

Jelölje A azt a teljes rangú mátrixot, melynek $i$-edik sora a $\Delta^k y_i$ differencia együtthatóit tartalmazza. (Azaz $(\mathbf{Ay})_i=\Delta^k y_i$.) Így A-nak $n-k$ sora és $n$ oszlopa van.

Jelöljük $\mathbf{a}^2$-tel az $\mathbf{a}\in R^n$ vektor önmagával vett skalárszorzatát, így feladatunk a következő formában írható fel:

$$\min\{\lambda^2(\mathbf{y}-\mathbf{c})^2+(\mathbf{Ay}-\mathbf{e})^2\}. \tag{3.2}$$

Így feladatunk egy kvadratikus függvény feltétel nélküli minimumának meghatározása. Ennek duálja a következő:

$$\max\left\{2\mathbf{xAc}-\frac{1}{\lambda^2}(\mathbf{xA})^2-2\mathbf{xe}-\mathbf{x}^2\right\} \tag{3.3}$$

ahol $\mathbf{x}\in R^{n-k}$.

Elemi egyenlőtlenségek segítségével látható be az alábbi lemma:

3.1. LEMMA. Tetszőleges $\mathbf{y}\in R^n$ és $\mathbf{x}\in R^{n-k}$ esetén

$$\lambda^2(\mathbf{y}-\mathbf{c})^2+(\mathbf{Ay}-\mathbf{e})^2 \geqq 2\mathbf{xAc}-\frac{1}{\lambda^2}\mathbf{xA}^2-2\mathbf{xe}-\mathbf{x}^2. \tag{3.4}$$

Egyenlőséggel akkor és csak akkor igaz, ha

$$\mathbf{x}=\mathbf{Ay}-\mathbf{e}, \tag{3.5}$$

$$\mathbf{y}=\mathbf{c}-\frac{1}{\lambda^2}\mathbf{xA}. \tag{3.6}$$

*Bizonyítás.* Nyilván fennáll az alábbi két egyenlőtlenség

$$0 \leqq [\mathbf{x}-(\mathbf{Ay}-\mathbf{e})]^2 = \mathbf{x}^2+(\mathbf{Ay}-\mathbf{e})^2-2\mathbf{xAy}+2\mathbf{xe}.$$

Egyenlőség akkor és csak akkor áll fenn, ha (3.5) igaz.

$$0 \leqq \lambda^2\left[(\mathbf{y}-\mathbf{c})+\frac{1}{\lambda^2}\mathbf{xA}\right]^2=\lambda^2(\mathbf{y}-\mathbf{c})^2+\frac{1}{\lambda^2}(\mathbf{xA})^2+2\mathbf{xAy}-2\mathbf{xAc}.$$

Egyenlőség akkor és csak akkor áll fenn, ha (3.6) igaz.

Adjuk össze a fenti két egyenlőtlenséget:

$$0 \leqq \lambda^2(\mathbf{y}-\mathbf{c})^2+(\mathbf{Ay}-\mathbf{e})^2+\mathbf{x}^2-2\mathbf{xAy}+2\mathbf{xe}+\frac{1}{\lambda^2}(\mathbf{xA})^2+2\mathbf{xAy}-2\mathbf{xAc}.$$

Átrendezve nyerjük a kívánt összefüggést:

$$\lambda^2(\mathbf{y}-\mathbf{c})^2+(\mathbf{Ay}-\mathbf{e})^2 \geqq 2\mathbf{xAc}-\frac{1}{\lambda^2}(\mathbf{xA})^2-2\mathbf{xe}-\mathbf{x}^2.$$

KÖVETKEZMÉNY. Ha valamely $\bar{\mathbf{y}} \in R^n$, $\bar{\mathbf{x}} \in R^{n-k}$ esetén egyenlőség áll fenn (3.4)-ben, akkor $\bar{\mathbf{y}}$ optimális megoldása (3.2) feladatnak, $\bar{\mathbf{x}}$ optimális megoldása a (3.3) feladatnak.

*Bizonyítás.* A lemma alapján nyilvánvaló.

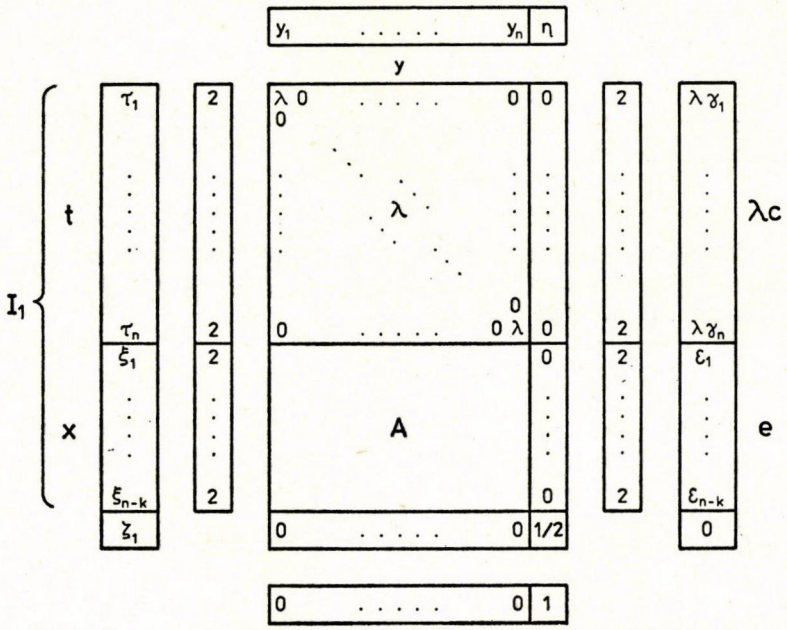
3.2. TÉTEL. A (3.2) és (3.3) feladatoknak létezik egy, és csak egy optimális megoldása úgy, hogy (3.4)-ben egyenlőség áll fenn.

*Bizonyítás.* A (3.2) feladat ekvivalens az alábbi $l_p$ programozási *primál* feladattal:

$$\max \eta,$$

$$\sum_{i=1}^{n} \frac{1}{2} |\lambda y_i - \lambda \gamma_i|^2 + \sum_{i=1}^{n-k} \frac{1}{2} |\Delta^k y_i - \varepsilon_i|^2 + \frac{1}{2} \eta \leqq 0.$$

A feladat szerkezetét az 1. ábra szemlélteti.




*1. ábra*

Az 1. ábra jelöléseit használva a *duál* feladat a következő:

$$\lambda \mathbf{t} + \mathbf{A}^T \mathbf{x} = \mathbf{0}$$

$$\frac{1}{2} \zeta_1 = 1$$

$$\min \left\{ \lambda \mathbf{t}\mathbf{c} + \mathbf{x}\mathbf{e} + \frac{1}{2\zeta_1} \mathbf{x}^2 + \frac{1}{2\zeta_1} \mathbf{t}^2 \right\}.$$

Ez pedig nyilván ekvivalens az alábbi feltétel nélküli feladattal:

$$\min\left\{-\mathbf{xAc}+\mathbf{xe}+\frac{1}{4}\mathbf{x}^2+\frac{1}{4\lambda^2}(\mathbf{xA})^2\right\}.$$

Melyből $\mathbf{x}=2\mathbf{x}$ helyettesítéssel nyerjük (3.3) duál feladatot.

A primál feladat nyilván *Slater reguláris* (l. Függelék), mert rögzített $\mathbf{y}$ mellett alkalmas $\eta<0$-t választva teljesül a *Slater regularitási feltétel*. A primál feladat célfüggvénye nyilván felülről korlátos (nempozitív), így létezik optimális megoldása (F1. tétel), ebből következik (F2. tétel), hogy a duál feladatnak is van optimális megoldása, valamint (3.4)-ben egyenlőség áll fenn az optimális megoldásokra. Könnyen látható, hogy az $l_p$ programozás egyensúlyi feltételei a (3.5) és (3.6) egyensúlyi feltételekkel ekvivalensek.

Az optimális megoldások egyértelműek, mivel véve egy tetszőleges $\bar{\mathbf{y}}$ optimális megoldását a primál feladatnak, erre (3.5) fenn áll tetszőleges $\bar{\mathbf{x}}$ optimális megoldására a duál feladatnak, így $\bar{\mathbf{x}}=\mathbf{A}\bar{\mathbf{y}}-\mathbf{e}$ egyértelmű. Hasonlóan (3.6) miatt $\bar{\mathbf{y}}=\mathbf{c}-\frac{1}{\lambda^2}\bar{\mathbf{x}}\mathbf{A}$ szintén egyértelmű.

*Megjegyzés.* Tételünket ebben az egyszerű esetben elemi úton is igazolhattuk volna, de az egységes tárgyalásmód kedvéért választottuk az $l_p$ programozás segítségével történő bizonyítást.

### *3.2. Az optimális megoldás meghatározása*

A fent bizonyított 3.1. lemma és 3.2. tétel ismeretében tudjuk, hogy feladatunknak pontosan egy megoldása létezik, amelyet a (3.5), (3.6) egyenletrendszer megoldása útján nyerhetünk. Ha (3.6)-ot (3.5)-be helyettesítjük, az alábbi képletet nyerjük $\mathbf{x}$-re:

$$\mathbf{x}=-\frac{1}{\lambda^2}\mathbf{A}\mathbf{A}^T\mathbf{x}+\mathbf{Ac}-\mathbf{e}. \tag{3.7}$$

Legyen $\mathbf{B}=\mathbf{A}\mathbf{A}^T$, nyilván valós, szimmetrikus, pozitív definit mátrix. ($\mathbf{A}$ sorai nyilván függetlenek, mivel tartalmaz olyan felsőháromszög mátrixot, melynek diagonálisában vagy $+1$ vagy $-1$-esek állnak.)

Így feladatunk a (3.7) lineáris egyenletrendszer megoldására redukálódott, amely $n-k$ ismeretlent tartalmaz. Az optimális $\bar{\mathbf{x}}$ értékből (3.6) képlet segítségével határozhatjuk meg az optimális $\bar{\mathbf{y}}$ értéket.

Az optimális $\bar{\mathbf{x}}$ értéket az alábbi módon határozhatjuk meg: Legyen $K$ a $\mathbf{B}$ mátrix legnagyobb sajátértékének egy felső becslése. (Erre $\mathbf{B}$ sornormája megfelelő értéket ad, ami a $2k$-adik differencia együtthatói abszolútértékének összege.) Ekkor (3.7) egyenletrendszer az alábbi ekvivalens alakra hozható, ahol $\mathbf{E}$ az $n-k$ dimenziós egységmátrixot jelöli:

$$\mathbf{x}=\frac{K}{K+2\lambda^2}\left(\mathbf{E}-\frac{2}{K}\mathbf{B}\right)\mathbf{x}+\frac{2\lambda^2}{K+2\lambda^2}(\mathbf{Ac}-\mathbf{e}).$$

Ezt a következő lineáris iterációval oldhatjuk meg:

$$\mathbf{x}^{(k+1)}=\frac{K}{K+2\lambda^2}\left(\mathbf{E}-\frac{2}{K}\mathbf{B}\right)\mathbf{x}^{(k)}+\frac{2\lambda^2}{K+2\lambda^2}(\mathbf{Ac}-\mathbf{e}). \tag{3.8}$$

A (3.8) iteráció tetszőleges $\mathbf{x}^{(0)}$ kezdeti értékből indulva konvergens, mivel $\mathbf{E}-\frac{2}{K}\mathbf{B}$ spektrálsugara kisebb, mint egy, valamint $\frac{K}{K+2\lambda^2}$ is kisebb, mint egy. Használhatjuk a jól ismert hibabecslő formulákat is, ezek megtalálhatók pl. SZIDAROVSZKY [6] és RALSTON [5] könyvében.

## 4. Simítás, amikor a $\gamma_i$ függvény értékek mérési pontossága ismert

### *4.1. Dualitás*

A második fejezetben megfogalmazott II. modell szerint feladatunk meghatározni azon „legsimább" ($k$-adik differenciában mérve) $y_1, \ldots, y_n$ értékeket, melyek euklideszi távolsága a mért $\gamma_1, \ldots, \gamma_n$ értékektől $\delta$-nál nem nagyobb. Azaz

$$\min (\mathbf{Ay}-\mathbf{e})^2 \tag{4.1}$$

$$(\mathbf{y}-\mathbf{c})^2 \leqq \delta^2.$$

Amennyiben $\delta^2=0$, akkor $\mathbf{y}=\mathbf{c}$, és feladatunkat megoldottuk. Így feltehető a továbbiakban, hogy $\delta^2>0$.

A (4.1) feladat egy kvadratikus feltételes kvadratikus programozási feladat, melynek duálja az alábbi:

$$\max\left(2\mathbf{xAc}-\mathbf{x}^2-2\mathbf{xe}-2\delta\sqrt{(\mathbf{xA})^2}\right). \tag{4.2}$$

ahol $\mathbf{x}\in R^{n-k}$.

Elemi úton igazolható a következő lemma:

4.1. LEMMA. Tetszőleges $\mathbf{y}\in R^n$-re, melyre $(\mathbf{y}-\mathbf{c})^2\leqq\delta^2$, és $\mathbf{x}\in R^{n-k}$-ra

$$(\mathbf{Ay}-\mathbf{e})^2 \geqq 2\mathbf{xAc}-\mathbf{x}^2-2\mathbf{xe}-2\delta\sqrt{(\mathbf{xA})^2} \tag{4.3}$$

egyenlőséggel, akkor és csak akkor, ha

$$\mathbf{x}=\mathbf{Ay}-\mathbf{e} \tag{4.4}$$

és vagy $\mathbf{xA}=\mathbf{0}$ (ami azt jelenti, hogy $\mathbf{x}=\mathbf{0}$ és $\mathbf{Ay}=\mathbf{e}$)

$$\text{vagy } \mathbf{y}=\mathbf{c}-\frac{\delta}{\sqrt{(\mathbf{xA})^2}}\mathbf{xA} \text{ (ami azt jelenti, hogy } (\mathbf{y}-\mathbf{c})^2=\delta^2). \tag{4.5}$$

*Bizonyítás.* Ha $\mathbf{xA}=\mathbf{0}$, akkor $\mathbf{x}=\mathbf{0}$, mivel az $\mathbf{A}$ teljes sorrangú mátrix, ekkor a duál feladat célfüggvény értéke 0, és mivel a primál feladat célfüggvény értéke nemnegatív, így a (4.3) egyenlőtlenség igaz. Egyenlőség nyilván akkor és csak akkor áll fenn, ha $\mathbf{Ay}=\mathbf{e}$.

Ha $\mathbf{xA}\neq\mathbf{0}$, akkor nyilván igaz az alábbi két egyenlőtlenség:

$$0 \leqq [\mathbf{x}-(\mathbf{Ay}-\mathbf{e})]^2 = (\mathbf{Ay}-\mathbf{e})^2+\mathbf{x}^2-2\mathbf{xAy}+2\mathbf{xe}. \tag{4.6}$$

Egyenlőséggel akkor és csak akkor, ha $\mathbf{x}=\mathbf{Ay}-\mathbf{e}$, azaz, ha (4.4) igaz.

(4.7)
$$0 \leqq \frac{\sqrt{(\mathbf{xA})^2}}{\delta}\left[(\mathbf{y}-\mathbf{c})+\frac{\delta}{\sqrt{(\mathbf{xA})^2}}\mathbf{xA}\right]^2 = \frac{\sqrt{(\mathbf{xA})^2}}{\delta}(\mathbf{y}-\mathbf{c})^2+\delta\sqrt{(\mathbf{xA})^2}+2\mathbf{xAy}-2\mathbf{xAc}.$$

Eygnəlőséggel akkor és csak akkor, ha $\mathbf{y}=\mathbf{c}-\frac{\delta}{\sqrt{(\mathbf{xA})^2}}\mathbf{xA}$, azaz, ha (4.5) igaz.

A (4.6) egyenlőtlenséget átrendezve:

$$(\mathbf{Ay}-\mathbf{e})^2 \geqq 2\mathbf{xAy}-\mathbf{x}^2-2\mathbf{xe}.$$

Ebből a (4.7) egyenlőtlenséget kivonva:

$$(\mathbf{Ay}-\mathbf{e})^2 \geqq 2\mathbf{xAy}-\mathbf{x}^2-2\mathbf{xe}-\frac{\sqrt{(\mathbf{xA})^2}}{\delta}(\mathbf{y}-\mathbf{c})^2-\delta\sqrt{(\mathbf{xA})^2}-2\mathbf{xAy}+2\mathbf{xAc} =$$

$$= 2\mathbf{xAc}-\mathbf{x}^2-2\mathbf{xe}-\delta\sqrt{(\mathbf{xA})^2}-\frac{\sqrt{(\mathbf{xA})^2}}{\delta}(\mathbf{y}-\mathbf{c})^2.$$

(4.1) feltételt alkalmazva — egyenlőség akkor és csak akkor áll fenn, ha $(\mathbf{y}-\mathbf{c})^2 = \delta^2$, ami (4.5) feltételből is következik — a kívánt egyenlőtlenséget nyerjük:

$$(\mathbf{Ay}-\mathbf{e})^2 \geqq 2\mathbf{xAc}-\mathbf{x}^2-2\mathbf{xe}-2\delta\sqrt{(\mathbf{xA})^2}.$$

KÖVETKEZMÉNY. Ha valamely $\mathbf{y}\in R^n$, $(\mathbf{y}-\mathbf{c})^2 \leqq \delta^2$ és $\mathbf{x}\in R^{n-k}$ esetén (4.3)-ban egyenlőség áll fenn, akkor $\mathbf{y}$ és $\mathbf{x}$ optimális megoldásai a (4.1), illetve a (4.2) feladatnak.

*Bizonyítás.* A lemma alapján kézenfekvő.

$l_p$ programozási ismereteinket felhasználva belátjuk, hogy a (4.1) primál és a (4.2) duál feladatnak is pontosan egy optimális megoldása létezik.

4.2. TÉTEL. A (4.1) és (4.2) feladatoknak egy és csak egy optimális megoldása létezik, melyekre (4.3)-ban egyenlőség áll fenn.

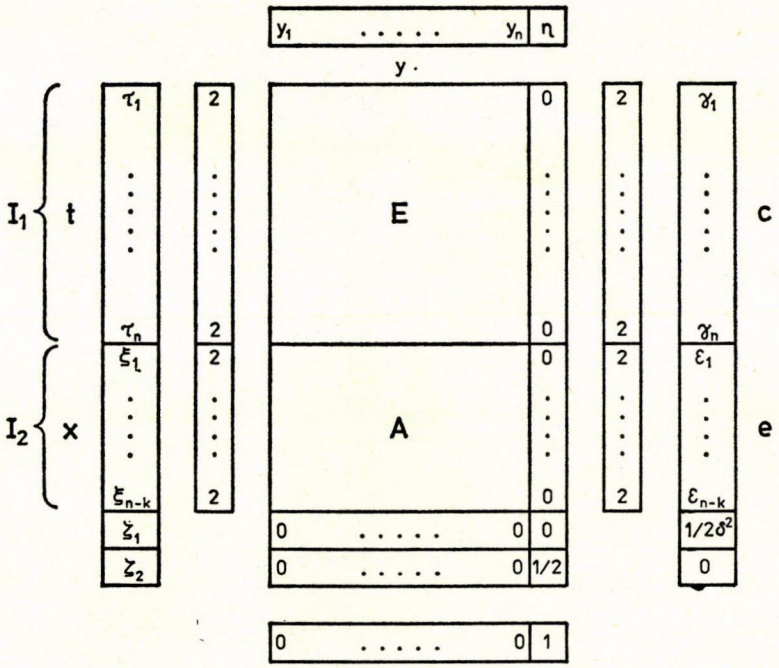
*Bizonyítás.* Az alábbi formában a (4.1) feladat egy $l_p$ programozási *primál* feladat:

$$\max \eta$$

$$\frac{1}{2}|\mathbf{Ay}-\mathbf{e}|^2+\frac{1}{2}\eta \leqq 0$$

$$\frac{1}{2}|\mathbf{y}-\mathbf{c}|^2-\frac{1}{2}\delta^2 \leqq 0.$$

A feladat szerkezetét a 2. ábra szemlélteti.




*2. ábra*

Az ábra jelöléseit felhasználva a *duál* feladat a következő:

$$\mathbf{t}+\mathbf{xA}=\mathbf{0}$$

$$\zeta_1 \geqq 0, \quad \frac{1}{2}\zeta_2 = 1$$

$$\zeta_1 = 0 \Rightarrow \mathbf{t} = \mathbf{0}$$

$$\min\left(\mathbf{tc}+\mathbf{xe}+\frac{\zeta_1\delta^2}{2}+\frac{1}{2\zeta_2}\mathbf{x}^2+\begin{cases}\dfrac{1}{2\zeta_1}\mathbf{t}^2, & \text{ha} \quad \zeta_1>0\\ 0, & \text{ha} \quad \zeta_1=0\end{cases}\right).$$

Átalakítva:

$$\zeta_1 \geqq 0, \quad \zeta_1 = 0 \Rightarrow \mathbf{xA} = \mathbf{0}$$

$$\min\left(-\mathbf{xAc}+\frac{\zeta_1\delta^2}{2}+\mathbf{xe}+\frac{1}{4}\mathbf{x}^2+\begin{cases}\dfrac{1}{2\zeta_1}(\mathbf{xA})^2, & \text{ha} \quad \zeta_1>0\\ 0, & \text{ha} \quad \zeta_1=0\end{cases}\right).$$

A primál feladat *Slater reguláris*, mivel $\mathbf{y}=\mathbf{c}$, és alkalmas $\eta<0$-t választva teljesül a *Slater regularitási feltétel*. A primál feladat célfüggvénye felülről korlátos, így létezik optimális megoldása (F1. és F2. tételek miatt) a primál és a duál feladatnak is.

Meg kell mutatnunk, hogy duál feladatunk ekvivalens a (4.2) feladattal. Tekintsük az $l_p$ programozás egyensúlyi feltételeit:

$$\mathbf{x} = 2(\mathbf{A}\mathbf{y}-\mathbf{c}) \tag{4.8}$$

$$\mathbf{t} = \zeta_1(\mathbf{y}-\mathbf{c}) \tag{4.9}$$

$$\zeta_1[(\mathbf{y}-\mathbf{c})^2-\delta^2] = 0. \tag{4.10}$$

A $\mathbf{t}=-\mathbf{x}\mathbf{A}$ és a (4.9) összefüggésből az $(\mathbf{y}-\mathbf{c})=\frac{-\mathbf{x}\mathbf{A}}{\zeta_1}$ összefüggést nyerjük ($\zeta_1>0$ esetén), amelyet a (4.10)-be helyettesítve a

$$\zeta_1 = \frac{\sqrt{(\mathbf{x}\mathbf{A})^2}}{\delta}$$

összefüggést nyerjük, amelyet a duál feladatba helyettesítve, egy vele ekvivalens feladatot nyerünk:

$$\min\left(-\mathbf{x}\mathbf{A}\mathbf{c}+\mathbf{x}\mathbf{e}+\frac{1}{4}\mathbf{x}^2+\frac{1}{2}\delta\sqrt{(\mathbf{x}\mathbf{A})^2}+\frac{1}{2}\delta\sqrt{(\mathbf{x}\mathbf{A})^2}\right).$$

(Azt is felhasználtuk, hogy $\mathbf{x}\mathbf{A}=\mathbf{0}$-ból következik, hogy $\mathbf{x}=\mathbf{0}$.) Az $\mathbf{x}=2\mathbf{x}$ helyettesítést elvégezve nyerjük az eredeti (4.2) duál feladatot:

$$\max\left(2\mathbf{x}\mathbf{A}\mathbf{c}-2\mathbf{x}\mathbf{e}-\mathbf{x}^2-2\delta\sqrt{(\mathbf{x}\mathbf{A})^2}\right).$$

Így beláttuk, hogy (4.1) és (4.2) feladatok $l_p$ programozási primál és duál feladatokkal ekvivalensek, melyeknek van optimális megoldásuk, és az optimális célfüggvény értékek megegyeznek, azaz (4.3)-ban egyenlőség áll fenn.

Az optimális megoldások egyértelműek, mivel tetszőleges optimális megoldáspár kielégíti a (4.8), (4.9) és (4.10) egyensúlyi feltételeket. Így tetszőleges, de rögzített $\bar{\mathbf{y}}$ optimális megoldás esetén az egyensúlyi feltételek, valamint a duál feltételek miatt $\bar{\mathbf{x}}$, $\bar{\mathbf{t}}$, $\bar{\zeta}_1$ megoldás egyértelmű. Ugyanezen feltételekből következik, hogy $\bar{\mathbf{y}}$ optimális megoldás egyértelműen állítható elő a duál feladat egyértelmű optimális megoldásával.

### *4.2. Az optimális megoldás meghatározása*

#### *a) Ha zérus az optimális célfüggvény érték*

Amennyiben $\mathbf{x}=\mathbf{0}$ optimális megoldása a duál feladatnak, akkor az eddigiek alapján a (4.1) feladat optimális megoldásához tartozó célfüggvény érték is zérus, tehát az

$$\mathbf{A}\mathbf{y} = \mathbf{e} \tag{4.11}$$

$$(\mathbf{y}-\mathbf{c})^2 \leqq \delta^2 \tag{4.12}$$

egyenlőtlenségrendszer megoldható.

Legyen $\mathbf{A}=[\mathbf{D}, \mathbf{N}]$, ahol $\mathbf{D}$: $(n-k)\times k$ mátrix, és $\mathbf{N}$: $(n-k)\times(n-k)$ reguláris alsó háromszög mátrix, valamint $\mathbf{y}=(\mathbf{y}^1, \mathbf{y}^2)$, és $\mathbf{c}=(\mathbf{c}^1, \mathbf{c}^2)$, ahol $\mathbf{c}^1, \mathbf{y}^1 \in R^k$, $\mathbf{c}^2, \mathbf{y}^2 \in R^{n-k}$. Ekkor (4.11)-ből

$$\mathbf{D}\mathbf{y}^1+\mathbf{N}\mathbf{y}^2 = \mathbf{e}$$

$$\mathbf{y}^2 = \mathbf{N}^{-1}(\mathbf{e}-\mathbf{D}\mathbf{y}^1).$$

Így feladatunk ellenőrizni, hogy $(\mathbf{y}^1-\mathbf{c}^1)^2+(\mathbf{N}^{-1}\mathbf{D}\mathbf{y}^1-\mathbf{N}^{-1}\mathbf{e}+\mathbf{c}^2)^2$ kvadratikus függvény minimuma $\delta^2$-nél nem nagyobb-e. Ezt könnyű ellenőrizni, mivel a változók száma kicsi ($k$), de a *Whittaker modell* megoldásakor bemutatott módszert is alkalmazhatjuk. Ha a minimum $\delta^2$-nél nem nagyobb, akkor az így nyert megoldás optimális. Amennyiben $\delta^2$-nél nagyobb a minimális érték, akkor ellentmondásra jutottunk, az $\mathbf{x}=\mathbf{0}$ nem optimális megoldása a duál feladatnak.

*b) Ha az optimális célfüggvény érték nem zérus*

Ebben az esetben $\mathbf{x}=\mathbf{0}$ nem optimális megoldása (4.2)-nek. A (4.4), (4.5) optimalitási kritériumokból nyerjük, hogy az optimális $\mathbf{x}$ vektor kielégíti az

$$\mathbf{x}=-\frac{\delta}{\sqrt{(\mathbf{x}\mathbf{A})^2}}\mathbf{A}\mathbf{A}^T\mathbf{x}+\mathbf{A}\mathbf{c}-\mathbf{e} \tag{4.13}$$

egyenletet, melyről 4.2. tétel alapján tudjuk, hogy pontosan egy megoldása létezik. Oldjuk meg a fenti egyenletet az

$$\mathbf{x}^{(k+1)}=-\frac{\delta}{\sqrt{(\mathbf{x}^{(k)}\mathbf{A})^2}}\mathbf{A}\mathbf{A}^T\mathbf{x}^{(k)}+\mathbf{A}\mathbf{c}-\mathbf{e} \tag{4.14}$$

iterációval.

A 3.2. fejezetben bevezetett jelölés szerint legyen most is $\mathbf{B}=\mathbf{A}\mathbf{A}^T$, szimmetrikus reguláris mátrix.

4.3. LEMMA. A (4.14) iteráció tetszőleges $\mathbf{x}^{(0)}\neq\mathbf{0}$ vektorból indítva alulról is és felülről is korlátos sorozatot állít elő.

*Bizonyítás.* Tudjuk (l. Függelék), hogy $\sqrt{(\mathbf{x}\mathbf{A})^2}=\|\mathbf{x}\|_\mathbf{A}$ vektornorma, valamint $\|\mathbf{B}\mathbf{x}\|_\mathbf{A}\leqq\|\mathbf{C}\|_\mathbf{E}\|\mathbf{x}\|_\mathbf{A}$, ahol $\mathbf{C}=\mathbf{A}^T\mathbf{A}$. Így

$$\|\mathbf{x}^{(k+1)}\|_\mathbf{A}=\left\|-\frac{\delta}{\|\mathbf{x}^{(k)}\|_\mathbf{A}}\mathbf{B}\mathbf{x}^{(k)}+\mathbf{A}\mathbf{c}-\mathbf{e}\right\|_\mathbf{A}\leqq\|\mathbf{A}\mathbf{c}-\mathbf{e}\|_\mathbf{A}+\frac{\delta}{\|\mathbf{x}^{(k)}\|_\mathbf{A}}\|\mathbf{B}\mathbf{x}^{(k)}\|_\mathbf{A}\leqq$$

$$\leqq\|\mathbf{A}\mathbf{c}-\mathbf{e}\|_\mathbf{A}+\frac{\delta}{\|\mathbf{x}^{(k)}\|_\mathbf{A}}\|\mathbf{C}\|_\mathbf{E}\|\mathbf{x}^{(k)}\|_\mathbf{A}=\|\mathbf{A}\mathbf{c}-\mathbf{e}\|_\mathbf{A}+\delta\|\mathbf{C}\|_\mathbf{E}.$$

Tehát az $\mathbf{x}^{(k)}$ sorozat felülről korlátos. Másrészt

$$\|\mathbf{x}^{(k+1)}\|_\mathbf{A}\geqq\left|\|\mathbf{A}\mathbf{c}-\mathbf{e}\|_\mathbf{A}-\frac{\delta\|\mathbf{B}\mathbf{x}^{(k)}\|_\mathbf{A}}{\|\mathbf{x}^{(k)}\|_\mathbf{A}}\right|=\|\mathbf{A}\mathbf{c}-\mathbf{e}\|_\mathbf{A}-\frac{\delta\|\mathbf{B}\mathbf{x}^{(k)}\|_\mathbf{A}}{\|\mathbf{x}^{(k)}\|_\mathbf{A}}.$$

Az utóbbi egyenlőség igaz, mivel $\delta=0$ esetén $\|\mathbf{x}^{(k+1)}\|_\mathbf{A}=\|\mathbf{A}\mathbf{c}-\mathbf{e}\|_\mathbf{A}$, valamint $\frac{\delta\|\mathbf{B}\mathbf{x}^{(k)}\|_\mathbf{A}}{\|\mathbf{x}^{(k)}\|_\mathbf{A}}$ monoton növekedő, folytonos függvénye $\delta$-nak, így ha

$$\delta\leqq\frac{\|\mathbf{A}\mathbf{c}-\mathbf{e}\|_\mathbf{A}\|\mathbf{x}^{(k)}\|_\mathbf{A}}{\|\mathbf{B}\mathbf{x}^{(k)}\|_\mathbf{A}},$$

akkor a fenti egyenlőség igaz, ha

$$\delta=\frac{\|\mathbf{A}\mathbf{c}-\mathbf{e}\|_\mathbf{A}\|\mathbf{x}^{(k)}\|_\mathbf{A}}{\|\mathbf{B}\mathbf{x}^{(k)}\|_\mathbf{A}},$$

akkor $\|\mathbf{x}^{(k+1)}\|=0$ lenne, ami azt jelenti, hogy az optimális célfüggvényérték zérus, ami ellent mond feltételezésünknek. Így

$$\|\mathbf{x}^{(k+1)}\|_A \geqq \|\mathbf{Ac}-\mathbf{e}\|_A - \frac{\delta\|\mathbf{Bx}^{(k)}\|_A}{\|\mathbf{x}^{(k)}\|_A} \geqq \|\mathbf{Ac}-\mathbf{e}\|_A - \delta\|\mathbf{C}\|_E.$$

*Megjegyzés.* A fenti lemmát úgy is bizonyíthattuk volna, hogy (4.14) iteráció egy $n-k$ dimenziós ellipszoid felületén állít elő pontokat, így az előállított sorozat nyilván alulról is és felülről is korlátos, mivel az a feltételünk, hogy az optimális célfüggvényérték pozitív (vagyis az $\mathbf{x}=\mathbf{0}$ nem optimális megoldása a duál feladatnak) azt jelenti, hogy az origó külső pontja a fent említett ellipszoidnak.

4.4. TÉTEL. Ha $\delta < \frac{\|\mathbf{Ac}-\mathbf{e}\|_A}{3\|\mathbf{C}\|_E}$, akkor a (4.14) iteráció a (4.13) egyenlet megoldásához konvergál.

*Bizonyítás.* A konvergencia elégséges feltétele a *Banach—Cacciopoli—Tyihonov fixpont-tétel* [5], [6] szerint, hogy az $\mathbf{F(t)}=-\frac{\delta}{\|\mathbf{t}\|_A}\mathbf{Bt}+\mathbf{Ac}-\mathbf{e}$ (ahol $\mathbf{B}=\mathbf{AA}^T$) operátor kontrakciós operátor legyen, azaz $\|\mathbf{F(t)}-\mathbf{F(u)}\|_A \leqq q\|\mathbf{t}-\mathbf{u}\|_A$, ahol $q<1$.

$$\|\mathbf{F(t)}-F(\mathbf{u})\|_A = \left\|-\frac{\delta\mathbf{Bt}}{\|\mathbf{t}\|_A}+\mathbf{Ac}-\mathbf{e}+\frac{\delta\mathbf{Bu}}{\|\mathbf{u}\|_A}-\mathbf{Ac}+\mathbf{e}\right\|_A =$$

$$= \left\|\delta\mathbf{B}\left(\frac{\mathbf{t}}{\|\mathbf{t}\|_A}-\frac{\mathbf{u}}{\|\mathbf{u}\|_A}\right)\right\|_A \leqq \delta\|\mathbf{C}\|_E\left[\left\|\frac{\mathbf{t}}{\|\mathbf{t}\|_A}-\frac{\mathbf{u}}{\|\mathbf{t}\|_A}\right\|_A+\left\|\frac{\mathbf{u}}{\|\mathbf{t}\|_A}-\frac{\mathbf{u}}{\|\mathbf{u}\|_A}\right\|_A\right] \leqq$$

$$\leqq \delta\|\mathbf{C}\|_E\left[\frac{\|\mathbf{t}-\mathbf{u}\|_A}{\|\mathbf{t}\|_A}+\frac{\|\mathbf{u}\|_A|\|\mathbf{u}\|_A-\|\mathbf{t}\|_A|}{\|\mathbf{t}\|_A\|\mathbf{u}\|_A}\right] \leqq \frac{2\delta\|\mathbf{C}\|_E}{\|\mathbf{t}\|_A}\|\mathbf{t}-\mathbf{u}\|_A.$$

Esetünkben $\|\mathbf{t}\|_A \geqq \|\mathbf{Ac}-\mathbf{e}\|_A-\delta\|\mathbf{C}\|_E$ minden, az iterációban előforduló $\mathbf{t}$ pont esetén, így

$$\frac{2\delta\|\mathbf{C}\|_E}{\|\mathbf{Ac}-\mathbf{e}\|_A-\delta\|\mathbf{C}\|_E} < 1,$$

vagyis

$$\delta < \frac{\|\mathbf{Ac}-\mathbf{e}\|_A}{3\|\mathbf{C}\|_E}$$

feltételnek kell teljesülni ahhoz, hogy $\mathbf{F}$ kontrakciós operátor legyen.

Mivel $R^{n-k}$ teljes metrikus tér, melyben az $\mathbf{Ac}-\mathbf{e}$ középpontú ellipszoid zárthalmaz, ami a (4.14) iteráció értékkészlete, így a fixpont-tétel szerint lemmánkat bebizonyítottuk.

Eddigi eredményeink biztosítják (4.13) egyenletrendszer megoldását elegendően kicsi $\delta$ esetén, (amikor a (4.14) iteráció kontrakció), illetve elég nagy $\delta$ esetén (amikor az optimális célfüggvény érték zérus). Feladatunk maradt az, hogy a közbülső intervallumba eső $\delta$-k esetén is megoldási eljárást adjunk (4.13) egyenletrendszerre, s így eredeti feladatunkra is. Mielőtt rátérnénk megoldási módszerünk ismertetésére, előbb a (4.13) egyenletrendszer $\bar{\mathbf{x}}$ megoldása és a (4.13) egyenlet jobb oldala által definiált ellipszoid kapcsolatát mutatjuk be.

4.5. LEMMA. A (4.13) egyenletrendszer $\bar{\mathbf{x}}$ megoldása az $F=\{\mathbf{v}\,|\,\mathbf{v}=\mathbf{Ac}-\mathbf{e}-\delta\mathbf{Bu},\ \|\mathbf{u}\|_A\leqq 1\}$ ellipszoidnak, ahol $\mathbf{u}, \mathbf{v}\in R^{n-k}$, az origóhoz legközelebbi pontja euklideszi normával mérve a távolságot.

*Bizonyítás.* Az $F$ ellipszoid origóhoz legközelebbi pontját az alábbi konvex programozási feladat optimális megoldása adja:

$$\min(\mathbf{Ac}-\mathbf{e}-\delta\mathbf{Bu})^2,$$

$$\mathbf{uBu}\leqq 1,$$

Ez pedig egy $l_p$ programozási primál feladat:

$$\max\eta$$

$$\frac{1}{2}(\delta\mathbf{Bu}-(\mathbf{Ac}-\mathbf{e}))^2+\frac{1}{2}\eta\leqq 0$$

$$\frac{1}{2}(\mathbf{A}^T\mathbf{u})^2-\frac{1}{2}\leqq 0,$$

Duálja a következő feladat (l. Függelék):

$$\min\left[\mathbf{t}(\mathbf{Ac}-\mathbf{e})+\frac{\zeta_2}{2}+\frac{1}{2\zeta_1}\mathbf{t}^2+\begin{cases}0, & \text{ha}\quad \zeta_2=0\\ \dfrac{1}{2\zeta_2}\mathbf{s}^2, & \text{ha}\quad \zeta_2>0\end{cases}\right]$$

$$\delta\mathbf{tB}+\mathbf{sA}^T=0$$

$$\zeta_1=2,\quad \zeta_2\geqq 0$$

$$\zeta_2=0\Rightarrow\mathbf{s}=\mathbf{0},$$

ahol $\mathbf{t}\in R^{n-k}$, $\mathbf{s}\in R^n$, $\zeta_1, \zeta_2\in R$.

A primál feladat *Slater reguláris* ($\mathbf{u}=\mathbf{0}$, $\eta<0$-t választva), célfüggvénye korlátos, így mindkét feladatnak létezik optimális megoldása, amely kielégíti az egyensúlyi feltételeket, vagyis

$$\delta\mathbf{tB}+\mathbf{As}=0 \tag{4.15}$$

$$\mathbf{s}=\zeta_2(\mathbf{A}^T\mathbf{u}) \tag{4.16}$$

$$\mathbf{t}=2[\delta\mathbf{Bu}-(\mathbf{Ac}-\mathbf{e})] \tag{4.17}$$

$$\zeta_2(\mathbf{uBu}-1)=0. \tag{4.18}$$

Felhasználva, hogy $\mathbf{B}$ szimmetrikus mátrix, valamint (4.16), (4.17)-et (4.15)-be helyettesítve a

$$2\delta^2\mathbf{B}^2\mathbf{u}-2\delta\mathbf{B}(\mathbf{Ac}-\mathbf{e})=\zeta_2\mathbf{Bu}$$

egyenletet kapjuk, amelyből az

$$\mathbf{Ac}-\mathbf{e}-\delta\mathbf{Bu}=\frac{\zeta_2}{2\delta}\mathbf{u}$$

egyenlethez jutunk. Mivel $\bar{\mathbf{x}} \neq \mathbf{0}$, azaz az origó az $F$ ellipszoidnak külső pontja, így $\mathbf{uBu}=1$, azaz $\|\mathbf{u}\|_A=1$, így $\left\|\frac{\zeta_2}{2\delta}\mathbf{u}\right\|_A=\frac{\zeta_2}{2\delta}$. Legyen $\bar{\mathbf{x}}=\frac{\zeta_2}{2\delta}\mathbf{u}$, így

$$\mathbf{Ac}-\mathbf{e}-\frac{\delta \mathbf{B}\bar{\mathbf{x}}}{\|\bar{\mathbf{x}}\|_A}=\bar{\mathbf{x}},$$

azaz $\bar{\mathbf{x}}$, amely egyetlen megoldása a (4.13) egyenletrendszernek, az $F$ ellipszoidnak az origóhoz legközelebbi pontja euklideszi normával mérve a távolságot.

A 4.5. lemma miatt, feladatunk ekvivalens egy ellipszoid origóhoz legközelebbi pontjának a meghatározásával.

Ennyi kitérő után térjünk vissza a (4.13) egyenletrendszer megoldásához.

Az alábbiakban tetszőleges $\delta>0$ esetén használható eljárást adunk a keresett megoldás meghatározására. Eljárásunk, amennyiben közelítő becslés nélkül alkalmazzuk, lényegesen több számolást igényel, mint (4.14) iteráció végrehajtása, így „kicsiny" $\delta$ esetén továbbra is (4.14) iterációval célszerű megoldani feladatunkat.

Egy egyismeretlenes egyenlet, és egy teljes rangú lineáris egyenletrendszer megoldására vezetjük vissza feladatunkat:

4.6. TÉTEL. Legyen $\tilde{\mathbf{x}}$ megoldása az $\tilde{\mathbf{x}}=-L\mathbf{B}\tilde{\mathbf{x}}+\mathbf{Ac}-\mathbf{e}$ egyenletrendszernek, ahol $L=\frac{\delta}{\|\tilde{\mathbf{x}}\|_A}$, akkor $\tilde{\mathbf{x}}$ megoldása (4.13) egyenletrendszernek is.

*Bizonyítás.* Mivel $\tilde{\mathbf{x}}=-L\mathbf{B}\tilde{\mathbf{x}}+\mathbf{Ac}-\mathbf{e}$, így $L=\frac{\delta}{\|\tilde{\mathbf{x}}\|_A}$-t helyettesítve (4.13) egyenletet kapjuk, ami bizonyítja állításunkat.

Eljárásunk 4.6. tétel alapján adott $\delta$ esetén a következő:

— Mivel $\tilde{\mathbf{x}}=-L\mathbf{B}\tilde{\mathbf{x}}+\mathbf{Ac}-\mathbf{e}$ így $\tilde{\mathbf{x}}=(\mathbf{E}+L\mathbf{B})^{-1}(\mathbf{Ac}-\mathbf{e})$.

Határozzuk meg a

$$\delta = L\|(\mathbf{E}+L\mathbf{B})^{-1}(\mathbf{Ac}-\mathbf{e})\|_A \tag{4.19}$$

egyenlet $L$ megoldását. Ez jól ismert numerikus módszerekkel (szakaszfelezési módszer, húrmódszer, stb. [5], [6]) meghatározható. Sajnos ez a rész nagyon számolás igényes, mivel minden lépés tartalmaz egy mátrixinvertálást.

— A fent meghatározott $L$-lel oldjuk meg az

$$\tilde{\mathbf{x}} = -L\mathbf{B}\tilde{\mathbf{x}}+\mathbf{Ac}-\mathbf{e} \tag{4.20}$$

lineáris egyenletrendszert. Ezt megoldhatjuk direkt módszerrel is, vagy a (3.8) iterációnak megfelelően $\frac{1}{\lambda^2}=L$ helyettesítéssel nyerjük az alábbi konvergens iteratív eljárást.

$$\mathbf{x}^{(k+1)}=\frac{KL}{KL+2}\left(\mathbf{E}-\frac{2}{K}\mathbf{B}\right)\mathbf{x}^{(k)}+\frac{2}{KL+2}(\mathbf{Ac}-\mathbf{e}).$$

*1. Megjegyzés.* Amennyiben a (4.19) egyenlet megoldását el akarjuk kerülni, és megelégszünk egy közelítő $L'$ megoldással, amely egy $\delta' \sim \delta$ paraméterhez tartozik, akkor használhatjuk az $L'=\frac{\delta}{\|\mathbf{Ac}-\mathbf{e}\|_A-\delta\|\mathbf{C}\|_E}$ képletet. Ezt a képletet (4.13)-

ból nyerjük, mivel onnan az $\|\tilde{\mathbf{x}}\|_A+\delta\|\mathbf{C}\|_E \geqq \|\mathbf{Ac}-\mathbf{e}\|_A$ egyenlőtlenséget kapjuk, melyből $\|\tilde{\mathbf{x}}\|_A=\frac{\delta}{L}$ és egyenlőség helyettesítéssel nyerjük a fenti közelítő képletet $L'$-re. $L'$-ből (4.19) képlet segítségével határozhatjuk meg, hogy ténylegesen milyen $\delta'$ értékre oldottuk meg feladatunkat. Ha ez nem kielégítő, akkor mivel $L\|\tilde{\mathbf{x}}\|_A=\delta$ és $\tilde{\mathbf{x}}$ az $F$ ellipszoid origóhoz legközelebbi pontja (euklideszi normában), így $\delta$ növekedtével $\|\tilde{\mathbf{x}}\|$ csökken, tehát $L$ nő. Így, ha $\delta$-t növeljük, $L$-t is növelnünk kell, ha $\delta$-t csökkentjük, akkor $L$ is csökken.

*2. Megjegyzés.* Ha (4.19) egyenletrendszert valamilyen iterációs módszerrel megoldjuk, akkor ennek eredményeképp az $(\mathbf{E}+L\mathbf{B})^{-1}(\mathbf{Ac}-\mathbf{e})=\tilde{\mathbf{x}}$ vektort is megkapjuk, amivel (4.20) egyenletrendszert is megoldottuk.

*3. Megjegyzés.* Az $(\mathbf{E}+L\mathbf{B})^{-1}$ inverz közelítő meghatározására sorfejtést is használhatunk. Mivel $(\mathbf{E}+L\mathbf{B})^{-1}=\frac{2}{KL+2}\left[\mathbf{E}-\frac{KL}{KL+2}\left(\mathbf{E}-\frac{2}{K}\mathbf{B}\right)\right]^{-1}$ és mint láttuk a $\frac{KL}{KL+2}\left(\mathbf{E}-\frac{2}{K}\mathbf{B}\right)$ mátrix spektrálsugara kisebb egynél, így használhatjuk az $(\mathbf{E}-\mathbf{D})^{-1}=\mathbf{E}+\mathbf{D}+\mathbf{D}^2+\ldots$ sorfejtést, amely konvergens, ha $\mathbf{D}$ spektrálsugara kisebb, mint egy. Amennyiben az $\left(\mathbf{E}-\frac{2}{K}\mathbf{B}\right)^i$, $i=0, 1, 2, \ldots$ hatványokat tárolni tudjuk, akkor az inverz meghatározására különböző $L$ értékek esetén csak mátrix és konstans szorzását, valamint mátrixok összegzését tartalmazza.

Végül belátjuk, hogy (4.19) egyenletnek akkor és csak akkor van megoldása, ha (4.1) feladat optimális célfüggvényértéke nem zérus, de $\delta>0$.

4.7. LEMMA. Ha $\delta>0$, akkor (4.19) egyenletnek akkor és csak akkor van megoldása, ha $0<\delta<\bar{\delta}$, ahol $\bar{\delta}$ az a legkisebb érték, melyre (4.1) feladat célfüggvényértéke zérus.

*Bizonyítás.* Elég azt belátnunk, hogy $\lim_{L\to\infty} L\|(\mathbf{E}+L\mathbf{B})^{-1}(\mathbf{Ac}-\mathbf{e})\|_A=\bar{\delta}$ és $\lim_{L\to 0} L\|(\mathbf{E}+L\mathbf{B})^{-1}(\mathbf{Ac}-\mathbf{e})\|_A=0$, mivel $L\|(\mathbf{E}+L\mathbf{B})^{-1}(\mathbf{Ac}-\mathbf{e})\|_A$ folytonos függvénye $L$-nek $L>0$ esetén. Egyszerű átalakításokkal könnyen bizonyíthatók a fenti összefüggések.

$$\lim_{L\to 0} L\|(\mathbf{E}+L\mathbf{B})^{-1}(\mathbf{Ac}-\mathbf{e})\|_A = \lim_{L\to 0} L \lim_{L\to 0} \|(\mathbf{E}+L\mathbf{B})^{-1}(\mathbf{Ac}-\mathbf{e})\|_A = 0\cdot\|\mathbf{Ac}-\mathbf{e}\|_A = 0$$

$$\lim_{L\to\infty} [L\|(\mathbf{E}+L\mathbf{B})^{-1}(\mathbf{Ac}-\mathbf{e})\|_A]^2 = \lim_{L\to\infty}\left[\left\|\left(\frac{\mathbf{E}}{L}+\mathbf{B}\right)^{-1}(\mathbf{Ac}-\mathbf{e})\right\|_A\right]^2 =$$

$$= [\|\mathbf{B}^{-1}(\mathbf{Ac}-\mathbf{e})\|_A]^2 = (\mathbf{Ac}-\mathbf{e})^T\mathbf{B}^{-1}\mathbf{B}\mathbf{B}^{-1}(\mathbf{Ac}-\mathbf{e}).$$

Legyen $\bar{\mathbf{y}}$ olyan, hogy $\mathbf{A}\bar{\mathbf{y}}=\mathbf{e}$ és $(\bar{\mathbf{y}}-\mathbf{c})^2=\bar{\delta}^2$, azaz $\bar{\mathbf{y}}$ optimális megoldása az

$$\mathbf{Ay}=\mathbf{e}$$

$$\min (\mathbf{y}-\mathbf{c})^2$$

$l_p$ programozási primál feladatnak. Az $l_p$ programozás egyensúlyi feltételeit használva belátható, hogy $\bar{\mathbf{z}}\in R^{n-k}$, akkor és csak akkor optimális megoldása a duál feladat-

nak, ha $(\bar{\mathbf{y}}-\mathbf{c})=\mathbf{A}^T\bar{\mathbf{z}}$, így mivel $\mathbf{A}\bar{\mathbf{y}}=\mathbf{e}$

$$\lim_{L\to\infty}[L\|(\mathbf{E}+L\mathbf{B})^{-1}(\mathbf{Ac}-\mathbf{e})\|_{\mathbf{A}}]^2=(\mathbf{Ac}-\mathbf{A}\bar{\mathbf{y}})^T\mathbf{B}^{-1}\mathbf{B}\mathbf{B}^{-1}(\mathbf{Ac}-\mathbf{A}\bar{\mathbf{y}})=$$

$$=(\bar{\mathbf{y}}-\mathbf{c})^T\mathbf{A}^T\mathbf{B}^{-1}\mathbf{B}\mathbf{B}^{-1}\mathbf{A}(\bar{\mathbf{y}}-\mathbf{c})=\bar{\mathbf{z}}\mathbf{B}\mathbf{B}^{-1}\mathbf{B}\mathbf{B}^{-1}\mathbf{B}\bar{\mathbf{z}}=\bar{\mathbf{z}}\mathbf{B}\bar{\mathbf{z}}=(\bar{\mathbf{y}}-\mathbf{c})^2=\bar{\delta}^2.$$

Így tételünket beláttuk.

## 5. Simítás, amikor a differenciák mérési pontossága ismert

### *5.1. Dualitás*

A második fejezetben konstruált III. modell szerint feladatunk meghatározni a $\gamma_1,\ldots,\gamma_n$ értékekhez legközelebbi $y_1,\ldots,y_n$ értékeket, melyekből számított $k$-adik differenciák euklideszi távolsága a mért $\varepsilon_1,\ldots,\varepsilon_{n-k}$ értékektől $\delta$-nál nem nagyobb. Azaz

$$\min(\mathbf{y}-\mathbf{c})^2$$

$$(\mathbf{Ay}-\mathbf{e})^2\leqq\delta^2. \tag{5.1}$$

Ha $\delta^2=0$, akkor $\mathbf{Ay}=\mathbf{e}$, és feladatunkat a 4.2. fejezet a., részében tárgyaltakkal megegyező módon oldhatjuk meg. Így a továbbiakban feltehető, hogy $\delta^2>0$.

Az (5.1) feladat egy kvadratikus feltételes kvadratikus programozási feladat, melynek duálja az alábbi:

$$\max\left(2\mathbf{xAc}-(\mathbf{xA})^2-2\mathbf{xe}-2\delta\sqrt{\mathbf{x}^2}\right), \tag{5.2}$$

ahol $\mathbf{x}\in R^{n-k}$.

Elemi úton igazolható az alábbi lemma.

5.1. LEMMA. Tetszőleges $\mathbf{y}\in R^n$-re, melyre $(\mathbf{Ay}-\mathbf{e})^2\leqq\delta^2$ és $\mathbf{x}\in R^{n-k}$-ra

$$(\mathbf{y}-\mathbf{c})^2\geqq 2\mathbf{xAc}-(\mathbf{xA})^2-2\mathbf{xe}-2\delta\sqrt{\mathbf{x}^2} \tag{5.3}$$

egyenlőséggel akkor és csak akkor, ha

$$\mathbf{y}=\mathbf{c}-\mathbf{xA} \tag{5.4}$$

és vagy $\mathbf{x}=\mathbf{0}$ (ami azt jelenti, hogy $\mathbf{y}=\mathbf{c}$),

$$\text{vagy } \mathbf{x}=\frac{\sqrt{\mathbf{x}^2}}{\delta}(\mathbf{Ay}-\mathbf{e}) \text{ (ami azt jelenti, hogy } (\mathbf{Ay}-\mathbf{e})^2=\delta^2). \tag{5.5}$$

*Bizonyítás.* Ha $\mathbf{x}=\mathbf{0}$, akkor a duál feladat célfüggvényértéke 0 és a primál feladat célfüggvénye mindig nem negatív, így (5.3) fennáll. Egyenlőséggel akkor és csak akkor, ha $\mathbf{y}=\mathbf{c}$, azaz (5.4) is igaz.

Ha $\mathbf{x}\neq\mathbf{0}$, akkor nyilván igaz az alábbi két egyenlőtlenség:

$$0\leqq[(\mathbf{y}-\mathbf{c})+\mathbf{xA}]^2=(\mathbf{y}-\mathbf{c})^2+(\mathbf{xA})^2+2\mathbf{xAy}-2\mathbf{xAc}. \tag{5.6}$$

Egyenlőséggel akkor és csak akkor, ha (5.4) igaz.

$$(5.7) \qquad 0 \leqq \frac{\delta}{\sqrt{\mathbf{x}^2}}\left[\mathbf{x}-\frac{\sqrt{\mathbf{x}^2}}{\delta}(\mathbf{Ay}-\mathbf{e})\right]^2 = \delta\sqrt{\mathbf{x}^2}+\frac{\sqrt{\mathbf{x}^2}}{\delta}(\mathbf{Ay}-\mathbf{e})^2-2\mathbf{xAy}+2\mathbf{xe}.$$

Egyenlőséggel akkor és csak akkor, ha (5.5) igaz.

Az (5.6) egyenlőtlenséget átrendezve:

$$(\mathbf{y}-\mathbf{c})^2 \geqq 2\mathbf{xAc}-2\mathbf{xAy}-(\mathbf{xA})^2.$$

Ebből az (5.7) egyenlőtlenséget kivonva:

$$(\mathbf{y}-\mathbf{c})^2 \geqq 2\mathbf{xAc}-2\mathbf{xAy}-(\mathbf{xA})^2-\delta\sqrt{\mathbf{x}^2}-\frac{\sqrt{\mathbf{x}^2}}{\delta}(\mathbf{Ay}-\mathbf{e})^2+2\mathbf{xAy}-2\mathbf{xe} =$$

$$= 2\mathbf{xAc}-(\mathbf{xA})^2-2\mathbf{xe}-\delta\sqrt{\mathbf{x}^2}-\frac{\sqrt{\mathbf{x}^2}}{\delta}(\mathbf{Ay}-\mathbf{e})^2.$$

Az (5.1) feltételt alkalmazva — egyenlőség akkor és csak akkor áll fenn, ha $(\mathbf{Ay}-\mathbf{e})^2=\delta^2$, ami (5.5) feltételből is következik — a kívánt egyenlőtlenséget nyerjük:

$$(\mathbf{y}-\mathbf{c})^2 \geqq 2\mathbf{xAc}-(\mathbf{xA})^2-2\mathbf{xe}-2\delta\sqrt{\mathbf{x}^2}.$$

KÖVETKEZMÉNY. Ha valamely $\mathbf{y}\in R^n$, $(\mathbf{Ay}-\mathbf{e})^2\leqq\delta^2$ és $\mathbf{x}\in R^{n-k}$ esetén (5.3)-ban egyenlőség áll fenn, akkor $\mathbf{y}$ és $\mathbf{x}$ optimális megoldásai az (5.1), illetve (5.2) feladatnak.

*Bizonyítás.* A lemma alapján nyilvánvaló.

$l_p$ programozási ismereteinket felhasználva belátjuk, hogy az (5.1) primál és az (5.2) duál feladatnak is pontosan egy optimális megoldása létezik.

4.2. TÉTEL. Az (5.1) és (5.2) feladatoknak egy és csak egy optimális megoldása létezik, melyekre (5.3)-ban egyenlőség áll fenn.

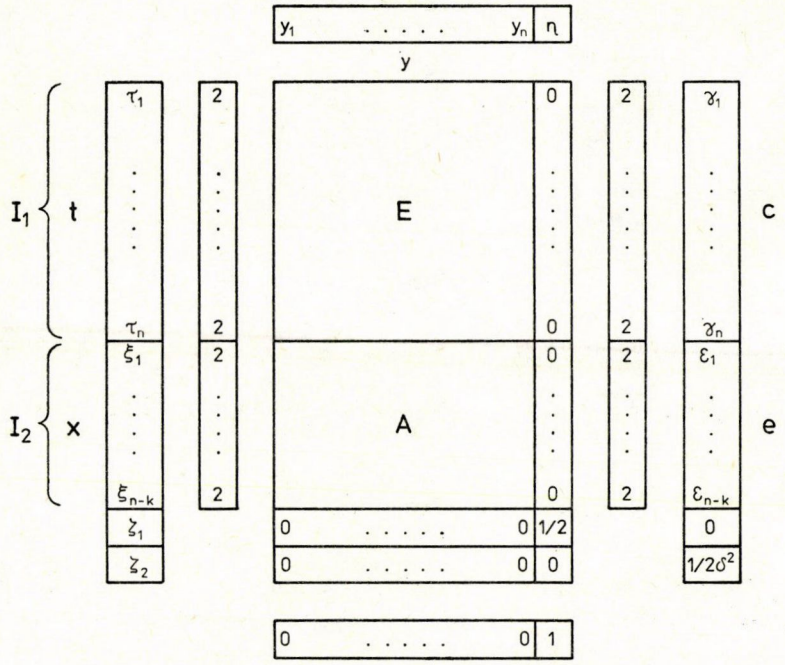
*Bizonyítás.* Az alábbi formában az (5.1) feladat egy $l_p$ programozási *primál* feladat:

$$\max \eta$$

$$\frac{1}{2}(\mathbf{y}-\mathbf{c})^2+\frac{1}{2}\eta \leqq 0$$

$$\frac{1}{2}(\mathbf{Ay}-\mathbf{e})^2-\frac{1}{2}\delta^2 \leqq 0.$$

A feladat szerkezetét a 3. ábra szemlélteti.




*3. ábra*

Az ábra jelöléseit felhasználva a *duál* feladat a következő:

$$\mathbf{t}+\mathbf{xA}=\mathbf{0}$$

$$\frac{1}{2}\zeta_1 = 1$$

$$\zeta_2 \geqq 0, \quad \zeta_2 = 0 \Rightarrow \mathbf{x} = \mathbf{0}$$

$$\min\left(\mathbf{tc}+\mathbf{xe}+\frac{1}{2}\zeta_2\delta^2+\frac{1}{2\zeta_1}\mathbf{t}^2+\begin{cases}\dfrac{1}{2\zeta_2}\mathbf{x}^2, & \text{ha} \quad \zeta_2 > 0\\ 0, & \text{ha} \quad \zeta_2 = 0\end{cases}\right).$$

Átalakítva:

$$\zeta_2 \geqq 0, \quad \zeta_2 = 0 \Rightarrow \mathbf{x} = \mathbf{0}$$

$$\min\left(-\mathbf{xAc}+\mathbf{xe}+\frac{1}{2}\zeta_2\delta^2+\frac{1}{4}(\mathbf{xA})^2+\begin{cases}\dfrac{1}{2\zeta_2}\mathbf{x}^2, & \text{ha} \quad \zeta_2 > 0\\ 0, & \text{ha} \quad \zeta_2 = 0\end{cases}\right).$$

A primál feladat *Slater reguláris,* mivel olyan **y**-t választva, melyre $\mathbf{Ay}=\mathbf{e}$ (ez nyilván megtehető) és alkalmas $\eta<0$-t, teljesül a *Slater regularitási feltétel.* A primál feladat célfüggvénye nyilván felülről korlátos, így létezik optimális megoldása (F1. és F2. tételek miatt) a primál és a duál feladatnak is.

Meg kell mutatnunk, hogy duál feladatunk ekvivalens (5.2) feladattal. Tekintsük az $l_p$ programozás egyensúlyi feltételeit:

$$\mathbf{t} = 2(\mathbf{y}-\mathbf{c}) \tag{5.8}$$

$$\mathbf{x} = \zeta_2(\mathbf{A}\mathbf{y}-\mathbf{e}) \tag{5.9}$$

$$\zeta_2[(\mathbf{A}\mathbf{y}-\mathbf{e})^2-\delta^2] = 0. \tag{5.10}$$

Az (5.9) összefüggésből az $(\mathbf{A}\mathbf{y}-\mathbf{e})=\frac{\mathbf{x}}{\zeta_2}$ összefüggést nyerjük ($\zeta_2>0$ esetén), amelyet (5.10)-be helyettesítve a

$$\zeta_2 = \frac{\sqrt{\mathbf{x}^2}}{\delta}$$

összefüggést kapjuk, amelyet a duál feladatba helyettesítve egy vele ekvivalens feladatot nyerünk:

$$\min\left(-\mathbf{x}\mathbf{A}\mathbf{c}+\mathbf{x}\mathbf{e}+\frac{1}{2}\delta\sqrt{\mathbf{x}^2}+\frac{1}{4}(\mathbf{x}\mathbf{A})^2+\frac{1}{2}\delta\sqrt{\mathbf{x}^2}\right).$$

Ahol azt is kihasználtuk, hogy $\zeta_2=0\Rightarrow\mathbf{x}=\mathbf{0}$, és így eltekinthetünk az esetszétválasztástól. Az $\mathbf{x}=2\mathbf{x}$ helyettesítést elvégezve nyerjük az eredeti (5.2) duál feladatot:

$$\max\left(2\mathbf{x}\mathbf{A}\mathbf{c}-2\mathbf{x}\mathbf{e}-(\mathbf{x}\mathbf{A})^2-2\delta\sqrt{\mathbf{x}^2}\right).$$

Így beláttuk, hogy (5.1) és (5.2) feladatok $l_p$ programozási primál és duál feladatokkal ekvivalensek, melyeknek van optimális megoldásuk és az optimális célfüggvényértékek megegyeznek, azaz (5.3)-ban egyenlőség áll fenn.

Az optimális megoldások egyértelműek, mivel tetszőleges optimális megoldáspár kielégíti (5.8), (5.9), (5.10) egyensúlyi feltételeket, így tetszőleges, de fix $\bar{\mathbf{y}}$ optimális megoldás esetén az egyensúlyi feltételek, valamint a duál feltételek miatt $\bar{\mathbf{x}}$, $\bar{\mathbf{t}}$, $\bar{\zeta}_2$ optimális megoldás egyértelmű, valamint ugyanezen feltételekből következik, hogy $\bar{\mathbf{y}}$ optimális megoldás egyértelműen állítható elő az egyértelmű duál optimális megoldással.

*5.2. Az optimális megoldás meghatározása*

*a) Ha zérus az optimális célfüggvényérték*

Amennyiben $(\mathbf{y}-\mathbf{c})^2=0$, akkor $\mathbf{y}=\mathbf{c}$. Így ha az $(\mathbf{A}\mathbf{c}-\mathbf{e})^2\leqq\delta^2$ feltétel teljesül, akkor feladatunkat megoldottuk (nem végeztünk simítást). Ha $(\mathbf{A}\mathbf{c}-\mathbf{e})^2>\delta^2$, akkor a feladat optimális célfüggvényértéke nem zérus, így a duál feladat optimális megoldása sem lehet az $\mathbf{x}=\mathbf{0}$ vektor.

*b) Ha az optimális célfüggvényérték nem zérus*

Ebben az esetben $\mathbf{x}=\mathbf{0}$ vektor nem optimális megoldása (5.2) duál feladatnak. Az (5.4), (5.5) optimalitási kritériumokból nyerjük, hogy az optimális $\mathbf{x}$ vektor kielégíti az

$$\mathbf{x} = \frac{\sqrt{\mathbf{x}^2}}{\delta}[-\mathbf{A}\mathbf{A}^T\mathbf{x}+\mathbf{A}\mathbf{c}-\mathbf{e}] \tag{5.11}$$

összefüggést, melyből átrendezéssel nyerjük (felhasználva, hogy $\mathbf{A}\mathbf{A}^T=\mathbf{B}$ reguláris mátrix):

$$\mathbf{x}=-\frac{\delta}{\sqrt{\mathbf{x}^2}}\mathbf{B}^{-1}\mathbf{x}+\mathbf{B}^{-1}(\mathbf{A}\mathbf{c}-\mathbf{e})$$

egyenletet, melyről 5.2. tétel alapján tudjuk, hogy pontosan egy megoldása létezik. Oldjuk meg a fenti egyenletet az

$$\mathbf{x}^{(k+1)}=-\frac{\delta}{\sqrt{(\mathbf{x}^{(k)})^2}}\mathbf{B}^{-1}\mathbf{x}^{(k)}+\mathbf{B}^{-1}(\mathbf{A}\mathbf{c}-\mathbf{e}) \tag{5.12}$$

iterációval.

5.3. LEMMA. Az (5.12) iteráció tetszőleges $\mathbf{x}^{(0)}\neq\mathbf{0}$ vektorból indítva alulról is és felülről is korlátos sorozatot állít elő.

*Bizonyítás.* Vizsgáljuk a vektorsorozat euklideszi normáját:

$$\|\mathbf{x}^{(k+1)}\|_E=\left\|-\frac{\delta}{\|\mathbf{x}^{(k)}\|_E}\mathbf{B}^{-1}\mathbf{x}^{(k)}+\mathbf{B}^{-1}(\mathbf{A}\mathbf{c}-\mathbf{e})\right\|_E\leqq$$

$$\leqq\frac{\delta}{\|\mathbf{x}^{(k)}\|_E}\|\mathbf{B}^{-1}\mathbf{x}^{(k)}\|_E+\|\mathbf{B}^{-1}(\mathbf{A}\mathbf{c}-\mathbf{e})\|_E\leqq\delta\|\mathbf{B}^{-1}\|_E+\|\mathbf{B}^{-1}(\mathbf{A}\mathbf{c}-\mathbf{e})\|_E.$$

Tehát az $\mathbf{x}^{(k)}$ sorozat felülről korlátos. Másrészt

$$\|\mathbf{x}^{(k+1)}\|_E\geqq\left|\|\mathbf{B}^{-1}(\mathbf{A}\mathbf{c}-\mathbf{e})\|_E-\frac{\delta\|\mathbf{B}^{-1}\mathbf{x}^{(k)}\|_E}{\|\mathbf{x}^{(k)}\|_E}\right|=\|\mathbf{B}^{-1}(\mathbf{A}\mathbf{c}-\mathbf{e})\|_E-\frac{\delta\|\mathbf{B}^{-1}\mathbf{x}^{(k)}\|_E}{\|\mathbf{x}^{(k)}\|_E}.$$

Az utóbbi egyenlőtlenség igaz, mivel $\delta=0$ esetén $\|\mathbf{x}^{(k+1)}\|_E=\|\mathbf{B}^{-1}(\mathbf{A}\mathbf{c}-\mathbf{e})\|_E$, valamint $\dfrac{\delta\|\mathbf{B}^{-1}\mathbf{x}^{(k)}\|_E}{\|\mathbf{x}^{(k)}\|_E}$ monoton növekedő folytonos függvénye $\delta$-nak, így ha

$$\delta<\frac{\|\mathbf{B}^{-1}(\mathbf{A}\mathbf{c}-\mathbf{e})\|_E\|\mathbf{x}^{(k)}\|_E}{\|\mathbf{B}^{-1}\mathbf{x}^{(k)}\|_E},$$

akkor a fenti egyenlőség igaz. Ha

$$\delta=\frac{\|\mathbf{B}^{-1}(\mathbf{A}\mathbf{c}-\mathbf{e})\|_E\|\mathbf{x}^{(k)}\|_E}{\|\mathbf{B}^{-1}\mathbf{x}^{(k)}\|_E},$$

akkor $\|\mathbf{x}^{(k+1)}\|=0$ lenne, ami azt jelentené, hogy az optimális célfüggvény érték zérus és ez ellentmond feltételezésünknek. Így

$$\|\mathbf{x}^{(k+1)}\|_E\geqq\|\mathbf{B}^{-1}(\mathbf{A}\mathbf{c}-\mathbf{e})\|_E-\frac{\delta\|\mathbf{B}^{-1}\mathbf{x}^{(k)}\|_E}{\|\mathbf{x}^{(k)}\|_E}\geqq\|\mathbf{B}^{-1}(\mathbf{A}\mathbf{c}-\mathbf{e})\|_E-\delta\|\mathbf{B}^{-1}\|_E.$$

Tehát sorozatunk alulról és felülről is korlátos.

*Megjegyzés.* A fenti lemmát, hasonlóan 4.3. lemmához, úgy is bizonyíthattuk volna, hogy az (5.12) iteráció egy $n-k$ dimenziós ellipszoid felületén állít elő pontokat. Az ellipszoidnak az origó külső pontja (mivel az optimális célfüggvény érték pozitív), így a sorozat alulról és felülről is korlátos.

5.4. TÉTEL. Ha $\delta < \dfrac{\|\mathbf{B}^{-1}(\mathbf{Ac}-\mathbf{e})\|_E}{3\|\mathbf{B}^{-1}\|_E}$, akkor az (5.12) iteráció az (5.11) egyenlet megoldásához konvergál.

*Bizonyítás.* A konvergencia elégséges feltétele a *Banach—Cacciopoli—Tyihonov fixponttétel* szerint, hogy az $\mathbf{F}(\mathbf{t}) = -\dfrac{\delta}{\|\mathbf{t}\|_E}\mathbf{B}^{-1}\mathbf{t} + \mathbf{B}^{-1}(\mathbf{Ac}-\mathbf{e})$ operátor kontrakciós operátor legyen, azaz

$$\|\mathbf{F}(\mathbf{t})-\mathbf{F}(\mathbf{u})\|_E \leqq q\|\mathbf{t}-\mathbf{u}\|_E, \quad \text{ahol} \quad q<1, \quad \mathbf{t},\mathbf{u}\in R^{n-k}.$$

$$\|\mathbf{F}(\mathbf{t})-\mathbf{F}(\mathbf{u})\|_E = \left\|-\frac{\delta\mathbf{B}^{-1}\mathbf{t}}{\|\mathbf{t}\|_E}+\mathbf{B}^{-1}(\mathbf{Ac}-\mathbf{e})+\frac{\delta\mathbf{B}^{-1}\mathbf{u}}{\|\mathbf{u}\|_E}-\mathbf{B}^{-1}(\mathbf{Ac}-\mathbf{e})\right\|_E =$$

$$= \left\|\delta\mathbf{B}^{-1}\left(\frac{\mathbf{t}}{\|\mathbf{t}\|_E}-\frac{\mathbf{u}}{\|\mathbf{u}\|_E}\right)\right\|_E \leqq \delta\|\mathbf{B}^{-1}\|_E\left\|\frac{\mathbf{t}}{\|\mathbf{t}\|_E}-\frac{\mathbf{u}}{\|\mathbf{u}\|_E}\right\|_E \leqq$$

$$\leqq \delta\|\mathbf{B}^{-1}\|_E\left[\left\|\frac{\mathbf{t}}{\|\mathbf{t}\|_E}-\frac{\mathbf{u}}{\|\mathbf{t}\|_E}\right\|_E+\left\|\frac{\mathbf{u}}{\|\mathbf{t}\|_E}-\frac{\mathbf{u}}{\|\mathbf{u}\|_E}\right\|_E\right] \leqq$$

$$\leqq \delta\|\mathbf{B}^{-1}\|_E\left[\frac{\|\mathbf{t}-\mathbf{u}\|_E}{\|\mathbf{t}\|_E}+\frac{\|\mathbf{u}\|_E|\|\mathbf{u}\|_E-\|\mathbf{t}\|_E|}{\|\mathbf{t}\|_E\|\mathbf{u}\|_E}\right] \leqq \frac{2\delta\|\mathbf{B}^{-1}\|_E}{\|\mathbf{t}\|_E}\|\mathbf{t}-\mathbf{u}\|_E.$$

Esetünkben $\|\mathbf{t}\|_E \geqq \|\mathbf{B}^{-1}(\mathbf{Ac}-\mathbf{e})\|_E - \delta\|\mathbf{B}^{-1}\|_E$ minden az iterációban előforduló $\mathbf{t}$ pont esetén, így a

$$\frac{2\delta\|\mathbf{B}^{-1}\|_E}{\|\mathbf{B}^{-1}(\mathbf{Ac}-\mathbf{e})\|_E-\delta\|\mathbf{B}^{-1}\|_E} < 1,$$

vagyis

$$\delta < \frac{\|\mathbf{B}^{-1}(\mathbf{Ac}-\mathbf{e})\|_E}{3\|\mathbf{B}^{-1}\|_E}$$

feltételnek kell teljesülni ahhoz, hogy $\mathbf{F}$ kontrakciós operátor legyen.

Mivel $R^{n-k}$ teljes metrikus tér, melyben a $\mathbf{B}^{-1}(\mathbf{Ac}-\mathbf{e})$ középpontú ellipszoid zárt halmaz, ami az (5.12) iteráció értékkészlete, így a fixponttétel szerint lemmánkat bizonyítottuk.

Eddigi eredményeink biztosítják feladatunk megoldását elegendően kicsi $\delta$ esetén, (amikor (5.12) iteráció kontrakció), illetve olyan nagy $\delta$ esetén, mikor az optimális célfüggvényérték zérus. Így feladatunk maradt a közbülső intervallumba eső $\delta$ esetén megoldási eljárást adni (5.11) egyenletrendszerre. Mielőtt megoldási módszerünk ismertetését elkezdenénk, előbb az (5.11) egyenletrendszer megoldása és az egyenletrendszer jobb oldala által definiált ellipszoid kapcsolatát mutatjuk meg.

5.5. LEMMA. Az (5.11) egyenletrendszer $\bar{\mathbf{x}}$ megoldása az $F=\{\mathbf{v}\,|\,\mathbf{v}=\mathbf{B}^{-1}(\mathbf{Ac}-\mathbf{e})-\delta\mathbf{B}^{-1}\mathbf{u},\ \|\mathbf{u}\|_E\leqq 1\}$ ellipszoidnak $(\mathbf{u},\mathbf{v}\in R^{n-k})$ az origóhoz legközelebbi pontja $\|\ \|_A$ normában mérve a távolságot.

*Bizonyítás.* Az $F$ ellipszoid origóhoz legközelebbi pontját ($\| \|_A$ normában mérve) az alábbi konvex programozási feladat optimális megoldása adja:

$$\min \left(\mathbf{B}^{-1}(\mathbf{Ac}-\mathbf{e})-\delta\mathbf{B}^{-1}\mathbf{u}\right)^T\mathbf{B}\left(\mathbf{B}^{-1}(\mathbf{Ac}-\mathbf{e})-\delta\mathbf{B}^{-1}\mathbf{u}\right)$$

$$\mathbf{u}^T\mathbf{u} \leqq 1.$$

Ez pedig egy $l_p$ programozási primál feladat:

$$\max \eta$$

$$\frac{1}{2}\left[\mathbf{A}^T\left(\mathbf{B}^{-1}(\mathbf{Ac}-\mathbf{e})-\delta\mathbf{B}^{-1}\mathbf{u}\right)\right]^2+\frac{1}{2}\eta \leqq 0$$

$$\frac{1}{2}(\mathbf{u})^2-\frac{1}{2} \leqq 0.$$

Ennek duálja a következő feladat (l. Függelék):

$$\min\left(\mathbf{t}^T\mathbf{A}^T\mathbf{B}^{-1}(\mathbf{Ac}-\mathbf{e})+\frac{\alpha_2}{2}+\frac{1}{4}\mathbf{t}^2+\begin{cases}0, & \text{ha } \alpha_2=0\\ \dfrac{1}{2\alpha_2}\mathbf{s}^2, & \text{ha } \alpha_2>0\end{cases}\right)$$

$$\delta\mathbf{t}^T\mathbf{A}^T\mathbf{B}^{-1}+\mathbf{s}=\mathbf{0}$$

$$\alpha_1=2, \quad \alpha_2 \geqq 0$$

$$\alpha_2=0 \Rightarrow \mathbf{s}=\mathbf{0},$$

ahol $\mathbf{t}\in R^n$, $\mathbf{s}\in R^{n-k}$, $\alpha_1, \alpha_2 \in R$.

A primál feladat *Slater reguláris,* ($\mathbf{u}=\mathbf{0}$, $\eta<0$-t választva), célfüggvénye korlátos, így a primál és a duál feladatnak is létezik optimális megoldása, amely kielégíti az egyensúlyi feltételeket (l. Függelék), vagyis

$$\delta\mathbf{t}^T\mathbf{A}^T\mathbf{B}^{-1}+\mathbf{s}=\mathbf{0} \tag{5.13}$$

$$\mathbf{t}=2[\delta\mathbf{A}^T\mathbf{B}^{-1}\mathbf{u}-\mathbf{A}^T\mathbf{B}^{-1}(\mathbf{Ac}-\mathbf{e})] \tag{5.14}$$

$$\mathbf{s}=\alpha_2\mathbf{u} \tag{5.15}$$

$$\alpha_2(\mathbf{u}^2-1)=0. \tag{5.16}$$

Felhasználva, hogy $\mathbf{B}^{-1}$ szimmetrikus mátrix, és így $\mathbf{t}^T\mathbf{A}^T\mathbf{B}^{-1}=\mathbf{B}^{-1}\mathbf{At}$, valamint, hogy $\mathbf{AA}^T=\mathbf{B}$, (5.13), (5.14), (5.15)-ből kapjuk, hogy

$$-\alpha_2\mathbf{u}=2\delta[\delta\mathbf{B}^{-1}\mathbf{u}-\mathbf{B}^{-1}(\mathbf{Ac}-\mathbf{e})],$$

amelyet $(-2\delta)$-val osztva az

$$\frac{\alpha_2}{2\delta}\mathbf{u}=-\delta\mathbf{B}^{-1}\mathbf{u}+\mathbf{B}^{-1}(\mathbf{Ac}-\mathbf{e})$$

egyenlethez jutunk. Mivel $\bar{\mathbf{x}}\neq\mathbf{0}$, azaz az origó külső pontja az $F$ ellipszoidnak, így $\|\mathbf{u}\|_E=1$ az optimális megoldásnál, tehát

$$\left\|\frac{\alpha_2}{2\delta}\mathbf{u}\right\|_E=\frac{\alpha_2}{2\delta}.$$

Legyen $\bar{\mathbf{x}}=\frac{\alpha_2}{2\delta}\mathbf{u}$, így

$$\bar{\mathbf{x}}=-\frac{\delta\mathbf{B}^{-1}\bar{\mathbf{x}}}{\|\bar{\mathbf{x}}\|_E}+\mathbf{B}^{-1}(\mathbf{Ac}-\mathbf{e}),$$

azaz az $\bar{\mathbf{x}}$, amely egyetlen megoldása az (5.11) egyenletrendszernek, az $F$ ellipszoidnak az origóhoz legközelebbi pontja $\|\ \|_A$ normával mérve a távolságot.

Az 5.5. lemma értelmében feladatunk ekvivalens egy ellipszoid origóhoz legközelebbi pontjának a meghatározásával.

Térjünk vissza az (5.11) egyenletrendszer megoldásához. A továbbiakban tetszőleges $\delta>0$ esetén használható eljárást adunk a keresett megoldás meghatározására. Eljárásunk lényegesen több számolást igényel, mint az (5.12) iteráció végrehajtása, így „kicsiny" $\delta$ esetén továbbra is célszerű azt alkalmaznunk.

Egy egyismeretlenes egyenlet, és egy lineáris egyenletrendszer megoldására vezetjük vissza feladatunkat.

5.6. TÉTEL. Legyen $\tilde{\mathbf{x}}$ megoldása az $\tilde{\mathbf{x}}=-L\mathbf{B}^{-1}\tilde{\mathbf{x}}+\mathbf{B}^{-1}(\mathbf{Ac}-\mathbf{e})$ lineáris egyenletrendszernek, ahol $L=\frac{\delta}{\|\tilde{\mathbf{x}}\|_E}$, akkor $\tilde{\mathbf{x}}$ megoldása (5.11) egyenletrendszernek is.

*Bizonyítás.* Mivel $\tilde{\mathbf{x}}=-L\mathbf{B}^{-1}\tilde{\mathbf{x}}+\mathbf{B}^{-1}(\mathbf{Ac}-\mathbf{e})$, így $L=\frac{\delta}{\|\tilde{\mathbf{x}}\|_E}$-t helyettesítve (5.11) egyenletet kapjuk, ami bizonyítja állításunkat.

Eljárásunk adott $\delta$ esetén 5.6. tétel alapján a következő:

— Mivel $\tilde{\mathbf{x}}=-L\mathbf{B}^{-1}\tilde{\mathbf{x}}+\mathbf{B}^{-1}(\mathbf{Ac}-\mathbf{e})$, így $\tilde{\mathbf{x}}=(\mathbf{E}+L\mathbf{B}^{-1})^{-1}\mathbf{B}^{-1}(\mathbf{Ac}-\mathbf{e})=$
$=(\mathbf{B}+L\mathbf{E})^{-1}(\mathbf{Ac}-\mathbf{e})$.

Határozzuk meg a

$$\delta = L\|(\mathbf{B}+L\mathbf{E})^{-1}(\mathbf{Ac}-\mathbf{e})\|_E \tag{5.17}$$

egyenlet $L$ megoldását. Ez jól ismert numerikus módszerekkel ugyanúgy, mint (4.19) egyenlet megoldható. Sajnos $L$ meghatározása nagyon számolásigényes, mivel minden lépés tartalmaz mátrixinvertálást.

— A fent meghatározott $L$-lel oldjuk meg az

$$\tilde{\mathbf{x}}=-L\mathbf{B}^{-1}\tilde{\mathbf{x}}+\mathbf{B}^{-1}(\mathbf{Ac}-\mathbf{e}) \tag{5.18}$$

lineáris egyenletrendszert. Ezt megoldhatjuk direkt módszerekkel, vagy a (3.8) iterációhoz hasonló módon konstruált konvergens iteratív eljárással, mely az alábbi alakba írható:

$$\mathbf{x}^{(k+1)}=\frac{K}{K+L}\left(\mathbf{E}-\frac{1}{K}\mathbf{B}\right)\mathbf{x}^{(k)}+\frac{1}{K+L}(\mathbf{Ac}-\mathbf{e}).$$

*1. Megjegyzés:* Amennyiben el akarjuk kerülni (5.17) egyenlet megoldását, és megelégszünk egy közelítő $L'$ megoldással, amely egy $\delta'\sim\delta$ paraméterhez tartozik, akkor használhatjuk az $L'=\frac{\delta}{\|\mathbf{B}^{-1}(\mathbf{Ac}-\mathbf{e})\|_E-\delta\|\mathbf{B}^{-1}\|_E}$ képletet. Ezt a képletet (5.11)-ből nyerjük, mivel onnan az $\|\tilde{\mathbf{x}}\|_E+\delta\|\mathbf{B}^{-1}\|_E\geqq\|\mathbf{B}^{-1}(\mathbf{Ac}-\mathbf{e})\|_E$ egyenlőséget kapjuk, melyből $\|\tilde{\mathbf{x}}\|_E=\frac{\delta}{L}$, és egyenlőség helyettesítéssel nyerjük a fenti közelítő képletet $L'$-re.

$L'$ értékből (5.17) képlet segítségével meghatározhatjuk, hogy ténylegesen milyen $\delta'$ értékre oldottuk meg feladatunkat. Ha ez nem kielégítő, akkor javíthatunk az $L'$ paraméteren. Mivel $L\|\tilde{\mathbf{x}}\|_{\mathbf{E}}=\delta$ és $\tilde{\mathbf{x}}$ az $F$ ellipszoid origóhoz legközelebbi pontja (igaz, hogy $\|\ \|_{\mathbf{A}}$ normában), így $\delta$ növekedtével $\|\tilde{\mathbf{x}}\|$ csökken, tehát $L$ nő, és ha $\delta$ csökken, akkor $\|\tilde{\mathbf{x}}\|$ nő, és így $L$-nek is csökkennie kell.

*2. Megjegyzés.* Amennyiben valamilyen numerikus módszerrel megoldjuk (5.17) egyenletet, akkor annál az iteratív lépésnél, mikor $L$-et elfogadjuk, rendelkezésünkre áll a $(\mathbf{B}+L\mathbf{E})^{-1}(\mathbf{Ac}-\mathbf{e})$ mennyiség, amiről láttuk, hogy egyenlő $\tilde{\mathbf{x}}$-mal, így (5.18)-at is megoldottuk egyidejűleg.

*3. Megjegyzés.* A $(\mathbf{B}+L\mathbf{E})^{-1}$ inverz meghatározására sorfejtést is használhatunk. Mivel $(\mathbf{B}+L\mathbf{E})^{-1}=\frac{1}{K+L}\left[\mathbf{E}-\frac{K}{K+L}\left(\mathbf{E}-\frac{1}{K}\mathbf{B}\right)\right]^{-1}$ és $\frac{K}{K+L}\left(\mathbf{E}-\frac{1}{K}\mathbf{B}\right)$ spektrálsugara kisebb, mint egy, így használhatjuk az $(\mathbf{E}-\mathbf{D})^{-1}=\mathbf{E}+\mathbf{D}+\mathbf{D}^2+\ldots$ konvergens sorfejtést. Amennyiben az $\left(\mathbf{E}-\frac{1}{K}\mathbf{B}\right)^i$, $i=0, 1, 2, \ldots$ mátrixhatványokat tárolni tudjuk, akkor az inverz meghatározása különböző $L$ értékek esetén csak mátrix és konstans szorzását, valamint mátrixok összegzését tartalmazza.

Végül belátjuk, hogy (5.17) egyenletnek akkor és csak akkor van megoldása, ha (5.1) feladat optimális célfüggvényértéke nem zérus.

5.7. LEMMA. Az (5.17) egyenlet akkor és csak akkor oldható meg, ha

$$0 < \delta^2 < (\mathbf{Ac}-\mathbf{e})^2.$$

*Bizonyítás.* Mivel $L\|(\mathbf{B}+L\mathbf{E})^{-1}(\mathbf{Ac}-\mathbf{e})\|_{\mathbf{E}}$ $L>0$ esetén folytonos függvénye $L$-nek, így elég azt belátnunk, hogy $\lim_{L\to 0} L\|(\mathbf{B}+L\mathbf{E})^{-1}(\mathbf{Ac}-\mathbf{e})\|_{\mathbf{E}}=0$ és $\lim_{L\to\infty} L\|(\mathbf{B}+L\mathbf{E})^{-1}(\mathbf{Ac}-\mathbf{e})\|_{\mathbf{E}}=(\mathbf{Ac}-\mathbf{e})^2$. Egyszerű átalakításokkal bizonyíthatjuk ezeket az összefüggéseket:

$$\lim_{L\to 0} L\|\mathbf{B}+L\mathbf{E})^{-1}(\mathbf{Ac}-\mathbf{e})\|_{\mathbf{E}} = \lim_{L\to 0} L \lim_{L\to 0}\|(\mathbf{B}+L\mathbf{E})^{-1}(\mathbf{Ac}-\mathbf{e})\|_{\mathbf{E}} = 0\|\mathbf{B}^{-1}(\mathbf{Ac}-\mathbf{e})\|_{\mathbf{E}} = 0.$$

$$\lim_{L\to\infty} L\|(\mathbf{B}+L\mathbf{E})^{-1}(\mathbf{Ac}-\mathbf{e})\|_{\mathbf{E}} = \lim_{L\to\infty}\left\|\left(\frac{\mathbf{B}}{L}+\mathbf{E}\right)^{-1}(\mathbf{Ac}-\mathbf{e})\right\|_{\mathbf{E}} = \|\mathbf{Ac}-\mathbf{e}\|_{\mathbf{E}}.$$

## 6. Számítási tapasztalatok

A három simítási eljárást mintapéldákkal hasonlítottuk össze, azok eredményét közöljük röviden az alábbiakban. Mielőtt erre rátérnénk, iterációs módszereink azon előnyös tulajdonságára mutatunk rá, hogy amennyiben új mérési alappontok felvételére kerül sor, akkor eddigi eredményünket megtartva folytathatjuk az iterációt. (Az eddig kapott közelítő vektort, kiegészítve az új koordinátákkal, használhatjuk induló vektornak.) Az **A** mátrix és a **c** és **e** vektorok bővítése új koordinátákkal szintén egyszerű feladat.

Példánkhoz a $k=2$ értéket választottuk. A valódi, a mért és a simított adatokat az 1. táblázatban közöljük.

## 1. TÁBLÁZAT

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|
| 0,81 | 0,72 | | | 0,7408 | | 0,7911 | |
| 0,51 | 0,63 | −0,23 | 0,02 | 0,5962 | −0,1682 | 0,5202 | 0,0011 |
| 0,24 | 0,3 | −0,1 | 0,03 | 0,2834 | −0,0426 | 0,2504 | 0,0195 |
| 0,0 | −0,13 | 0,41 | 0,04 | −0,072 | −0,0547 | 0,0001 | 0,0495 |
| −0,2 | −0,17 | −0,07 | 0,06 | −0,1941 | −0,197 | −0,2007 | 0,0577 |
| −0,34 | −0,28 | 0,04 | 0,28 | −0,3191 | 0,1775 | −0,3438 | 0,2734 |
| −0,2 | −0,35 | 0,56 | 0,06 | −0,2666 | 0,2698 | −0,2135 | 0,0961 |
| 0,0 | 0,14 | −0,6 | −0,08 | 0,0557 | −0,4476 | −0,0171 | −0,0949 |
| −0,12 | 0,03 | 0,0 | −0,24 | 0,0696 | −0,1477 | 0,0844 | −0,2359 |
| 0,0 | −0,08 | 0,01 | −0,08 | −0,0642 | −0,0236 | −0,05 | −0,0699 |
| −0,2 | −0,18 | −0,17 | 0,06 | −0,2216 | −0,062 | −0,2543 | 0,0689 |
| −0,34 | −0,45 | 0,38 | 0,28 | −0,441 | 0,3543 | −0,3897 | 0,3066 |
| −0,2 | −0,34 | 0,33 | 0,06 | −0,3061 | 0,25 | −0,2185 | 0,0286 |
| 0,0 | 0,1 | | | 0,0788 | | −0,0187 | |
| iterációszám | | | | 17 | | 394 | |

| 9 | 10 | 11 | 12 | 13 | 14 |
|---|---|---|---|---|---|
| 0,7282 | | 0,7597 | | 0,7653 | |
| 0,6172 | −0,2147 | 0,566 | −0,0989 | 0,5575 | −0,0794 |
| 0,2914 | −0,069 | 0,2734 | −0,0176 | 0,2703 | −0,0115 |
| −0,1034 | 0,3171 | −0,0368 | 0,1427 | −0,0284 | 0,1217 |
| −0,1811 | −0,0415 | −0,2043 | 0,0365 | −0,2054 | 0,044 |
| −0,3003 | 0,1126 | −0,3353 | 0,2325 | −0,3384 | 0,2437 |
| −0,3069 | 0,41 | −0,2338 | 0,1531 | −0,2277 | 0,1308 |
| 0,0965 | 0,4486 | 0,0208 | −0,1919 | 0,0138 | −0,1697 |
| 0,0513 | −0,0798 | 0,0835 | −0,2026 | 0,0856 | −0,2128 |
| −0,0737 | 0,0002 | −0,0564 | −0,0614 | −0,0554 | −0,0577 |
| −0,1985 | −0,1226 | −0,2487 | 0,0078 | −0,2541 | 0,0239 |
| −0,4459 | 0,367 | −0,4332 | 0,3525 | −0,4289 | 0,3476 |
| −0,3263 | 0,2984 | −0,2652 | 0,1474 | −0,2561 | 0,1254 |
| 0,0917 | | 0,0502 | | 0,0421 | |
| 6 | | 43 | | 56 | |

| 15 | 16 | 17 | 18 | 19 | 20 |
|---|---|---|---|---|---|
| 0,8087 | | 0,7315 | | 0,7378 | |
| 0,5153 | 0,0192 | 0,6115 | −0,2025 | 0,601 | −0,1791 |
| 0,2411 | 0,0293 | 0,289 | −0,606 | 0,2851 | −0,0476 |
| −0,0038 | 0,0403 | −0,0941 | 0,2922 | −0,0784 | 0,2504 |
| −0,2084 | 0,0597 | −0,185 | −0,0305 | −0,1915 | −0,011 |
| −0,3533 | 0,2799 | −0,3064 | 0,1341 | −0,3156 | 0,1657 |
| −0,2183 | 0,0604 | −0,2937 | 0,3642 | −0,274 | 0,2956 |
| −0,0229 | −0,0801 | 0,0832 | −0,4027 | 0,0632 | −0,3371 |
| 0,0924 | −0,2391 | 0,0574 | −0,1025 | 0,0633 | −0,1295 |
| −0,0314 | −0,0789 | −0,0709 | −0,0061 | −0,0661 | −0,0213 |
| −0,2341 | 0,0612 | −0,2053 | −0,1047 | −0,2168 | −0,0745 |
| −0,3756 | 0,2816 | −0,4444 | 0,3627 | −0,442 | 0,3566 |
| −0,2355 | 0,0613 | −0,3208 | 0,2853 | −0,3106 | 0,2608 |
| −0,0341 | | 0,0881 | | 0,0816 | |
| 16 | | 3 | | 17 | |

A táblázat egyes oszlopai a következő adatokat tartalmazzák:

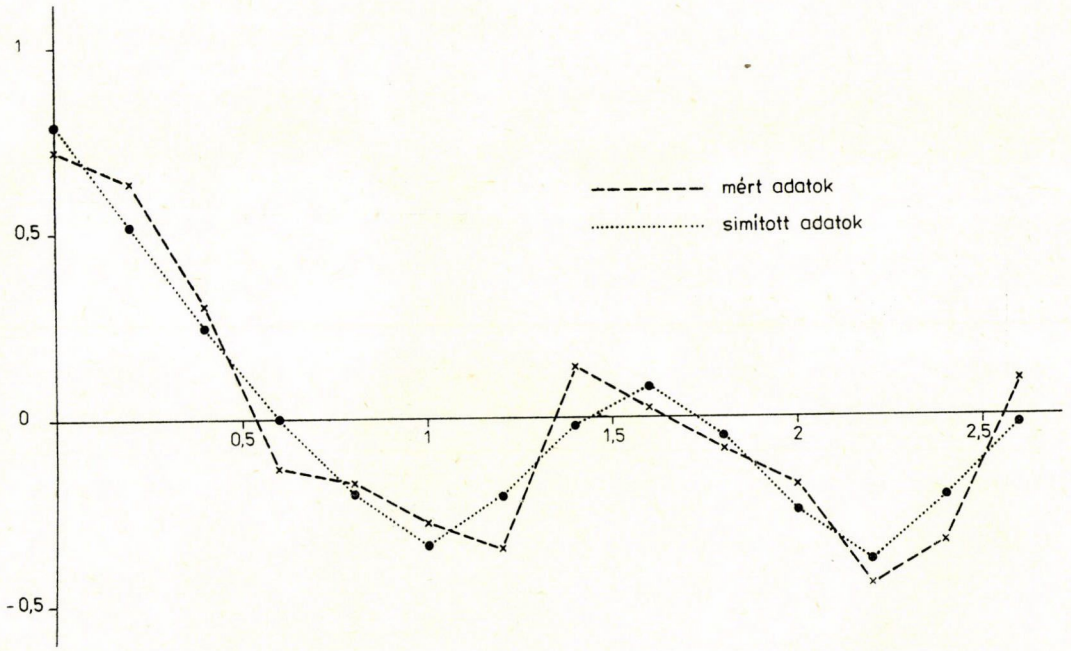
1. oszlop: A valódi $y_i$ értékek
2. oszlop: A mért $\gamma_i$ értékek
3. oszlop: A $\gamma_i$ értékekből számított 2. differenciák
4. oszlop: A mért, ismert $\varepsilon_i$ 2. differenciák
5. oszlop: Az I. modellel nyert $y_i$ értékek $\lambda=3$ paraméterrel
6. oszlop: Az 5. oszlopból számított 2. differenciák
7. oszlop: Az I. modellel nyert $y_i$ értékek $\lambda=0{,}5$ paraméterrel
8. oszlop: A 7. oszlopból számított 2. differenciák
9. oszlop: A II. modellel $\delta=0{,}08$ paraméterérték mellett, a (4.14) iterációval nyert $y_i$ értékek
10. oszlop: A 9. oszlopból számított 2. differenciák
11. oszlop: A II. modellel $\delta=0{,}252$ paraméterérték mellett, (4.19) egyenlet megoldásával nyert $y_i$ értékek $(L=0{,}3364)$
12. oszlop: A 11. oszlopból számított második differenciák
13. oszlop: A II. modellel $\delta=1$ paraméterrel és a közelítő formulával számított $L=0{,}4951$ értékhez tartozó $y_i$ értékek. Valójában $\delta=0{,}2723$
14. oszlop: A 13. oszlopból számított 2. differenciák
15. oszlop: A III. modellel $\delta=0{,}003$ paraméterérték mellett az (5.12) iterációval nyert $y_i$ értékek
16. oszlop: A 15. oszlopból számított 2. differenciák
17. oszlop: A III. modellel $\delta=0{,}76$ paraméterérték mellett az (5.17) egyenlet megoldásával nyert $y_i$ értékek $(L=19{,}1269)$
18. oszlop: A 17. oszlopból számított 2. differenciák
19. oszlop: A III. modellel $\delta=0{,}275$ paraméterre és a közelítő formulával számított $L=11{,}044$ értékhez tartozó $y_i$ értékek. Valójában $\delta=0{,}681$
20. oszlop: A 19. oszlopból számított 2. differenciák.

Mint számítási eredményeinkből is látható, az első modellel végzett számítások esetén, ha $\lambda>1$, akkor iterativ eljárásunk gyorsan konvergál, viszont simításunk csak kis mértékű. Nagyobb simítást érhetünk el „kis" $\lambda$ érték választással, ekkor viszont iterációnk konvergenciája lesz lassúbb.

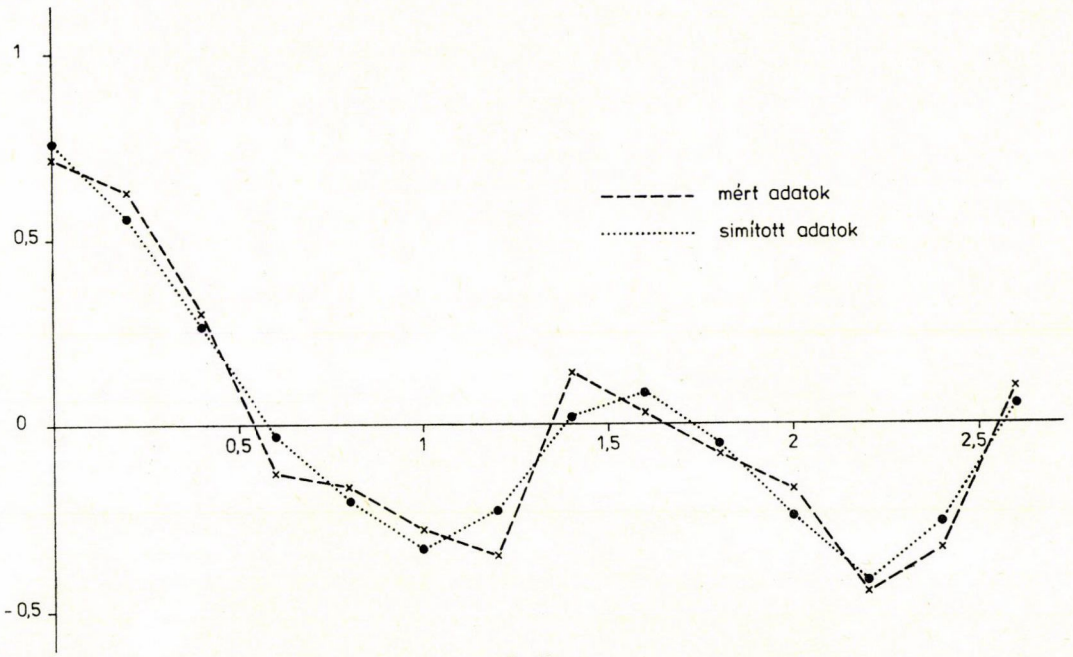
Ez az észrevétel nyilvánvaló következménye annak, hogy a $\frac{K}{K+\lambda^2}$ együttható értéke közel egy, ha $\lambda$ kicsi, és $\lambda$ növekedtével értéke csökken.

Második modellünk esetén, ha $\delta$ „kicsi", akkor a simítás mértéke kicsi, míg $\delta$ növekedtével a simítás mértéke nő. A harmadik modellnél végzett simítások éppen fordított képet mutatnak, ott „kicsiny" $\delta$ esetén végzünk nagymértékű simítást, míg „nagy" $\delta$ esetén a simítás mértéke csökken. Így ha a gyorsan konvergáló direkt iterációs eljárást szeretnénk alkalmazni, akkor attól függően választhatjuk a II. vagy a III. modellt, hogy kisebb vagy nagyobb mértékű simítást akarunk végezni.

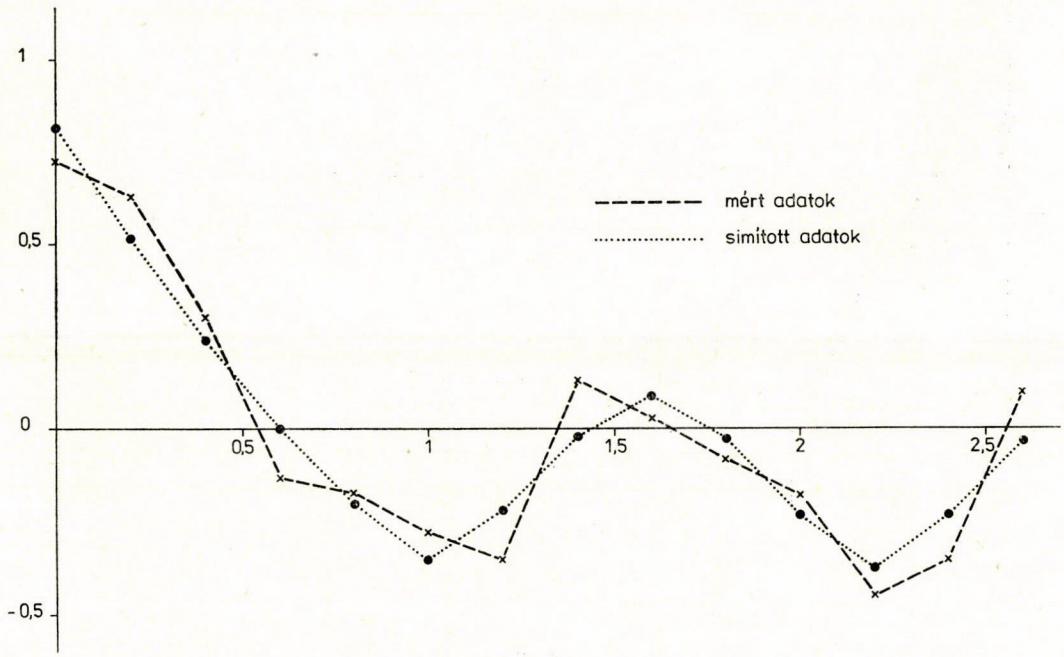
A táblázat utolsó sorában található számok azt mutatják, hogy hány iterációra volt szükség a feladat megoldásához. Ezek közül külön kell választani a 11. és 17. oszlop alatt levő iterációszámokat, mivel a többi iteráció csak mátrixvektor szorzást és vektor-vektor összeadát tartalmaz, ezek pedig iterációs lépésenként egy mátrix-invertálást.


1
0,5
0
-0,5
0,5
1
1,5
2
2,5
mért adatok
simított adatok

*4. ábra*


1
0,5
0
-0,5
0,5
1
1,5
2
2,5
mért adatok
simított adatok

*5. ábra*




*6. ábra*

Összességében megállapíthatjuk, hogy a különböző paraméterek ismeretében mindig kiválaszthatjuk az alkalmas modellt, az alkalmas, gyorsan konvergáló iterációs megoldási módszerrel, amellyel feladatunkat gyorsan megoldhatjuk.

Végül néhány ábrával illusztráljuk számítási eredményeinket.

Mindhárom ábrán a mért és a simított függvényértékeket ábrázoljuk. A 4. ábrán a mért és az I. modellel $\lambda=0{,}5$ paraméterértékkel nyert függvényértékeket ábrázoljuk (az 1. táblázat 2. és 7. oszlopa). Az 5. ábrán a mért és a II. modellel $\delta=0{,}252$ paraméterrel számított függvényértékek láthatók (az 1. táblázat 2. és 11. oszlopa). A 6. ábrán a mért és a III. modellel $\delta=0{,}003$ paraméterrel számított függvényértékek láthatók (az 1. táblázat 2. és 15. oszlopa).

## 7. Függelék

### *7.1. Vektor és mátrix normák*

Legyenek $\mathbf{t}\in R^{n-k}$ és $\mathbf{A}$: $(n-k)\times n$ teljes sorrangú mátrix, és legyen $\mathbf{B}=\mathbf{A}\mathbf{A}^T$.

a)
$$\|\mathbf{t}\|_E = \sqrt{\mathbf{t}^2} = \Big(\sum_{i=1}^{n-k} (\tau_i)^2\Big)^{1/2}$$

euklideszi vektornorma,

$$\|\mathbf{B}\|_E = \Big(\sum_{i=1}^{n-k}\sum_{j=1}^{n-k} (b_{ij})^2\Big)^{1/2}$$

euklideszi mátrixnorma.

Ismert, hogy $\|\mathbf{Bt}\|_E \leqq \|\mathbf{B}\|_E \|\mathbf{t}\|_E$

b) $$\|\mathbf{t}\|_A = \sqrt{(\mathbf{tA})^2} = \sqrt{\mathbf{tBt}}$$

vektornorma.

ÁLLÍTÁS. $\|\mathbf{Bt}\|_A \leqq \|\mathbf{A}^T\mathbf{A}\|_E \|\mathbf{t}\|_A$.

*Bizonyítás:*

$$\|\mathbf{Bt}\|_A = (\mathbf{t}^T\mathbf{AA}^T\mathbf{AA}^T\mathbf{AA}^T\mathbf{t})^{1/2} \leqq [(\mathbf{tA})^2]^{1/2} \|\mathbf{A}^T\mathbf{A}\|_E = \|\mathbf{A}^T\mathbf{A}\|_E \|\mathbf{t}\|_A,$$

ahol a *Cauchy—Schwarc—Bunyakovszky egyenlőtlenséget* alkalmaztuk.

Vektor és mátrix normákról részletes információk találhatók YUNG [9] könyvében.

*7.2. Az $l_p$ programozási feladatpár és legfontosabb tulajdonságai*

Az $l_p$ programozási feladatpár részletes vizsgálata TERLAKY [7] dolgozatában található. Itt csak a dolgozatunkban felhasznált tételeket közöljük, bizonyítás nélkül.

Legyen $\mathbf{y}, \mathbf{f} \in R^m$, $\mathbf{x}, \mathbf{c} \in R^n$, $\mathbf{z}, \mathbf{d} \in R^r$, $\mathbf{A}$: $n \times m$, $\mathbf{B}$: $r \times m$ mátrixok, $p_i, q_i > 1$, úgy, hogy $\frac{1}{p_i} + \frac{1}{q_i} = 1$, $i = 1, \ldots, n$, valamint $I_k$, $k = 1, \ldots, r$ indexhalmazok úgy, hogy $I_k \cap I_j = \emptyset$, $k \neq j$ esetén, valamint $\bigcup_{k=1}^{r} I_k = \{1, \ldots, n\}$.

*Az $l_p$ programozás primál feladata:*

$$\max \mathbf{fy}$$

$$G_k(\mathbf{y}) = \sum_{i \in I_k} \frac{1}{p_i} |\mathbf{a}_i\mathbf{y} - \gamma_i|^{p_i} + \mathbf{b}_k\mathbf{y} - \delta_k \leqq 0. \tag{7.1}$$

*Az $l_p$ programozás duál feladata:*

$$\inf \left\{ \mathbf{cx} + \mathbf{dz} + \sum_{\substack{k=1 \\ \zeta_k > 0}}^{r} \zeta_k \sum_{i \in I_k} \frac{1}{q_i} \left| \frac{\xi_i}{\zeta_k} \right|^{q_i} \right\}$$

$$\mathbf{xA} + \mathbf{zB} = \mathbf{f} \tag{7.2}$$

$$\mathbf{z} \geqq \mathbf{0}$$

$$\zeta_k = 0 \Rightarrow \xi_i = 0, \quad i \in I_k, \quad k = 1, \ldots, r.$$

F.1. LEMMA. Ha $\mathbf{y}$ megengedett megoldása (7.1) és $(\mathbf{x}, \mathbf{z})$ megengedett megoldása (7.2) feladatnak, akkor

$$\mathbf{fy} \geqq \mathbf{cx} + \mathbf{dz} + \sum_{\substack{k=1 \\ \zeta_k > 0}}^{r} \zeta_k \sum_{i \in I_k} \frac{1}{q_i} \left| \frac{\xi_i}{\zeta_k} \right|^{q_i}.$$

Egyenlőséggel akkor és csak akkor, ha

$$\zeta_k G_k(\mathbf{y}) = 0, \quad k = 1, \ldots, r$$

és

$$\zeta_k = 0, \quad \text{vagy} \quad \mathbf{a}_i\mathbf{y} - \gamma_i = \operatorname{sgn} \xi_i \left| \frac{\xi_i}{\zeta_k} \right|^{q_i - 1}.$$

F.1. DEFINÍCIÓ. A (7.1) feladat *Slater reguláris,* ha van olyan $\mathbf{y} \in R^m$, hogy $G_k(\mathbf{y}) < 0$ azon $k$-kra, mikor $G_k$ nemlineáris és $G_k(\mathbf{y}) \leqq 0$ azon $k$-kra, mikor $G_k$ lineáris.

F.1. TÉTEL. Ha a (7.1) feladatnak létezik megengedett megoldása, és célfüggvénye felülről korlátos, akkor van optimális megoldása.

F.2. TÉTEL. Ha a (7.1) feladat *Slater reguláris,* és célfüggvénye felülről korlátos, akkor a (7.2) feladatnak létezik optimális megoldása és a primál és a duál feladat optimális célfüggvényértékei megegyeznek.

## IRODALOM

[1] NYIRI, A., „Tapasztalati függvények simítása", *Alkalmazott Matematikai Lapok* **6** (1980), 237—286.
[2] PETERSON, E. L. and ECKER, J. G., "Geometric programming: duality in quadratic programming and $l_p$ approximation, I. "(Proceedings of the International Symposion on Methematical Programming, ed. *H. W. Kuhn* and *A. W. Tucker,* Princeton University Press, 1970).
[3] PETERSON, E. L. and ECKER, J. G., "Geometric programming: duality in quadratic programming and $l_p$ approximation, II.", *SIAM Journal on Applied Mathematics* **13** (1967) 317—340.
[4] PETERSON, E. L. and ECKER, J. G., "Geometric programming: duality in quadratic programming and $l_p$ approximation, III.", *Journal of Mathematical Analysis and Applications* **29** (1970) 365—383.
[5] RALSTON, A., *Bevezetés a numerikus analízisbe* (Műszaki Könyvkiadó, Budapest, 1969)
[6] SZIDAROVSZKY, F., *Bevezetés a numerikus módszerekbe* (Közgazdasági és Jogi Könyvkiadó, Budapest, 1974).
[7] TERLAKY, T., „Az $l_p$ programozásról", *Alkalmazott Matematikai Lapok* **6** (1980) 27—63.
[8] WHITTAKER, E., *The Calculus of Observations* (Blackie Son Ltd., London Glasgow, 1954.)
[9] YUNG, D. M., *Nagy lineáris rendszerek iterációs megoldása* (Műszaki Könyvkiadó, Budapest, 1979.)

(Beérkezett: 1983. május 9.)

TERLAKY TAMÁS
ELTE TTK OPERÁCIÓKUTATÁSI TANSZÉK
1088 BUDAPEST, MÚZEUM KRT. 6—8.

## SMOOTHING EMPIRICAL FUNCTIONS BY $l_p$ PROGRAMMING

T. TERLAKY

This paper presents three models for smoothing empirical functions. The first model is the well known *Whittakers graduating model,* the second and the third model is new.

New iterative procedures are presented for solving these models. These iterative procedures are based on the equilibrum conditions of $l_p$ programming, so our solution procedure for solving the *Whittakers model* is also new.

*Alkalmazott Matematikai Lapok* **10** (1984) 353—357

# FUZZY LEKÉPEZÉSEK ÉS TULAJDONSÁGAIK

FULLÉR RÓBERT

Budapest

A dolgozat a fuzzy leképezések néhány tulajdonságával foglalkozik. Előbb két speciális tulajdonságú fuzzy leképezés képterére adunk összefüggést, majd néhány fixponttétel fuzzy kiterjesztésével foglalkozunk.

## 1. Bevezetés

A műszaki, gazdasági tervezésnél, a természeti folyamatok kutatásánál, szociológiai, orvos-pszichológiai vizsgálatoknál stb. a matematikai modell kialakítása mindig számos bizonytalan elem figyelembevételét igényli. Ilyen bizonytalan elem adódhat pl. a nem egyértelmű meghatározottságból, a véletlenszerűen bekövetkező események előfordulásából, mérések pontatlanságából, a vélemények, döntések szubjektív voltából. A feladat jellegéből adódó bizonytalan információk kezelésére a matematikai modellezésben különböző módszerek használatosak, pl. a valószínűségelméleti és statisztikai módszerek, az intervallumaritmetika módszerei, a nem korrekt felállítású feladatok kezelésére vonatkozó regularizálási technikák stb. Ezek a módszerek azonban kevésbé alkalmasak akkor, ha a modellben lényeges szerepet játszanak a szubjektív döntések és szakértői vélemények, amikor a feladatot jellemző paraméterhalmaz határai nem adhatók meg éles konturral. E nehézségek áthidalására jól használható az L. A. ZADEH által 1965-ben bevezetett fuzzy halmaz-fogalom [1] és a rá épülő fuzzy leképezések elmélete.

## 2. Alapfogalmak, jelölések

Legyen $X \neq \emptyset$, $I=[0, 1] \subset \mathbf{R}$. Ekkor az

$$I^X = \{\mu | \mu\colon X \to I\}$$

elemeit $X$ fuzzy halmazainak nevezzük. Speciálisan, az $A \subset X$ halmaz $\chi_A$ karakterisztikus függvényére $\chi_A \in I^X$. A

$$\operatorname{supp} \mu = \{x \in X\colon \mu(x) > 0\}$$

halmaz a $\mu$ fuzzy halmaz tartója.

Ha $\alpha \in I$, akkor a $\mu \in I^X$ $\alpha$-nivóhalmaza

$$\omega_\alpha(\mu) = \{x \in X\colon \mu(x) \geqq \alpha\}.$$

A $|\mu|=\sup_{x\in X}\mu(x)$ jelöléssel

$$\omega_{|\mu|} = \arg\sup_{x\in X}\mu(x) = \omega_{|\mu|}(\mu)$$

a $\mu$ fuzzy halmaz maximális nivóhalmaza. Ha $X$ vektortér $\mathbf{R}$ felett, akkor a $\mu\in I^X$ fuzzy halmaz konvex, ha minden $r\in I$-re és $x, y\in X$ esetén

$$\mu(rx+(1-r)y) \geqq \min\{\mu(x), \mu(y)\}.$$

Ismeretes, hogy ha $X$ lineáris tér $\mathbf{R}$ felett, akkor $\mu\in I^X$ konvex akkor és csak akkor, ha minden $\alpha\in I$ esetén $\omega_\alpha(\mu)$ konvex.

Ha $(X, \tau)$ topologikus tér, akkor a $\mu\in I^X$ fuzzy halmaz

*a)* folytonos, ha $\mu(x)$ folytonos $X$-en a $\tau$ topológiában;

*b)* felülről félig folytonos, ha $\mu(x)$ felülről félig folytonos $X$-en a $\tau$ topológiában, vagyis ha $\omega_\alpha(\mu)$ zárt a $\tau$ topológiában minden $\alpha\in I$-re;

*c)* kompakt, ha $\omega_\alpha(\mu)$ kompakt a $\tau$ topológiában minden $\alpha\in I$-re. Legyen $\mu\in I^X$ és $\nu\in I^X$. Azt mondjuk, hogy $\mu\subset\nu$, ha $\mu(x)\leqq\nu(x)$ minden $x\in X$-re.

## 3. Fuzzy leképezések

Legyenek $X$ és $Y$ nemüres halmazok, $f$: $X\to Y$ klasszikus értelmezésű függvény $R_f\subset Y$ értékkészlettel. Ekkor az $f$ által generált

$$\hat{f}\colon I^X \to I^Y$$

fuzzy leképezés az $X$ tér egy $\mu$ fuzzy halmazához az $Y$ tér

$$(\hat{f}(\mu))(y) = \begin{cases} \sup\limits_{x\in f^{-1}(y)}\mu(x), & \text{ha} \quad y\in R_f \\ 0, & \text{ha} \quad y\notin R_f \end{cases}$$

formulával adott fuzzy halmazát rendeli, ahol

$$f^{-1}(y) = \{x\in X\colon f(x) = y, \quad y\in R_f\}.$$

Pl. legyen $X=\mathbf{R}$, $Y=\mathbf{R}$, $f(x)=x^3$. Ekkor

$$(\hat{f}(\mu))(y) = \sup_{x\in f^{-1}(y)}\mu(x) = \sup_{x=\sqrt[3]{y}}\mu(x) = \mu(\sqrt[3]{y}).$$

Az $\hat{f}$ fuzzy leképezések egy speciális tulajdonságát mutatja az alábbi tétel:

3.1. TÉTEL. Ha egy $\mu\in I^X$-re igaz, hogy

$$f(\omega_1(\mu)) = R_f,$$

akkor

$$\hat{f}(\mu) = \chi_{R_f},$$

ahol $\chi_{R_f}$ az $R_f$ karakterisztikus függvénye.

*Bizonyítás.* $\hat{f}$ definíciójából következik, hogy

$$(\hat{f}(\mu))(y) = 0, \quad \text{ha} \quad y\notin R_f.$$

Másrészt, a feltétel szerint

$$R_f = f(\mu^{-1}(1)) = \{f(x)\colon \mu(x) = 1,\ x \in X\},$$

tehát, ha $y \in R_f$, akkor létezik olyan $z \in X$, melyre $f(z) = y$ és $\mu(z) = 1$. Ekkor az

$$1 \geqq (\hat{f}(\mu))(y) = \sup_{x \in f^{-1}(y)} \mu(x) = \sup_{f(x)=y} \mu(x) \geqq \mu(z) = 1$$

összefüggésből következik, hogy

$$(\hat{f}(\mu))(y) = 1, \quad \text{ha} \quad y \in R_f,$$

azaz $\hat{f}(\mu)$ valóban $R_f$ karakterisztikus függvénye.

A fenti tétel modellezés-technikai szempontból a következőképpen interpretálható. Ha az $X$ halmaz elemeire vonatkozó bizonytalansági információt a $\mu(x)$ tartalmazási függvény hordozza, azaz $x \in X\ \mu(x)$ szinten tekinthető a modell alapadataként elfogadhatónak, akkor a képtér biztosságát nem befolyásolják a bizonytalan elemek, hacsak a biztos elemek képe kitölti $f$ értékkészletét.

Vezessük be a következő definíciót:

*3.1. Definíció.* Azokat a fuzzy halmazokat, amelyek egyelemű halmazok karakterisztikus függvényei, valódi fuzzy számoknak nevezzük.

Jelölje $V(X)$ az $X$ valódi fuzzy számainak halmazát és vizsgáljuk az $\hat{f}$ fuzzy leképezés $V(X)$-re való $\hat{f}|V(X)$ leszűkítését. Igaz a következő

3.2. TÉTEL. Ha $f\colon X \to X$, akkor

$$\hat{f}|V(X)\colon V(X) \to V(X),$$

azaz $\hat{f}$ a valódi fuzzy számok halmazát önmagában képezi le.

*Bizonyítás.* Tekintsük $\hat{f}$ értékeit a $\chi_{\{x\}}(x \in X)$ valódi fuzzy számokon:

$$(\hat{f}(\chi_{\{x\}}))(y) = \sup_{t \in f^{-1}(y)} \chi_{\{x\}}(t) = \begin{cases} 1, & \text{ha} \quad t = x \\ 0, & \text{ha} \quad t \neq x, \end{cases}$$

azaz

$$(\hat{f}(\chi_{\{x\}}))(y) = \begin{cases} 1, & \text{ha} \quad x \in f^{-1}(y), \quad \text{azaz ha} \quad f(x) = y \\ 0, & \text{ha} \quad x \notin f^{-1}(y), \quad \text{azaz ha} \quad f(x) = y, \end{cases}$$

vagyis

$$\hat{f}(\chi_{\{x\}}) = \chi_{\{f(x)\}}.$$

E tétel fontos szerepet fog játszani a fuzzy leképezések fixpontvizsgálatánál.

## 4. Fixponttételek fuzzy kiterjesztése

Leképezések fixponttulajdonságának fuzzy kiterjesztése kétféleképpen is lehetséges.

*4.1. Definíció.* Az $f\colon X \to X$ klasszikus leképezésnek $x$ fuzzy fixpontja a $\mu \in I^X$ fuzzy halmazon, ha $f(x) = x$ és $x \in \omega_{|\mu|}$.

*4.2. Definíció.* Az $f\colon X \to X$ által generált $\hat{f}\colon I^X \to I^X$ fuzzy leképezésnek $\mu \in I^X$ fixpontja, ha $\hat{f}(\mu) = \mu$.

Klasszikus leképezések fuzzy fixpontjának létezését vizsgálta [2—4]. Egy ide vonatkozó alaptételnek tekinthető a következő

4.1. TÉTEL ([2]). Legyen *X lokálisan konvex Hausdorff típusú topologikus vektor, tér,* $\mu \in I^X$ nemüres, kompakt, konvex fuzzy halmaz. Ha $f: X \to X$ folytonos és az általa generált $\hat{f}$ fuzzy leképezésre teljesül az $\hat{f}(\mu) \subset \mu$ tartalmazás, akkor létezik $f$-nek fuzzy fixpontja $\mu$-n.

Nem nehéz belátni, hogy a fenti tételnek a feltételei enyhíthetők, nevezetesen $\mu$ konvexitása helyett elegendő megkívánni az $\omega_{|\mu|}$ maximális nívóhalmaz konvexitását.

Ha $X$ metrikus tér, akkor az $f$-re tett kontrakciós feltétel mellett a fuzzy fixpont egzisztenciája és unicitása is biztosítható.

A *Banach tétel* fuzzy kiterjesztésével kapjuk:

4.2. TÉTEL. Legyen $X$ teljes metrikus tér és $f: X \to X$ kontrakció, továbbá $\mu \in I^X$ nemüres felülről félig folytonos fuzzy halmaz, melyre az $\omega_{|\mu|}$ maximális nívóhalmaz nem üres és az $f$ által generált $\hat{f}$ fuzzy leképezésre $\hat{f}(\mu) \subset \mu$. Akkor $f$-nek egy és csak egy fuzzy fixpontja van $\mu$-n.

*Bizonyítás.* Legyen

$$W = \bigcap_{i=2} \omega_{\alpha_i}(\mu), \quad \alpha_i = \left(1 - \frac{1}{i}\right) |\mu|.$$

Nyilvánvalóan $W = \omega_{|\mu|}$. Mivel $\mu$ felülről félig folytonos, így $\omega_{\alpha_i}(\mu)$ zárt, következésképpen $W$ is az. Felhasználva, hogy $\hat{f}(\mu) \subset \mu$, bármely $w \in W$-re igaz, hogy

$$\sup_{x \in X} \mu(x) \geqq \mu(f(w)) \geqq (\hat{f}(\mu))(f(w)) = \sup_{x \in f^{-1}(f(w))} \mu(x) \geqq \mu(w) = \sup_{x \in X} \mu(x),$$

azaz

$$\mu(f(w)) = \sup_{x \in X} \mu(x), \quad \text{vagyis} \quad f(w) \in W.$$

Minthogy $w \in W$ tetszőleges volt, így $f(W) \subset W$. $f$ tehát a zárt $W$ önmagába való kontrakciója, így *Banach tétele* értelmében létezik egyetlen $w \in W = \omega_{|\mu|}$, melyre $f(w) = w$.

A fuzzy fixpont létezését vizsgálhatjuk olyan leképezésekre is, amelyek ugyan nem kontrakciók, de egymáshoz közeli argumentumokra kontrakcióként viselkednek.

*4.3. Definíció.* Legyen $(X, \varrho)$ metrikus tér. Az $f: X \to X$ leképezést $(\varepsilon, \lambda)$-paraméterű lokális kontrakciónak nevezzük, ha

$$\varrho(f(x), f(y)) < \lambda \varrho(x, y),$$

hacsak $\varrho(x, y) < \varepsilon$ és $0 < \lambda < 1$, $x, y \in X$.

Ismeretes EDELSTEIN tétele [5], miszerint, ha $X$ $\varepsilon$-láncosítható metrikus tér, azaz minden $x, y \in X$ $(x \neq y)$ pontpárhoz létezik olyan véges $X$-beli $\{x_i\}_{i=\overline{0,n}}$ pontsorozat úgy, hogy $x_0 = x$, $x_n = y$ és $\varrho(x_i, x_{i+1}) < \varepsilon$, $i = \overline{0, n-1}$, akkor az $X$ egy zárt részhalmazának minden önmagába való lokális kontrakciójának egy és csak egy fixpontja van. E tétel fuzzy kiterjesztése a következő:

4.3. TÉTEL. Legyen $(X, \varrho)$ $\varepsilon$-láncosítható metrikus tér, $Z \subset X$ zárt halmaz, $f\colon Z \to Z$ $(\varepsilon, \lambda)$-paraméterű lokális kontrakció, $\mu \in I^X$ felülről félig folytonos nem üres fuzzy halmaz, melyre $\operatorname{supp}\mu \subset Z$ és $\omega_{|\mu|} \neq \emptyset$, továbbá az $f$ által generált $\hat{f}$ fuzzy leképezésre teljesüljön az $\hat{f}(\mu) \subset \mu$ tartalmazás. Akkor $f$-nek egy és csak egy fuzzy fixpontja van $\mu$-n.

*Bizonyítás.* Az előző tétel bizonyításához hasonlóan belátható, hogy $f$ a nem üres zárt $W = \omega_{|\mu|} \subset Z$ önmagába való lokális kontrakciója, így a tétel állítása Edelstein tételéből következik.

A továbbiakban az $f$ által generált $f$ fuzzy leképezés fixpontját vizsgáljuk.

4.4. TÉTEL. Legyen $(X, \|.\|)$ *Banach tér,* $f\colon X \to X$ *Frechet-deriválható* és $\|f'(x)\| \leqq W < 1$ minden $x \in X$-re. Ekkor az $f$ által generált $\hat{f}$ fuzzy leképezésnek a valódi fuzzy számok körében van fixpontja.

*Bizonyítás.* A 3.2. tétel alapján $\hat{f}$ a valódi fuzzy számok halmazát önmagába képezi le, így tetszőleges $P_0 \in X$ kezdőértékkel képezhetjük a valódi fuzzy számoknak a $v_n = \chi_{\{P_n\}}$ sorozatát a következő szabállyal:

$$v_{n+1} = \hat{f}(v_n), \quad n = 0, 1, 2, \ldots .$$

Ekkor

$$\begin{aligned} v_1 &= \hat{f}(v_0) = \chi_{\{f(P_0)\}} = \chi_{\{P_1\}} \\ &\vdots \\ v_{n+1} &= \hat{f}(v_n) = \chi_{\{f(P_0)\}} = \chi_{\{P_{n+1}\}}. \end{aligned}$$

A $\|f'(x)\| \leqq W < 1$ feltétel mellett $f\colon X \to X$ kontrakció, így a $P_{n+1} = f(P_n)$ sorozat konvergens: $\lim_{n\to\infty} P_n = P^*$ és $P^*$ az $f$ fixpontja, azaz $f(P^*) = P^*$. Felhasználva a 2.2. tétel bizonyításában kapott eredményt

$$\hat{f}(\chi_{\{P^*\}}) = \chi_{\{f(P^*)\}} = \chi_{\{P^*\}}.$$

vagyis $\chi_{\{P^*\}} \in V(X)$ fixpontja $\hat{f}|V(X)$-nek.

## IRODALOM

[1] ZADEH, L. A., "Fuzzy sets", *Information and Control* **8** (1965) 338—353.
[2] WEISS, M. D., "Fixed points, separation and induced topologies for fuzzy sets", *J. Math. Anal.* **50** (1979) 142—150.
[3] STANISLAW, H., "Fuzzy mapping and fixed point theorem", *J. Math. Anal. Appl.* **83** (1981) 556—569.
[4] BUTNARIU, D., "Fixed points for fuzzy mappings", *Fuzzy Sets and Systems* 7 (1982) 191—207.
[5] HEGEDŰS, M., Fixponttételek metrikus terekben és néhány alkalmazásuk, Kandidátusi értekezés, Budapest, 1978.

(Beérkezett: 1984. március 5.)

FULLÉR RÓBERT
ELTE SZÁMÍTÓKÖZPONT
1117 BUDAPEST, BOGDÁNFFY ÚT 10/B.

## FUZZY MAPPINGS AND THEIR PROPERTIES

R. FULLÉR

The paper deals with some properties of fuzzy mappings. First it is determined the image of two special mappings, then the fuzzy extension of some fixed point theorems is examined.

*Alkalmazott Matematikai Lapok* **10** (1984) 359—372

# KATONA G. O. H. EGY PROBLÉMÁJÁNAK ÁLTALÁNOSÍTÁSÁRÓL

VARECZA ÁRPÁD

Nyíregyháza

Legyen $H$ egy teljesen rendezett $n$ elemű halmaz, amelynek rendezését nem ismerjük. KATONA G. O. H. vetette fel azt a kérdést, hogy ha $A=\{p, q\}$ $(1\leqq p<q\leqq n)$ és $x$, $y$ a $H$ halmaz két tetszőleges eleme, akkor mennyi összehasonlítást kell legalább végeznünk annak eldöntéséhez, hogy az $x$, $y$ elemek indexei elemei-e $A$-nak vagy sem?

Dolgozatunkban $A=\{r_1, r_2, \ldots, r_k\}$ $(1\leqq r_1<r_2<\ldots<r_k\leqq n)$ esetet vizsgáljuk és pontos alsó korlátot adunk az összehasonlítások számára $r_k=k$, illetve $r_k\neq k$, $r_k\leqq\left\lceil\frac{n}{2}\right\rceil$ esetekben.

## 1. Bevezetés

Legyen adott egy véges, $n$ számosságú teljesen rendezett $H$ halmaz, amelynek rendezését nem ismerjük. A $H$ halmaz elemeinek páronkénti összehasonlításával meg kívánjuk határozni a $H$ halmaz bizonyos elemét (elemeit), vagy a $H$ halmaz bizonyos elemeiről el kívánjuk dönteni, hogy rendelkeznek-e egy adott tulajdonsággal. Például megkívánjuk határozni a legnagyobb és az utána következő elemeket vagy tetszőleges elemről el kívánjuk dönteni, hogy — mondjuk csökkenő sorrendet tekintve — $i$-edik-e $H$-ban? Kérdés az, hogy ehhez legalább mennyi összehasonlítást kell végeznünk. Ismert, hogy a $H$ halmaz legnagyobb elemének meghatározásához legalább $n-1$ összehasonlítást kell végeznünk és az is, hogy ennél kevesebb összehasonlítással nem is érhetünk célba ([7]). Ha a legnagyobb és az utána következő elem kiválasztása a célunk, akkor ehhez legalább $n-2+\lceil\log_2 n\rceil$ összehasonlítás szükséges ([3], [4] 211—212. old.) és ha két tetszőleges egymás után következő elem meghatározása a célunk, akkor is legalább ennyi összehasonlítást kell végeznünk ([9], [8]). ($\lceil x\rceil$ ($\lfloor x\rfloor$) jelöli az $x$-nél nem kisebb (nagyobb) legkisebb (legnagyobb) egészet).

IRA POHL bizonyította először ([1]), hogy a $H$ halmaz legnagyobb és legkisebb elemének egyszerre való kiválasztásához legalább $n+\left\lceil\frac{n}{2}\right\rceil-2$ összehasonlítás szükséges (más bizonyítások [5] és [8]-ban).

KATONA G. O. H. vetette fel a következő problémát ([2]): Legyen

$$A=\{r_1, r_2, \ldots, r_k\} \qquad (1\leqq r_1<r_2<\ldots<r_k\leqq n)$$

és legyen $x$ a $H$ halmaz egy tetszőleges eleme. Kérdés az, hogy — mondjuk csökkenő sorrendet tekintve — az $x$ elem $r_i$-edik-e $H$-ban $(i\in\{1, 2, \ldots, k\})$? Bizonyította ([2]), hogy ennek eldöntéséhez legalább $n-1$ összehasonlítás szükséges (más bizonyítás [6] és [8]-ban).

KATONA G. O. H. vetette fel azt a problémát is, hogy ha $A=\{p, q\}$ és $x, y$ a $H$ halmaz két tetszőleges eleme, akkor mennyi összehasonlítást kell legalább végeznünk annak eldöntéséhez, hogy az $x, y$ elemek indexei elemei-e $A$-nak vagy sem?

KATONA az $1 \leqq p < q \leqq \left\lceil \frac{n}{2} \right\rceil$ esetben megadott egy olyan stratégiát ([2]), amely $n+q-3$ lépésben elvégzi feladatát és azt sejtette, hogy a stratégiája optimális. [6]-ban bizonyítást nyert, hogy a sejtése igaz. Ha $p=1$, $q=n$, akkor a szükséges összehasonlítások száma $n+\left\lceil \frac{n-1}{2} \right\rceil -2$ és ebből következik, hogy tetszőleges $p, q$-ra az $n+q-3$ nem áll ([5], [8]).

Nyitott kérdés, hogy mit mondhatunk tetszőleges $p, q$ esetben?

A következőkben azzal a problémával foglalkozunk, hogy ha $A=\{r_1, r_2, \ldots, r_k\}$ $(k<n$ és $1 \leqq r_1 < \ldots < r_k \leqq n)$ és $x, y$ két tetszőleges eleme $H$-nak, akkor legalább mennyi összehasonlítást kell végeznünk annak eldöntéséhez, hogy az $x, y$ elemek indexei — mondjuk csökkenő sorrendet tekintve — elemei-e $A$-nak? Bizonyítjuk, hogy $r_k=k$ esetben legalább $n-1$ és $r_k \neq k$ esetben, ha $r_k \leqq \left\lceil \frac{n}{2} \right\rceil$, akkor legalább $n+r_k-3$ összehasonlítást kell végeznünk.

## 2. Jelölések, definíciók

Legyen $x, y$ a $H$ halmaz két tetszőleges eleme és legyen $A=\{r_1, r_2, \ldots, r_k\}$ $(1 \leqq r_1 < r_2 < \ldots < r_k \leqq n)$. Tegyük fel, hogy a $H$ elemeinek páronkénti összehasonlításával el kívánjuk dönteni, hogy az $x, y$ elemek indexei — mondjuk csökkenő sorrendet tekintve — elemei-e $A$-nak.

Legyen az első összehasonlítandó elempár — mondjuk — $a, b$ és jelölje $S_0$ $(S_0=(a, b))$. Az $\varepsilon_1$ legyen 1 vagy 0 aszerint, hogy $a>b$ vagy $a<b$. Az $\varepsilon_1$ választól függően választunk egy $S_1(\varepsilon_1)$ párt, mondjuk $c(\varepsilon_1)$, $d(\varepsilon_1)$-et és $\varepsilon_2$-t 1-nek definiáljuk, ha $c(\varepsilon_1)>d(\varepsilon_1)$ és 0-nak, ha $c(\varepsilon_1)<d(\varepsilon_1)$.

Ugyanígy folytatva, bizonyos

$$\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{i-1}$$

sorozatokra ($\varepsilon_j=0$ vagy 1, $j=1, 2, \ldots, i-1$) megadjuk az

$$S_{i-1}(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{i-1}) \tag{2.1}$$

párt azzal a kikötéssel, hogy ha az

$$S_{i-1}(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{i-1})$$

definiálva van, akkor az

$$S_{i-2}(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{i-2})$$

is definiálva van. Az $\varepsilon_i$ értéke 1 vagy 0 aszerint, hogy a (2.1) pár első vagy második tagja a nagyobb.

A kérdéseknek így — a közbeeső válaszok függvényeként — megadott sorozatát annak eldöntésére alkalmas stratégiának nevezzük, hogy az $x, y$ elemek indexei

elemei-e $A$-nak (a továbbiakban egyszerűen csak stratégiának), ha minden $\varepsilon_1, \varepsilon_2, \ldots$ $\ldots, \varepsilon_l$ sorozat esetén, amikor

(2.2) $S_{l-1}(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{l-1})$ meg van határozva, de

(2.3) $S_l(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_l)$ már nincs, akkor az

(2.4) $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_l$ válaszok (az $S_0, \ldots, S_{l-1}(\varepsilon_1, \ldots, \varepsilon_{l-1})$ kérdésekkel együtt) egyértelműen választ adnak arra, hogy az $x, y$ elemek indexei elemei-e $A$-nak vagy sem.

A (2.2), (2.3), (2.4) feltételeket kielégítő $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_l$ sorozatra azt mondjuk, hogy erre a stratégia befejeződik. A maximális hosszúságú $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_l$ sorozat hosszát, amelyre a stratégia befejeződik a stratégia hosszának fogjuk nevezni. Jelölje a továbbiakban $\mathscr{S}(A, 2)$ a stratégiát és $L(\mathscr{S}(A, 2))$ a stratégia hosszát. Az $\mathscr{S}(A, 2)$ stratégia $(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_i)$ állapotán, az $S_{i-1}(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{i-1})$ kérdésekre adott válasz utáni helyzetet fogjuk érteni. A rövidebb írásmód kedvéért — ha az félreértésre nem ad okot — az $\varepsilon$-okat elhagyjuk.

Jelölje $T_i(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_i)$ azt az egyenlőtlenséget, amit az $S_{i-1}(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{i-1})$ párból készíthetünk az $\varepsilon_i$ válasz alapján.

A (2.4) feltételt ezek után úgy is mondhatjuk, hogy a

$$T_1(\varepsilon_1), T_2(\varepsilon_1, \varepsilon_2), \ldots, T_l(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_l) \tag{2.5}$$

egyenlőtlenségek alapján egyértelműen választ tudunk adni arra, hogy az $x, y$ elemek indexei elemei-e $A$-nak vagy sem.

A bizonyítások során gyakran hivatkozunk majd egy tetszőleges $(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_i)$ állapothoz tartozó

$$T_1(\varepsilon_1), \quad T_2(\varepsilon_1, \varepsilon_2), \ldots, T_i(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_i)$$

egyenlőtlenségrendszerre. Jelölje ezt a következőkben $\mathscr{E}_i$.

Gyakran lesz szükségünk azon egyenlőtlenségekre, amelyek $\mathscr{E}_i$-ből következnek. Kérdés tehát, hogy az $\mathscr{E}_i$ alapján mely egyenlőtlenségek adódnak?

Legyen $H = \{h_1, h_2, \ldots, h_n\}$ egy teljesen rendezett halmaz és legyen adott egy

$$x_i < x_j \quad (i \neq j,\ i, j \in \{1, 2, \ldots, n\}) \tag{2.6}$$

egyenlőtlenség rendszer. Keressük az egyenlőtlenségrendszer megoldását a $H$ halmazon!

Ha

$$x_{i_1} < x_{i_2},\ x_{i_2} < x_{i_3}, \ldots, x_{i_{k-1}} < x_{i_k} \quad (1 < k)$$

egyenlőtlenségek léteznek, akkor azt mondjuk, hogy az $x_{i_1} < x_{i_k}$ egyenlőtlenség tranzitive következik (2.6)-ból. Azon egyenlőtlenségeket, amelyek tranzitive következnek (2.6)-ból a (2.6) következményeinek nevezzük. A (2.6) ellentmondásmentes, ha $x_i < x_i$ nem következménye (2.6)-nak.

Igaz a következő, alapvető lemma:

2.1. LEMMA ([8], 2.1. tétel). Ha (2.6) ellentmondásmentes és $x_v < x_w$, $x_v > x_w$ nem következménye (2.6)-nak, akkor van olyan megoldása is $H$-ban, hogy $x_v > x_w$ és olyan is, hogy $x_v < x_w$.

A lemma bizonyításával itt nem foglalkozunk.

Az $\mathcal{S}(A, 2)$ stratégia tetszőleges $(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_i)$ állapotához egyértelműen rendelhetünk egy irányított gráfot a következőképpen:

Legyenek a gráf szögpontjai a $H$ halmaz elemei. Az $S_{i-1}(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{i-1}) = (c, d)$ kérdésnek feleljen meg a $c, d$ pontokat összekötő él és ha $c > d$, akkor irányítsuk az élet $c$-ből $d$-be és ha $c < d$, akkor $d$-ből $c$-be.

Jelölje az $\mathcal{S}(A, 2)$ stratégia $(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_i)$ állapotához így rendelt gráfot $\vec{G}^i$. A $\vec{G}^i$ — mivel a $H$ teljesen rendezett — nem tartalmaz irányított kört és ha $\mathcal{E}_i$-ből — mondjuk — $e < f$ következik, akkor van $\vec{G}^i$-ben irányított út $f$-ből $e$-be. Jelölje $G^i$ azt a gráfot, amit $\vec{G}^i$-ből az élek irányításának megszüntetésével kapunk.

## 3. Eredmény

3.1. TÉTEL.

$$\min_{\mathcal{S}(A,2)} L(\mathcal{S}(A, 2)) = n-1, \tag{3.1}$$

ha $|A| < n$ és $r_k = k$;

$$\min_{\mathcal{S}(A,2)} L(\mathcal{S}(A, 2)) = n + r_k - 3, \tag{3.2}$$

ha $r_k \neq k$ és $r_k \leqq \left\lceil \frac{n}{2} \right\rceil$.

*Bizonyítás.* Megadunk egy olyan stratégiát, amely a tétel feltételeit kielégíti. (Az alábbi stratégia lényegében KATONA $|A| = 2$ esetre adott stratégiája.)

Legyen $S_0(x, y)$. Feltehetjük, hogy $\varepsilon_1 = 1$. Hasonlítsuk össze $H - \{x, y\}$ elemeit rendre $y$-nal. Ha $r_k = k$ és $y$ ezen összehasonlításokban legalább $n - r_k$-szor volt nagyobb, akkor $x$ és $y$ indexe eleme $A$-nak és $n-1$ összehasonlítást végeztünk. Ha $r_k \neq k$ és $y$ $n - r_i$ $(r_i \in A)$ esetben nagyobb, akkor az $y$ $r_i$-edik eleme $H$-nak. Ha $r_i = i$, akkor — nyilván — $x$ indexe is eleme $A$-nak. Ha $r_i \neq i$, akkor az $y$-nál nagyobbnak adódott elemeket rendre $x$-szel összehasonlítva választ kapunk arra, hogy az $x$ elem indexe eleme-e $A$-nak vagy sem. Így legfeljebb $n + r_k - 3$ összehasonlítást végeztünk.

Ezzel igazoltuk, hogy

$$\min_{\mathcal{S}(A,2)} L(\mathcal{S}(A, 2)) \leqq n-1, \quad \text{ha} \quad r_k = k$$

és

$$\min_{\mathcal{S}(A,2)} L(\mathcal{S}(A, 2)) \leqq n + r_k - 3, \quad \text{ha} \quad r_k \neq k,\ r_k \leqq \left\lceil \frac{n}{2} \right\rceil.$$

Bizonyítjuk, hogy tetszőleges $\mathcal{S}(A, 2)$ stratégiára az

$$L(\mathcal{S}(A, 2)) \geqq n-1, \quad \text{ha} \quad r_k = k \tag{3.3}$$

és

$$L(\mathcal{S}(A, 2)) \geqq n + r_k - 3, \quad \text{ha} \quad r_k \neq k,\ r_k \leqq \left\lceil \frac{n}{2} \right\rceil \tag{3.4}$$

egyenlőtlenségek is teljesülnek.

A két esetet külön-külön vizsgáljuk.

*Tegyük fel először, hogy* $r_k = k$.

Legyen $\mathcal{S}(A, 2)$ egy tetszőleges olyan stratégia, amely az $x, y$ elemekről képes eldönteni, hogy indexei elemei-e $A$-nak vagy sem (csökkenő sorrendet tekintve). A következőkben megadjuk az $\mathcal{S}(A, 2)$ stratégia egy ágát és ennek a hosszáról fogjuk belátni, hogy $\geqq n-1$. Feltehetjük, hogy a definiálásra kerülő ágon $y$ nem szerepel előbb, mint $x$, mivel ha előbb szerepelne, akkor az $x, y$ szerepét felcseréljük. Az $\mathcal{S}(A, 2)$ stratégia egy tetszőleges $(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_i)$ állapotában a $H-\{x, y\}$ elemeit — rekurzív módon — $N^i, K^i, C^i$ részhalmazokba soroljuk és az $\varepsilon$-okat is rekurzíven definiáljuk.

Az $\mathcal{S}(A, 2)$ stratégia indításakor legyen $N^0, K^0=\emptyset, C^0=H-\{x, y\}$. Tetszőleges $(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_i)$ állapotban $N^i(K^i)$ elemei legyenek azok az elemek $(\neq x)$, amelyek $\mathcal{E}_i$ alapján $y$-nál nagyobbak (kisebbek) és legyen $C^i=H-(N^i \cup K^i \cup \{x, y\})$.

Tegyük fel, hogy az $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_i$ és az $N^i, K^i, C^i$ már definiálva van és legyen

$$S_i(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_i) = (g, h).$$

Megadjuk az $\varepsilon_{i+1}$ értékét és az $N^{i+1}, K^{i+1}, C^{i+1}$ halmazokat. A definiálásra kerülő ág létezése a 2.1. lemmából következik. Mivel az $\mathcal{E}_{i+1}$-ből az $N^{i+1}, K^{i+1}, C^{i+1}$ halmazok következnek, ezért ezek írásától a következőkben eltekintünk. Azon esetek felsorolását is mellőzzük, amelyek a szereplőkből a $g, h$ elemek szerepének felcserélésével adódnak.

1. $g, h \in C^i$; $g, h \in N^i$; $g, h \in K^i$ — $\varepsilon_{i+1}$ értéke tetszőleges, de olyan, hogy $\mathcal{E}_i$-vel ne kerüljünk ellentmondásba.
2. $g \in N^i$, $h \in C^i \cup K^i$ — $\varepsilon_{i+1}=1$
3. $g \in C^i$, $h \in K^i$ — $\varepsilon_{i+1}=1$
4. $g=y$, $h \in N^i(K^i)$ — $\varepsilon_{i+1}=0(1)$
5. $g=y$, $h \in C^i$ — $\varepsilon_{i+1}=0$, ha $|N^i|+|C_1^i| \leqq k-3$, ahol $C_1^i$ elemei azon $C^i$-beli elemek halmaza, amelyek $\mathcal{E}_i$ alapján $h$-nál nagyobbak.
   $\varepsilon_{i+1}=1$, ha $|N^i|+|C_1^i|>k-3$.
6. $g=x$, $h \neq x$ — $\varepsilon_{i+1}=1$.

Ezzel az $\varepsilon_{i+1}$ értékét és az $N^{i+1}, K^{i+1}, C^{i+1}$ halmazokat definiáltuk. Tegyük fel, hogy a definiált ágon az $\mathcal{S}(A, 2)$ stratégia az $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_l$ válaszsorozatra befejeződött. Jelölje ezt az ágat $P(A, 2)$, és az ág hosszát $|P(A, 2)|$. Igazoljuk, hogy $|P(A, 2)| \geqq n-1$.

*Igazoljuk először, hogy* $|N^l| \leqq k-2$.

Az állítást indirekt bizonyítjuk. Tegyük fel, hogy $|N^l|>k-2$. Legyen $(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_i)$ az állapot, amelyre $|N^i| \leqq k-2$, de $|N^{i+1}|>k-2$. Ilyen állapot a feltevésünk szerint nyilván létezik.

Legyen

$$S_i(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_i) = (a, b).$$

A $P(A, 2)$ definícióból következik, hogy $a=y$ vagy $b=y$.

Tegyük fel, hogy $a=y$. Az $|N^i| \leqq k-2$ és $|N^{i+1}|>k-2$ feltevésből következik, hogy $b \in C^i$ és $\varepsilon_{i+1}=0$. Viszont 5. szerint $\varepsilon_{i+1}=1$. Az ellentmondás igazolja állításunkat.

*Igazoljuk, hogy* $|K^l| \leqq n-k$.

Ezt az állítást is indirekt bizonyítjuk.

Tegyük fel, hogy $|K^l|>n-k$. Legyen $(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_j)$ az az állapot, amelyre $|K^j|\leqq n-k$, de $|K^{j+1}|>n-k$. Legyen

$$S_j(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_j) = (c, d).$$

A $P(A, 2)$ definíciójából következik, hogy $c=y$ vagy $d=y$. Tegyük fel, hogy $c=y$. Könnyen látható, hogy $d\in C^j$ és így 5. kerül alkalmazásra. Mivel feltevésünk szerint $\varepsilon_{j+1}=1$, ezért $|N^j|+|C_1^j|>k-3$ és ha $C_2^j$ jelöli a $d$-nél $\mathscr{E}_j$ alapján kisebb $C^j$-beli elemek halmazát, akkor $|K^j|+|C_2^j|+1>n-k$.

Így

$$|C_1^j|+|N^j| \geqq k-2$$
$$|C_2^j|+|K^j| \geqq n-k$$

és a két egyenlőtlenség megfelelő oldalait összeadva

$$|C_1^j|+|C_2^j|+|N^j|+|K^j| \geqq n-2$$

adódik. Viszont

$$C_1^j \cup C_2^j \cup N^j \cup K^j \subseteq H-\{z, y, d\},$$

azaz

$$|C_1^j|+|C_2^j|+|N^j|+|K^j| \leqq n-3.$$

Az ellentmondás igazolja állításunkat.

*Igazoljuk, hogy* $|K^l|=n-k$.

Ezt az állítást is indirekt bizonyítjuk. Tegyük fel, hogy $|K^l|\neq n-k$. Az előzőekből következik, hogy akkor $|K^l|<n-k$. Mivel $|N^l|\leqq k-2$, ezért

$$|K^l|+|N^l| < n-2,$$

azaz $C^l\neq\emptyset$.

Legyen $a\in C^l$. A $P(A, 2)$ definíciójából következik, hogy sem $a>y$ sem $a<y$ nem következik $\mathscr{E}_l$-ből. Ezek szerint $C^l$ elemei nagyobbak és kisebbek is lehetnek $y$-nál. Ha $C^l$ elemei $y$-nál kisebbek, akkor $y$ indexe eleme $A$-nak és ha $C^l$ elemei $y$-nál nagyobbak, akkor $y$ indexe nem eleme $A$-nak. Ez azt jelenti, hogy a stratégia nem fejeződött be.

Az ellentmondás igazolja állításunkat.

Ezek szerint, ha az $\mathscr{S}(A, 2)$ stratégia az $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_l$ válaszsorozatra befejeződik, akkor $|K^l|=n-k$, $|N^l|\leqq k-2$. Tekintsük azt az állapotot, amelyre $|K^i|<n-k$, $|K^{i+1}|=n-k$ teljesül. Legyen

$$S_i(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_i) = (a, b).$$

Az $a=y$ vagy $b=y$ $P(A, 2)$ definíciója miatt. Legyen — mondjuk — $a=y$. Ekkor $b\in C^i$ és 5. kerül alkalmazásra. Mivel $\varepsilon_{i+1}=1$, ezért

$$|N^i|+|C_1^i| \geqq k-2,$$
$$|K^i|+|C_2^i| = n-k-1,$$

azaz

$$|N^i|+|C_1^i|+|K^i|+|C_2^i| \geqq n-3.$$

Viszont

$$N^i \cup C_1^i \cup K^i \cup C_2^i \subseteq H-\{x, y, b\}$$

és így

$$|N^i|+|C_1^i|+|K^i|+|C_2^i| \leqq n-3$$

is teljesül, azaz

$$|N^i|+|C_1^i|+|K^i|+|C_2^i| = n-3.$$

*Bizonyítjuk, hogy a* $G^{i+1}$ *gráf összefüggő.*

A $P(A,2)$ definíciója szerint minden $N^i(K^i)$ elemből (elemhez) vezet irányított út $\bar{G}^i$-ben $y$-hoz ($y$-ból) és a $C_1^i(C_2^i)$-beli elemekből (elemekhez) $b$-hez ($b$-ből) és $y>b\in\mathscr{E}_{i+1}$. Ebből következik, hogy a $H-\{x\}$ elemek összefüggő gráfot feszítenek.

Az $x$ is szerepel legalább egy egyenlőtlenségben, mivel feltettük, hogy az $y$ nem szerepel előbb, mint $x$. Ezekből pedig az állítás következik. Mivel a $G^{i+1}$ összefüggő így legalább $n-1$ élet tartalmaz, azaz $\mathscr{E}_{i+1}$-ben van legalább $n-1$ egyenlőtlenség. Ebből következik, hogy $\mathscr{E}_l$ is tartalmaz legalább ennyi egyenlőtlenséget. Ezzel igazoltuk, hogy

$$L(\mathscr{S}(A,2)) \geqq n-1$$

teljesül tetszőleges $\mathscr{S}(A,2)$ stratégiára, ha $r_k=k$. Ezzel (3.3)-at bizonyítottuk.

Igazoljuk (3.4)-et.

*Tegyük fel, hogy* $r_k\neq k$, $r_k\leqq\left\lceil\frac{n}{2}\right\rceil$.

Legyen $\mathscr{S}(A,2)$ egy tetszőleges stratégia, amely az $x,y$ elemekről képes eldönteni, hogy indexei-e $A$-nak vagy sem. Feltehetjük, hogy $k>2$, mivel a $k=2$ eset elintézett ([6]).

Két esetet fogunk megkülönböztetni:

I. Van olyan $r_{k-i}$ $(i\in\{1,2,\ldots,k-1\})$, hogy

$$r_{k-i}\neq r_k-i.$$

II. Minden $i$-re $r_{k-i}=r_k-i$ $(i\in\{1,2,\ldots,k-1\})$.

Legyen $r_k=q$.

Az I. esetben legyen $p$ értéke $r_{k-i}$, ha $r_{k-i}\neq r_k-i$, de $r_{k-s}=r_k-s$, ha $s<i$.

A II. esetben legyen $p$ értéke $r_1$. (Az $r_1\neq 1$ az $r_k\neq k$ feltevésből következik.)

Ezek szerint

$$1\leqq p<q\leqq\left\lceil\frac{n}{2}\right\rceil$$

egyenlőtlenségsorozat teljesül.

A következőkben megadjuk az $\mathscr{S}(A,2)$ stratégia egy ágát és ennek az ágnak a hosszát fogjuk becsülni alulról.

Most is feltehetjük, hogy a definiálásra kerülő ágon $y$ nem szerepel előbb, mint $x$. Az ágat rekurzív módon definiáljuk és a $H-\{x,y\}$ halmaz elemeit — szintén rekurzíven — $B_1, B_2, B_3, C$ részhalmazokba soroljuk.

A stratégia indításakor $B_1^0, B_2^0, B_3^0=\emptyset$, $C^0=H-\{x,y\}$. A definiálásra kerülő halmazok heurisztikus jelentése a következő:

$C^i$ elemeiről nincs információnk, még nem szerepeltek.

$B_1^i$ elemei a stratégia befejezésekor az $x,y$ elemeknél nagyobbak lesznek.

$B_2^i$ elemei a stratégia befejezésekor $x$-nél ($y$-nál) kisebbek, $y$-nál ($x$-nél) nagyobbak lesznek.

$B_3^i$ elemei a stratégia befejezésekor az $x,y$ elemeknél kisebbek lesznek.

Tegyük fel, hogy az $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_i, B_1^i, B_2^i, B_3^i, C^i$ már definiálva van és legyen

$$S_i(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_i) = (g, h).$$

Megadjuk az $\varepsilon_{i+1}$ értékét és a $B_1^{i+1}, B_2^{i+1}, B_3^{i+1}, C^{i+1}$ halmazokat. Azon esetek felsorolásától eltekintünk, amelyek a szereplőkből az elemek felcserélésével adódnak és azon halmazokat soroljuk csak fel, amelybe valamely elem kerül, mivel nyilvánvaló lesz, hogy mely halmazból jutott oda, a többi pedig nem változik.

1. $g, h \in C^i$ $\varepsilon_{i+1}=1$.
   *Ha* $|B_3^i|<n-q$, akkor $h \in B_3^{i+1}$ és
   *I. esetben,*
   ha $|B_1^i|<p-1$, akkor $g \in B_1^{i+1}$;
   ha $|B_1^i|=p-1$, $|B_2^i|<q-1$,
   akkor $g \in B_2^{i+1}$;
   ha $|B_1^i|=p-1$, $|B_2^i|=q-p-1$,
   akkor $g \in B_3^{i+1}$;
   *II. esetben*
   ha $|B_2^i|<q-p-1$, akkor $g \in B_2^{i+1}$;
   ha $|B_2^i|=q-p-1$, $|B_1^i|<p-1$,
   akkor $g \in B_1^{i+1}$;
   ha $|B_2^i|=q-p-1$, $|B_1^i|=p-1$,
   akkor $g \in B_3^{i+1}$.
   *Ha* $|B_3^i|=n-q$, *akkor*
   *I. esetben*
   ha $|B_1^i|<p-2$, akkor $g, h \in B_1^{i+1}$;
   ha $|B_1^i|=p-2$, akkor $g \in B_1^{i+1}$, $h \in B_2^{i+1}$;
   ha $|B_1^i|=p-1$, akkor $g, h \in B_2^{i+1}$;
   *II. esetben*
   ha $|B_2^i|<q-p-2$, akkor $g, h \in B_2^{i+1}$;
   ha $|B_2^i|=q-p-2$, akkor $h \in B_2^{i+1}$, $g \in B_1^{i+1}$;
   ha $|B_2^i|=q-p-1$, akkor $g, h \in B_1^{i+1}$.

2. $g \in B_r^i$, $h \in B_s^i$ $(r<s)$, akkor $\varepsilon_{i+1}=1$.

3. $g \in B_1^i$, $h \in C^i$ $\varepsilon_{i+1}=1$ és
   ha $|B_3^i|<n-q$, akkor $h \in B_3^{i+1}$;
   ha $|B_3^i|=n-q$, akkor
   *I. esetben*
   ha $|B_1^i|<p-1$, akkor $h \in B_1^{i+1}$;
   ha $|B_1^i|=p-1$, akkor $h \in B_2^{i+1}$;
   *II. esetben*
   ha $|B_2^i|<q-p-1$, akkor $h \in B_2^{i+1}$;
   ha $|B_2^i|=q-p-1$, akkor $h \in B_1^{i+1}$.

4. $g \in B_2^i$, $h \in C^i$, ha $|B_3^i|<n-q$, akkor $\varepsilon_{i+1}=1$, $h \in B_3^{i+1}$.
   Ha $|B_3^i|=n-q$, akkor
   *I. esetben*
   ha $|B_1^i|<p-1$, akkor $\varepsilon_{i+1}=0$, $h \in B_1^{i+1}$;
   ha $|B_1^i|=p-1$, akkor $\varepsilon_{i+1}=1$, $h \in B_2^{i+1}$;
   *II. esetben*
   ha $|B_2^i|<q-p-1$, akkor $\varepsilon_{i+1}=1$, $h \in B_2^{i+1}$;
   ha $|B_2^i|=q-p-1$, akkor $\varepsilon_{i+1}=0$, $h \in B_1^{i+1}$.

5. $g\in B_3^i$, $h\in C^i$ $\varepsilon_{i+1}=0$ és
*I. esetben*
ha $|B_1^i|<p-1$, akkor $h\in B_1^{i+1}$;
ha $|B_1^i|=p-1$, $|B_2^i|<q-p-1$,
akkor $h\in B_2^{i+1}$;
ha $|B_1^i|=p-1$, $|B_2^i|=q-p-1$,
akkor $h\in B_3^{i+1}$.
*II. esetben*
ha $|B_2^i|<q-p-1$, akkor $h\in B_2^{i+1}$;
ha $|B_2^i|=q-p-1$, $|B_1^i|<p-1$,
akkor $h\in B_1^{i+1}$;
ha $|B_2^i|=q-p-1$, $|B_1^i|=p-1$,
akkor $h\in B_3^{i+1}$.
6. $g=x$, $h=y$, ha $x$, $y$ még nem szerepelt, akkor $\varepsilon_{i+1}=1$.
Ha $x$ már szerepelt és először nagyobb (kisebb) volt, akkor $\varepsilon_{i+1}=1(0)$.
7. $g=x$, $h\in C^i$, ha $x$ még nem szerepelt, akkor (feltevésünk szerint $y$ sem),
ha $|B_3^i|<n-q$, akkor $\varepsilon_{i+1}=1$, $h\in B_3^{i+1}$;
ha $|B_3^i|=n-q$, akkor
*I. esetben*
Ha $|B_1^i|<p-1$, akkor $\varepsilon_{i+1}=0$, $h\in B_1^{i+1}$;
ha $|B_1^i|=p-1$, akkor $\varepsilon_{i+1}=0$, $h\in B_2^{i+1}$;
*II. esetben*
Ha $|B_2^i|<q-p-1$, akkor $\varepsilon_{i+1}=0$, $h\in B_2^{i+1}$;
ha $|B_2^i|=q-p-1$, akkor $\varepsilon_{i+1}=0$, $h\in B_1^{i+1}$.
Ha $x$ már szerepelt és az első szereplésekor nagyobb (kisebb) volt, akkor az
*I. esetben*
ha $|B_3^i|<n-q$ ($|B_1^i|<p-1$), akkor $\varepsilon_{i+1}=1(0)$, $h\in B_3^{i+1}(h\in B_1^{i+1})$;
ha $|B_3^i|=n-q$ ($|B_1^i|=p-1$), akkor
ha $|B_1^i|<p-1$ ($|B_2^i|<q-p-1$), akkor $\varepsilon_{i+1}=0(0)$, $h\in B_1^{i+1}(h\in B_2^{i+1})$;
ha $|B_3^i|=n-q(|B_1^i|=p-1)$, $|B_1^i|=p-1$ ($|B_2^i|=q-p-1$), akkor $\varepsilon_{i+1}1(1)$, $h\in B_2^{i+1}$ ($h\in B_3^{i+1}$);
*II. esetben*
ha $|B_3^i|<n-q$ ($|B_2^i|<q-p-1$), akkor $\varepsilon_{i+1}1(0)$, $h\in B_3^{i+1}$ ($h\in B_2^{i+1}$);
ha $|B_3^i|=n-q$ ($|B_2^i|=q-p-1$), $|B_2^i|<n-q-1$
($|B_1^i|<p-1$), akkor $\varepsilon_{i+1}=1(0)$, $h\in B_2^{i+1}$ ($h\in B_1^{i+1}$);
ha $|B_3^i|=n-q$ ($|B_2^i|=q-p-1$), $|B_2^i|=q-p-1$
($|B_1^i|=p-1$), akkor $\varepsilon_{i+1}=0(1)$, $h\in B_1^{i+1}$ ($h\in B_3^{i+1}$).
8. $g=y$, $h\in C^i$, mivel feltettük, hogy $y$ nem szerepel előbb, mint $x$, ezért $x$ már szerepelt. Ha $x$ első szereplésekor nagyobb (kisebb) volt, akkor
*I. esetben*
ha $|B_1^i|<p-1$ ($|B_3^i|<n-q$), akkor $\varepsilon_{i+1}=0(1)$, $h\in B_1^{i+1}(h\in B_3^{i+1})$;
ha $|B_1^i|=p-1$, $|B_2^i|<q-p-1$ ($|B_3^i|=n-q$, $|B_1^i|<p-1$), akkor
$\varepsilon_{i+1}=0$, $h\in B_2^{i+1}$ ($\varepsilon_{i+1}=0$, $h\in B_1^{i+1}$);
ha $|B_1^i|=p-1$, $|B_2^i|=q-p-1$ ($|B_3^i|=n-q$, $|B_1^i|=p-1$),
akkor $\varepsilon_{i+1}=1$, $h\in B_3^{i+1}$ ($\varepsilon_{i+1}=1$, $h\in B_2^{i+1}$);
*II. esetben*
ha $|B_2^i|<q-p-1$ ($|B_3^i|<n-q$), akkor
$\varepsilon_{i+1}=0$, $h\in B_2^{i+1}$ ($\varepsilon_{i+1}=1$, $h\in B_3^{i+1}$);

ha $|B_2^i|=q-p-1$, $|B_1^i|<p-1$ ($|B_3^i|=n-q$, $|B_2^i|<q-p-1$), akkor $\varepsilon_{i+1}=0$, $h\in B_1^{i+1}$ ($\varepsilon_{i+1}=1$, $h\in B_2^{i+1}$);
ha $|B_2^i|=q-p-1$, $B_1^i=p-1$ ($|B_3^i|=n-q$, $|B_2^i|=q-p-1$), akkor $\varepsilon_{i+1}=1$, $h\in B_3^{i+1}$ ($\varepsilon_{i+1}=0$, $h\in B_1^{i+1}$).

9. $g\in\{x,y\}$, $h\in B_1^i(B_3^i)$ $\varepsilon_{i+1}=1(0)$.
10. $g=x$, $h\in B_2^i$, ha $x$ még nem szerepelt (feltevésünk szerint akkor $y$ sem) $\varepsilon_{i+1}=0$. Ha $x$ már szerepelt és első szereplésekor nagyobb (kisebb) volt, akkor $\varepsilon_{i+1}=1(0)$.
11. $g=y$, $h\in B_2^{i+1}$, ha $x$ első szereplésekor nagyobb (kisebb) volt — feltevésünk szerint $x$ már szerepelt —, akkor $\varepsilon_{i+1}=0(1)$.
12. $g,h\in B_r^i$ ($r\in\{1,2,3\}$) $\varepsilon_{i+1}$ értéke tetszőleges, de olyan, hogy $\mathscr{E}_i$-vel ne kerüljünk ellentmondásba.

Ezzel az $\varepsilon_{i+1}$ értékét és a $C^{i+1}$, $B_1^{i+1}$, $B_2^{i+1}$, $B_3^{i+1}$ halmazokat definiáltuk. A fentiekben definiált ág létezését a 2.1. lemma biztosítja. Jelölje az $\mathscr{S}(A,2)$ stratégia egy így meghatározott ágát $P(A,2)$ és az ág hosszát $|P(A,2)|$. Bizonyítjuk, hogy

$$l=|P(A,2)|\geqq n+q-3.$$

3.1. LEMMA. Ha valamely elem — mondjuk $b$ — az $\varepsilon_i$ válasz eredményeként $B_r^i$ halmazba kerül ($r\in\{1,2,3\}$), akkor $b\in B_r^j$, ha $j>i$.
$C^i$-beli elemek még nem szerepeltek.

*Bizonyítás.* A lemma állításai a $P(A,2)$ ág definíciójának egyszerű következményei.

3.2. LEMMA. Az I. esetben, ha $|B_1^i|<p-1$, akkor $|B_2^i|=0$ és az II. esetben, ha $|B_2^i|<q-p-1$, akkor $|B_1^i|=0$.

*Bizonyítás.* Ennek a lemmának az állításai is a $P(A,2)$ ág definíciójának egyszerű következményei.

3.3. LEMMA. Ha $a>b\in\mathscr{E}_i$, $a\in B_r^i$, $b\in B_s^i$, akkor $s\geqq r$ ($r,s\in\{1,2,3\}$).

*Bizonyítás.* Tegyük fel, hogy $a>b\in\mathscr{E}_i$. Legyen $S_j(\varepsilon_1,\varepsilon_2,\ldots,\varepsilon_j)$ az a pár, amelyben $a$, $b$ szerepel; $S_j(\varepsilon_1,\varepsilon_2,\ldots,\varepsilon_j)=(a,b)$.
Ha $a,b\notin C^j$, akkor 2. vagy 12. kerülhet alkalmazásra, és a 3.1. lemma alapján a lemma állítása teljesül. Ha $a,b\in C^j$, akkor 1. kerül alkalmazásra és — könnyen látható — a lemma állítása teljesül. Ha $a$, $b$ közül valamelyik eleme $C^j$-nek — mondjuk $a\in C^j$ — és a másik nem ($b\notin C^j$), akkor 3., 4., 5. kerülhet alkalmazásra és — könnyen látható — hogy a lemma állítása teljesül.
Ezzel a lemma bizonyítását befejeztük.

3.4. LEMMA. Ha $x$ első szereplésekor nagyobb (kisebb) volt és $x>a\in\mathscr{E}_i$, akkor $a\in B_2^i\cup\{y\}\cup B_3^i$ ($a\in B_3^i$) és ha $x<a\in\mathscr{E}_i$, akkor $a\in B_1^i(a\in B_1^i\cup B_2^i\cup\{y\})$. Ha $x$ első szereplésekor nagyobb (kisebb) volt és $y>a\in\mathscr{E}_i$, akkor $a\in B_3^i$ ($a\in B_3^i\cup B_2^i\cup\{x\}$) és ha $y<a\in\mathscr{E}_i$, akkor $a\in B_2^i\cup B_1^i\cup\{x\}$ ($a\in B_1^i$).

*Bizonyítás.* Tegyük fel, hogy $x$ első szereplésekor nagyobb volt — legyen mondjuk $x>e$ — és $x>a\in\mathscr{E}_i$. Legyen $S_j(\varepsilon_1,\varepsilon_2,\ldots,\varepsilon_j)$ az a pár, amelyben $x$, $a$ szerepelt:

$$S_j(\varepsilon_1,\varepsilon_2,\ldots,\varepsilon_j)=(x,a).$$

Két esetet különböztetünk meg aszerint, hogy $e=a$ vagy $e\neq a$.

Az $e=a$ esetben, ha $a=y$, akkor 6. kerül alkalmazásra és $x>y$. Ha $a\neq y$, akkor — mivel $x>a\in\mathscr{E}_i$ — 7., 9. vagy 10. kerülhet alkalmazásra és a 7., 9., 10. definíciója alapján $a\in B_2^{i+1}\cup B_3^{i+1}$, amiből — 3.1. lemma szerint — $a\in B_2^i\cup B_3^i$, azaz a lemma állítása teljesül.

Az $e\neq a$ esetben, ha $a=y$, akkor 6. kerül alkalmazásra és $x>y$. Ha $a\neq y$, akkor — mivel $x>a\in\mathscr{E}_i$ — 7., 9. vagy 10. kerülhet alkalmazásra és az előzőekben látottakhoz hasonlóan adódik, hogy a lemma állítása teljesül.

Ezzel igazoltuk, hogy ha $x$ első szereplésekor nagyobb volt és $x>a\in\mathscr{E}_i$, akkor $a\in B_2^i\cup\{y\}\cup B_3^i$. Hasonlóan egyszerűen látható be a lemma többi állítása. Ezzel a lemma bizonyítását befejeztük.

3.5. LEMMA. Ha $x$ első szereplésekor nagyobb (kisebb) volt és ha

(3.5) $x<a$ levezethető $\mathscr{E}_i$-ből, akkor $a\in B_1^i$ $(a\in B_1^i\cup\{y\}\cup B_2^i)$;

(3.6) ha $x>a$ levezethető $\mathscr{E}_i$-ből, akkor $a\in B_3^i\cup\{y\}\cup B_2^i$ $(a\in B_3^i)$;

(3.7) ha $y>a$ levezethető $\mathscr{E}_i$-ből, akkor $a\in B_3^i$ $(a\in B_3^i\cup\{x\}\cup B_2^i)$;

(3.8) ha $y<a$ levezethető $\mathscr{E}_i$-ből, akkor $a\in B_2^i\cup\{x\}\cup B_1^i$ $(a\in B_1^i)$.

*Bizonyítás.* Tegyük fel először, hogy $x$ első szereplésekor nagyobb volt és $x<a$ levezethető $\mathscr{E}_i$-ből. Bizonyítjuk, hogy $a\in B_1^i$.

Ha $x<a$ levezethető $\mathscr{E}_i$-ből, akkor léteznek az

$$x=a_0<a_1, a_1<a_2, \ldots, a_{k-1}<a_k=a$$

egyenlőtlenségek $\mathscr{E}_i$-ben a 2.1. lemma alapján. A 3.4. lemma szerint $a_1\in B_1^i$ és a 3.3. lemma szerint $a_2, a_3, \ldots, a_k\in B_1^i$, azaz $a\in B_1^i$. Ezzel (3.5) első felét igazoltuk.

Hasonlóan egyszerűen igazolható (3.5) másik fele és a lemma többi állítása. Ezzel a lemma bizonyítását befejeztük.

3.6. LEMMA. Ha az $\mathscr{S}(A,2)$ stratégia a $P(A,2)$ ágon az $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_l$ válaszsorozatra befejeződik, akkor $|B_1^l|=p-1$, $|B_2^l|=q-p-1$, $|B_3^l|=n-q$, $|C^l|=0$.

*Bizonyítás.* Tegyük fel, hogy az $\mathscr{S}(A,2)$ stratégia az $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_l$ válaszsorozatra befejeződött és — mondjuk — $|B_3^l|\neq n-q$. A $P(A,2)$ definíciója szerint az $\mathscr{S}(A,2)$ stratégia tetszőleges $(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_i)$ állapotában a $|B_1^i|\leqq p-1$, $|B_2^i|\leqq q-p-1$, $|B_3^i|\leqq n-q$ egyenlőtlenségek teljesülnek. Ezek szerint, ha $|B_3^l|\neq n-q$, akkor $|B_3^l|<n-q$ és $C^l\neq\emptyset$.

A 2.1. lemma szerint $C^l$ elemei bármely elemnél nagyobbak és kisebbek is lehetnek. Mivel $q\in A$, $p\in A$, de $q+1\notin A$ és az I. esetben $p+1\notin A$, II. esetben $p-1\notin A$, ezért ha $x$ első szereplésekor nagyobb volt, akkor a 3.5. lemma szerint $y$ még lehet a $q$-adik is és nem is, hasonlóan $x$ lehet a $p$-edik is meg nem is, azaz a stratégia még nem fejeződhetett be. Hasonlóan látható be az az eset, amikor $x$ első szereplésekor kisebb volt, valamint a lemma többi állítása.

Ezzel a lemma bizonyítását befejeztük.

3.7. LEMMA. Ha az $\mathscr{S}(A,2)$ stratégia a $P(A,2)$ ágon az $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_l$ válaszsorozatra befejeződött, akkor ha $x$ első szereplésekor nagyobb (kisebb) volt, akkor $x$ ($y$) a $p$-edik, $y$ ($x$) a $q$-adik eleme $H$-nak.

*Bizonyítás.* A lemma állításai a $P(A, 2)$ ág definíciója és a 3.4., 3.5., 3.6. lemmákból egyszerűen következik.

A 3.6. lemma szerint az $(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_l)$ állapotban a $H=\{x, y\}$ elemei $B_1^l, B_2^l, B_3^l$ elemei. Legyen $(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_i)$ az az állapot, amelyre $|B_1^r \cup B_2^r| < q-2$, $|B_3^r| < n-q$, ha $r<i$ és $|B_1^i \cup B_2^i| = q-2$ vagy $|B_3^i| = n-q$. Ilyen állapot az előzőek szerint létezik.

Két esetet különböztetünk meg:

$\alpha$) $$|B_1^i \cup B_2^i| = q-2$$

$\beta$) $$|B_3^i| = n-q.$$

Igazoljuk a következőt:

3.8. LEMMA. Ha $x$ első szerep'ésekor nagyobb (kisebb) volt és az $\alpha$ eset következett be, akkor minden $B_1^i \cup B_2^i$-beli elem szerepel $\mathscr{E}_i$-ben $B_3^i \cup \{y\}$ $(B_3^i \cup \{x\})$-beli elemel. Ha az $x$ első szereplésekor nagyobb (kisebb) volt és a $\beta$ eset következett be, akkor minden $B_3$-beli elem szerepel $\mathscr{E}_i$-ben $B_1^i \cup B_2^i \cup \{x\}$ $(B_1^i \cup B_2^i \cup \{y\})$-beli elemmel.

*Bizonyítás.* Tegyük fel először, hogy az $\alpha$ eset következett be és $x$ első szereplésekor nagyobb volt. Legyen $a \in B_1^i \cup B_2^i$. Igazoljuk, hogy $a$ szerepel $\mathscr{E}_i$-ben $B_3^i \cup \{y\}$-beli elemmel.

Tekintsük $a$ első szereplését. Legyen — mondjuk —

$$S_j(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_j) = (a, b)$$

és $i>j$, $a \in C^j$. Ha $b \in C^j$, akkor 1. kerül alkalmazásra és mivel $|B_3^j| < n-q$, ezért $b \in B_3^{j+1}$, $b \in B_3^i$ és $a > b \in \mathscr{E}_i$. Ha $b \in B_3^j$, akkor 5. kerül alkalmazásra és $b \in B_3^i$, $a > b \in \mathscr{E}_i$.

A $b \notin B_1^i \cup B_2^i \cup \{x, y\}$ mivel akkor — könnyen látható — $a \notin B_1^i \cup B_2^i$ ellentmondáshoz jutunk. Ezzel igazoltuk, hogy ha $x$ első szereplésekor nagyobb volt és $\alpha$ következett be, akkor minden $B_1^i \cup B_2^i$-beli elem szerepel $B_3^i \cup \{y\}$-beli elemmel $\mathscr{E}_i$-ben. Hasonlóan egyszerűen igazolható a lemma többi állítása. Ezzel a lemma bizonyítását befejeztük.

Rátérünk ezek után (3.5) bizonyítására.

Tegyük fel, hogy az $\mathscr{S}(A, 2)$ stratégia a $P(A, 2)$ ágon az $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_l$ válaszsorozatra befejeződött. Nyilvánvaló, hogy az $\alpha$ vagy a $\beta$ eset bekövetkezett.

A két esetet külön-külön vizsgáljuk.

Tegyük fel először, hogy az $\alpha$ eset következett be és $x$ az első szereplésekor nagyobb (kisebb) volt. Ha $(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_i)$ azaz állapot, amelyre $|B_1^i \cup B_2^i| = q-2$, de $|B_1^r \cup B_2^r| < q-2$, $|B_3^r| < n-q$, ha $r<i$, akkor a 3.7. lemma szerint minden $B_1^i \cup B_2^i$-beli elem szerepel $B_3^i \cup \{y\}$ $(B_3^i \cup \{x\})$-beli elemmel $\mathscr{E}_i$-ben, és így a 2.1. lemma szerint $\mathscr{E}_i$ azon $a>b$ egyenlőtlenségeinek száma, amelyre $a \in B_1^i \cup B_2^i$, $b \in B_3^i \cup \{y\}$ $(b \in B_3^i \cup \{x\})$ legalább $q-2$. A 2.1. lemma szerint minden $B_3^i$-beli elemről $\mathscr{E}_i$ alapján be tudjuk látni, hogy $y$-nál ($x$-nél) kisebbek, ezért $G^i$-ben a $B_3^i \cup \{y\}$ $(B_3^i \cup \{x\})$ elemek által feszített gráf összefüggő és így legalább $|B_3^i|$ élet tartalmaz. Ez azt jelenti, hogy a $B_3^i \cup \{y\}$ $(|B_3^i| \cup \{x\})$ elemei közötti egyenlőtlenségek száma $\mathscr{E}_i$-ben legalább $n-q$.

Hasonlóan látható be, hogy a $B_1^i \cup B_2^i \cup \{x\}$ $(B_1^i \cup B_2^i \cup \{y\})$-beli elemek által feszített gráf is összefüggő $G^i$-ben és így a $B_1^i \cup B_2^i \cup \{x\}$ $(B_1^i \cup B_2^i \cup \{y\})$-beli elemek

közötti egyenlőtlenségek száma $\mathscr{E}_l$-ben legalább $q-2$. Az eddig figyelembe vett egyenlőtlenségeket összeadva adódik, hogy $\mathscr{E}_l$-ben van legalább $n-q-4$ egyenlőtlenség.

Igazoljuk, hogy ha az $x$ első szereplésekor nagyobb volt és az $\alpha$ eset következett be, akkor $x$ szerepel $B_3^l \cup \{y\}$-beli elemmel. Tegyük fel, hogy $x$ az $S_j(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_j)$ kérdésben szerepelt először.

Legyen — mondjuk —

$$S_j(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_j) = (x, d).$$

A $P(A, 2)$ definíciója szerint 6., 7., 9. vagy 10. kerülhet alkalmazásra és — könnyen látható — $d \in B_3^{j+1} \cup \{y\}$, $d \in B_3^l \cup \{y\}$, $x > d \in \mathscr{E}_l$. Ezt az egyenlőtlenséget is figyelembe véve

$$l \geqq n-q-3$$

adódik. Ebben az esetben tehát (3.4) teljesül. Tegyük fel, hogy $x$ első szereplésekor kisebb volt. A 2.7. lemma szerint $x$ a $q$-adik, $y$ a $p$-edik eleme $H$-nak. Ha $y$ szerepel $B_3^l \cup \{x\}$-beli elemmel, akkor — az előzőket figyelembe véve —

$$l \geqq n+q-3,$$

Tegyük fel, hogy $y$ nem szerepel $B_3^l \cup \{x\}$-beli elemmel $\mathscr{E}_l$-ben.

Tekintsük $x$ első szereplését. Legyen — mondjuk —

$$S_j(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_i) = (x, c).$$

Tegyük fel, hogy $c$ először az $S_r(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_r)$ kérdésben szerepel és

$$S_v(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_v) = (c, d).$$

Mivel feltettük, hogy $x < c$, ezért — a 2.4. lemma szerint — $c \in B_1^{j+1} \cup B_2^{j+1}$ és — milve az $\alpha$ eset következett be — $v < i$, $|B_3^r| < n-q$. Ezek szerint 1. vagy 5. kerülhetett alkalmazásra és $d \in B_3^{r+1}$, $d \in B_3^l$. Így $x < c$, $d < c \in \mathscr{E}_l$ és az előzőeket figyelembe véve

$$l \geqq n+q-3$$

adódik.

Ezzel az $\alpha$ eset vizsgálatát befejeztük.

Tegyük fel, hogy a $\beta$ eset következett be. Az $\alpha$ eset vizsgálatánál látottakhoz hasonlóan belátható, hogy ha $x$ az első szereplésekor nagyobb (kisebb) volt, akkor a $B_1^l \cup B_2^l \cup \{x\}$ $(B_1^l \cup B_2^l \cup \{y\})$-beli és a $B_3^l \cup \{y\}$ $(B_3^l \cup \{x\})$-beli elemek összefüggő gráfokat feszítenek $G^l$-ben és így ezen elemek közötti egyenlőtlenségek száma $\mathscr{E}$-ben legalább $q-2$, $n-q$.

A 2.7. lemma szerint minden $B_3^l$-beli elem szerepel $B_1^l \cup \{x\} \cup B_2^l$ $(B_1^l \cup B_2^l \cup \{y\})$-beli elemmel és így azon $a > b$ egyenlőtlenségek száma $\mathscr{E}_l$-ben, amelyben

$$a \in B_1^l \cup B_2^l \cup \{x\} \quad (a \in B_1^l \cup B_2^l \cup \{y\}), \qquad b \in B_3^l$$

legalább $n-q$.

Ezek szerint

$$l \geqq 2(n-q)+q-2 = 2n-q-2.$$

Mivel feltettük, hogy $q \leqq \left\lceil \frac{n}{2} \right\rceil$, ezért

$$l \geqq 2n-q-2 \geqq 2n-\left\lceil \frac{n}{2} \right\rceil -2 \geqq n+\left\lceil \frac{n}{2} \right\rceil -3 \geqq n+q-3.$$

Ezzel a $\beta$ eset vizsgálatát is befejeztük és a tételt bizonyítottuk.

Végül köszönetemet fejezem ki KATONA G. O. H. professzornak a dolgozat megírása során tett értékes megjegyzéseiért.

## IRODALOM

[1] IRA POHL, "A sorting problem and its complexity", *Com. of the ACM* **15** (1972) 462—464.
[2] KATONA, G. O. H., személyes közlés.
[3] KISZLICYN, S. S., On the selection of $k$ the element of an ordered set of pairwise comparison (oroszul) *Sib. Math. Z.* **5** (1964) 557—564.
[4] KNUTH, D. E., *The Art of Computer Programming vol. 3., Sorting and Searching* (Addison—Wesley, New York, 1975).
[5] VARECZA, Á., "On the smallest and the largest elements", *Annales Univ. Sci.*, Budapest Tom. IV (1983), 1—10.
[6] VARECZA, Á., "On a conjecture of G. O. H. Katona", *Studia Sci. Math. Hung.* (megjelenés alatt).
[7] VARECZA, Á., "Módszerek a rendezési algoritmusok elvi korlátainak meghatározására", *Alk. Mat. Lapok* **5** (1979) 191—202.
[8] VARECZA, Á., Optimális rendezési algoritmusok, Kandidátusi értekezés (1981).
[9] VARECZA, Á., "Finding two consecutiv elements", *Stud. Sci. Math. Hung.* **17** (1982), 291—302.

(Beérkezett: 1983. június 24.)

VARECZA ÁRPÁD
BESSENYEI GYÖRGY TANÁRKÉPZŐ FŐISKOLA
4400 NYÍREGYHÁZA, PF. 166.

## ON GENERALIZATION OF A PROBLEM OF G. O. H. KATONA

Á. VARECZA

Let $H$ be a finite ordered set ($|H|=n$) however their ordening is unknown for us. G. O. H. KATONA raised the following problem. If $A=\{p, q\}$ $(1 \leqq p < q \leqq n)$ and $x, y$ are arbitrary elements of $H$ then the minimal number of comparisons needed to decide whether the indexes of the elements $x, y$ are in $A$ or not (say in decreasing order). In this paper we show that if

$$A = \{r_1, r_2, \ldots, r_k\} \quad (1 \leqq r_1 < r_2 < \ldots < r_k \leqq n)$$

then the number of comparisons is at least

$$n-1 \quad \text{if} \quad r_k = k, \quad k < n$$

and

$$n+r_k-3 \quad \text{if} \quad r_k \neq k, \quad r_k \leqq \left\lceil \frac{n}{2} \right\rceil.$$

*Alkalmazott Matematikai Lapok* **10** (1984) 373—388

# DIGITÁLIS KÉPEK GEOMETRIAI KORREKCIÓI

HEGEDŰS GY. CSABA

Budapest

A digitális képfeldolgozásban sűrűn előforduló, nagy számításigényű feladat a képek geometriai korrekciója.

Ez általános esetben globálisan (a teljes képre vonatkoztatva) nem lineáris. A feladat számításigénye jelentősen csökkenthető a globálisan tetszőleges — illetve bizonyos simasági feltételeknek eleget tevő — korrekciók lokálisan lineáris approximációjával.

A javasolt eljárás mind a teoretikus úton megadott, mind pedig a kísérleti úton meghatározott geometriai korrekciók esetében hatékonyan alkalmazható, azaz a megadott pontosságot a lineáris korrekciókkal azonos nagyságrendű számításigény mellett valósítja meg.

## 1. Bevezetés

A digitális képfeldolgozásban sűrűn előforduló, nagy számításigényű feladat a képek geometriai korrekciója. Ez általános esetben egy (vagy több, geometriailag összetartozó) INPUT kép OUTPUT kép(ek)re való átvitelét jelenti oly módon, hogy a képi információ a megadott mértéknél kevésbé sérüljön, míg a kívánt geometriai összefüggés a megadott mértéknél pontosabban teljesüljön.

A digitális kép általában a folytonos, háromdimenziós világ valamilyen képfelvevő rendszer segítségével történő méréseként jön létre.

Geometriai korrekciókra pl. a következő okok miatt lehet szükség:

— a képfelvevő rendszer torzít, illetve pontatlan;
— a háromdimenziós világ objektumainak kétdimenziós képre vetítése torzít (perspektivikus torzítás);
— a mérés folyamata közben az objektum és a mérőeszköz közti geometriai összefüggés változó;
— az objektum különböző képei közt kell megfeleltetést létrehozni.

A geometriai korrekciók megadása két fő módszerrel lehetséges:

— Elméleti úton.

Ismert torzítású képfelvevő rendszer, ismert összefüggések az ábrázolt objektumok elmozdulásáról, illetve egyéb teoretikus ismeret esetén a geometriai összefüggés képletszerűen megfogalmazható.

Ekkor a korrigálandó kép(ek) minden egyes pontjára egzakt módon megadható a korrigálás utáni pozíció.

— Kísérleti úton.

A geometriai korrekciót olyan pontok segítségével is definiálhatjuk, melyeknek mind INPUT, mind pedig OUTPUT pozíciója ismeretes. Ezek az ún. *azonosítási pontok*, pozíciójuk kísérleti úton (mérésekkel) határozható meg. Ez a megadási

mód nem egzakt (az azonosítási pontok pozíciói mérési hibával ismeretesek, a többi pont pozíciója ismeretlen).

Az elméleti és a kísérleti megadási mód keveredhet egymással, pl. az elméleti úton meghatározott korrekció paraméterezhető azonosítási pontok segítségével.

A geometriai korrekciók általános esetben globálisan (a teljes képre vonatkoztatva) nem lineárisak. Bonyolult esetben a korrekciós összefüggés minden egyes képpontra való alkalmazása korrekt, de nagy számításigényű megoldás. Szerencsére bizonyos belső összefüggések (pl.: invertálható korrekciók esetén a szomszédos képpontok szomszédosak maradnak, lineáris korrekció esetén egyenesek képe is egyenes stb.) gyorsabb algoritmusok alkalmazását teszik lehetővé.

A 2. részben röviden áttekintjük a digitális kép létrehozásának és geometriai korrekciójának folyamatát.

A 3. részben számítástechnikai megfontolások alapján az általános geometriai korrekciók hatékonyabb megoldására teszünk javaslatot. A 4. részben az elméleti, az 5. részben a kísérleti úton való korrekciómegadás esetét vizsgáljuk, a 6. részben pedig a megvalósítással kapcsolatos eredményeket mutatjuk be. A témához kapcsolódó rövid irodalomjegyzék található a 7. részben.

## 2. A digitális képalkotás és képkorrekció modellje

A digitális képfeldolgozás azon alapul, hogy a folytonos képet megfelelő digitális képpé lehet alakítani, illetve a digitális képből megfelelő folytonos képet lehet előállítani. (Az átalakítást akkor mondjuk megfelelőnek, ha az információveszteség a megadott mértéknél nem nagyobb.)

Legyenek $(x, y, z, t)$ folytonos téridő koordináták, $\lambda$ a kölcsönhatás hullámhossza. Nevezzük az $F(x, y, z, t, \lambda, \mathbf{n})$ skalárfüggvényt fényességfüggvénynek (ahol az $\mathbf{n}$ egységvektor az irányfüggést tartalmazza). $(F)$ folytonos, pozitív definit és korlátos.

A *digitális képalkotás* a következő folyamatábrával modellezhető [8]:

$$\boxed{F(x, y, z, t, \lambda, \mathbf{n})} \xrightarrow{\mathscr{K}} \boxed{C_F(x, y)} \xrightarrow{\mathscr{M}} \boxed{C_M(m, n)} \xrightarrow{\mathscr{Q}} \boxed{C_D(m, n)}.$$

A $\mathscr{K}$: $F(x, y, z, t, \lambda, \mathbf{n}) \to C_F(x, y)$ leképezést *képalkotásnak* nevezzük; az $(F)$ fényességfüggvénynek a szintén folytonos, pozitív definit és korlátos $C_F(x, y)$ *képfüggvényt* felelteti meg. $(\mathscr{K})$ magában foglalja a képalkotáshoz használt kölcsönhatás valamennyi jellemzőjét (geometriai, spektrális, idő-jellemzőket).

Az $\mathscr{M}$: $C_F(x, y) \to C_M(m, n)$ leképezés *(mintavételezés)* a folytonos értelmezési tartományú $(C_F)$ függvényhez a diszkrét értelmezési tartományú $C_M \leqq C_{M\,\mathrm{MAX}}$ korlátos függvényt rendeli. A mintavételezés utáni kép a $(C_F)$ képfüggvény és a meghatározott rácson értelmezett *Dirac-delta függvények* konvolúciójaként is felfogható:

$$C_M(m, n) = C_F(x, y) \cdot S(x, y) = \sum_{i=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} C_F(i \cdot \Delta x, j \cdot \Delta y) \cdot \delta(x - i\Delta x, y - j\Delta y).$$

(Ekkor a rácsállandó: $\Delta x$, ill. $\Delta y$).

A $\mathcal{Q}$: $C_M(m, n) \to C_D(m, n)$ leképezés *(kvantálás)* a folytonos értékkészletű ($C_M$) projekciója a diszkrét értékkészletű ($C_D$) függvényre:

$$\{0 \leqq C_M(m, n) \leqq C_{M\,\mathrm{MAX}}\} \to \{C_D(m, n) | \{0, C_{D1}, C_{D1}, \ldots, C_{D\,\mathrm{MAX}}\}\}.$$

Pl.: $$C_M(m, n) \to C_{Di} \Leftrightarrow (C_{Di} \leqq C_M(m, n) \prec C_{D,i+1}).$$

A mintavételezést és kvantálást együttesen *digitalizálásnak* nevezik, ($C_D$) pedig a *digitális kép.*

A *geometriai korrekció* a következő folyamatábrával modellezhető [1, 2]:

$$\boxed{C_D(m, n)} \xrightarrow{\mathcal{R}} \boxed{C'_F(x, y)} \xrightarrow{\mathcal{M}'} \boxed{C'_M(m^*, n^*)} \xrightarrow{\mathcal{Q}'} \boxed{C'_D(m^*, n^*)}.$$

Az $\mathcal{R}$: $C_D(m, n) \to C'_F(x, y)$ leképezést *rekonstrukciónak* nevezik. Ezzel a digitális képkorrekció visszavezethető a folytonos képfüggvényből való új mintavételezésre ($\mathcal{M}'$) és új kvantálásra ($\mathcal{Q}'$). A rekonstrukció kétdimenziós interpolációs függvény segítségével valósítható meg:

$$C'_F(x, y) = \sum_{i=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} C_D(i \cdot \Delta x, j\Delta y) \cdot R(x - i \cdot \Delta x, y - j \cdot \Delta y).$$

Az $R(x, y)$ interpolációs függvény megválasztásától függően a rekonstrukció kisebb-nagyobb hibával állítja elő a folytonos képfüggvényt. Elméletileg igen jó közelítést ad, de gyakorlatban a jelentős számításigény miatt nem használatos pl. a következő:

$$R(x, y) = 2\pi\omega_0 \frac{J_1\{\sqrt{x^2+y^2}\}}{\sqrt{x^2+y^2}},$$

(ahol $J_1$ elsőrendű *Bessel-függvény*). A másik véglet a legegyszerűbben számítható *legközelebbi szomszéd-módszer:*

$$C'_F(x, y) = C_D([x], [y]),$$

(ahol $[x]$ az $x$-hez legközelebbi egészt jelöli, $\Delta x = \Delta y = 1$). A különböző interpolációs függvények [1, 2] közül a számításigény és a rekonstrukció minősége közti kompromisszummal kell választani.

## 3. Számítástechnikai megfontolások

*a)* Az általánosság megsértése nélkül fel lehet tételezni, hogy kis tartományban minden szóba jöhető korrekció lineárisan viselkedik (pl.: nem túl nagyok a korrigálandó hibák). A digitális képek lineáris, illetve affin korrekciója jól kidolgozott témakör. Különböző gyors algoritmusok ismeretesek [2, 9], melyek a digitális vonalábrázolás alapvető törvényszerűségein alapulnak [3, 5, 6]. A globálisan nemlineáris korrekciók megvalósításánál is célszerű az ismert gyors algoritmusokat használni, alkalmas módosítás segítségével.

Ez a módosítás történhetne pontról-pontra is, de sima korrekciós függvény esetén ennél lényegesen ritkábban is. Elvileg jó megoldás lenne, ha a konkrét korrek-

ciónak megfelelően a lassan változó helyeken ritkábban, a gyorsan változó helyeken sűrűbben kerülne sor módosításra. A szükséges módosítások számának szempontjából szuboptimális, de esetünkben célszerűbb a következő megoldás alkalmazása:

— Fedjük le az egyik ($A$) rendszerbeli képet szabályos rácsozatban elhelyezkedő négyzetekkel (ezeket a továbbiakban *rácselemnek* nevezzük), melyeken belül a korrekciós összefüggés feltehetően lineáris; (minél közelebb van a korrekciós összefüggés a lineárishoz, annál több képpont tartozhat egy rácselemhez).

— Számítsuk ki külön-külön a rácselemek sarokpontjainak ($B$) rendszerbeli pozícióját (ezeket a továbbiakban *vezéradatoknak* nevezzük).

— A tényleges korrekciót a korrigálandó kép minden egyes rácselemére külön-külön hajtsuk végre a sarokpontok ($B$) rendszerbeli pozíciójának ismeretében, mint általános négypontos transzformációt. Invertálható korrekciós összefüggés esetén ezek a ($B$) rendszerben is diszjunktak, és teljes lefedést biztosítanak.

A következők szólnak e megoldás mellett:

— A változó sűrűségű módosítás vezérléséhez külön számításokra (gradiens-számításra) lenne szükség, ez szabályos esetben elmarad;

— a vezéradatok létrehozására ugyan a szabályos esetben többször van szükség, mint amennyiszer valóban indokolt lenne (hiszen a leggyorsabban változó részre kell felkészülni), de a korrekció tényleges számításigényét úgyis elsősorban a sokkal nagyobb mennyiségű, képpontokkal kapcsolatos műveletek szabják meg;

— az egyik ($A$) rendszerben az egy sorban levő rácselemek együttes kezelésével sorfolytonos adatkezelés valósítható meg, ami illeszkedik a hagyományos háttértárak szervezéséhez.

*b)* Geometriai transzformációknál az egyik rendszerben felvett (általában szabályos) rács sarokpontjainak pozícióját határozzuk meg a másik rendszerben.

Ha a szabályos rácsot az INPUT rendszerben vesszük fel, akkor *direkt transzformációról,* ha az OUTPUT rendszerben, akkor *inverz transzformációról* beszélünk.

Feltételezve, hogy a korrekciós összefüggés invertálható, a két eljárás elvileg azonos eredményt adhat.

Több szempontból [8] célszerűbb a második megoldás választása (pl.: ekkor a korrigált rendszerbeli kép minden pozíciója definiált lesz, azaz nincs szükség inicializálásra).

*c)* A 2. részben leírtak szerint a korrigált és korrigálatlan képrács egymáshoz rendelése valamilyen $R$ interpolációs függvénnyel képzett konvolúciót is igényel. Homogén képtartalom mellett nem jelentkezik a köztük levő különbség, kontúros-vonalas ábra esetén viszont jelentős eltérések vannak. A gyors algoritmusokban sokszor felhasznált legközelebbi szomszéd-módszer összetett képeken nem alkalmazható jó eredménnyel, hamis kontúrt eredményez. Jobb eredményt adnak a bilineáris, esetleg ívdarabokkal interpoláló függvények, ezek azonban jelentősen nagyobb számításigényűek.

Célszerű kompromisszum, hogy az interpoláció megválasztását maga a képtartalom vezérelje. Erre a vezérlésre azok a viszonylag gyorsan létrehozható adatok alkalmazhatók, melyeket ún. *élkiemeléssel* kaphatunk a korrigálandó képből. (Élkiemelés a kép-függvény és gyorsan lecsengő hatáskörzetű, a lokális képi változást kiemelő függvények konvolúciójával képezhető; fixpontos aritmetika alkalmazásával).

Így a rácselemek transzformációjánál

— a homogén képrészeknél módosítás nélkül használható a fixpontos aritme-

tikát alkalmazó gyors algoritmus (ekkor a legközelebbi szomszéd módszernek megfelelő interpoláció érvényesül);
— a kevésbé homogén részeken a pozícióhoz az algoritmus maradéktagjából képezhető korrekció járul, és e pontos pozíció körül pontosabb eredményt adó interpolációs formula alkalmazható.

Így a geometriai korrekciónak ezen a szintjén is megvalósulhat az a célszerű elv, hogy csak a kívánt pontosság eléréséhez feltétlenül szükséges számításokat hajtsuk végre.

## 4. Geometriai korrekció megadása elméleti úton

Ha ismert az $f$: $A \to B$ korrekciós összefüggés, a korrekció célszerűen a következő lépésekben hajtható végre:

*a)* A korrekciós összefüggés simaságának függvényében a rácsállandó megválasztása. (Ez történhet a gradiensfüggvény szélsőértékének vizsgálatával, vagy eleve megadható előzetes ismeretek alapján.)

*b)* A korrekció végrehajtása a rácspontokra, és ezzel a vezéradatok előállítása.

*c)* A vezéradatok alapján a korrekció sorfolytonos elvégzése:
— az egyes elemi négyzetekre vonatkozó általános négypontos transzformáció végrehajtása fixpontos aritmetikával;
— az aktuálisan feldolgozott sort tartalmazó elemi négyzetek egyidejű feldolgozása;
— a képtartalommal vezérelt interpoláció alkalmazása (a korrekcióval egyidejűleg alkalmazott gyors élkiemelés segítségével).

## 5. Geometriai korrekció megadása kísérleti úton

A geometriai korrekciót az INPUT és az OUTPUT képen ugyanolyan képi jelentésű pontok (ún. *azonosítási pontok*) pozícióinak megadásával is definiálhatjuk.

Mivel ezek a pozíciók mérések eredményei, korlátozott pontosságúak. (Még a feltűnő képi alakzatok kiemelt pontjai sem azonosíthatóak eléggé egyértelműen, másrészt a pontosságnak elvi korlátot szab a pozíció kvantálása.)

A korrekciós összefüggést ezek figyelembevételével logikus a következő feltételek alapján keresni:
— az azonosítási pontok összerendelésénél elkövetett hiba legyen minimális;
— a meghatározott leképezés a teljes képsíkon a lehető legsimábban viselkedjék;
— a korrekció megadása a szükségesnél nagyobb számú azonosítási pont felhasználásával legyen pontosítható.

A korrekciós összefüggést előállító algoritmussal kapcsolatos igények:
— az azonosítási pontok pozícióin kívül ne legyen szükség más összefüggés ismeretére, de ha ilyenek rendelkezésre állnak, azokat fel lehessen használni (pl.: elméleti úton meghatározott korrekció paraméterezése azonosítási pontok segítségével);
— egyszerűbb és bonyolultabb közelítéseket egyaránt tartalmazzon az algoritmus, de a szükséges bonyolultságot elő lehessen írni;
— nagy azonosítási ponthalmazra is elég gyorsan konvergáljon;
— jól kézbentartható legyen (az elkövetett hibákat lehessen ellenőrizni).

Legyen $(N)$ pontból álló azonosítási ponthalmazunk, mely az $(A)$ és a $(B)$ rendszer közt kölcsönösen egyértelmű leképezést definiál:

$$P_s\colon (X_{SA}, Y_{SA}) \leftrightarrow (X_{SB}, Y_{SB}) \quad (s = 1, 2, \ldots, N), \tag{5.1}$$

ahol $P_s$ az $s$-edik azonosítási pont, $(X_{SA}, Y_{SA}, X_{SB}, Y_{SB})$ pedig az $(A)$, illetve a $(B)$ rendszerbeli pozíció koordinátái. Mivel ezek mérések eredményei, tegyük fel, hogy normális eloszlásúak.

Keressük külön az

$$f_x\colon \{X_{SA}, Y_{SA} | s = 1, 2, \ldots, N\} \to X_{SB} \tag{5.2}$$

és

$$f_y\colon \{X_{SA}, Y_{SA} | s = 1, 2, \ldots, N\} \to Y_{SB} \tag{5.3}$$

folytonos függvényeket, melyek a $P_s$ $(s=1, 2, \ldots, N)$ pontbeli megfeleltetést a lehető legkisebb hibával valósítják meg, és elegendően „sima" függvényként viselkednek. (Így az elkövetett hiba nem izotróp, viszont a számításigényt a szeparálás jelentősen csökkenti.)

Tételezzük fel, hogy az $f$ ($f_x$ vagy $f_y$) függvény közelíthető $(m)$ db. $g_i(X_A, Y_A)$ $(i=1, 2, \ldots, m)$ lineárisan független, folytonos függvény összegzésével:

$$f_a(X_A, Y_A) = \sum_{i=1}^{m} a_i g_i(X_A, Y_A),$$

ahol $f_a$ az $f_x$ vagy $f_y$ függvény közelítő függvénye. Ekkor az azonosítási pontokra a következő összefüggést írhatjuk fel:

$$\mathbf{f}_a = \sum_{i=1}^{m} a_i \bar{\mathbf{g}}_i, \tag{5.4}$$

ahol:

$\mathbf{f}_a$: $N$ elemű vektor (elemei az $X_{SB}$ vagy $Y_{SB}$ függő változók közelítő értékei, vagyis az $f_a(X_{SA}, Y_{SA})$ közelítő függvény helyettesítési értékei az $(X_{SA}, Y_{SA})$ pontban, $i=1, 2, \ldots, N$);

$a_i$: az $(m)$ elemű együttható-vektor $i$-edik eleme ($a_i$ a $g_i$ függvényhez tartozó együttható);

$\mathbf{g}_i$: $N$ elemű függvényérték-vektor $(g_i(X_{SA}, Y_{SA})$ az $i$-edik függvény $(X_{SA}, Y_{SA})$ pontban felvett értéke; $s=1, 2, \ldots, N)$; $(i=1, 2, \ldots, m)$.

Az első (pontossági) feltétel kielégíthető az eltéréseket alkalmas módon tartalmazó $\Phi(\mathbf{a})$ funkcionál minimalizálásával. Így az $a_i$ paramétereket a következő feltételből határozzuk meg:

$$\min_{a_1, \ldots, a_m} \Phi(\mathbf{a}), \tag{5.5}$$

ahol szokásos például a következő:

$$\Phi(\mathbf{a}) = \sum_{s=1}^{N} \varphi_s D(\mathbf{a}), \quad \text{és} \quad D(\mathbf{a}) = \left(\mathbf{f} - \sum_{i=1}^{m} a_i \mathbf{g}_i\right)^2 \tag{5.6}$$

($\mathbf{f}$ tényleges függvényérték; $\varphi_s$ súlyok).

A $\varphi_s$ súlyokat célszerű a $P_s$ pontok eloszlási sűrűségétől függően választani; ahol a $P_s$ pontok sűrűbben helyezkednek el, ott $\varphi_s$ legyen kisebb.

A második (simasági) feltétel teljesülését bizonyos mértékben elősegíti az $m \ll N$ választás, ekkor ui. az $f_a$ függvénynek viszonylag kevés paramétere van, és sima $g_i$ függvények esetén — kisebb a lehetősége, hogy $f$-től különbözzék.

A simaságot alkalmasan választott regularizációval is elő lehet segíteni. Ekkor $\Phi(\mathbf{a})$ helyett a $\Phi'(\lambda, \mathbf{a})$ kifejezés minimumából határozhatók meg az $a_i$ együtthatók:

$$\Phi'(\lambda, \mathbf{a}) = \Phi(\mathbf{a}) + \lambda \cdot \Psi(\mathbf{a}). \tag{5.7}$$

A $\psi(\mathbf{a})$ funkcionál a függvény simaságát kell, hogy kifejezze, így lehet pl. az

$$\int_K |\operatorname{grad} \mathbf{y}_a|^2 dP \quad (K\text{: képsik}, \quad P \in K) \tag{5.8}$$

integrál valamilyen közelítése.

A $\lambda$ paraméter megválasztásától függ, hogy a közelítés pontosabb lesz a megadott pontokban, vagy simább.

Sajnos a $\Psi(\mathbf{a})$ funkcionál számítása ebben a megfogalmazásban igen műveletigényes (gradiensképzés a teljes értelmezési tartományban) és átfogalmazása (pl. a max. gradienst tartalmazó alakra) általános bázisfüggvények esetén úgyszintén.

Esetünkben célszerűbb más megoldás alkalmazása:

— az approximációt $\lambda = 0$ beállításban (tehát simasági feltétel nélkül) hajtjuk végre;

— a $P_s$ azonosítási pontok egyenletes elosztásával, és sima $g_i$ függvények választásával elősegítjük az approximáció simaságát;

— a globális vizsgálatokat tartalmazó — tehát lassú — simasági optimalizálás helyett a gyorsabb megoldást választjuk:

— több, különböző bonyolultságú közelítést hajtunk végre, ezek pontosságáról listát készítünk;

— az egyes közelítések simaságát utólag, gyors (vizuális) módszerrel ellenőrizzük, és a pontosság és a simaság közti kompromisszumot igénylő választást emberi döntésre bízzuk.

(*Megjegyzés:* ez közel azonos eredményre vezethet, mint az (5.7) kifejezés alkalmazása, viszont nagyságrendekkel gyorsabb, ezenkívül összekapcsolható a korrekciót közvetlenül vezérlő adatok előállításával.)

Az (5.4) approximációra vonatkozó (5.6) feltétel a következőképpen alakítható át:

$$D(\mathbf{a}) = \mathbf{f}^T\mathbf{f} - 2 \cdot \sum_{i=1}^{m} a_i(\mathbf{f}^T\mathbf{g}_i) + \sum_{i=1}^{m}\sum_{j=1}^{m} a_i a_j(\mathbf{g}_i^T\mathbf{g}_j), \tag{5.9}$$

ahol a $T$ felső index-szel a transzponált vektort jelöltük.

Vezessük be a

$$\mathbf{z} = \begin{bmatrix} \mathbf{f}^T\mathbf{g}_1 \\ \mathbf{f}^T\mathbf{g}_2 \\ \vdots \\ \mathbf{f}^T\mathbf{g}_m \end{bmatrix} (m \times 1)\text{-es}, \quad \text{és} \quad \mathbf{G} = [\mathbf{g}_i^T\mathbf{g}_j] (m \times m)\text{-es mátrixokat},$$

és transzponáltjaikat. Ekkor az (5.9) egyenlet így írható fel:

$$D(\mathbf{a}) = \mathbf{f}^T\mathbf{f} - 2\mathbf{z}^T\mathbf{a} + \mathbf{a}^T\mathbf{G}\mathbf{a}. \tag{5.10}$$

$D(\mathbf{a})$-t a minimalizáláshoz fejtsük sorba tetszőleges ($m$-dimenziós térbeli) $\mathbf{a}^*$ pont körül. Ez:

$$D(\mathbf{a}) = D(\mathbf{a}^*) + \left[\frac{\partial D(\mathbf{a})}{\partial \mathbf{a}}\right]_{\mathbf{a}=\mathbf{a}^*}(\mathbf{a}-\mathbf{a}^*) + \frac{1}{2}(\mathbf{a}-\mathbf{a}^*)\left\{\frac{\partial}{\partial \mathbf{a}}\left[\frac{\partial D(\mathbf{a})}{\partial \mathbf{a}}\right]\right\}_{\mathbf{a}=\mathbf{a}^*}(\mathbf{a}-\mathbf{a}^*) + \tag{5.11}$$

$$+ \ldots(\text{magasabbfokú tagok}).$$

Az (5.10) kifejezés segítségével:

$$\left[\frac{\partial D(\mathbf{a})}{\partial \mathbf{a}}\right] = -2\cdot\mathbf{z}^T + 2\mathbf{a}^T\mathbf{G}, \tag{5.12}$$

és

$$\left\{\frac{\partial}{\partial \mathbf{a}}\left[\frac{\partial D(\mathbf{a})}{\partial \mathbf{a}}\right]\right\} = 2\mathbf{G}. \tag{5.13}$$

(5.13)-ból eleve következik, hogy az összes magasabbfokú parciális derivált értéke zérus, mivel [**G**] független **a**-tól. Így:

$$D(\mathbf{a}) = D(\mathbf{a}^*) + (-2\mathbf{z}^T + 2\mathbf{a}^{*T}\mathbf{G})(\mathbf{a}-\mathbf{a}^*) + \frac{1}{2}[(\mathbf{a}-\mathbf{a}^*)^T 2\mathbf{G}(\mathbf{a}-\mathbf{a}^*)]. \tag{5.14}$$

Így az (5.5) feltétel teljesülésének feltétele, hogy

$$\left[\frac{\partial D(\mathbf{a})}{\partial \mathbf{a}}\right] = -2\mathbf{z}^T + 2\mathbf{a}^T\mathbf{G} = \mathbf{0}, \tag{5.15}$$

azaz:

$$\mathbf{a}^T\mathbf{G} = \mathbf{z}^T \Rightarrow \sum_{j=1}^{m} a_j(\mathbf{g}_i^T\mathbf{g}_j) = \mathbf{f}^T\mathbf{g}_i \quad (i = 1, 2, \ldots, m)$$

legyen (5.9) és (5.16) segítségével az approximáció hibája:

$$D(\mathbf{a}) = \mathbf{f}^T\mathbf{f} - \sum_{i=1}^{m} a_i\mathbf{f}^T\mathbf{g}_i.$$

A leírt megoldás műveletigénye erősen függ ($m$)-től. (Az (5.16) mátrixegyenlet szerint ($m$) approximációs függvény alkalmazása esetén $m\times m$-es mátrixot kell invertálnunk.) Így nagy ($m$) esetén a $\mathbf{g}_i$ bázisfüggvények cserélgetése, illetve az approximációban résztvevő függvények számának változtatása nehézkes, lassú.

A probléma áthidalására keressünk az ($m$) tagú és az ($m+1$) tagú approximáció közt rekurzív kapcsolatot. Ezenkívül az adott ($m$) tagú approximációban részt vevő függvényeket egy $M \gg m$ függvényt tartalmazó ún. *bázisfüggvény-könyvtárból* válassza ki a program az adott választásokhoz tartozó (5.17) hiba minimális értékének megfelelően.

Azt várjuk, hogy így egyrészt nem kell előre meghatározni az approximáció tagszámát, másrészt minden szinten a legjobb eredményt adó bázisfüggvények kerülnek felhasználásra [4]. A bázisfüggvények célszerűen alacsony fokszámú kétváltozós polinomok lehetnek; ezek egyrészt egyszerűen és gyorsan számíthatóak, másrészt a

geometriai torzítást okozó fizikai jelenségek leírására kielégítő pontossággal alkalmazhatóak. A rekurzív algoritmustól ezenkívül azt is várjuk, hogy az $(m+1)$ tagú approximáció számításainak nagy része támaszkodjék az $(m)$ tagú számítások részeredményeire, azaz az eljárás folytatásakor a műveletigény nagyságrendje ne nőjön. (Természetesen a számábrázolás véges pontossága, kerekítési hibák miatt a magasabb tagszám alkalmazási lehetősége korlátos. Az algoritmustól a hiba jól kézbentarthatóságát is elvárjuk.)

Írjuk fel az (5.16) egyenlet felhasználásával a különböző tagszámú approximációt:

$m = 1$

$$(\mathbf{g}_1^T \mathbf{g}_1)(a_{11}) = \mathbf{f}^T \mathbf{g}_1$$

$$\Rightarrow a_{11} = \frac{\mathbf{f}^T \mathbf{g}_1}{\mathbf{g}_1^T \mathbf{g}_1}.$$

Bevezetve a $B_{11} = \mathbf{g}_1^T \mathbf{g}_1$ jelölést: $a_{11} = \dfrac{\mathbf{f}^T \mathbf{g}_1}{B_{11}}$.

A bázisfüggvénykönyvtár $(M)$ függvényéből az lesz a választott $\mathbf{g}_1$ függvény, melyre az (5.17) hiba minimális, az $(m=2)$ szinten pedig a maradék $(M-1)$ függvény közül lehet választani.

$m = 2$

$$\begin{bmatrix} \mathbf{g}_1^T \mathbf{g}_1 & \mathbf{g}_1^T \mathbf{g}_2 \\ \mathbf{g}_2^T \mathbf{g}_1 & \mathbf{g}_2^T \mathbf{g}_2 \end{bmatrix} \begin{bmatrix} a_{21} \\ a_{22} \end{bmatrix} = \begin{bmatrix} \mathbf{f}^T \mathbf{g}_1 \\ \mathbf{f}^T \mathbf{g}_2 \end{bmatrix}.$$

A $B_{ij}$ jelölés bevezetésével a megoldás:

$$B_{12} = \frac{\mathbf{g}_1^T \mathbf{g}_2}{B_{11}}, \quad B_{21} = \mathbf{g}_2^T \mathbf{g}_1$$

$$B_{22} = \mathbf{g}_2^T \mathbf{g}_2 - B_{21} B_{12}$$

$$a_{22} = \frac{\mathbf{f}^T \mathbf{g}_2 - \mathbf{a}_{11} B_{21}}{B_{22}}, \quad a_{21} = a_{11} - B_{12} \cdot a_{22}.$$

(Látható, hogy az $m=1$ szintről $a_{11}$ és $B_{11}$ felhasználható.)

$m=3$

A mátrixegyenlet megoldása a következő eredményre vezet:

$$B_{13} = \frac{\mathbf{g}_1^T \mathbf{g}_3}{B_{11}}, \quad B_{23} = \frac{\mathbf{g}_2^T \mathbf{g}_3 - B_{21} B_{13}}{B_{22}}$$

$$B_{31} = \mathbf{g}_3^T \mathbf{g}_1, \quad B_{32} = \mathbf{g}_3^T \mathbf{g}_2 - B_{31} B_{12}$$

$$B_{33} = \mathbf{g}_3^T \mathbf{g}_3 - B_{31} B_{13} - B_{32} B_{23}$$

$$a_{33} = \frac{\mathbf{f}^T \mathbf{g}_3 - B_{31} a_{11} - B_{32} a_{22}}{B_{33}}$$

$$a_{32} = a_{22} - B_{23} a_{33}, \quad a_{31} = a_{11} - B_{12} a_{32} - B_{13} a_{33}.$$

(Látható, hogy $m \leqq 2$ approximációs szintek részeredményei közül az $a_{11}, a_{22}, B_{11}, B_{21}, B_{12}, B_{22}$ felhasználható.)

Általánosítva, a rekurziós algoritmust az $m \geqq 3$ esetekre:

$$\left.\begin{aligned} B_{i1} &= \mathbf{g}_i^T \mathbf{g}_1, \quad B_{1i} = \frac{B_{i1}}{B_{11}} \\ B_{ij} &= \mathbf{g}_i^T \mathbf{g}_j - \sum_{k=1}^{j-1} B_{kj} B_{ik}, \quad B_{ji} = \frac{B_{ij}}{B_{jj}} \end{aligned}\right\} \quad (j = 2, \ldots, i-1)$$

$$B_{ii} = \mathbf{g}_i^T \mathbf{g}_k - \sum_{k=1}^{i-1} B_{ki} B_{ik}$$

$$a_{ii} = \frac{\mathbf{f}^T \mathbf{g}_i - \sum_{k=1}^{i-1} a_{kk} B_{ik}}{B_{ii}}$$

$$a_{i,i-j} = a_{i-j,i-j} - \sum_{k=1}^{j} a_{i,i-j+k} B_{i-j,i-j+k} \quad (j = 1, \ldots, i-1).$$

(Az adott ($m$) szinten a korábbi szinteken fel nem használt ($M-m+1$) függvény közül az (5.17) feltétel szerint optimális lesz a $\mathbf{g}_m$ függvény.)

Az algoritmus végrehajtásának eredménye ($m$) db különböző approximációs függvény, melyek közül pontossági és simasági feltételek teljesülésének figyelembevételével kell választanunk.

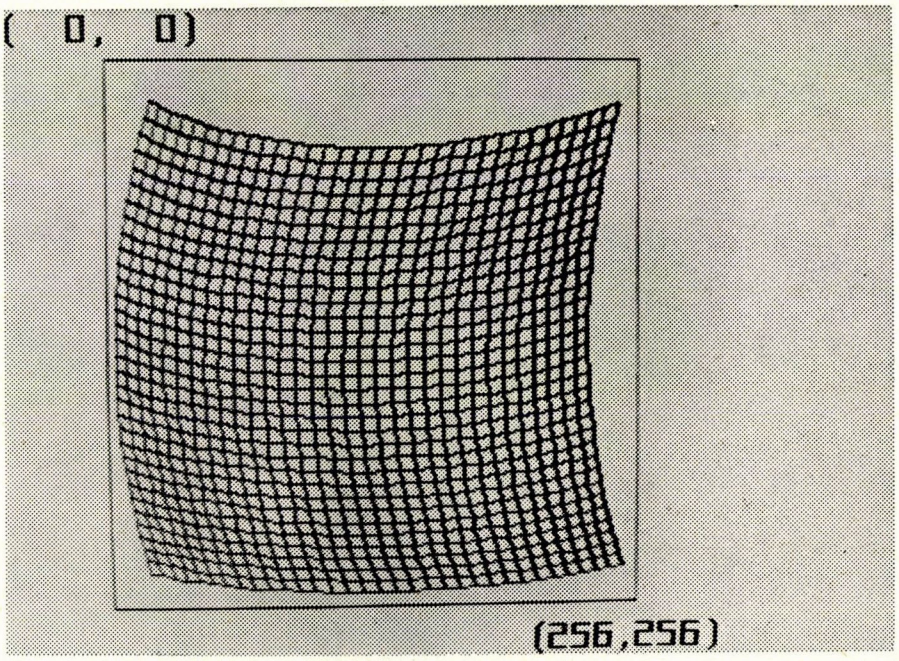
A pontossági feltételek teljesülése az algoritmust megvalósító program informatív üzenetei alapján ismeretes. A simasági feltételek teljesülésének ellenőrzését célszerű összekapcsolni a korrekciós vezéradatok előállításával; azaz a szóbajövő approximációkra a 4. rész b) pontjában leírtak közben célszerű lehetővé tenni a vizuális ellenőrzést. Ez történhet pl. a korrigált rács tv-monitoron való megjelenítésével. A korrekció tényleges végrehajtása a 4. rész c) pontjában leírtak szerint történhet.

## 6. Megvalósítás, tapasztalatok

Közvetlen módon max. másodfokú polinommal leírható korrekciók adhatók meg. (Ez a szokásos esetekben elegendő.) Az approximációs eljárás ötödfokú racionális törtfüggvénnyel való közelítést tesz lehetővé, így az alkalmazott bázisfüggvénykönyvtár 20+20 elemű. A rekurzív algoritmus ennek megadható részhalmazából megadható számú (max. 15) elemet használ fel a közelítéshez. Az azonosítási pontok száma max. 255 lehet, numerikus és képi úton (megjelenített képen ellenőrizhető pozíciómegadással) egyaránt megadhatóak.

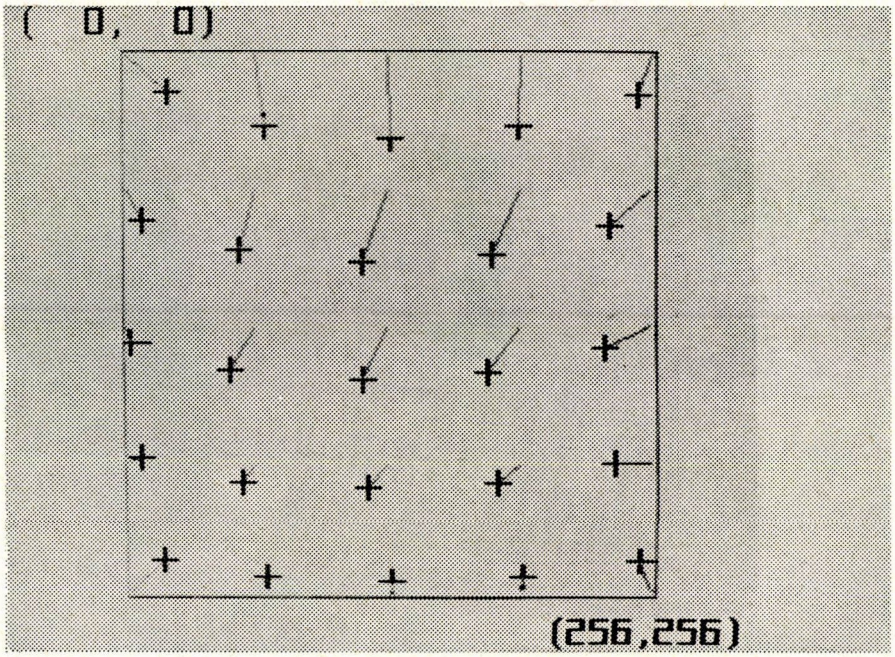
Kb. tízes nagyságrendű azonosítási pontszám esetén az approximáció egy-két percet igényel.

Az egyes közelítések közti választás a korrigált rács képi megjelenítésének segítségével történik, a korrekciót közvetlenül vezérlő adatok egyidejű létrehozása mellett. Egy-egy közelítés vizsgálata így perc nagyságrendű. (1. ábra). A tényleges korrekció 256×256 pontos képek esetén a háttértár igénytől és az interpolációváltás küszöbértékének megválasztásától függően 0,5—5 perc alatt készül el.




*1. ábra*

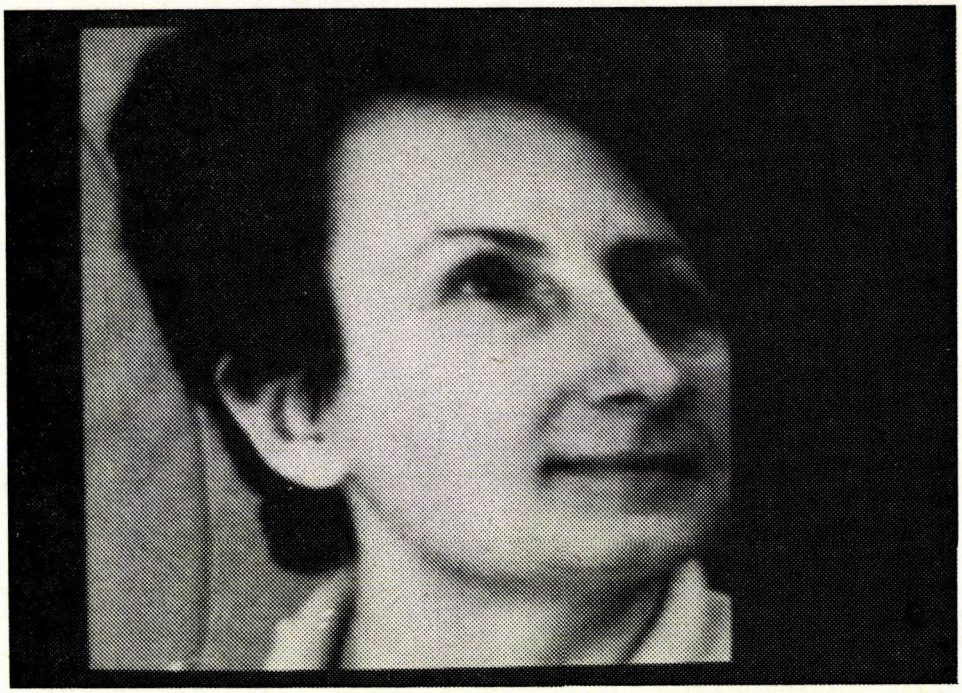
*2. ábra*




*3. ábra*

*4. ábra*

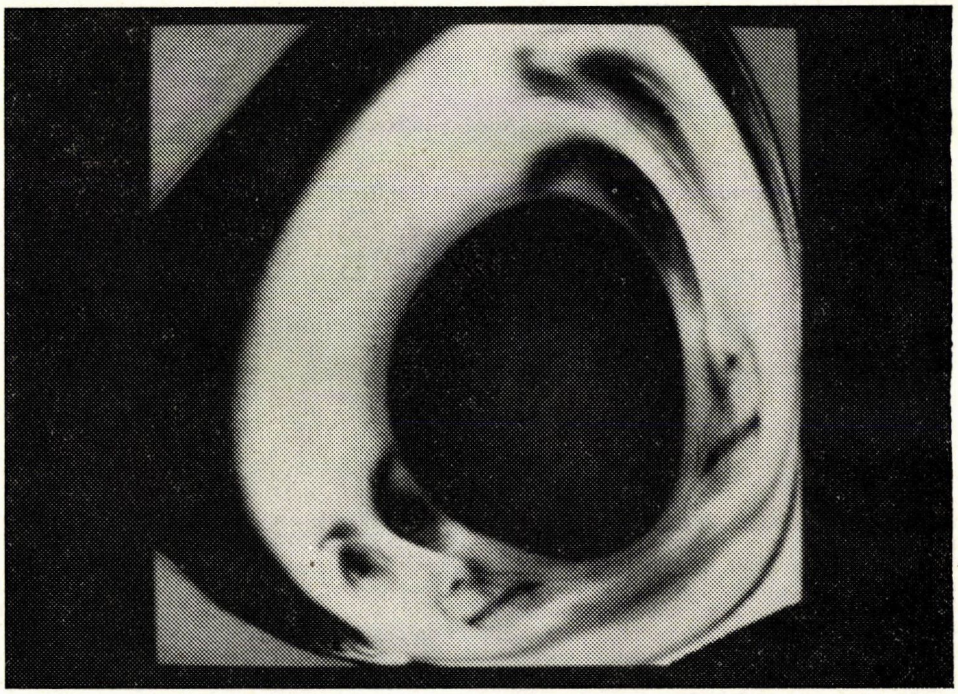
*5. ábra*



*6. ábra*

*7. ábra*

*8. ábra*



*9. ábra*

A szükséges idő független az approximált korrekció összetettségétől, és azok tényleges megvalósításánál több nagyságrenddel kisebb, közel azonos eredmény mellett.

A módszer szemléltetésére a 2. ábrán látható képet a 3. ábrán látható azonosítási pontok felhasználásával transzformáltuk; az eredmény a 4. ábrán látható. További eredményképek láthatóak a 6—9. ábrákon; ezeket az 5. ábrán látható kép nemlineáris torzításával kaptuk.

Köszönetnyilvánítás

Ezúton szeretnék köszönetet mondani dr. ÁLLÓ GÉZA lelkiismeretes korrektúrájáért és értékes útmutatásaiért.

IRODALOM

[1] ANDREWS, H. C. and PATTERSON, C. L., "Digital interpolation of discrete images", *IEEE Transaction on Computers* **C—25** (1976) 196—202.

[2] BRACCINI, C. and MARINO, G., "Fast geometrical manipulations of digital pictures", *Computer Graphics and Image Processing* **13** (1980) 127—141.

[3] DANIELSSON, P. E., "Incremental curve generation", *IEEE Transaction on Computers* **C—19** (1970) 783—793.

[4] INGRAM, H. L. and HOOKER, R., "The selection of approximating functions for tabulated numerical", Technical Report, X-64658, (George C. Marshall Space Flight Center, Alabama).

[5] JORDAN, B. W., LEMON, W. J. and HOLM, B. D., "An improved algorithm for the generation of nonparametric curves", *IEEE Transaction on Computers* **C—22** (1973) 1052—1060.
[6] LOCEFF, M., „The line", *Computer* (1980) 57—65.
[7] PERNY, D., GANGNET, M. and COUEIGNOUX, PH., "Perspective mapping of planar textures", *Computer Graphics* **16** (1982) 72—89.
[8] PRATT, W. K., *Digital Image Processing* (A Wiley-interscience publication, John Wiley & Sons, N. Y., 1978).
[9] WEIMAN, C. F. R., "Highly parallel digitized geometric transformations without matrix multiplication", Proceedings of the 1976 International Conference on Parallel Proceedings, Detroit, Michigan, 1—10.

(Beérkezett: 1983. június 2.)

HEGEDŰS GY. CSABA
SZKI MATEMATIKAI LABORATÓRIUM
1016 BUDAPEST, DONÁTI U. 35—45.

## FAST GEOMETRIC CORRECTION OF DIGITAL IMAGES

GY. CS. HEGEDŰS

Computationally demanding geometric correction of pictures are often used in digital image processing.

In the general case the geometric correction globally is nonlinear and the proper solution is to use the correcting relation at each point of the picture. However this requires much computation.

Fortunately one can use faster algorithms with the locally linear approximation. We suggest an efficient solution of the general geometric correction problem. The time needed to investigate a correction is essentially independent of the complexity of the correction approximated and although the approximation is several magnitudes faster than the precise correction that yield almost identical results.

*Alkalmazott Matematikai Lapok* **10** (1984) 389—403

# LEHETŐSÉGEK ÉS KORLÁTOK A SZŐLŐÁGAZAT 2000-IG TARTÓ FEJLESZTÉSÉBEN

FERENCZY ANTAL—PAPP ZSOLT—SZIDAROVSZKY FERENC—URBÁN ANDRÁS

Budapest

A magyar élelmiszergazdaság fontosabb ágazatai közül a szőlő—bor vertikum jóval a termelésből való részesedést meghaladó arányban vesz részt az ország exportjában. A megtermelt bor 40%-a exportra kerül, s ez az arány — Algériát és Bulgáriát kivéve — a legmagasabb a többi bortermelő országhoz viszonyítva. Ebből a körülményből és az ország gazdaságában végbement változásokból viszont az is adódik, hogy ebben az ágazatban az export további növeléséhez fűződő társadalmi érdekeink mindenképpen konfliktusba kerülnek a gazdaságosság növelését szorgalmazó követelményekkel. Ennek a súlyos ellentmondásnak rövid távú feloldása elképzelhetetlen. Bonyolítja a helyzetet, hogy a rövid távú döntések — az ültetvények jellegéből adódóan — hosszú távon is alapvető szerkezeti változásokat, esetleg aránytalanságokat okozhatnak. Mindezek a termelőalapok összetételére, a telepítések és kivágások arányára, a telepítés-politikára irányítják a figyelmet. A kedvezőtlen ültetvény összetétel, az elmúlt 20 év hullámzó telepítései egyrészt előidézői az utóbbi évekre is jellemző nagyarányú termésingadozásoknak, másrészt azonban meg is szabják a jelen és a jövő telepítési feladatait. Kérdés tehát: milyen következményekkel jár az egyenletes szükségletkielégítés, ami a későbbiek során — időszakonként — eltérő telepítéseket feltételez és hosszabb távon sem segíti az arányos korösszetételű ültetvények kialakulását. Ki kell tehát alakítani azt a hosszabb távra szóló termeléspolitikát, amely töretlenül érvényesíti a felhasználói igények folyamatos kielégítését, de várhatóan nem mérsékli a telepítési feladatok időszakonkénti különbségeit. E telepítéspolitika megalapozására összeállítottunk egy olyan többperiódusos lineáris programozási modellt, amellyel választ kaphatunk a különböző időszakok telepítéseinek és kivágásainak nagyságára, feltételként elfogadva a különböző felhasználási célok érvényesítését.

Vizsgálatunk alapjában véve módszertani munka, mégpedig kettős értelemben. Egyrészt olyan matematikai eljárás kertészeti témájú adaptálására törekszünk, amely módszer lehetővé teszi az időtényező, valamint az egymás mellett, illetve az egymással szemben érvényesülő célok komplex figyelembe vételét. Másrészt pedig azért tekintjük módszertaninak a tanulmányt, mert következtetéseink nem valamely konkrét, megvalósítható vagy megvalósítandó termelésfejlesztési programot jelentenek, hanem csak tendenciákat, szempontokat egy ilyen kidolgozásához.

Mi ebben a tanulmányban az ezredfordulóig tartó két évtized[1] szőlőtelepítési politikájához kívánunk néhány adalékkal szolgálni, korántsem a teljesség igényével, hanem inkább a vita szándékával. A kialakítandó telepítési elképzelésekben alapvetően három szempontot érvényesítettünk:

— a pénzügyi szervek a fejlesztési forrásokkal való maximális takarékosságot hangsúlyozzák elsődleges célkitűzésként;
— az ágazati irányítók egyrészt az újratermelés oldalának technikai zavartalanságát teszik az első helyre, másrészt a szolídan megfogalmazott felhasználói igények mellett a lehető legalacsonyabb termelési költségek elérését tartják kívánatos célkitűzésnek;
— a népgazdasági tervvel foglalkozók viszont a hazai gazdaság jelenlegi helyzete alapján a maximális exportbevételnek adnak prioritást.

E hármas célrendszerből természetesen konfliktusos helyzet adódik, de ez az ellentmondásosság hazai gazdaságunk egészének jellemzője. Első közelítésre ez a három nézőpont szándékosan kiélezettnek tűnhet, amelyek közül tiszta formájában egyik sem juthat érvényre. Kompromisszumos elegyük viszont megfelel a valóságnak, így — a módszer nyújtotta lehetőség alapján — egyszerre a három célkitűzést is figyelembe vettük. De megvizsgáltuk azt is, hogy egyikük vagy másikuk előtérbe kerülése milyen következményekkel jár.

[1] Vizsgálatunkban az 1981—1985. közötti időszakot is figyelembe vettük.

## 1. A matematikai modell verbális megfogalmazása

Vizsgálatunk módszere a lineáris programozás. Ami új ebben, az az, hogy a módszert a szőlőültetvény telepítés hosszú távú programozásához használjuk. Új elem az is, hogy e témakörben először a többcélú programozást alkalmazzuk.

A modell belső taglalására — a középtávú tervezés gyakorlata alapján, valamint a modell méretét tekintve — az ötéves periódusok alkalmazása látszott a legcélszerűbbnek. Ennek megfelelően az 1980—2000. közötti telepítési és kivágási ütem meghatározására 6 ötéves periódusból álló (2010-ig terjedő) részmodellt állítottunk össze. Az időszak ilyen meghosszabbítására azért volt szükség, mert a még vizsgált utolsó évtized szőlőtelepítéseire (1990—2000) csak az azt követő évtized felhasználói igényeinek figyelembe vételével kaphatunk reális eredményeket. Ennek megfelelően viszont az ezredforduló utáni évekre kapott eredmények már nem értelmezhetők.

*a) A változók rendszerének összeállítása*

A változók rendszere egy-egy időszakon belül egyrészt termelési, másrészt szükségleti változókból tevődik össze. Ezek:

— területi változók,
— telepítési-kivágási változók,
— termelési változók,
— a felhasználói célokat jelölő változók.

A területi változók egy-egy időszakban a különböző társadalmi szektorok és korösszetétel szerinti szőlőterület nagyságát jelölik. A korcsoportok száma az állami szektorban és a termelőszövetkezeti közösnél megegyező, az egyéb szektornál viszont lényegesen hosszabb termőkort feltételeztünk.[2]

Az új telepítéseket időszakonként és szektoronként egy változó szimbolizálja. A kivágásoknál viszont a termőkor vége előtti kivágásokra is lehetőséget adtunk. Az úgynevezett „idő előtti" kivágásokhoz 2, illetve 3 változó is tartozhat.

A termelési változók szektoronként és időszakonként a termelés nagyságát mutatják. Egy-egy időszakban szektoronként egy-egy, illetve a szektorok összes termelését szimbolizáló termelési változó szükséges.

A felhasználói célokat jelölő változók a következők:

— a hazai étkezési szőlő felhasználást, illetve az üdítőital gyártás alapanyag igényét jelölő változók;
— a hazai borfogyasztást szimbolizáló változók;
— a rubel, illetve dollár elszámolású borexport mennyiségét meghatározó változók;
— s végül a felhasználói igényektől elmaradó termelést (a hiányt), vagy a felhasználást meghaladó termelést (a többletet) jelölő változók.

[2] Az állami szektorban és a termelőszövetkezeti közösnél 30 éves, az egyéb gazdaságoknál pedig 45 éves termőkort feltételeztünk.

*b) A korlátozó feltételek megfogalmazása*

A feltételi rendszerünkben a felhasználók oldaláról jelentkező igények és a kielégítésüket szolgáló termelési erőforrások kapcsolatát fogalmaztuk meg.

A területre vonatkozó korlátozó feltételeink megfogalmazásakor a szőlőültetvények 1980. évi tényleges állapotából indultunk ki, melyet a modellben folyamatosan korosbítottunk. Adott időszak telepítéseinél külön korlátot nem írtunk elő, ehhez határt csupán a későbbi időszakok szükségletei, illetőleg az ott megfogalmazott felhasználói célok jelentenek. A kivágásoknál — az egyéb szektort kivéve — általában három korcsoport kivágását tettük lehetővé, ebből egy meghatározott életkort elérve kötelező kivágást írtunk elő, míg két megelőző időszakban a különböző mérlegelések tárgyává, a modell szabadságává tettük ezek mértékét és arányát. Az egyéb szektornál — tekintettel a kisüzemi ültetvények jellegére és a termelést meghatározó körülményekre — négy korcsoport kivágását is lehetővé tettük.

A feltételrendszer második nagyobb csoportját a termelés, valamint a termelés és felhasználás összefüggése jelenti. A felhasználás különböző irányai közül az étkezési szőlő termelést és a hazai borfogyasztást szigorú egyenlőség formájában írtuk elő. A rubel, illetve a dollár elszámolású borexportnál viszont alsó és felső korlátot adtunk meg.

*c) A célfüggvény meghatározása*

A távlati telepítési prognózis meghatározásához különböző szempontokat tarthatunk fontosnak. A kialakítható célkitűzések különbözősége elsősorban a vállalati és a népgazdasági érdekek esetenkénti ellentmondásaiból, ütközéséből adódik. E közelítés nyilvánvalóan szükségessé is teszi az említett érdekeket megtestesítő célkitűzések egyértelmű megfogalmazását. Így az általunk legfontosabbnak ítélt tényezők figyelembe vételével 3 célfüggvényt állítottunk be a modellbe:

— a folyó termelési költségeket minimalizáló célfüggvényt, amikor is a termékfelhasználás előzetesen meghatározott, vagy alsó és felső határral korlátozott mérete mellett az egyes időszakokat, vagy a teljes időhorizontot érintő termelési döntéseknél a költségek szerepe válik meghatározóvá;
— a telepítési költségek minimalizálását előirányzó célfüggvény szerinti változat a szükségletek kielégítése szempontjából feltétlenül szükséges, de a legkisebb pénzeszközöket igénylő ültetvény-beruházásokra ad útmutatást;
— az exportbevételt maximalizáló változatnál arra a kérdésre kerestünk választ, hogy az exportlehetőségek maradéktalan kihasználása mellett milyen telepítési és kivágási program fogalmazható meg.

Ezen célfüggvények külön-külön való kiragadása csak egyoldalú megközelítése az ültetvénytelepítéseket befolyásoló tényezőknek. Együttes kezelése, mindhárom szempont együttes érvényesülése viszont már olyan lehetőségeket kínál, melyeket egy hosszú távra kialakítandó termeléspolitikánál mindenképpen érdemes figyelembe venni.

## 2. A modell matematikai megfogalmazása

A következőkben a modell matematikai leírását foglaljuk össze, figyelembe véve a jelölés rendszerét, a feltételeket, valamint a célfüggvényeket.

Matematikai jellegüknél fogva a változókat több csoportra osztottuk.

*a) Döntési változók*

$y_{ijk}$ = $i$-edik társadalmi szektorhoz tartozó, $j$-edik korcsoportú szőlőből a $k$-adik időszakban kivágott összterület (ha)
$z_{ik}$ = $i$-edik társadalmi szektorban a $k$-adik időszakban telepített szőlő összmennyisége (ha)
$u1_k$ = $k$-adik időszakban étkezési célra és üdítőital készítésére használt szőlőmennyiség
$u3_k$ = $k$-adik időszakban hazai bor készítésére használt szőlőmennyiség
$u4_k$ = $k$-adik időszakban szocialista borexportra használt szőlőmennyiség
$u5_k$ = $k$-adik időszakban tőkés borexportra használt szőlőmennyiség
$u6_k$ = $k$-adik időszakban termelt többlet szőlőmennyiség.

*b) Állapotváltozók*

$x_{ijk}$ = $i$-edik társadalmi szektorhoz tartozó, $j$-edik korcsoportú szőlő összterülete (ha) a $k$-adik időszakban.

*c) Származtatott változók*

$u_{ik}$ = $i$-edik társadalmi szektorban a $k$-adik időszakban termelt összes szőlőmennyiség
$u_k$ = az összes társadalmi szektorban a $k$-adik időszakban együttesen termelt szőlőmennyiség
$v_k$ = $k$-adik időszakban importált bornak megfelelő szőlőhiány mennyisége.

*d) Korlátozó feltételek*

$x_{ij1}$ = adott értékek, a jelenlegi telepítéssel meghatározottak

(2.1) $$y_{ir_ik} = x_{ir_ik},$$

ahol

(2.2) $$r_i = \begin{cases} 6, & \text{ha } i = 1, 2 \\ 9, & \text{ha } i = 3, \end{cases}$$

vagyis a legrégebbi ültetvényeket kivágják.

(2.3) $$y_{ijk} = 0 \quad (j < r_i - 1),$$

azaz fiatal ültetvény nem vágható ki.

A szőlőtermelés és felhasználás mérlegegyenletei:

(2.4) $$u_k + v_k - u1_k - u3_k - u4_k - u5_k - u6_k = 0 \quad (k = 1, 2, 3),$$

valamint

(2.5) $$\sum_j x_{ijk} T_j - u_{ik} = 0 \quad (i = 1, 2, 3)$$

(2.6) $$\sum_{i=1}^{3} u_{ik} - u_k = 0,$$

$T_j$ = $j$-edik korcsoporthoz tartozó szőlő átlagtermése.

Az állapotátmeneti összefüggéseket az

$$x_{ijk} = \begin{cases} x_{i,j-1,k-1} - y_{ijk}, & \text{ha } j \neq 1 \\ z_{ik}, & \text{ha } j = 1 \end{cases} \tag{2.7}$$

egyenletek írják le, azaz a következő időszakban az előzőből megmaradó ültetvény is egy korcsoporttal öregebb lett, valamint a legfiatalabb ültetvényeket most kellett telepíteni.

A fentieken kívül még az

$$u3_k,\ u4_k,\ u5_k,\ u6_k$$

változókra alsó és felső korlátok adottak minden $k$ esetén.

*e) Célfüggvények*

A *folyó termelési költség* teljes mennyisége

(2.8)

$$\varphi_1 = \sum_i \sum_j \sum_k x_{ijk} \cdot c_{ijk} + \sum_i \sum_k u_{ik} \cdot C_{ik} + \sum_k y_{i,r_i-1,k} \cdot K_{ik} + \sum_k v_k L_k + \sum_k u6_k \cdot l_k,$$

ahol

$c_{ijk}$ = 1 ha művelési költsége az $i$-edik társadalmi szektor, $j$-edik korcsoport és $k$-adik időszak esetén

$C_{ik}$ = egységnyi betakarítási költség az $i$-edik társadalmi szektorban és $k$-adik időszakban

$K_{ik}$ = idő előtti kivágás egységköltsége az $i$-edik társadalmi szektorban és a $k$-adik időszakban

$L_k$ = $k$-adik időszakban borimport egységköltsége

$l_k$ = $k$-adik időszakban többlettárolási egységköltség.

*A telepítés beruházási költsége:*

$$\varphi_2 = \sum_i \sum_k z_{ik} B_{ik}, \tag{2.9}$$

ahol

$B_{ik}$ = $i$-edik társadalmi szektorban és a $k$-adik időszakban 1 ha szőlő telepítési költsége.

Az *exportbevételt* a következő kifejezés adja meg:

$$\varphi_3 = \sum_k u4_k \cdot h_k + \sum_k u5_k H_k - \sum_k v_k L_k - \sum_k u6_k l_k, \tag{2.10}$$

ahol

$h_k, H_k$ = borexport egységnyi exportra vonatkozó bevétele a szocialista és tőkés relációban a $k$-adik időszakban.

Feladatunk tehát egy lineáris feltételrendszerrel korlátozott, három célfüggvénynyel leírt optimalizálási feladat megoldása. Vegyük észre, hogy esetünkben $\varphi_1$ és $\varphi_2$ minimumra, valamint $\varphi_3$ maximumra törekszünk.

## 3. A modell matematikai megoldása

A fenti probléma megoldásánál két módszert választhatunk.

a) A *súlyozásos módszer* esetén a három célfüggvénynek egy-egy súlyt adtunk és a

$$c_1\varphi_1 + c_2\varphi_2 - c_3\varphi_3 \to \min \tag{3.1}$$

feladatot így oldottuk meg. A súlyokat három különböző módon is megválasztottuk, ezeket az összeállítás adja meg. Az egyes súlyszámok a célfüggvények relatív súlyosságát, fontosságát adják meg.

| | $c_1$ | $c_2$ | $c_3$ |
|---|---|---|---|
| 1. változat | 2 | 3 | 5 |
| 2. változat | 2 | 5 | 3 |
| 3. változat | 5 | 2 | 3 |

*b)* A *korlátok módszerénél* $\varphi_1$ és $\varphi_2$-re felső korlátokat adtunk meg, amelynél nagyobb költségeket semmiképpen sem fogadunk el. E két újabb korlátozó feltétel mellett maximalizáltuk a $\varphi_3$ célfüggvényt. Tehát a

$$\begin{aligned} \varphi_1 &\leqq \varepsilon_1 \\ \varphi_2 &\leqq \varepsilon_2 \\ \varphi_3 &\to \max \end{aligned} \tag{3.2}$$

egycélú feladatot oldottuk meg. Az $\varepsilon_1, \varepsilon_2$ korlátokat is változtattuk a futtatások során.

## 4. Kiinduló feltételezések

Kiinduló feltételezéseink közül elsősorban a termelőalapokra vonatkozó ismereteink voltak meghatározóak. A hazai szőlőterület az elmúlt évtizedben ugyanis több, mint 60 ezer ha-ral csökkent. Gond az is, hogy a szőlőültetvények korösszetétele — a korábbi rapszódikus telepítési és kivágási politika következtében — szélsőséges aránytalanságot tükröz (1. táblázat).

1. TÁBLÁZAT

*A hazai szőlőültetvények korösszetétele (1980. év)*

Me.: %

| | Korcsoportok (év) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | −5 | 6−10 | 11−15 | 16−20 | 21−25 | 26−30 | 31−35 | 35− | Összesen* |
| Állami szektor | 18 | 7 | 7 | 32 | 30 | 6 | — | — | 100 |
| Termelőszövetkezeti közös | 23 | 3 | 4 | 35 | 32 | 3 | — | — | 100 |
| Egyéb gazdaságok | 4 | 2 | 3 | 6 | 8 | 5 | 12 | 60 | 100 |

* Állami szektor, termelőszövetkezeti közös és egyéb gazdaságok.

Az 1. táblázat alapján nyilvánvaló, hogy a nagyarányú kivágások ellenére változatlanul jelentős a kiöregedő, vagy a már nem termőkorú ültetvények hányada. A telepítési feladatok egyenetlensége tehát továbbra is jellemző marad és ha ennek a telepítési politikában nem sikerül érvényt szerezni, gyorsan csökkenő területre, de alacsony termelési színvonalra is fel kell készülni.

A jelenlegi 159 ezer ha-os szőlőterület társadalmi szektoronkénti megbontását a termelési színvonalban és a költségekben jelentkező különbségek indokolják. Az említett eltérések alapján három csoportot[3] alakítottunk ki, melyek azonban a kérdésfeltevések sarkosságára is utalnak, s így a szektorbontás szerepe a jelenlegi, de még inkább a jövőbeli körülmények alapján módosulhat. Nyilvánvaló, hogy a legmagasabb színvonalú, de a területegységre jutó költségek alapján a legdrágább termelés az állami szektorban folyik. Mindez ellenkező előjellel — jelenleg még — az egyéb gazdaságokra is igaz. Hangsúlyozandó viszont, hogy az utóbbi időben telepített úgynevezett kisüzemi „tömbös" telepítések hozamai jóval meghaladják az állami szektorét is. Megítélésünk szerint azonban — legalábbis az ezredfordulóig — jellemző marad egyéb gazdaságaink átlagosan alacsonyabb termelési színvonala, de a nagyüzemekétől elmaradó termelési költsége is. Ezt igazolja a korösszetétel csak lassan változtatható arányai, a területi elhelyezkedés különbségei, valamint a közvetett költségek nagyságának szektoronkénti igen számottevő eltérései. Az említett szempontok a termésgörbék meghatározásakor és a termelési, valamint a telepítési költségek számbavételekor voltak fontosak.

Mai ismereteink szerint és körülményeink alapján gépi betakarítással nem számoltunk még az ezredforduló körül sem. A kézi szedés területenként változó lehetőségeit, illetőleg az ebből adódó magas szintű költségeket és a szektoronkénti eltéréseket a modell összeállításakor is figyelembe vettük.

Nem számoltunk a ráfordítás árak növekvő tendenciájával. Ezen igény kielégítése ugyanis lehetőségeinket meghaladó szakmai és módszertani problémákat eredményezett volna. Így az adott ráfordításárak mellett csak a hatékonyság növelésében rejlő tartalékok feltárására vállalkoztunk. Mindez természetesen nemcsak a folyó termelésre, de a beruházási tevékenységre is vonatkozik.

A szőlő-bor ágazat hazai fejlesztésében jelenleg az exportérdekeltség a meghatározó. A borkivitel — 1983-ban — meghaladta a 2,8 millió hl-t, benne több, mint 200 ezer hl pezsgővel. A szocialista országokba irányuló kivitel a meghatározó mennyiségű, a tőkés export részesedése mindössze 21% (22 millió $). Ennek ellenére a továbbiakban is változatlanul fontos feladat a tőkés export növelése, illetőleg a várhatóan bővülő szocialista kontingensek kielégítése. Mindez indokolja az export mennyiségi növelését ösztönző célfüggvény beállítását. Az exportbevétel maximalizálására kialakított célfüggvény természetesen felveti ár-, illetőleg az árfolyamalakulás módszertanilag nehezen közelíthető problémáit is. Megoldására hagyományos, kalkulatív módszert választottunk, figyelembe véve az elmúlt évtized ártendenciáit, a szocialista árképzés elveit és a nemzetközi borpiac várható helyzetét.

Kalkulációt igényelt a felhasználói igényektől való lehetséges eltérés mértékének és költségének a megállapítása is. Abból indultunk ki, hogy a felhasználás és a termelés tartós egyensúlya feltételezi, illetőleg megköveteli az eltérés lehetőségét is. Fontos hangsúlyozni, hogy a termelés egyik évről a másikra számottevően[4] ingadozik, ami

[3] Állami szektor, termelőszövetkezeti közös és egyéb gazdaságok.

[4] 1981-ben 3,9 millió hl volt a bortermés, 1982-ben pedig 6,8 millió hl.

vagy termékhiányt okoz, vagy termékfeleslegként jelentkezik. Feltételeinkben az összes lehetséges felhasználástól való eltérés maximális mértéke — 1980. és 1990. között — 30% volt, melyet a következő időszakban már 25%-ra mérsékeltünk. Az úgynevezett „termékhiány" költségét a jelenlegi importárral közelítettük, a terméstöbbletnél viszont a kényszertárolás borminőséget rontó hatását vettük figyelembe.

## 5. A szükséglet és a termelés összefüggése

A modell összeállítása során a megválaszolandó kérdések közül első helyre a következő került: mikor és mennyi szőlőt kell telepíteni, ha elsődleges cél a szolídan megfogalmazható társadalmi igények kielégítése, ugyanakkor — változó mértékben, de — fontos törekvés az export mennyiségi növelése is.

A hazai étkezési szőlő és borfogyasztás meghatározására — a távlati fejlesztésekkel foglalkozó kutatásaink alapján — a következő összeállítást készítettük:

| | Hazai borfogyasztás l/fő | Üdítőital alapanyag-igény és étkezési szőlő-felhasználás, ezer t |
|---|---|---|
| 1976—1980. | 34 | 65 |
| 1985. | 35 | 80 |
| 1990. | 35 | 80 |
| 1995. | 37 | 80 |
| 2000. | 37 | 80 |

Az így meghatározott szükségletek kielégítését a modellben egyenlőség formájában elő is írtuk.

Az export mennyiségének az alakulására viszont a számítások eredményei adnak választ. A kapott eredmények értékelésekor hangsúlyozni kell, hogy egyértelműen a súlyozásos módszer bizonyult használhatónak és szakmai szempontból kevésbé volt sikeres a korlátok módszerének az alkalmazása. Mindez elsősorban abból adódik, hogy a telepítési feladatok, legalábbis az első tíz évben — 1980—1990 — olyan mértékben behatároltak — elsősorban a korábbi telepítéspolitika következtében —, hogy a korlátok módszere a lehetséges megoldások tartományainak meghatározó részében nem eredményezett értelmezhető megoldásokat. Így a továbbiakban csak a súlyozásos módszer eredményeit ismertetjük, noha az igények által megszabott feladatok ennél a módszernél is szigorú korlátként jelentkeztek.

A súlyozásos módszer változatait[5] a modell matematikai leírásánál már ismertettük, így a továbbiakban csak a változatok számaira hivatkozunk.

[5]

| Célok: | 1. Termelési költség min. | 2. Telepítési költség min. | 3. Export bev. max. |
|---|---|---|---|
| 1. változat súlyai | 2 | 3 | 5 |
| 2. változat súlyai | 2 | 5 | 3 |
| 3. változat súlyai | 5 | 2 | 3 |

2. TÁBLÁZAT

*A rubel és a nem rubel elszámolású borexport alakulása*

| Megnevezés | 1985 | 1990 | 1995 | 2000 |
|---|---|---|---|---|
| Rubel elszámolású borexport (ezer hl) | | | | |
| 1. változat | 2000 | 2000 | 2200 | 2700 |
| 2. változat | 2000 | 2000 | 2000 | 2000 |
| 3. változat | 2000 | 2000 | 2500 | 2700 |
| Nem rubel elszámolású borexport (ezer hl) | | | | |
| 1. változat | 650 | 650 | 650 | 900 |
| 2. változat | 650 | 650 | 650 | 650 |
| 3. változat | 650 | 650 | 650 | 1150 |

A borexport lehetséges mennyiségeit mutató 2. táblázat egyrészt abból a szempontból érdemel figyelmet, hogy a 80-as években többféle változat kialakítására esélyünk sincs. Másrészt a 90-es években annak ellenére, hogy az 1. változatban jobban preferált az exportösztönzés, mégis a 3. változat exportelőirányzatai a legmagasabbak. Mindez összefügg a termelés alakulásával, illetőleg azzal, hogy még a 90-es években is a keresleti pozíció a meghatározó. Hiába ösztönözzük tehát a felhasználás bővülését, jelentős korlát még az ezredforduló táján is termőalapjaink helyzete, összetétele.

A jelenlegi ültetvényösszetétel és a várható termések kapcsolatát mutató 3. táblázat elsősorban arra hívja fel a figyelmet, hogy a 80-as évtized második felére vala-

3. TÁBLÁZAT

*Szőlőtermelés alakulása*

Me.: %

| | 1. változat | | 2. változat | | 3. változat | |
|---|---|---|---|---|---|---|
| | 1985 = 100 % | 1. vált. = 100 % | 1985 = 100 % | 1. vált. = 100 % | 1985 = 100 % | 1. vált. = 100 % |
| 1985 | 100 | 100 | 100 | 98 | 100 | 100 |
| 1990 | 99 | 100 | 89 | 88 | 100 | 100 |
| 1995 | 126 | 100 | 101 | 78 | 134 | 105 |
| 2000 | 144 | 100 | 107 | 73 | 154 | 107 |

milyen mértékű termelés visszaesés várható. (Ez természetesen a hatékonyság jelentős javítását is feltételezi.) Ugyanakkor arról is meggyőződhetünk, hogy a telepítéssel való takarékosság mértékétől függően esélyeink kedvezőek, vagy kedvezőtlenek a termelés gyorsütemű növelésére (2.—3. változat).

A termelés és a termékfelhasználás eltérései más oldalról szintén az előbbi összefüggést igazolják, főképpen azt, hogy a telepítésekkel való túlzott takarékosság hosszú távon komoly egyensúlyi problémákhoz vezethet.

A termelés és a felhasználás összefüggése alapján megállapíthatjuk tehát, hogy viszonylag szolídan megfogalmazott felhasználói célok mellett is várhatóan továbbra is gond lesz a jelenlegit jóval meghaladó exportárualap megteremtése, más szóval a kiugró évjáratoktól eltekintve továbbra sem kell tartós termékfeleslegekkel számolnunk.

4. TÁBLÁZAT

*A felhasználói célok és a szőlőtermelés eltérésének alakulása*

Me.: %

| | 1985 | 1990 | 1995 | 2000 |
|---|---|---|---|---|
| Hiány (−) többlet (+) az összes termelés %-ában | | | | |
| 1. változat | −19 | −20 | — | — |
| 2. változat | −21 | −36 | −24 | −15 |
| 3. változat | −19 | −20 | — | — |

## 6. A telepítések és a terület alakulása

A vizsgálat során a telepítéseknél nem teszünk különbséget a pótló és a bővítő jellegű beruházások között. Az új telepítés új ültetvényfelület létrehozását jelenti, függetlenül attól, hogy az a kivágásra kerülő ültetvények pótlását vagy a bővített újratermelés céljait szolgálja. A szükséges adatok híján nem foglalkozunk a meglevő ültetvények korszerűsítésének és részleges felújításának kérdéseivel.

Elsőként — a súlyozásos módszer eredményei alapján — a telepítési feladatok időszakonkénti megoszlását mutatjuk be (5. táblázat).

5. TÁBLÁZAT

*Az összes telepítés időszakonkénti megoszlása*

Me.: %

| | 1981—1985 | 1986—1990 | 1991—1995 | 1996—2000 | Együtt |
|---|---|---|---|---|---|
| 1. változat | 30 | 35 | 15 | 20 | 100 |
| 2. változat | 28 | 30 | 10 | 32 | 100 |
| 3. változat | 25 | 30 | 20 | 25 | 100 |

Egyértelmű tehát a feladatok megoszlásának egyenetlensége, ami különösen a 80-as és a 90-es évtized fordulóján jelentkezik nagy aránytalanságokkal.

Más oldalról ezt igazolja az az összeállítás is, amelyben csak egy-egy célkitűzés érvényesítésének hatását vizsgáltuk.

6. TÁBLÁZAT

*Az összes telepítés időszakonkénti megoszlása*
(egy célfüggvénnyel való optimalizálás mellett)

Me.: %

| | 1981—1985 | 1986—1990 | 1991—1995 | 1996—2000 | Együtt |
|---|---|---|---|---|---|
| Termelési költség minimalizálás | 22 | 44 | — | 34 | 100 |
| Telepítési költség minimalizálás | 17 | 32 | — | 51 | 100 |
| Export bevétel maximalizálás | 21 | 50 | 10 | 19 | 100 |

## 7. TÁBLÁZAT

*A telepítések és a kivágások szektoronkénti alakulása*

Me.: %

| Megnevezés | 1. változat | | | | 2. változat | | | | 3. változat | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1981—1985 | 1986—1990 | 1191—1995 | 1996—2000 | 1981—1985 | 1986—1990 | 1991—1995 | 1996—2000 | 1981—1985 | 1986—1990 | 1991—1995 | 1996—2000 |
| Állami szektor | | | | | | | | | | | | |
| Telepítés: | | | | | | | | | | | | |
| 1981—2000=100% | 25 | 37 | 12 | 26 | 25 | 38 | 13 | 24 | 35 | 35 | 20 | 10 |
| Összes telepítés=100% | 30 | 50 | 40 | 40 | 50 | 65 | 60 | 50 | 30 | 25 | 20 | 15 |
| Kivágás: | | | | | | | | | | | | |
| 1981—2000=100% | 2 | 32 | 32 | 34 | 2 | 32 | 40 | 26 | 17 | 30 | 30 | 23 |
| Összes kivágás=100% | 5 | 15 | 25 | 20 | 5 | 15 | 30 | 30 | 5 | 15 | 25 | 20 |
| Termelőszövetkezeti közös | | | | | | | | | | | | |
| Telepítés: | | | | | | | | | | | | |
| 1981—2000=100% | 29 | 29 | 14 | 28 | 24 | 28 | 10 | 38 | 20 | 30 | 20 | 30 |
| Összes telepítés=100% | 30 | 30 | 30 | 40 | 15 | 15 | 20 | 25 | 40 | 45 | 40 | 50 |
| Kivágás: | | | | | | | | | | | | |
| 1981—2000=100% | 1 | 35 | 35 | 29 | 8 | 40 | 40 | 12 | 10 | 35 | 35 | 20 |
| Összes kivágás=100% | 5 | 25 | 45 | 30 | 5 | 25 | 40 | 30 | 5 | 25 | 45 | 40 |
| Egyéb gazdaságok | | | | | | | | | | | | |
| Telepítés: | | | | | | | | | | | | |
| 1981—2000=100% | 36 | 36 | 14 | 14 | 40 | 30 | 5 | 25 | 25 | 25 | 25 | 25 |
| Összes telepítés=100% | 40 | 20 | 30 | 20 | 35 | 20 | 20 | 25 | 30 | 30 | 40 | 35 |
| Kivágás: | | | | | | | | | | | | |
| 1981—2000=100% | 28 | 40 | 10 | 22 | 30 | 42 | 10 | 18 | 30 | 40 | 10 | 20 |
| Összes kivágás=100% | 90 | 60 | 30 | 50 | 90 | 60 | 30 | 40 | 50 | 60 | 30 | 40 |

Az eredmények értékelésekor hangsúlyozni kell, hogy ilyen szélsőséges helyzet, amikor egyetlen célkitűzés határozná meg az ágazat hosszútávú fejlesztésének irányát és mértékét, a valóságban nem fordulhat elő. Abból a szempontból viszont érdekes, hogy egyik vagy másik célkitűzés előtérbe kerülése — a többi rovására — jól látható torzulásokat okoz a telepítéspolitikában és ez gátolja a termelés-felhasználás egyensúlyának a megteremtését. Nem képzelhető el ugyanis — vagy szükséges lenne elkerülni —, hogy az egyik ötéves időszak több tízezer ha-os telepítéseit követően a következő öt évben semmit, vagy csak nagyon szerény mértékben telepítsünk szőlőültetvényeket. Az aránytalan korösszetétel komoly gondokat okoz a bővített újratermelés folyamatában (megoldhatatlan pótlási feladatok elé állítva a beruházókat), gátolja a hatékonyság folyamatos javítását és felerősíti az amúgy is sok problémát okozó termésingadozásokat.

A telepítések és a kivágások szektoronkénti összetételére csak a súlyozásos módszer adott értékelhető eredményeket (7. táblázat). A telepítési feladatok időszakonkénti megoszlása társadalmi szektoronként nem különbözik lényegesen. Jellemző a 80-as évek végére növekvő telepítési igény és az az összefüggés, hogy a 80-as évek telepítései minden esetben meghaladják a 90-es évekét. A kivágásoknál az állami szektor és a termelőszövetkezeti közös gazdaságoknál 1985-től jelentkeznek a komolyabb feladatok, ami sajnos előidézője a már említett telepítési „hullámnak". Az egyéb gazdaságok kivágásai viszont már a 80-as évek első felében is jelentősek. Az összes telepítés szektoronkénti arányait a nagyüzemek növekvő feladatai szabják meg. Következik ez a magasabb termelési színvonalból és a feltételezhetően javuló hatékonyságból is.

A telepítések és a kivágások eldöntik a szőlőterület és a termelési színvonal alakulását is (8., 9. táblázat).

## 8. TÁBLÁZAT

*Az eltérő fejlesztési követelmények hatása a szőlőterület alakulására*

| Megnevezés | Összes szőlőterület ezer ha | Ebből: | | |
|---|---|---|---|---|
| | | állami szektor | termelőszövetkezeti közös | egyéb gazdaság |
| | | ezer ha | | |
| 1985* 1. változat | 165,6 | 33,0 | 47,3 | 85,3 |
| 2. változat | 155,6 | 33,0 | 40,3 | 82,3 |
| 3. változat | 158,6 | 33,0 | 40,3 | 85,3 |
| 1990 1. változat | 138,0 | 40,4 | 42,3 | 55,3 |
| 2. változat | 116,5 | 40,4 | 28,8 | 47,3 |
| 3. változat | 143,0 | 40,4 | 47,3 | 55,3 |
| 1995 1. változat | 119,4 | 37,8 | 32,3 | 49,3 |
| 2. változat | 90,9 | 37,8 | 14,8 | 38,3 |
| 3. változat | 135,4 | 37,8 | 42,3 | 55,3 |
| 2000 1. változat | 135,5 | 46,1 | 41,1 | 48,3 |
| 2. változat | 103,0 | 46,1 | 18,6 | 38,3 |
| 3. változat | 150,5 | 34,1 | 56,1 | 60,3 |

* 1982-ben az összes szőlőterület 159 ezer ha volt.

9. TÁBLÁZAT

*A termelési színvonal és az eltérő fejlesztési követelmények kapcsolata*

Me.: t/ha

| Megnevezés | 1981*—1985 | 1986—1990 | 1991—1995 | 1996—2000 |
|---|---|---|---|---|
| Állami szektor | | | | |
| 1. változat | 7,9 | 8,1 | 10,1 | 10,8 |
| 2. változat | 7,9 | 8,1 | 10,1 | 10,8 |
| 3. változat | 7,9 | 7,8 | 10,1 | 10,4 |
| Termelőszövetkezeti közös | | | | |
| 1. változat | 6,3 | 5,7 | 8,9 | 9,1 |
| 2. változat | 5,3 | 5,9 | 8,4 | 8,4 |
| 3. változat | 6,3 | 7,4 | 8,7 | 9,3 |
| Egyéb gazdaságok | | | | |
| 1. változat | 4,6 | 6,1 | 7,2 | 8,0 |
| 2. változat | 4,6 | 6,1 | 6,9 | 7,3 |
| 3. változat | 4,6 | 6,1 | 7,2 | 7,9 |

* 1980—1982. években a szőlőtermelés átlaghozamai:
állami szektorban 6,7 t/ha
termelőszövetkezeti közösben 6,0 t/ha
és az egyéb gazdaságokban 4,4 t/ha.

A szőlőterületre általában jellemző csökkenő tendencia arra utal, hogy a 80-as évek második felére összegyűlő pótlási feladatainkat a következő évtized alatt sem tudjuk maradéktalanul teljesíteni. A területcsökkenős elsősorban az egyéb gazdaságokat érinti, ami azonban a termelési színvonal jelentős javulását ígéri.

Az elmondottak alapján megállapítható tehát, hogy a csak szerényen növekvő igények folyamatos kielégítése érdekében a jövőben gyorsabb ültetvényrotációra lesz szükség. Ugyanakkor a mai korösszetételből adódóan a telepítések és a kivágások nagyságának időszakok közötti jelentős mértékű ingadozásával is számolni kell. E két tényező együttes hatásaként a következő évtizedekben nehezen megoldható telepítési feladatok várnak a szőlőtermelő gazdaságokra. Gond az is, hogy az ingadozó telepítési igény közvetítésére megfelelő gazdaságszabályozási gyakorlattal sem rendelkezünk.

## 7. Összefoglalás

A hazai szőlőtermelés fejlesztési irányainak a meghatározása a piaci igények ismerete mellett egy sor olyan kérdést is felvet, ami összefügg a termelőalapok jelenlegi összetételével, alakításának problémáival. Vizsgálatunk elsősorban ezekre a kérdésekre kereste a választ a három célfüggvénnyel leírt optimalizálási feladat súlyozásos módszerrel való megoldása révén.

Elemzésünk főbb szakmai tanulságait a következőkben foglalhatjuk össze:

— Az 1981. és 2000. közötti időszakban 50—80 ezer ha szőlőt kell telepíteni, ha a mérsékelten növekvő szükségletek folyamatos kielégítését biztosítani akarjuk. Ugyanakkor a kivágások 90—100 ezer ha körül alakulhatnak. A telepítések nagyságát elsősorban a nagyüzemi feladatok szabják meg, míg a nagyarányú kivágásokra a 80-as évek első felében csak az egyéb gazdaságoknál, ezt követően azonban minden szektorban fel kell készülni.

— Általános tendencia a szőlőterület további jelentős csökkenése. Ezzel különösen akkor kell számolni, ha telepítésekre fordítható erőforrásokkal való takarékosság az elsődleges célkitűzés. A terület csökkenéssel párhuzamosan feltételezhető, hogy a hatékonyág, különösen a következő évtizedben jelentősen javul. Ugyanakkor az is beigazolódott, hogy a telepítésekkel való túlzott takarékosság egyben a termelési színvonal (átlagtermések) legalacsonyabb szintjét is eredményezi.

— Némileg módosulnak a társadalmi szektorok arányai is. Az ültetvényterületek alapján az egyéb gazdaságok jelenlegi 60%-os részesedése várhatóan 40—45%-ra mérséklődik. Ugyanakkor attól függően, hogy melyik célkitűzés preferáltabb érvényesítéséről van szó, növekedhet az állami szektor, vagy a termelőszövetkezeti gazdaság szerepe.

— A súlyozásos módszer egyértelműen bizonyította, hogy a 80-as években a borexport jelentős mennyiségi fejlesztésére nincs lehetőség. A termelőalapok helyzete, a lökésszerűen jelentkező telepítési — és kivágási — igény a termelés oldaláról korlátozza, illetve behatárolja a felhasználói célokat. A 90-es években viszont már jelentős különbségek tervezhetők az export mennyiségében és némileg javítható a hazai fogyasztás szintje is.

— A termelés növekedése természetesen az összes felmerülő költségeket is módosítja, melyek időszakonkénti alakulását a következő összeállításunk mutatja (me.: 1985=100%):

| | 1. változat | 2. változat | 3. változat |
|---|---|---|---|
| 1985. | 100 | 100 | 100 |
| 1990. | 100 | 100 | 100 |
| 1995. | 182 | 177 | 182 |
| 2000. | 133 | 200 | 133 |

A termelési költségek a termék „hiány" költségeit is tartalmazzák, s ez választ ad a 2. változat gyorsütemű költségnövekedésére. Az alacsony termelési színvonal, a drága termelés ez a következő évtized legfőbb öröksége a telepítéssel való túlzott takarékosság esetén.

## IRODALOM

[1] HWANG, C. I. and MASUD, A. S. M., *Multiple Objective Decision Making—Methods and Applications* (Lecture Notes in Economics and Mathematical Systems, No. **164**, Springer Verlag, Berlin—Heidelberg—New York, 1979).

[2] SZIDAROVSZKY, F. and YAKOWITZ, S., *Principles and Procedures of Numerical Analysis* (Plenum, New York, 1978).

(Beérkezett: 1984. május 23.)

FERENCZY ANTAL ÉS SZIDAROVSZKY FERENC
KERTÉSZETI EGYETEM, MATEMATIKAI ÉS SZÁMÍTÁSTECHNIKAI TANSZÉK
1502 BUDAPEST, VILLÁNYI ÚT 29—35.

URBÁN ANDRÁS
KERTÉSZETI EGYETEM, BORÁSZATI TANSZÉK
1502 BUDAPEST, VILLÁNYI ÚT 29—35.

PAPP ZSOLT
AGRÁRGAZDASÁGI KUTATÓ INTÉZET
1093 BUDAPEST, ZSIL U. 3—5.

# POSSIBILITIES AND BOUNDS FOR GRAPE PRODUCING UP TO 2000

A. FERENCZY, ZS. PAPP, F. SZIDAROVSZKY, A. URBÁN

In this paper an optimization model for developing and operating a large scale agricultural system is examined. The elements of the system are grape producing farms and wine making facilities. Three objective functions are considered: operating cost, investment cost and export profit. The mostly known solution algorithms (such as weighting, $\varepsilon$-constraint, goal programming methods) are applied and the numerical solutions are compared.

*Alkalmazott Matematikai Lapok* **10** (1984) 405—411

# A FA BIOGEOCÖNÓZISÁNAK SZIMULÁCIÓS MODELLJE

RACSKÓ PÉTER

Budapest

A tanulmány a fa, mint az erdei biogeocönózis elemének szimulációs modelljét írja le és ismertet néhány, számítógépes kísérletből levonható következtetést. Az ismertetett modell lényegében más, mint a növényi kultúrák fejlődésének vizsgálatára készült, biológiai megfigyelések, mérések formalizált eredményeit tartalmazó leíró jellegű modellek. Itt a növekedés dinamikáját a biogeocönózisos anyagkörforgás egyensúlyi egyenletei és a fa adaptációképességét realizáló, a képzett asszimilátumok elosztását szabályozó hipotetikus mechanizmus határozzák meg.

A modell — jellegénél fogva — alkalmas a természetes környezeti paraméterektől való lényeges eltérés esetén is (pl. mesterséges művelés) a fejlődés dinamikájának leírására.

## 1. Bevezetés

A világon komoly erőfeszítéseket folytatnak a Föld bioszférájának, mint az emberiség életterének kutatására. Ezeket a kutatásokat az motiválja, hogy jelenleg az emberiség rendelkezik olyan erőforrásokkal, amelyek felhasználásának hatására a bioszféra kimozdulhat az évmilliók során kialakult homeosztatikus állapotából és az emberiség számára jóval kedvezőtlenebb állapotba mehet át. A széndioxid és hőkibocsátás emelheti a Föld átlaghőmérsékletét, ennek hatására csökkenhet a sarki hó és jégtakaró területe, csökken a visszavert napenergia mennyisége, ami további felmelegedést okozhat, stb. [6].

Az intenzív mezőgazdasági monokultúrák gazdaságilag indokolt terjedése, a környezetszennyezés, urbanizáció a kialakult faji egyensúlyt megbonthatja és ez a teljes biomassza csökkenését eredményezheti. Az utóbbi évek kutatási eredményei megmutatták, hogy a természetes biogeocönózisok bizonyos mértékig stabilizáló hatást mutatnak a bioszféra egyensúlyát megbontó törekvésekkel szemben. Például, az atmoszféra széndioxidtartalma nem növekszik olyan mértékben, mint az az antropogén eredetű kibocsátásból következne. Ennek az a magyarázata, hogy a nagyobb $CO_2$ koncentráció kedvezőbb feltételeket teremt a fotoszintetikus folyamatokhoz, ami a széndioxid lekötés fokozódásához vezet. Ebben a stabilizáló folyamatban (de egyéb folyamatokban is) igen jelentős szerepet játszik a Föld erdőkészlete.

Nagy jelentősége van ezért azoknak a kutatásoknak, amelyek a fának, mint az erdei biogeocönózis alapelemének megismerésére irányulnak. A fentiek mellett nem elhanyagolható szempont az sem, hogy a fontos ipari nyersanyag és a fakészletek fenntartása, sőt növelése gazdasági érdek.

A világon szinte mindenütt folyó statisztikai leíró jellegű vizsgálatok, amellett, hogy igen gazdag tényanyagot állítanak elő, nem alkalmasak arra, hogy választ adjanak egy sor kérdésre, ok-okozati összefüggésre.

Az erdei biogeocönózist számos endogén és exogén hatás befolyásolja, mint pl. a fotoszintetikailag aktív sugárzás mennyisége, klimatikus viszonyok (csapadék és hőmérséklet nagysága és időbeni megoszlása) tápanyagok jelenléte a talajban, stb. Ezek a paraméterek adott környezetben viszonylag szűk határok közé szorulnak és így a naturális megfigyelések nem adhatnak választ arra, hogy hogyan reagál egy egyed ezeknek a paramétereknek a „megszokottnál" lényegesen nagyobb eltolódására, melyek a külső és belső paraméterek összefüggései, stb. Szép számmal léteznek analitikus modellek, amelyek bizonyos feltételezések mellett — korlátlan tápanyag rendelkezésre állása, a gyökér, törzs, levélzet tömegének arányos növekedése, a környezeti feltételek állandósága — fenomenologiai leírást adnak. Ezek összefoglalása megtalálható például [7]-ben. Ross [8] az alábbi differenciálegyenletrendszerrel írja le a fa részeinek növekedését:

$$\dot{m}_j = k_F \sum_{i=1}^{4} \alpha_{ij}\varphi_i - k_R R_j - V_j + M \sum_{i=1}^{4} \beta_{ij} \quad j = 1, \ldots, 4, \tag{1.1}$$

ahol:

$m_j$ — a fa részeinek biomassza tömege (gyökér, törzs, lombozat, reproduktív szervek) $j=1, \ldots, 4$

$M$ — az egész növény száraz tömege

$\varphi_i$ — a növény egyes szervei által egy nap alatt felhasznált széndioxid mennyisége, $k_F\varphi_i$ — az ennek megfelelő száraz biomassza $i=1, \ldots, 4$

$R_i$ — a fa $i$-edik szerve által egy nap alatt kiválasztott (kilélegzett) széndioxid tömege, $k_R R_i$ — az ennek megfelelő száraz biomassza, $i=1, \ldots, 4$

$V_i$ — a fa $i$-edik szervének egy nap alatti veszteség tömege (elhullás, elhalás következtében) $i=1, \ldots, 4$

$\alpha_{ij}$ — az egy nap alatt az $i$-edik szervben keletkezett „friss" asszimilátum mennyisége, amely a $j$-edik szervbe épül be

$\beta_{ij}$ — a „régi" asszimilátumok átrendezését leíró paraméterek (azaz az egy nap alatt a $j$-edik szervből az $i$-edik szervbe átfolyó asszimilátumok mennyisége egy egységnyi biomasszára vonatkoztatva) $i=1, \ldots, 4$.

A modell, minőségileg különbözik a tisztán a megfigyelési eredményeket approximáló analitikus leírásoktól, nem veszi azonban figyelembe a külső tényezők növekedést befolyásoló hatását, az $\alpha_{ij}$ és $\beta_{ij}$ paraméterek értékét pedig (vagyis a tényleges növekedési törvényeket) kísérletileg kell meghatározni.

Az alábbiakban olyan szimulációs modellt írunk le, amely lehetőséget ad a fa, mint az erdei biogeocönózis rendszerelemének vizsgálatára.

A modellel végzett kísérletek lehetővé teszik, hogy az egyedfejlődést a külső paraméterek tetszőlegesen változó értékei mellett leírjuk, megállapítsuk az adaptációs határokat (stabilitási tartományt), illetve beavatkozási stratégiákat szimuláljunk.

## 2. A modell leírása

Az alábbi alapvető környezeti tényezők befolyásolják az erdei biogeocönózis működését:

— fotoszintetikailag aktív sugárzás (FAR)
— a levegő széndioxid koncentrációja
— a csapadék mennyisége és időbeni eloszlása

— a levegő hőmérséklete

— a talaj tápanyagkészlete (nitrogén, foszfor, kálium, nátrium, stb.).

Ezek közül az első négy közvetlenül az idő függvénye, míg az utolsó általában véges mennyiségű „készletként" van jelen, amely csak az élő biomassza elhullásából és bomlásából termelődik újjá.

A tápanyagok növekedésszabályozó szerepét a biológiában általánosan elfogadott elv szerint, az ún. Liebig törvény [1] alapján modellezzük. Ez a törvény azt tételezi, hogy a növekedést mindig a minimális mennyiségben jelen levő elem korlátozza. (Itt elemen a tápanyagokat értjük.)

A növekedés szabályaira, az új asszimilátumok elosztására biológiai szempontból igazolható hipotéziseket kellett felállítanunk, hogy a modell zárt legyen. Ezek az alábbiak:

H1: Az új biomassza mennyisége, amely $y(t)$ sebességgel képződik, függ mind a $\lambda_k(t)$ környezeti — mind az egyes szervek $t$ időpontban mért $x_i(t)$ tömegétől és a fa szervei között bizonyos $\tau$ késleltetéssel oszlik szét.

ahol: $\lambda_1(t)$ — FAR intenzitás, $\lambda_2(t)$ — a csapadékeloszlást leíró paraméter, $\lambda_3(t)$ — $CO_2$ koncentráció, $\lambda_4(t)$ — levegőhőmérséklet.

$\lambda_2(t)$ pontos leírásától és a mértékegységek ismertetésétől eltekintünk (l. [9]).

$x_1(t)$ — a levélzet tömege, $x_2(t)$ — törzstömege, $x_3(t)$ — gyökérzet tömege.

A reproduktív szervek tömegét a modellben nem tárgyaljuk, mert az erdei ökológiai rendszereknél ez a mi szempontunkból jelentéktelen.

Most már felírhatjuk az anyagmegmaradás elvén alapuló egyensúlyi differencia egyenletrendszert:

$$(2.1)\quad x_i(t+\tau) = x_i(t) + e_i(t) y[x_1(t), x_2(t), x_3(t), \lambda_1(t), \lambda_2(t), \lambda_3(t), \lambda_4(t)]\tau \quad i = 1, 2, 3,$$

ahol

$\tau$ — az alap időintervallum

$$t = m\tau; \quad m = 1, \ldots, N$$

$e_i(t) \geqq 0$ minden $i = 1, 2, 3$ és $t$ esetén

$$\sum_{i=1}^{3} e_i(t) = 1.$$

A kezdeti feltételek:

$$(2.2)\qquad x_i(0) = x_i^0 \quad (i = 1, 2, 3).$$

A fenti egyenletrendszer megoldásához szükséges az $e_i(t)$ $i = 1, 2, 3$ vezérlési függvények megadása a $t = m\tau$, $m = 1, \ldots, N$ pontokban. Ezek a vezérlési függvények határozzák meg a „növekedési szabályt".

Az új asszimilátumok elosztását meghatározó $e_i(t)$ $(i = 1, 2, 3)$ függvények megadása központi kérdés, erre vonatkozóan mérési eredmények nem állnak rendelkezésre.

Az alábbiakban megfogalmazzuk azt a biológiai szempontból indokolt hipotézist, amely közvetlenül vezet $e_i(t)$ $(i = 1, 2, 3)$ meghatározásához.

H2: Az újonnan képződött biomassza olymódon oszlik szét a fa gyökerei, törzse és levélzete között, hogy a következő időszakban a fa összes biomasszájának maximális növekedését biztosítsa, amennyiben a külső feltételek nem változnak.

A hipotézist a maximális primér produktivitás elvének nevezzük.

Megjegyezzük, hogy hasonló elvek alkalmazása a természettudományokban meglehetősen elterjedt. (Pl. *Fermat-elv* az optikában, *Maupertuis—Lagrange elv* a klasszikus mechanikában, vagy az optimális konstrukció elve a biológiában.) Az alábbiakban formalizáljuk a H2 hipotézist.

A $t+\tau$ időpillanatban a környezetet a $(\lambda_j(t+\tau))=(\lambda_1(t+\tau), \lambda_2(t+\tau), \lambda_3(t+\tau), \lambda_4(t+\tau))$ vektor írja le, így, egyelőre feltételezve, hogy tápanyag korlátlan mennyiségben áll rendelkezésre a talajban, az új biomassza képződésének sebessége a $t+\tau$ időpontban

$$(2.3)\quad y(t+\tau) = y[x_1(t+\tau), x_2(t+\tau), x_3(t+\tau), \lambda_1(t+\tau), \lambda_2(t+\tau), \lambda_3(t+\tau), \lambda_4(t+\tau)].$$

Feltételezve, hogy a környezeti feltételek nem változnak, azaz $\lambda_i(t+\tau)=\lambda_i(t)$, $i=1, \ldots, 4$, ez a sebesség:

$$(2.4)\quad \tilde{y}(t+\tau) = y[x_1(t+\tau), x_2(t+\tau), x_3(t+\tau), \lambda_1(t), \lambda_2(t), \lambda_3(t), \lambda_4(t)].$$

Az $\tilde{y}(t+\tau)$ jelölést azért vezettük be, mert ez nem a tényleges primér produktivitás sebessége, hanem csupán a $(\lambda_j(t+\tau))=(\lambda_j(t))$ feltétel fennállása esetén fellépő sebesség. A feltétel általában nem teljesül, azonban nem tételezhetjük fel, hogy a fa az új biomassza képzésének időpontjában „prognosztizálni tudja" $(\lambda_j(t+\tau))$-t. Használva az $(x_j(t))=(x_1(t), x_2(t), x_3(t))$; $(\lambda_j(t))=(\lambda_1(t), \lambda_2(t), \lambda_3(t), \lambda_4(t))$ jelölést, behelyettesítve (2.1)-et a (2.4)-be kapjuk, hogy:

$$(2.5)\quad \tilde{y}(t+\tau) = y\{(x_j(t)+e_j(t)\, y[x_j(t), \lambda_k(t)]\tau), (\lambda_k(t))\} \quad (j=1, 2, 3;\ k=1, 2, 3, 4).$$

A H2 hipotézis alapján az $e_j(t)$ $(i=1, 2, 3)$ értéke a $t$ időpontban az

$$(2.6)\quad \tilde{y}^*(t+\tau) = \max_{e_j(t)} y\{(x_j(t)+e_j(t)\, y[x_j(t), \lambda_k(t)]\tau), (\lambda_k(t))\}$$

$$\sum_{j=1}^{3} e_j(t) = 1;\ e_j(t) \geqq 0 \quad (j=1, 2, 3)$$

feltételekből határozható meg.

Ezek után, ha az $y((\lambda_j(t)), (x_k(t)))$ $(j=1, 2, 3, 4,\ k=1, 2, 3)$ függvényt explicit formában felírjuk, akkor a (2.1) egyenletrendszer a (2.2) kezdeti feltételekkel és a (2.6) előírással együtt meghatározza a rendszer dinamikus viselkedését egy vegetációs időszakban. Az $y$ függvény explicit formájának megadását itt nem részletezzük, ez megtalálható [9]-ben.

A reális modellben természetesen nem feltételezhetjük a tápanyagok korlátlan rendelkezésre állását.

Legyen $z_1(t), z_2(t), z_3(t), z_4(t)$ a fa számára elérhető nitrogén, foszfor, kálcium és nátrium mennyisége a talajban a $t$ időpontban. Legyen $\delta_i$ — az $i$-edik tápanyagnak az elhalt szervesanyag bomlása útján történő „visszatérési ideje" a rendszerbe. (A forgási sebesség reciproka.)

Legyen $w_1$ — a vegetációs periódus végén a lombozat biomasszájának csökkenését, $w_3$ — a gyökérzet csökkenését leíró koefficiens, $\varepsilon$ pedig a rendszerből eltűnő tápanyag hányad. (Pl. a talajvízzel.) A $\delta_i$, $w_i$, $\varepsilon$ értékeket jó közelítéssel konstansnak tekinthetjük. Legyen $p_{il}$ $(i=1, 2, 3,\ l=1, 2, 3, 4)$ az $l$-edik tápanyag mennyisége a fa $i$-edik részében, $q_l$ pedig az $l$-edik tápanyag mennyisége az új, még el nem osztott

biomasszában. Ekkor a tápanyagkörforgás egyensúlyi egyenletrendszere:

$$z_l(t+\tau) = z_l(t)+x_1(t-\delta_l)w_1 p_{1l} u(t)+x_3(t-\delta_l)w_3 p_{3l} u(t)- \tag{2.7}$$

$$-\tilde{y}[(x_j(t)),(\lambda_k(t))]\tau q_l-\varepsilon z_l(t)u(t)$$

$$i=1,2,3;\quad k=1,2,3,4;\quad l=1,2,3,4;\quad t>\delta_l;\quad t=m\tau$$

$$u(t)=\begin{cases}0, & \text{ha} \quad t\neq n\cdot n_1\tau\\ 1, & \text{ha} \quad t=n\cdot n_1\tau\end{cases}$$

$n_1\tau$ — a vegetációs periódus hossza, $n=1,2,\ldots$.
Az $\tilde{y}$ függvényt a már ismertetett *Liebig-elv* alapján definiáljuk:

$$\tilde{y}(t)=\min\left[y(t),\frac{z_1(t)}{q_1},\frac{z_2(t)}{q_2},\frac{z_3(t)}{q_3},\frac{z_4(t)}{q_4}\right]. \tag{2.8}$$

A (2.7) rendszer kezdeti feltételei:

$$z_l(0)=z_l^0,\ \ldots,\ z_l(N_l n_1\tau)=z_l^{(N_l)}\quad (l=1,2,3,4). \tag{2.9}$$

Ha még a (2.1) egyenleteket kiegészítjük egy, a biomassza elhalását leíró tényezővel, akkor ezek az egyenletek a növekedési dinamikát több vegetációs perióduson keresztül is leírják:

$$x_i(t+\tau)=\{(x_i(t)+e_i(t)\tilde{y}[x_j(t),\lambda_k(t)]\tau)\}[1-w_i u(t)] \tag{2.10}$$

$$t=m_\tau;\quad m=1,\ldots,N;\quad i=1,2,3;\quad j=1,2,3;\quad k=1,2,3,4;$$

$$\sum_{i=1}^{3} e_i(t)=1,\quad e_i(t)\geqq 0\quad \text{minden } t\text{-re.}$$

Ezzel a modell felállítását befejeztük, a (2.10) és (2.7) egyenletek a (2.2) és (2.9) kezdeti feltételekkel, az $e_i(t)$-t meghatározó (2.6), valamint az $\tilde{y}$-t meghatározó (2.8) összefüggésekkel zárt modellt alkotnak.

### 3. A modell identifikálása

A modell identifikálását csak kellő mélységben ismert valós rendszeren lehet elvégezni. E célból a karéliai fenyveserdők környezetét választottuk, mert itt álltak rendelkezésre megfelelő mérési adatok [4], [5].

Az $y(t)$ függvény analitikus kifejezésekhez szükséges paraméterek (légzést kifejező koefficiensek, molekuláris diffúziós koefficiensek stb.) értékét kísérleti eredményeket leíró monográfiákból vettük [2], ahol pedig nem álltak rendelkezésre adatok, a paraméterértékeket biológiailag „valószínűsíthető" intervallumokból választottuk. A mintegy 40 biológiai jellegű paraméter meghatározása után különböző kezdeti feltételrendszereknél végeztünk számítógépes szimulációs kísérleteket két, alapvetően eltérő feltételezéssel. Az első esetben a tápanyagok mennyisége nem korlátozott, a másodikban igen.

A napi bológiai ciklus alapján a $\tau$ lépésközt 1 napnak választottuk, ennél rövidebb, vagy hosszabb intervallum használatát semmi nem indokolta.

Minthogy a (2.6) optimalizálási feladatot minden egyes lépésnél meg kell oldani és a szimuláció célja több vegetációs periódus, azaz többszáz vagy inkább többezer napi ciklus leírása, az optimumszámítást csak rendkívül egyszerű és gyors eljárással lehet elvégezni. Megjegyezzük, hogy az optimumot adó $e_i^*(t)$ $(i=1, 2, 3)$ és az optimális $\tilde{y}^*(t)$ számításánál egyébként sem törekedhetünk nagy pontosságra, mert $y(t)$ explicit kifejtésében többször tíz olyan paraméter szerepel, amelyek értéke csak közelítően ismert. $y(t)$ ezzel szemben többször differenciálható, mind a paraméterek, mind a változók szempontjából „jól viselkedő" függvény, így indokolt, hogy $\tilde{y}^*(t)$ maximumának egy kellő sűrűségű háló pontjaiban számított maximális értékét tekintsük.

A modellt sikeresen identifikáltuk, 3 vegetációs periódus (év) alatt a mért adatokból való eltérés nem haladja meg a 10%-ot pozitív irányban. (Ez az eltérés is jórészt arra vezethető vissza, hogy az identifikáció során végzett számítógépes kísérleteknél a tápanyagot nem korlátoztuk, mert a reál-rendszerből származó adatok ehhez nem álltak rendelkezésre.)

## 4. A modellkísérletek

A számítógépes kísérletek gépidőigénye viszonylag jelentős, egy-egy több évre végzett számítás kb. 5—10 perc nagyszámítógépes gépidőt igényel. Ezért elsősorban nem pontos prognózisok készítését tűztük ki célul, hanem „minőségi" kísérleteket végeztünk.

### *4.1. A modell kezdeti feltételek szerinti stabilitása*

Itt azt vizsgáltuk, hogy hogyan folyik le a fa növekedése különböző $x_1^0, x_2^0, x_3^0$ kezdeti feltételeknél, míg a többi paraméter változatlan. A kísérletek azt — a nem is meglepő — eredményt adták, hogy létezik olyan $\Omega_A(x_1^0, x_2^0, x_3^0)$ kezdeti feltétel halmaz, amelyen kívül a fa már nem képes növekedni — elhal — míg a halmazon belül a fa életképes marad és egyedfejlődése során egy „átlagos" fához konvergál. Például, ha egy fa gyökértömegének felét eltávolítjuk, a fa életképes marad és ezt néhány év alatt helyreállítja.

Érdekes volna megfigyelni, hogy $\Omega$ hogyan függ a környezeti tényezőktől, $\Omega=\Omega(\lambda_1(t), \lambda_2(t), \lambda_3(t), \lambda_4(t))$. A pontos függés leírása igen gépidőigényes lenne.

### *4.2. A modell paraméterérzékenysége*

A kísérletekkel sikerült kijelölni azoknak a paramétereknek a körét, amelyek kis változása is lényegesen befolyásolja a fa fejlődését. Ide tartoznak pl. a $w_1, w_2, w_3$ koefficiensek. Ezek pontosabb meghatározása feltétele a pontos prognosztikai jellegű kísérleteknek.

### *4.3. A külső feltételek hatásának mérése*

Több kísérletet végeztünk mind a klimatológiai, mind a rendelkezésre álló tápanyagmennyiség változásából következő hatások megfigyelésére. Ezek leírása megtalálható [9]-ben. Itt csak egy-két eredményt említünk az érdekesség kedvéért.

A megnövekedett széndioxid koncentráció gyorsabb növekedést vált ki. (Megjegyezzük, hogy így több széndioxidot is köt le a fa, ami némiképp ellensúlyozza az iparilag kibocsátott széndioxidmennyiség növekedését.) Ez az eredmény megegyezik a [6] monográfiában leírtakkal, bár ott egészen más a megközelítési mód. A növekedés szempontjából lényeges, hogy adott mennyiségű vizet milyen „ütemezésben" kap a fa. A gyakori, periodikus utánpótlás sokkal kedvezőbb, mint az egyszeri, nagymennyiségű víz bevezetése a talajba. (Csepegtetéses öntözés!)

A külső átlag hőmérséklet kismértékű csökkenése nagyobb törzstömegű, de kisebb lombtömegű fákat eredményez. A hőmérséklet további csökkenése kis méretet, lassú növekedést idéz elő. Még további csökkenése a fa elhalásához vezet.

## IRODALOM

[1] Liebig, J., *Chemistry in its Application to Agriculture and Physiology* (Taylor and Walton, London, 1847).
[2] Moldau, H., "Model of plant productivity at limited water supply considering adaptation", *Photo-synthetica* **5** (1971) 16—21.
[3] Rashevsky, N., *Mathematical Biophysics* (Dover, NY. vol. 2. 1960, 292).
[4] Казимиров, Н. И., Ельники, Карелии. Наука, Ленинград, 1971.
[5] Казимиров, Н. И., Морозова, Р. М., Биологический круговорот веществ в ельниках Карелии. Наука, Ленинград, 1973.
[6] Крапивин, В. Ф., Свирежев, Ю. М., Тарко, А. М., Математическое моделирование глобальных биосферных процессов. Москва, Наука, 1982.
[7] Кузьмичев, В. В., Закономерности роста древостоев. Наука, 1977, Новосибирск.
[8] Росс, Ю. К., Система уравнений для описания количественного роста растений. В. сб. фитоактинометрические исследования растительного покрова. Валгус, Таллин, 1967.
[9] Рачко, П., Имитационная модель динамики роста дерева, как элемента лесного биогеоценоза. В сб. Вопросы кибернетики, АНСССР, 1979, Москва.

(Beérkezett: 1984. március 19.)

RACSKÓ PÉTER
ELTE SZÁMÍTÓKÖZPONT
1117 BUDAPEST, XI. BOGDÁNFFY U. 10/B.

## SIMULATION MODEL OF THE TREE GROWTH DYNAMICS AS A PART OF THE FOREST BIOGEOCENOSIS

P. Racskó

Forest ecosystems are very important subjects of environmental research. A tree growth simulation model is suggested and studied in the report. The formal model is a system of difference equations representing the mass balance in the system for different nutritives and the biomass. In order to "close" the model a hypothesis of maximal primary productivity was postulated, providing the distribution law of new assimilates between the leaves, bole and roots. The following exogenous factors were taken into account: intensity of photosynthetically active radiation, air temperature, dynamics of the water supply, $CO_2$ concentration in the atmosphere and the limited amount of nutrition.

Some of the most intriguing results of the computer experiments are discribed in the conclusion.

*Alkalmazott Matematikai Lapok* **10** (1984) 413—426

# KIVÁLASZTÁSI ALGORITMUSOK ÉS ALKALMAZÁSUK AZ AGRÁRGAZDASÁGBAN

BÁN ISTVÁN

Budapest

A mezőgazdasági gyakorlatban előforduló igen nagyszámú függő és független változók számokkal kifejezhető és részben számokkal ki nem fejezhető ismérveket is tartalmazó adathalmazok elemzésének sajátos, a gyakorlati felhasználó által könnyen áttekinthető és számítástechnikailag kivitelezhető megoldását adja a kiválasztási algoritmusok alkalmazása.

A kiválasztási algoritmusok lehetővé teszik, hogy a rendelkezésre álló matematikai és számítástechnikai apparátus a gyakorlat számára minden esetben elfogadható, illetve további finomításra kerülő megoldást tudjon ajánlani. A kiválasztási algoritmusok alkalmazásának további előnye, hogy a matematikai és számítástechnikai módszereket kevésbé ismerő gyakorlati szakemberek számára is könnyen felhasználhatóvá teszi azokat.

A feladat a kiválasztás célját tekintve, mint ahogy ezt később részletesen látni fogjuk, két fő kérdéscsoportra különíthető el:

— az adott célállapotjellemzőknek határozza meg a gyakorlat számára legkedvezőbb értékeit a kiválasztási állapotjellemzők értékeinek a kiválasztási azonossági relációk által megadott környezetében,

vagy

— a célállapotjellemzők gyakorlat által legkedvezőbbnek tartott értékeihez határozza meg a kiválasztási állapotjellemzők értékkörnyezetét.

A kiválasztási algoritmusokat növénytermesztési, agrokémiai, növényvédelmi, erdészeti és vadgazdálkodási adatfeldolgozási feladatok megoldására alkalmaztuk.

## 1. Bevezetés

A mezőgazdasági, erdészeti és vadgazdálkodási gyakorlat döntéselőkészítésében ma már igen jelentős segítséget adnak a matematikai és számítástechnikai módszerek.

A feladatok jelentős része olyan többváltozós feladat, ahol fel kell deríteni a mezőgazdasági gyakorlatban állapotjellemzőnek nevezett változók közötti kapcsolat sajátosságait, meg kell határozni a számunkra lényeges változókat, azok sorrendjét, bizonyos szempontok szerint osztályokat kell képezni, meg kell határozni a törvényszerűségek matematikai alakját, meg kell keresni a legkedvezőbb megoldásokat és hibabecslést kell végezni. A feladatok nagyrészét az operációkutatás ismert eljárásaival oldottuk meg. Munkánk során a mezőgazdasági gyakorlat olyan problémákat is felvetett, amelyeket a későbbiekben ismertetendő sajátosságaik miatt az általunk kiválasztási algoritmusnak nevezett módszerekkel oldottunk meg.

A feladat természetszerűségéhez, az alkalmazandó matematikai módszerek sokféleségéhez, az emberi beavatkozás igényéhez és a megoldás gyakorlati mérlegelésének fontosságához történő alkalmazkodás igénye tette szükségessé a kiválasztási algoritmusok alkalmazását. A gyakorlatban eddig megoldott feladatok logikáját

és az alkalmazott matematikai módszereket előzetesen az első kiválasztási és a második kiválasztási feladat ábráján láthatjuk.

Az ember ismereteinek gyarapítására, döntéselőkészítésére és a lehető legkedvezőbb döntés meghozatalára az operációkutatás jelenlegi módszerei mellett régen használja a számára lényeges információk kiválasztását. A kiválasztás a természet egyik legrégebbi és legáltalánosabb biológiai jelensége. A megválaszolandó kérdés, mint ahogy később látni fogjuk, saját természetes környezetében vetődik fel, így a megoldás az információk kiválasztásában, csoportosításában és következtetéseiben ragaszkodik a természetszerűséghez.

A megoldandó feladatok közül számunkra tipikusak voltak azok, amelyekben
— a figyelembe veendő függő és független változók száma ezres nagyságrendű,
— a rendelkezésre álló megfigyelések száma igen nagy,
— igen sok változó értékkészlete tulajdonságok, minőségek halmaza, amelyek számokkal nem vagy csak közvetve, igen nehezen lennének kifejezhetőek,
— más változók értékkészlete a természetes számok halmaza, amelyek nagyságrendje igen eltérő,
— egy-egy változó értékkészlete olyan részhalmazokból áll, amelyeknek metszete üres,
— az egyes változók között szakmai kapcsolat van, de az összefüggés jellege ismeretlen.

A gyakorlat számára olyan kedvező megoldást kell adni, amely megadási módjában, a meghatározás menetének egyes szakaszaiban a gyakorlati szakember által bármikor ellenőrizhető, a megoldások természetes környezetéből ki nem szakított, értelmezéséhez nem kell matematikai ismeret. Ezt a szempontot emberi tényezőnek neveztük, és akármilyen furcsán is hangzik, de a megoldás gyakorlati alkalmazásának ez volt a legeslegfontosabb feltétele. A kiválasztási algoritmusok lehetővé teszik, hogy a rendelkezésre álló matematikai és számítástechnikai apparátus a gyakorlat számára minden esetben elfogadható, illetve további finomításra kerülő megoldást tudjon ajánlani. A kiválasztási algoritmusok alkalmazásának további előnye, hogy a matematikai és számítástechnikai módszereket kevésbé ismerő gyakorlati szakemberek számára is könnyen felhasználhatóvá teszi azokat.

## 2. Kiválasztási feladatok és megoldó algoritmusaik

A továbbiakban ismertetem a kiválasztási algoritmusoknak nevezett módszereket. Ragaszkodva a gyakorlati problémafelvető forráshoz, ill. az alkalmazók szempontjaihoz, az alkalmazott matematika szöveges feladatmegfogalmazási módját választottam.

A továbbiakban az alkalmazói és matematikai együttes megértést és mindenek felett az alkalmazhatóságot, ill. megvalósíthatóságot tekintem fő tárgyalási szempontnak. A nevezéktan így egyes esetekben óhatatlanul a mezőgazdasági elnevezéseket használja.

A gyakorlatunkban előforduló és itt tárgyalandó feladatok közös vonásai, mint feltételek az alábbiak:
— igen nagyszámú változó, továbbiakban állapotjellemző, fordul elő,
— az állapotjellemzők közül egyesek felvehetnek:
diszkrét értékeket,

folytonos értékkészletet,
számmal ki nem fejezhető diszkrét minőségi tulajdonságokat,
folytonos minőségi tulajdonságokat,

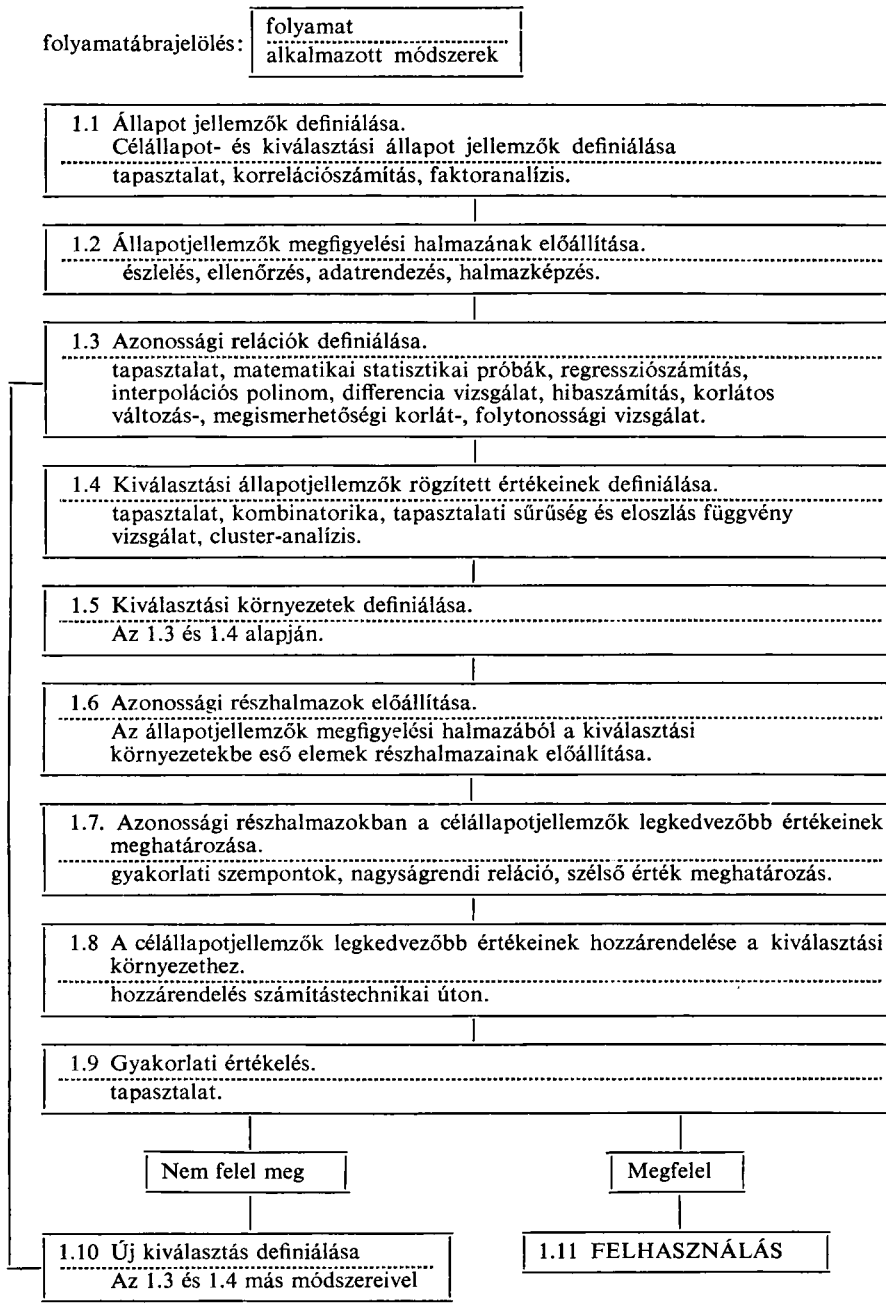
— a megfigyelések eredményeképpen az állapotjellemzők értékei ismertek,
— az állapotjellemzők között szakmai kapcsolat áll fenn,
— szakmai szempontból egyes állapotjellemzők kitüntetett szerepet töltenek be, a gyakorlati hasznuknál fogva leglényegesebb állapotjellemzőket ezért célállapotjellemzőknek, értékeiket pedig célállapotjellemző értékeknek nevezzük,
— a célállapotjellemzők mellett előforduló többi állapotjellemzőt, amelyek a szakmai környezetet adják meg, kiválasztási állapotjellemzőknek nevezzük,
— a gyakorlat minden egyes megfigyelését az állapotjellemzők és azok értékei jellemzik,
— két vagy több megfigyelés a gyakorlat számára azonos, ha kiválasztási állapotjellemzőinek mindegyikére teljesül valamely azonossági reláció,
— az azonossági relációk közül az alábbiak fordultak elő:

1. „vagy-vagy" reláció egy állapotjellemzőre: a megfigyelések adott kiválasztási állapotjellemzőjének megfigyelési értékei megegyeznek a szakmailag előre megadott értékek valamelyikével.
2. „vagy-vagy" reláció több állapotjellemzőre: a megfigyelések előre megadott kiválasztási állapotjellemzőinek megfigyelési értékei megegyeznek a szakmailag előre rögzített állapotjellemzők értékeinek valamelyikével.
3. „és-és" reláció több állapotjellemzőre: a megfigyelések előre megadott kiválasztási állapotjellemzőinek megfigyelési értékei megegyeznek a szakmailag előre rögzített valamennyi állapotjellemző értékeivel.
4. „tól-ig" reláció egy állapotjellemzőre: a megfigyelések adott kiválasztási állapotjellemzőjének megfigyelési értékei beleesnek a szakmailag előre megadott értékintervallumba.
5. „E-különbség" egy állapotjellemzőre: a megfigyelések adott kiválasztási állapotjellemzőjének megfigyelési értékei és a szakmailag előre megadott érték közti eltérés kisebb mint $E$.
6. „Egyenlő" egy állapotjellemzőre: az 5-ben leírt eset $E=0$ és az 1-ben leírt eset egyetlen szakmailag előre megadott érték esetén.
7. Előzőek együttese: a gyakorlatban ugyanazon megfigyelések más-más állapotjellemzőire (illetve ilyenek csoportjaira) az előző azonossági relációk közül más-más fordulhat elő.

A továbbiakban elvégzendő kiválasztáshoz szakmai szempontok alapján megadjuk a célállapotjellemzőket, a kiválasztási állapotjellemzőket és a kiválasztás utasításrendszerét.

A kiválasztás célja a bevezetőben felvetett feladat szerint két fő kérdéscsoportra különíthető el.

1. Az adott állapotjellemzőknek kell meghatároznunk a gyakorlat számára legkedvezőbb értékeit a kiválasztási állapotjellemzők értékeinek a kiválasztási azonossági relációk által megadott környezetében. Ezt nevezzük első kiválasztási feladatnak.
2. A célállapotjellemzők gyakorlat által legkedvezőbbnek tartott értékeihez kell meghatározni a kiválasztási állapotjellemzők értékkörnyezetét. Ezt nevezzük második kiválasztási feladatnak.




*1. ábra*
Az első kiválasztási feladat

Az alábbiakban megadjuk az első kiválasztási feladat megoldó algoritmusának blokkdiagramját (1. ábra), valamint ennek magyarázó leírását.

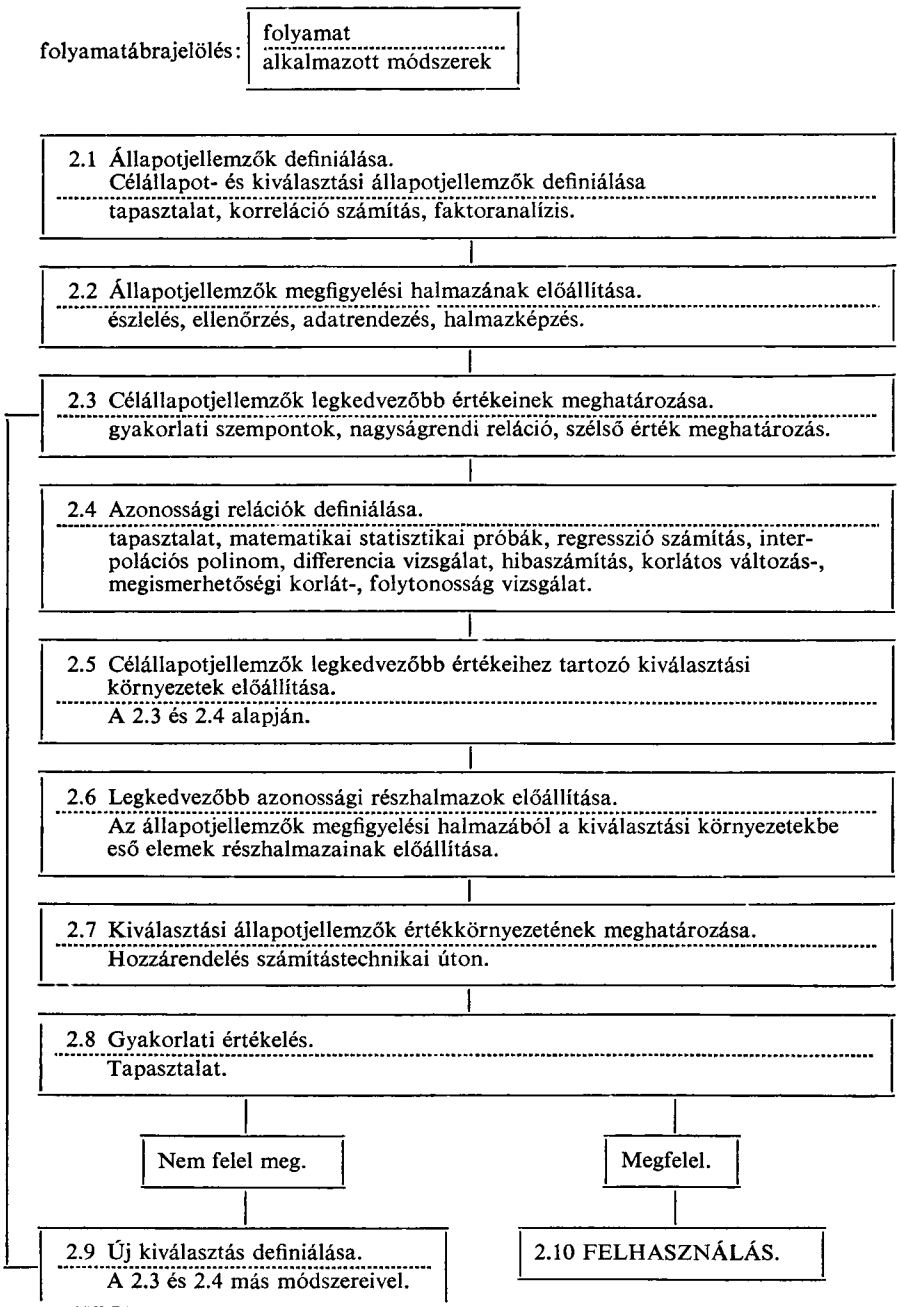
*Az első kiválasztási feladatban* megadjuk először az összes állapotjellemzőt (1. ábra 1.1). A szakmai gyakorlat által megfigyelt állapotjellemzők közül felsoroljuk a kiválasztási állapotjellemzőket. A szakmai gyakorlatban megfigyelt és így ismert események kiválasztási állapotjellemzőinek értékei jelölik ki azt a környezetet és környezeteket, amelyben, ill. amelyekben meg kell határoznunk a célállapotjellemzők legkedvezőbb értékeit. (1. ábra 1.2). Ezeket a környezeteket, továbbiakban kiválasztási környezeteket, a kiválasztási állapottjellemzők értékeivel és a rájuk vonatkozó azonossági relációk segítségével határozzuk meg (1. ábra 1.3). Egy adott kiválasztási környezet az összes többi megfigyelési események halmazából kiválasztja a vele azonos események halmazát úgy, hogy a kiválasztási állapotjellemzőket, értékeit és a rájuk megadott azonossági relációkat tekintve valamennyi megfigyelési eseményt sorbavéve kiválasztjuk azokat a megfigyeléseket, amelyekre teljesülnek a megadott azonossági relációk (1. ábra 1.4, 1.5). Az így kapott részhalmazt a továbbiakban *azonossági részhalmaznak nevezzük.*

Valamennyi környezethez meghatározzuk az azonossági részhalmazt. Az így előálló azonossági részhalmazok elemei az azonossági relációnak elegettevő valamennyi megfigyelés mindenegyes kiválasztási és célállapotjellemzőjének értékei. Az azonossági részhalmazokat a továbbiakban elemeikkel adjuk meg (1. ábra 1.6).

Ezután határozzuk meg azonossági részhalmazonként a célállapotjellemzőknek a gyakorlat számára legkedvezőbb értékeit (1. ábra 1.7). Az azonossági részhalmazokat sorbavéve minden egyes részhalmaz valamennyi eleme, azaz célállapotjellemzőinek értékei alapján meghatározzuk a gyakorlat számára legkedvezőbb értékeket. A továbbiakban ezeket a kedvező értékeket az azonossági részhalmazhoz tartozó kiválasztási környezet gyakorlat számára legkedvezőbb értékeinek fogadjuk el (1. ábra 1.8). Ezen értékeket a gyakorlat szempontjából értékeljük (1. ábra 1.9), és ha nem felelnek meg, további vizsgálatot végzünk és új kiválasztást definiálunk ezen vizsgálatok és gyakorlati kívánalmaink ismeretében (1. ábra 1.10). Amennyiben gyakorlatilag megfelelt, az eredményt felhasználjuk (1. ábra 1.11).

A továbbiakban megadjuk a második kiválasztási feladat megoldó algoritmusának blokkdiagramját (2. ábra), valamint ennek magyarázó leírását.

*A második kiválasztási feladatban* a célállapotjellemzők gyakorlat által kedvezőnek tartott értékeihez kell meghatározni a kiválasztási állapotjellemzők értékkörnyezetét. Bár a feladat logikája hasonló az első feladatéhoz, de a kiválasztás ellenkező irányú. Az előző feladathoz hasonlóan először az állapotjellemzőket adjuk meg, külön-külön felsorolva a cél- és külön a kiválasztási állapotjellemzőket (2. ábra 2.1). A szakmai gyakorlat által megfigyelt állapotjellemzők és értékeik adják a megfigyelési alaphalmazt (2. ábra 2.2). A szakmai szempontok alapján megadjuk a célállapotjellemzők legkedvezőbb értékeit (2. ábra 2.3). A célállapotjellemzőkre megadott azonossági relációk (2. ábra 2.4) segítségével a megfigyelési alaphalmazból kiválasztjuk a velük azonos gyakorlat szempontjából kedvezőbb azonossági részhalmazokat (2. ábra 2.5 és 2.6). Ezen részhalmazok elemei célállapotjellemzőiknek értékeit és az azonossági relációkat tekintve azonosak, kiválasztási állapotjellemzőiket és értékeiket nézve viszont nem. A kiválasztási állapotjellemzőiknek az értékei adják azt a környezetet, amelyekben a gyakorlat számára kedvező események bekövetkeztek (2. ábra 2.7). A felhasználó a környezetbe eső célállapotjellemzők értékeit a gyakorlat szempontjából értékeli (2. ábra 2.8) és ha megfelel közvetlenül hasznosítja, vagy ha




*2. ábra*
A második kiválasztási feladat

nem felel meg tovább vizsgálja, és ennek eredményeképpen új kiválasztást definiál (2. ábra 2.10) mindaddig, amíg a kapott eredmények felhasználhatóak (2. ábra 2.10).

Gyakran találkozunk a két kiválasztási feladat együttesével, amikor is a célállapotjellemzők által megadott bővebb környezetekben keressük a célállapotjellemzők legkedvezőbb értékeihez tartozó szűkebb környezetet.

Látható, hogy az előző kiválasztási feladatok megoldási stratégiát jelölnek ki. Egyszerűbb esetben elég az állapotjellemzőknek az értékeit, számmal közvetlenül ki nem fejezhető sajátosságait, ill. tulajdonságait, valamint az azonossági relációkat tekinteni, s a feladat az előzőekben leírt módon már programozható. Bonyolultabb esetekben a kiválasztási környezet megadására, vagy a legkedvezőbb értékek meghatározására a szakmai ismeretek mellett matematikai módszereket használunk, mint pl. cluster-analízis, regressziós analízis stb.

Jelenleg a biológiával és alkalmazási területeivel foglalkozunk, de ez nem jelenti egyéb alkalmazási területek kizárását. A bevezetőben leírt biológiai és gyakorlati adottságok, valamint a kiválasztási algoritmusok sajátosságai alapján a továbbiakban belátható, hogy a megoldás javasolt menete használható eredményhez vezet.

## 3. A kiválasztási feladatok általános matematikai modellje

Az előzőekben szövegesen leírt kiválasztási feladatok közös matematikai modellje az alábbi: (Az alkalmazott jelölések és elnevezések az [1] irodalomban leírtakkal egyeznek meg.)

DEFINÍCIÓ. Biológiai folyamat a biológiai állapotok összessége.

Jele: $F$.

A biológiai állapoton a vizsgált biológiai objektum összes észlelhető és nem észlelhető jellemzői értékeinek, ill. tulajdonságainak összességét értjük. A továbbiakban feltesszük, hogy ez nem üres halmaz. Jele: $B$ és $B \neq \emptyset$. A biológiai állapot egy adott jellemzőjén egy adott észlelési formának megfelelő mennyiséget, minőséget vagy tulajdonságot értünk. Jele: $b_i$, ahol $i=1, 2, \ldots$ index megmutatja, hogy melyik állapotjellemzőről van szó.

Egy adott biológiai állapotjellemző értékein az adott jellemző konkrét számértékeit, minőségeit, ill. tulajdonságait értjük. Jele $b_i^m$, az $i$-edik állapotjellemző, $m$-edik megfigyelésére, $i=1, 2, \ldots$; $m=1, 2, \ldots$.

Az előző jelölésekkel:

$$B = \bigcup_{\forall i \forall m} b_i^m.$$

A biológiai állapot tulajdonságai miatt feltesszük, hogy $B$ olyan félig rendezett halmaz, amelyben $b_1, b_2, \ldots, b_i$ állapotjellemzők által felvett

$$b_1^1, b_1^2, \ldots, b_1^m$$

$$b_2^1, b_2^2, \ldots, b_2^m$$

$$\vdots$$

$$b_i^1, b_i^2, \ldots, b_i^m$$

biológiai állapotjellemző értékek $i$ és $m$ szerint egyértelműen rendezettek.

A kiválasztási feladatok és megoldó algoritmusaik cím alatt 1—7 pontban szövegesen megfogalmazott azonossági relációk alapján az alábbi azonossági reláció definíciókat vezetjük be:

DEFINÍCIÓ. Egy $b_i$ biológiai állapotjellemző tetszőleges két $b_i^1$ és $b_i^2$ értékét a felhasználó azonosnak tekinti jele $b_i^1 \sim b_i^2$, ha az alábbi relációk közül az általa kiválasztott teljesül.

— $b_i^1 \sim b_i^2$ akkor és csak akkor, ha $b_i^1 \equiv b_i^2$

— $b_i^1 \sim b_i^2$ akkor és csak akkor, ha $aK \leqq b_i^1 \leqq fK$, $aK \leqq b_i^2 \leqq fK$, ahol $aK$ a felhasználó által megadott alsókorlát és $fK$ a felhasználó által megadott felső korlát.

— $b_i^1 \sim b_i^2$ akkor és csak akkor, ha a felhasználó által megadott rögzített $b_i^{m0}$ és $E$ esetén $b_i^1 \in G(b_i^{m0}; E)$

$$b_i^2 \in G(b_i^{m0}; E)$$

— $b_i^1 \sim b_i^2$ akkor és csak akkor, ha $b_i^1$ és $b_i^2$ közül a felhasználó által rögzített egyikére pl. $b_i^{1,0}$ és a felhasználó által megadott $E$-re

$$b_i^2 \in G(b_i^{1,0}; E).$$

DEFINÍCIÓ. A $b_1, b_2, \ldots, b_i$ biológiai állapotjellemzők által felvett értékek legyenek $(b_1^{1,2,\ldots,q}, b_2^{1,2,\ldots,q}, \ldots, b_i^{1,2,\ldots,q}) = S_q$, ahol $q < m$ és $S^q \subset B$. Az $S_q$ részhalmazt azonossági részhalmaznak tekintjük, akkor és csak akkor, ha a felhasználó által megnevezett $i$ állapotjellemzőkre és ezeknek tetszőleges két $b_i^q$ és $b_i^{q1}$ értékeire fennáll $b_i^q \sim b_i^{q1}$.

DEFINÍCIÓ. Két biológiai állapot egyenlő, ha a két biológiai állapot minden egyes állapotjellemzőjének megfelelő minden egyes $m$ megfigyelési értékpárjára, értékhármasára ..., érték $i$-esére fennáll $b_i^m \sim p_i^m$ $\forall_i$ és $\forall_m$, ahol $b_i^m$ az egyik, $p_i^m$ a másik biológiai állapot $i$-edik jellemzőjének $m$-edik megfigyelési értéke.

A továbbiakban jelöljük a felhasználó által definiált állapotjellemzőket $s_1, s_2, \ldots, s_j$-vel, a célállapotjellemzőket pedig $a_1, a_2, \ldots, a_f$-fel, $j = 1, 2, \ldots, i$, $f = 1, 2, \ldots, i$.

Az $m$ megfigyelési értékeket tekintve fennáll:

$$A = \bigcup_{\forall f, m} a_f^m \qquad S = \bigcup_{\forall j, m} s_j^m \qquad A \neq \emptyset \qquad S \neq \emptyset$$

$$B = A \cup S$$

Ahol $B$ biológiai állapotjellemzői értékeinek rendezettségi tulajdonságai miatt a kiválasztási állapotjellemzők $a_f^m$ értékeinek és a célállapotjellemzők $s_j^m$ értékeinek rendezettsége egyértelmű $j + f = i$ és $m$ szerint.

$$\begin{array}{l} a_1^1, a_1^2, \ldots, a_1^m \\ a_2^1, a_2^2, \ldots, a_2^m \\ \vdots \\ a_f^1, a_f^2, \ldots, a_f^m \\ s_1^1, s_1^2, \ldots, s_1^m \\ s_2^1, s_2^2, \ldots, s_2^m \\ \vdots \\ s_j^1, s_j^2, \ldots, s_j^m. \end{array}$$

A biológiai állapotjellemzők tulajdonságaiból következik, hogy $m$ szerint nemcsak azonos értékeket vesznek fel.

DEFINÍCIÓ. Az $A$-nak bármely részhalmazában *lokális optimumnak* nevezzük az őt alkotó állapotjellemzők értékei közül $a_1^m, a_2^m, \ldots, a_j^m$-nek azt az $m$-edik értékét ($m=1, 2, \ldots$), amely a gyakorlati felhasználó szempontjából a legkedvezőbb.

*Feltételek.* A megfigyelt biológiai folyamat adottságaiból és $B=A\cup S$ elemeinek rendezettségéből következik, hogy a kiválasztási állapotjellemzők értékei és a célállapotjellemzők értékei között kölcsönösen egyértelmű $f$ leképezés létezik. Ezen $f$ leképezés legfontosabb tulajdonsága, hogy minden $s_j^m \in S$, $j=1, 2, \ldots$ kiválasztási állapotjellemző $m=1, 2, \ldots$-edik megfigyelt értékéhez, minőségéhez, tulajdonságához kölcsönösen egyértelműen hozzárendeli a vele együtt észlelt $a_f^m \in A$, $f=1, 2, \ldots$ célállapotjellemző $m=1, 2, \ldots$ értékét, minőségét, tulajdonságát. Előfordulhat surjektív leképezés vagy ráképezés, de az adatok kezelésekor mindenkor gondoskodunk az együtt észlelt és így összetartozó kiválasztási célállapotjellemzők egymáshoz rendeltségéről.

$$s_j^m \overset{f}{\longleftrightarrow} a_f^m.$$

Az $f$ leképezés matematikai alakját nem kell meghatározni. Az $f$ az általa leképezett biológiai állapotjellemzők által felvett értékekkel adott. Az $f$ leképezés matematikai alakja ismeretének szükségtelensége igen nagy könyebbséget jelent a modellezés során, különösen nagyon nagyszámú biológiai állapotjellemzők esetén.

Ugyancsak a biológiai adottságokból következik, hogy a kiválasztási állapotjellemzők valamennyi megfigyelt értékét tekintve $S$-nek létezik az azonossági relációnak elegettevő $S_1, S_2, \ldots$, részhalmaz felbontása:

$S=S_1\cup S_2\cup \ldots \cup R_s$, ahol $S_1, S_2, \ldots$ azonossági részhalmazok, amelyekhez tartozó kiválasztási állapotjellemző értékei egy-egy azonossági részhalmazon belül eleget tesznek az azonossági reláció követelményeinek és ebből az is következik, hogy $S_1\cap S_2\cap \ldots = \emptyset$. Az $R_s$ maradék részhalmaz, amelybe tartozó kiválasztási állapotjellemzők értékei nem tesznek eleget az azonossági relációnak.

A biológiai sajátosságokból következik az is, hogy a biológiai állapotjellemzők $S_1, S_2, \ldots$, azonossági részhalmazaihoz $f$ kölcsönösen egyértelmű leképezés miatt tartozó $A_1, A_2, \ldots$, célállapotjellemző-részhalmazokban a *célállapotjellemzők értékei* nem mind azonosak. Az $R_s$ kiválasztási állapotjellemző-értékek maradék részhalmazhoz $f$ leképezéssel hozzátartozó $R_a$ célállapotjellemző-értékek részhalmazában a célállapotjellemzők értékei, szintén nem mind azonosak.

$$A = A_1\cup A_2\cup \ldots \cup R_a.$$

1. ÁLLÍTÁS. Az $S$-nek $S_1, S_2, \ldots, R_s$ azonossági részhalmaz-felosztásainak minden részhalmazához létezik lokális optimum.

*Bizonyítás.* Az $f$ leképezés feltételbeli kölcsönösen egyértelmű tulajdonságaiból következik, hogy $S=S_1\cup S_2\cup \ldots \cup R_s$-hez létezik $A$-nak $A=A_1\cup A_2\cup \ldots \cup R_a$ felosztása, ahol

$$\begin{aligned} S_1 &\overset{f}{\longleftrightarrow} A_1 \\ S_2 &\overset{f}{\longleftrightarrow} A_2 \\ &\vdots \\ R_s &\overset{f}{\longleftrightarrow} R_a. \end{aligned}$$

Az $A_1, A_2, \ldots, R_a$ feltételbeli biológiai sajátosságából következik, hogy a részhalmazok mindegyikének elemei nem mind azonosak, így mindegyik részhalmazból kiválasztható a definíció szerinti lokális optimum, azaz léteznek az alábbi lokális optimumok:

$$a_f^{\text{opt.}1} \in A_1;\ a_f^{\text{opt.}2} \in A_2;\ \ldots\ a_f^{\text{opt.}R_a} \in R_a.$$

2. ÁLLÍTÁS. Az $s_j, j=1,2,\ldots,i$ kiválasztási állapotjellemzők tetszőleges $m$-edik $s_j^m \in S$ értékére létezik lokális optimum.

*Bizonyítás.* A feltételbeli biológiai adottságokból következik, hogy az $S$-nek létezik az azonossági relációnak elegettevő $S=S_1 \cup S_2 \cup \ldots \cup R_s$ azonossági részhalmazfelosztása, ahol $S_1 \cap S_2 \cap \ldots \cap R_s = \emptyset$. Mivel $s_j^m$ maga is $\forall m$-re eleme $S$-nek, $s_j^m \underset{\forall m}{\in} S$, ezért

$$s_j^m \underset{\forall m}{\in} \begin{cases} S_1, S_2, \ldots \text{ valamelyikének} \\ \text{vagy} \\ R_s\text{-nek.} \end{cases}$$

Az 1. állítás miatt $S_1, S_2, \ldots, R_s$ mindegyikében létezik lokális optimum, ezáltal $s_j^m \in S$ tetszőleges $m$ megfigyelési értékére is létezik lokális optimum.

3. ÁLLÍTÁS. A definícióban lokális optimumnak nevezett $\text{opt}\, a_f^{m0} \in A$-hoz létezik olyan részhalmaz, amely része $S$-nek.

*Bizonyítás.* Az $f$ leképezés feltételbeli kölcsönösen egyértelmű tulajdonsága miatt minden $s_j^m \in S$, $j=1,2,\ldots$ kiválasztási állapotjellemző $m=1,2,\ldots$-edik megfigyelt értékéhez, minőségéhez, tulajdonságához kölcsönösen egyértelműen hozzárendeli a vele együtt észlelt $a_f^m \in A$, $f=1,2,\ldots$ állapotjellemző $m=1,2,\ldots$ értékét, minőségét, tulajdonságát.

Az opt $a_f^{m0} \in A$-hoz létezik $s_j^{m0} \in S$, $j=1,2,\ldots$ kiválasztási állapotjellemzőknek $m_0$-adik megfigyelési értéke.

A biológiai adottságokból következő azon feltétel, mely szerint a kiválasztási állapotjellemzők valamennyi megfigyelt értékét tekintve $S$-nek létezik az azonossági relációnak elegettevő $S_1, S_2, \ldots, R_s$ részhalmaz-felbontása $S=S_1 \cup S_2 \cup \ldots \cup R_s$ és $S_1 \cap S_2 \cap \ldots \cap R_s = \emptyset$.

Ebből következik, hogy

$$s_j^{m0} \in \begin{cases} S_1 & \text{vagy} \\ S_2 & \text{vagy} \\ \vdots & \\ R_s & \text{vagy} \end{cases}$$

azaz opt $a_j^{m0} \in A$-hoz létezik $S$-beli részhalmaz.

KÖVETKEZMÉNY. Az első kiválasztási feladatnak a 2. állítás alapján létezik megoldása, a második kiválasztási feladatra ugyanez a 3. állítás miatt igaz.

A felhasználó meghatározza a kiválasztási és állapotjellemzőket, valamint ezek megfigyelésével, ill. értékelésével az összetartozó állapotjellemző értékeket. Definiálja az egyes állapotjellemzőkre érvényes azonossági relációt s ezek alapján az azonossági részhalmaz felosztást. Definiálja az egyes célállapotjellemzők lokális optimumainak

kritériumait. Az előzőekben leírtak és a 2. állítás miatt létezik a felhasználó által elsőkiválasztási feladatnak nevezett feladat megoldása.

A felhasználó megadja a lokális optimum értékeket és a hozzájuk tartozó biológiai állapotjellemzőket, valamint az azonossági relációk kritériumait. A felhasználó által észlelt állapotjellemzők értékeinek halmazából az előzőekben leírtak és a 3. állítás miatt létezik a második kiválasztási feladatnak nevezett feladat megoldása.

*A kiválasztási feladatok főbb sajátosságai:*

A kiválasztási feladatok főbb sajátosságai, amelyek alkalmazásukat más ismert matematikai módszerek mellett is lehetővé tették, az alábbiak:

— igen nagyszámú biológiai állapotjellemzőt képes kezelni,
— a biológiai állapotjellemzők értékei igen változatosak: számok, tulajdonságok, minőségi és egyéb ismérvek,
— az $f$ leképezés matematikai alakjának ismerete nem szükséges,
— nagyfokú alkalmazói szabadság az azonossági reláció definiálásában,
— a feladatok valós a gyakorlatban felmerült feladatok megoldására alkalmasak,
— a megoldások valós megfigyelési állapotjellemző értékeken alapulnak,
— a megoldási stratégia közel áll a gyakorlati fogalmakhoz, ezáltal áttekinthető a gyakorlati szakember részére is,
— interaktív megvalósítást tesz lehetővé.

## 4. Gyakorlati alkalmazások és számítógépes megvalósítás

*Növénytermesztési szaktanácsadás*

A mezőgazdasági gyakorlatban a nagyüzemi növénytermesztés technológiai és műtrágyázási szaktanácsadására alkalmas az 1 és 2 típusú kiválasztási algoritmus.

A növénytermesztési szakember célja elsősorban az, hogy adott termőhelyen a lehető legtöbb és legértékesebb termést adó növényt termesszen és a rendelkezésére álló módszerekkel a termést befolyásoló tényezőket úgy alakítsa, hogy a számára legkedvezőbb eredményt érje el. A szaktanácsadási modellnek ezt a célt kell elérnie. Lehetőséget kell teremteni a modellben arra is, hogy a számítástechnikai megvalósítás interaktív hozzáférést biztosítson. A gyakorlati szakembernek továbbá szüksége van a kapott eredmények ismeretében újabb és újabb döntési stratégiák kipróbálására.

Az állapot adathalmazt mint megfigyelési alaphalmazt a mezőgazdasági táblák sokasága alkotja. Minden táblának és a rajta termesztett növénynek, valamint a termesztés technológiai folyamatának állapotjellemzőit feljegyezzük, mint pl. meteorológiai, talajelemtartalmi, növényfaj-, fajta, munkák időpontja, fajtája, műtrágya- és növényvédőszer-hatóanyagok és mennyiségük, valamint az elért termés mennyisége és minősége állapotjellemzőket, amelyek együttesen a szaktanácsadás állapot-alaphalmazát adják.

Az állapotjellemzők definiálása során az első lényeges feladat a szakemberek által felsorolt többezer állapotjellemző közül a „lényegesek" kiválasztása. A rendelkezésre álló biológiai állapotjellemzők értékei alapján elvégzett korrelációanalízissel (pl. BMDP 1 programcsomag) és regresszióanalízissel kapott eredmények, valamint a szakemberek gyakorlati tapasztalata alapján, mintegy 1400-ra csökkentettük a biológiai állapotjellemzők számát. Az egyes növénytermesztési technológiai minőségi

biológiai állapotjellemzők hatását statisztikai próbákkal vizsgáltuk. Az előző vizsgálatok eredményeképpen, figyelembevéve a felhasználók igényeit, meghatároztuk a célállapotjellemzőket és a kiválasztási állapotjellemzőket. Lényeges feladat volt a még hiányos értékekkel rendelkező állapotjellemzők feltöltése adatokkal. Ehhez optimális mintavételi módszerekkel kellett meghatároznunk az elérni kívánt pontossághoz szükséges mintavételi arányt. A hiányzó értékek feltöltésével előállítottuk az állapotjellemzők megfigyelési halmazát.

Minden egyes kiválasztási állapotjellemzőre definiálltuk az azonossági rellációt, amelyek jelentős részében segítségünkre volt a többváltozós regressziós függvények meghatározása és deriváltjainak vizsgálata, valamint a szakmai tapasztalat. Igen lényeges szempont volt, hogy a modell megoldásának interaktívvá tétele miatt gondoskodni kellett arról, hogy az ekvivalencia relációkat, illetve az azokban szereplő paramétereket (pl. $E$, $aK$, $fK$ stb.) a felhasználó kívánsága szerint változtatni tudja, s ezek hatását láthassa a későbbi megoldás célállapotjellemzők optimum értékében.

Szaktanácsadáskor az adott még csupasz szaktanácsolandó tábla meteorológiai, talaj állapotjellemzői értékeinek alapján kiválasztjuk az állapot-alaphalmazból a vele azonos, de már megfigyelt táblák halmazát. A megfigyelt azonossági részhalmaz táblákból kiválasztjuk azokat, ahol a termesztendő növények a legnagyobb termést adták. Ezeken a legnagyobb termésű azonossági részhalmaz táblákon meghatároztuk a hatóanyag-termés polinomiális összefüggéseket. Ezek alapján adjuk meg az alkalmazandó műtrágyahatóanyagértékeket, valamint az ott alkalmazott technológiai állapotjellemző értékek-termés összefüggés alapján tudjuk meg a kívánt technológiát (pl. talajművelés-, vetés időpontja stb.).

Az első 1977-ben bemutatott kiválasztási matematikai modell felhasználását a trágyázási szaktanácsadásban [2] irodalom részletesen ismerteti. További alkalmazásokat a [3], [4], [5] irodalmak tartalmaznak.

*Fafajmegválasztás*

Az erdészet egyik legnagyobb problémája a fafajmegválasztás. Az erdőgazdálkodás legkisebb egysége a mintegy 400 000 db erdőrészlet, amelynek ismerjük termőhelyi és erdőállapotjellemzőit. Feladatunk az üressé vált erdőrészletekre olyan fafajösszetétel megadása, amely maximális fatermést ad az adott termőhelyen. Az ország valamennyi erdőrészletéből, mint állapothalmazból kiválasztjuk az adott tervezendő erdőrészlet termőhelyi állapotjellemzőivel azonos termőhelyü erdőrészleteit, s ezek közül kiválasztjuk az optimális fatermést adó erdőrészleteket. Ezen erdőrészletek fafajösszetételét fogadjuk el adott üres erdőrészlet tervezendő fafajösszetételének. Erdősült erdőrészletek esetén az előzőekhez hasonlóan határozzuk meg az ún. fatermő értéket. Ekkor minden erdőrészletet a vele azonos termőhelyű erdőrészlet-részhalmazba soroljuk be, megnézzük, hogy ott mekkora az elérhető legnagyobb fatermés, és ezt viszonyítjuk az adott azonossági részhalmaz minden erdőrészletén található valóságos fatömeghez. Ez a százalékos viszonyszám a gazdálkodás egyik értékmérője.

*Vadeltartóképesség*

A mezőgazdaság, erdészet és vadgazdálkodás egymásközti legvitatottabb kérdése, hogy egy adott területen mekkora a *maximálisan eltartható vadlétszám*. A mintegy 8 millió ha-on gazdálkodó 800 vadgazdálkodási egység állapotjellemzői azok az erdé-

szeti és mezőgazdasági állapotjellemzők, amelyek leginkább befolyásolják a vad életét, amelyek egyúttal kiválasztási állapotjellemzők is. Számunkra leglényegesebbek mint célállapotjellemzők a vadlétszám (szarvasegység) és az általa okozott vadkár mértéke. A vadgazdálkodási egységeket kiválasztási állapotjellemzőik alapján azonos részhalmazokba csoportosítva, minden egyes részhalmazon belül kiválasztjuk azokat a vadgazdálkodási egységeket, ahol magas a vadlétszám és kicsi a vadkár. Ez az a vadlétszám, amit a többi vele azonos adottságú vadgazdálkodási egységen is fenn lehet tartani olyan feltételek mellett, mint az optimális vadgazdálkodási egységen. A maximálisan eltartható vadlétszám feletti vadállomány jelentős mezőgazdasági és erdészeti károkat okoz.

*Számítástechnikai megvalósítás*

Igen nagy segítséget és könnyebbséget jelentett az adatbáziskezelő rendszerekkel rendelkező számítógépek megjelenése, amelyek az ilyen modellezési lehetőségeket könnyen és gyorsan biztosítják.

Kiválasztási algoritmust először mintegy 13 évvel ezelőtt alkalmaztunk különböző növényekre, a növényvédőszerek fajtájának és dózisainak meghatározására. A feladatot a Fővárosi Növényvédő Állomás megbízása alapján a MÉM STAGEK IBM 1030-as gépén FORTRAN programozással szerveztük. Később hasonló feladatokat oldottunk meg a MÉM STAGEK IBM 360/40-es gépén. A növényvédelmi és agrokémiai kísérletek értékelését mindkét előző gépen végrehajtottuk.

A nagyüzemi növénytermesztés műtrágya és technológiai szaktanácsadásának feladatát az OTSZK szakembereivel először mintegy öt évvel ezelőtt az ICL System 4 gépén COBOL és FORTRAN programozással szerveztük, amit később interaktív rendszerré fejlesztettünk. Az országos adatbázisra épülő műtrágya és technológiai szaktanácsadást a KSH ÁSZSZ szakembereivel a HwB gépen az IDSI adatbáziskezelő rendszer felhasználásával oldottuk meg, ami később interaktív üzemben is megvalósult. A kiválasztási algoritmusra alapozott szaktanácsadást még a Cukoripari Tröszt is. A kiválasztási algoritmus erdészeti és vadgazdálkodási alkalmazása ugyancsak a KSH HwB számítógépén, és a MÉM Erdőrendezési Szolgálat R 20-as gépén valósult meg.

Az elmúlt 15 évben a számítástechnika nagymérvű előretörése az agrár gyakorlatban megteremtette a lehetőséget a kiválasztási algoritmusok széles körű gyakorlati alkalmazására. A nagy országos adathalmazok megjelenése, az adatbáziskezelőrendszerek elterjedése és a távadatfeldolgozás tovább növeli a kiválasztási algoritmusok jelentőségét.

A kiválasztási algoritmusok alkalmazásának egy új lehetőségét teremti meg a hazánkban kifejlesztett kisszámítógépek megjelenése. Ezek a gépek alkalmasak egy-egy mezőgazdasági üzem, erdőgazdaság, tervezőintézet saját feladatainak ellátására az adott alkalmazási költség teherviselési lehetőségén belül. Ezen gépeket ellátva felhasználói *kiválasztási algoritmus* optimalizálási *programcsomaggal*, a gyártó bővebb piacra talál, a felhasználó pedig saját gyakorlati feladatainak megoldására tudja rövid időn belül használni a gépet.

## IRODALOM

[1] BÁN, I., *Biomatematika és alkalmazása a növénytermesztésben* (Mezőgazdasági Kiadó, Budapest, 1977).
[2] BÁN, I., „A kiválasztási matematikai modell felhasználása a trágyázási szaktanácsadáshoz", *Agrokémia és Talajtan* **28** (1979).
[3] BÁN, I., „Talajerő-utánpótlási szaktanácsadás számítógéppel", *Magyar Mezőgazdaság* **33** (1978) 33. szám.
[4] BÁN, I. és RISKÓ, L., „Interaktív számítógépes talajerő-utánpótlási szaktanácsadó rendszer. (Előadás ismertetés) *Számítástechnika* (1979).
[5] BÁN, I. és RISKÓ, L., „Számítógépes talajerő-utánpótlás", *Magyar Mezőgazdaság* **34** (1979) 7. szám.
[6] BÁN, I., „A növénytermesztés szolgálatában", *Számítástechnika* (1980).
[7] BURINGTOM, R. S. and MAY, D. C., *Handbook of Probability and Statistics with Tables* (Sandersby, Handbook Publ., 1953).
[8] DIEDONNÉ, *Foundations of Modern Analysis* (Academic Press, London, 1961).
[9] ÉLTETŐ, Ö. és ZIERMANN, M., *Matematikai statisztika* (Budapest, 1961).
[10] LINDGREN, B. W., and MC ELRATH, A. W. *Introduction to Probability and Statistics* (Macmillan, New York, 1958).
[11] PRÉKOPA, A. és ÉLTETŐ, Ö., *Matematikai statisztika* (Kézirat, Budapest, 1961).
[12] PESARAN, M. H. and SLATER, L. J., *Dynamic Regression: Theory and Algorithms* (Ellis Horwood Limited, Chichester, 1980).
[13] PRÉKOPA, A., *Valószínűségelmélet* (Műszaki Könyvkiadó, Budapest, 1962).
[14] SMITH, C. A. B., *Biomathematics, the Principles of Mathematics for Students of Biological Science* (London, 1954).
[15] SZŐKEFALVI-NAGY, B., *Valós függvények és függvénysorok* (Egyet. Tankönyv, Budapest, 1965).
[16] VINCZE, I., *Matematikai statisztika ipari alkalmazásokkal* (Műszaki Könyvkiadó, Budapest, 1968).
[17] YULE, G. U. és KENDALL, M. G., *Bevezetés a statisztika elméletébe* (Budapest, 1964).

(Beérkezett: 1980. július 16.)

(Átdolgozva beérkezett: 1984. március 2.)

BÁN ISTVÁN
MÉM ERSZ
1054 BUDAPEST, SZÉCHENYI ÚT 14.

## EMPLOYING PLANNED METHOD OF SELECTION (PMS) IN AGRICULTURE

I. BÁN

The search of optimum in agricultural great mass of facts which involves numerous quantities and qualities can be solved by PMS (planned method of selection).

This mathematical method's data processing can be easy to survey by users who collects the quantitites and qualities.

Tasks of PMS are

— the optimums have to be determined in troops of quantities and qualities
— the troops of quantities and qualities have to be determined by given optimums. These optimums are taken by agricultural users.

PMS has been applied in plant cultivation, agro-chemistry, plant-protection and forestry.

*Alkalmazott Matematikai Lapok* **10** (1984) 427—451

# *A külföldi szakirodalomból*

# EGYSZERŰ MECHANIKAI JÁTÉKOK ELEMI DINAMIKÁJA[1]

W. BÜRGER

Karlsruhe

## 0. Bevezetés

Ez a dolgozat részben egy bevezető jellegű kurzuson alapszik, amit „A játékok mechanikája" címmel tartok a karlsruhei egyetemen, mint egy problémaorientált megközelítését a mechanikának, részben pedig egy előadáson, amit a nottinghami egyetemen tartottam. Kísérleti bemutatókból kiindulva haladok — matematikai modellek és a játékok észlelhető mozgásának megjövendölése felé. Így a tárgy folyamán a mechanika valamennyi alapvető fogalmát és elvét be kell vezetnem, mint a sebességet, az erőt, a mozgásmennyiséget, a forgatónyomatékot, a szabadságfokokat, Newton második törvényét, az impulzusnyomaték megmaradását, az energia megmaradását. Az eredeti előadás számos, a játékok és a klasszikus mechanika fogalmainak történetére vonatkozó megjegyzést tartalmazott. Sajnos nem foglalkozhatom sokat ezekkel a kérdésekkel egy ilyen rövid dolgozat keretében. Csaknem valamennyi játék, amit tárgyalni fogok, egészen szabványos, amit bármely játéküzletben meg lehet vásárolni.

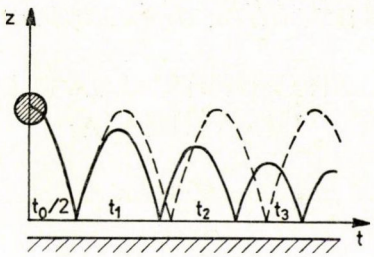
Köszönetet mondok ERNST BECKER professzornak *(Darmstadt)* inspiráló érdeklődéséért és azért a nagy segítségért, amit a dolgozat elkészítésében nyújtott nekem.

## 1. Visszapattanó labda

A labdát bizonyára valamennyi antik kultúra ismerte. A *British Múzeumban* látható egyiptomi labdák egy kb. i. e. 1400-tól kezdődő gyűjteménye. Ezek részben agyagból, részben afrikkal vagy papírral töltött bőrből készültek. A görögök és a rómaiak számos labdajátékot ismertek, melyek szabályai az idők folyamán elvesztek. A régi kelták egyfajta labdarúgást űztek juh vagy kecske hólyagjából készült rugalmas labdával. Ezt a sportot a kínaiak már i. e. 1000 körül kedvelték. Érdemes lenne azt is megvizsgálni, hogy vajon a rómaiak is annyira rajongtak-e a „*harpastum*"-nak nevezett hasonló játékukért, mint amennyire kortársaink lelkesednek a mai labdarúgásért a világbajnokság idején.

Még a középkorból is, amely pedig szegény volt játékokban, rendelkezünk labdajátékokról szóló beszámolókkal. Az utcán való labdázás szokása olyannyira terhessé vált *Londonban* a XIV. században, hogy III. *Edward* hatnapi börtön terhe alatt megtiltotta azt [22].

[1] Ez a dolgozat fordítása a következőnek: W. BÜRGER, "Elementary dynamics of simple mechanical toys", *Hefte der Gesellschaft für Angewandte Mathematik und Mechanik*, 1980/2 21—60. A fordítás közléséhez a szerző és a szerkesztőség hozzájárult.




*1. ábra*

A legegyszerűbb esettel kezdem, amikor egy labdát ledobnak a padlóra. A megfelelő modell ekkor egy tömegpont a labda középpontjában. A labda visszaverődik a földről (1. ábra). Túl egyszerű lenne azonban a dolog, ha visszatérne az eredeti magasságba. Azt tapasztaljuk, hogy minden visszaverődés után veszít valamit a magasságból. Minél kevesebb a veszteség, annál rugalmasabbnak mondjuk a labdát. A labda rugalmatlanságának a technikai mértéke azaz az ütközési együttható, az ütköző test ütközés utáni és ütközés előtti $v'$ és $v$ sebességének negatív hányadosa, vagyis

$$0 \leqq \varepsilon = -\frac{v'}{v} \leqq 1.$$

Ennél a pontnál nem tudom megállni, hogy ne tegyek egy történeti megjegyzést. A mérnökök alapvető kézikönyvének a „*die Hütte*"-nek, 1955. évi 28. kiadásában találtam egy táblázatot az ütközési együtthatókról. Így például

acél *(chalybs)* = 5/9

üveg *(vitrum)* = 15/16.

Elég különös, hogy az értékeket egyszerű törtekben adták meg tizedesek helyett. A számok eredetét kutatva eljutottam ISAAC NEWTON 1686-ban kiadott *„Mathematical Principles of Natural Philosophy"* c. művéhez:

> tendo Pendula & menſurando reflexionem, inveni quantitatem vis Elaſticæ ; deinde per hanc vim determinavi reflexiones in aliis caſibus concurſuum, & reſpondebant experimenta. Redibant ſemper pilæ ab invicem cum velocitate relativa,quæ eſſet ad velocitatem relativam concurſus ut 5 ad 9 circiter. Eadem fere cum velocitate redibant pilæ ex chalybe: aliæ ex ſubere cum paulo minore. In vitreis autem proportio erat 15 ad 16 circiter. Atq; hoc pacto Lex tertia quoad ictus & reflexiones per Theoriam com-

Ebből az alapvető munkából származó információ több, mint 269 évig élt a műszaki irodalomban.

Térjünk vissza a pattogó labdához. A $t_0, t_1, t_2, \ldots$ időintervallumok könnyen meghatározhatók a szabadesés törvényéből és a $v' = -\varepsilon v$ ütközési feltételekből.

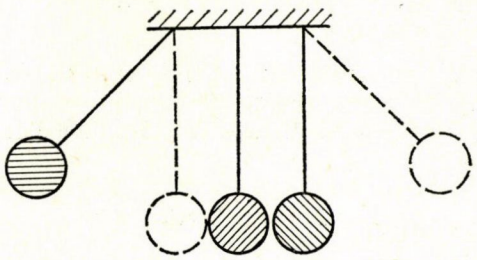
A rugalmatlan labdával kapcsolatban az az érdekes, hogy az időintervallumok sora konvergens, ugyanis

$$t = \frac{t_0}{2} + t_1 + t_2 + \ldots = \frac{1+\varepsilon}{1-\varepsilon} t_0 .$$

Ez azt jelenti, hogy a mozgás véges idő eltelte utáni megállását már a jelenlegi modell alapján is megmagyarázhatjuk. Továbbá $\varepsilon$-t a $t$ és $t_0$ időtartamok segítségével mérhetjük meg.

$$\varepsilon = \frac{t - t_0}{t + t_0} .$$

Most tekintsünk több labdát. Figyelmünket korlátozzuk a nagyon rugalmas labdákra, azaz tegyük fel egy pillanatig, hogy $\varepsilon = 1$. Mindenki ismeri a *Newton-féle bölcsőt* vagy más néven *Mariotti-féle ütköző apparátust* (2. ábra). A mozgás a szom-



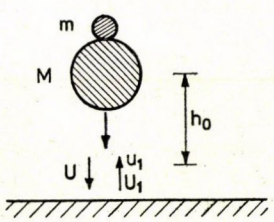
*2. ábra*

szédos tömegek között váltakozik. A szerkezet csak akkor működik helyesen, ha három feltétel teljesül

(i) tökéletesen rugalmas az ütközés ($\varepsilon = 1$),
(ii) a tömegek egyenlők,
(iii) nincs átfedés az ütközési intervallumok között.

*Labda egy másik labdán*

Az egyenlő tömegek jólismert esetének tanulmányozása helyett kísérletezzünk két nagyon különböző tömegű labdával (3. ábra). Ha külön-külön leejtik őket, akkor az eredeti magasságnak mintegy a feléig emelkednek fel ismét ($\varepsilon \approx 0{,}7$). Azonban mi történik akkor, ha a kicsit a nagy tetejére tesszük és hagyjuk, hogy együtt essenek le.




*3. ábra*

Nagy hablabdákat használunk a kísérlethez, mert ezeket könnyebb kezelni. Az $\frac{m}{M} \to 0$ határesetet tárgyalom, amikor is a nagy labda mozgását a kisebb labda egyáltalán nem befolyásolja. Tegyük fel, hogy $\varepsilon$ ugyanaz az érték mind a padlóval, mind az egymással való ütközés esetében, és nincs időbeli különbség az ütközések között. A nagy labda verődik vissza a földről és róla, mint egy teniszütőről a kicsi. Ha $U$ a szabadesés során elért sebességük, akkor az ütközés utáni sebesség

$$U_1 = \varepsilon U \qquad \text{a nagy labda esetén,}$$
$$u_1 = \varepsilon U + \varepsilon(U + \varepsilon U) = \varepsilon/2 + \varepsilon/U \qquad \text{a kicsi labda esetén.}$$

A magasság, amit az energiamegmaradás törvénye alapján elérnek

$$H_1 = \varepsilon^2 h_0 \approx 0{,}5 h_0 \qquad \text{a nagy labda esetén,}$$
$$h_1 = \varepsilon^2 (2 + \varepsilon)^2 h_0 \approx 3{,}6 h_0 \qquad \text{a kicsi labda esetén,}$$

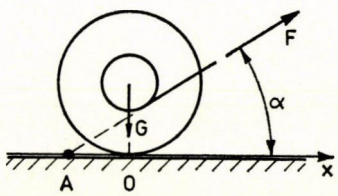
ha az ütközési együttható megfigyelt értéke 0,7. Tökéletesen rugalmas ütközés esetén, azaz, ha $\varepsilon = 1$, $h_1 = 9h_0$ lenne [1].

*Megjegyzés:* Egy diák, aki megértette ezt a folyamatot, az ismeri a nyitját a sokkal bonyolultabb ütközési folyamatok megértésének, mint mondjuk az amerikai *Pioneer* és *Voyager* űrszondáknak a *Jupiter bolygón* történt „himbálódzása".

## 2. Forgó kerék

*Engedelmes orsó*

Mindenki csinálhat magának egy egyszerű játékot egy orsó és egy darab fonal segítségével (4. ábra). A „gyere ide" és a „menj el" parancsokat adhatjuk ki, amelyeket az orsó végre is hajt. Világos, hogy a zsineget nagyon közel kell tartanunk az egyensúlyi irányhoz, amelynél az $F$ erő egyenesen az $O$ érintkezési ponton (a mozgás



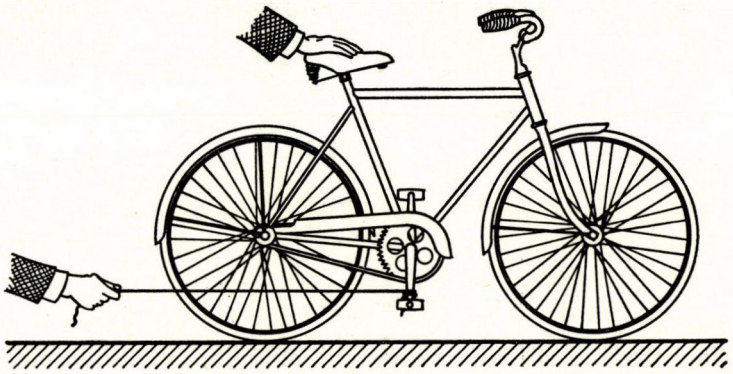

*4. ábra*

pillanatnyi középpontján) halad keresztül. Ha egy picit (csaknem észrevehetetlenül) meredekebben tartjuk a zsinórt, akkor az óra járásával ellentétes irányú forgatónyomaték eltávolítja az orsót. Ellenkező esetben az óra járásával megegyező irányú forgatónyomaték a kísérletező keze felé gyorsítja fel az orsót. A forgatónyomatékot az $\alpha$ szög meredekségével és az $A$ pont $x_A$ koordinátájával lehet kifejezni.

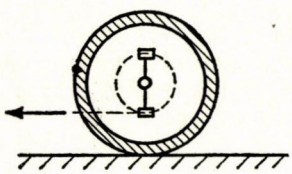
$$M = Fx_A \sin\alpha \lessgtr 0.$$

*A bicikli probléma*

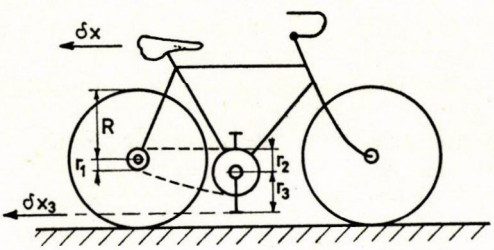
Most már az olvasó nyilván meg tudja oldani a következő problémát. Milyen irányban fog a bicikli elmozdulni abban a helyzetben, amit az 5. ábra mutat. (A nyergen levő kéz csak azt akadályozza meg, hogy a bicikli oldalra dűljön[1]. A probléma az engedelmes orsóra vezethető vissza, amikor nincs fogaskerék és a lánckerékről



*5. ábra*



*6. ábra*




*7. ábra*

és a váztól eltekinthetünk (6. ábra). Rövid kinematikai meggondolás után látható, hogy a pedál $\delta x_3$ és a váz $\delta x$ elmozdulása az alábbi

$$\delta x_3 = \left(1 - \frac{r_1 r_3}{r_2 R}\right) \delta x$$

összefüggésben van egymással (7. ábra). Mindkét elmozdulásnak ugyanaz az előjele, ha csak a bicikli nem rendelkezik valamilyen igen nagy fogaskerék áttétellel. A kerékpár visszafelé gurul.

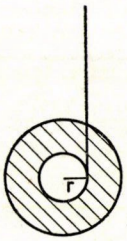
Visszatérve az orsóhoz megfigyelhetjük, hogy a forgás lehetetlenné válik, ha az $F$ erő nagysága meghalad egy bizonyos korlátot és csúszás vagy az orsó felemelése következik be. Az orsó felemelkedik, ha $P > G/\sin \alpha$.

[1] A kéz viszonylag kis erővel húzza a fonalat. (ford.)

A csúszás feltételének meghatározása akadémikusabb kérdés és sokkal bonyolultabb is, így ez a hagyományos mechanika előadásokra maradhat.

*Jojó* (mászó görgettyű)

Ennek a játéknak (8. ábra) hosszú és igen érdekes története van. Nem tudjuk, hogy *Kínában* fedezték-e fel vagy több különböző helyen egymástól függetlenül. A régi görögöktől ránk maradt egy tányér, amelyen egy fiú jojóval játszik. Beszámolókból tudjuk, hogy a *Fülöp szigeteken* mint vadászfegyvert használták. Európában a jojó körül hihetetlen hírverés folyt az 1790-es években *Franciaországban*. A *Párizs-*




*8. ábra*

*ból* külföldre menekült nemesek, akik értékes jojóikat magukkal vitték az emigrációba, „l'émigrette" vagy „coblentz" néven említették. A napóleoni idők egyik leghíresebb játékosa *Wellington* volt. Azonban magát a jojó nevet csak a mi századunk elején alkotta meg és jegyeztette be egy amerikai férfi, név szerint LUIS MARX. Bizonyosra vehető, hogy a név a régebbi francia „joujou" kifejezésből származik [22]. Napjainkban a játék időről időre feléled, mint az 1920 körüli jojó járvány vagy a világbajnokság 1975-ben *Londonban*. Ez a jojó kampány, amit nagy reklámhadjárat is kísért, sikeres vol *Angliában*, de alig lelt visszhangra *Németországban*, ami talán azt tükrözi, hogy a két országban egészen különbözőek a hagyományok az egyszerű játékok területén. A gyakorlott játékosok számos figurát és trükköt ismernek (pl. a „hurkolás", „vízesés", „kutyasétáltatás"). Egynémely jojó esetében a fonál nincs rögzítve a belső palásthoz, hanem hurkot alkot akörül. Az ilyen jojók tudnak „aludni", azaz azonnal pörögni a legalsó helyzetben és megindulhatnak felfelé a kéz egy hírtelen rántására.

Hogy egyszerűen kezelhessük a kérdést, a következő feltételezésekkel fogunk élni:

(i) síkban történő mozgás,
(ii) nem nyúló, tömeggel nem rendelkező, tökéletesen hajlékony és nagyon vékony fonal
(iii) igen hosszú fonal,
(iv) a fonal alsó végét a belső palásthoz erősítettük,
(v) az energia megmarad.

*Megjegyzések*

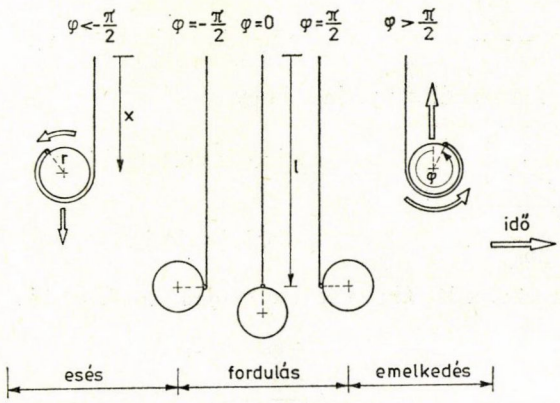
(i) A valóságban görcsös forgásokat észlelhetünk a függőleges tengely körül a felső fordulópontnál (és miért nem a mozgás során?).
(ii) A fonal megnyúlása eggyel növeli a szabadságfokok számát; a tömege és hajlítási merevsége növeli a helyzeti energiát; véges vastagsága pedig az orsó kerületének megváltozását eredményezi.

(iii) A végtelen hosszú fonal mindig függőleges marad és így a fonal irányú erő is függőlegesen hat. A jojó nyugalmi állapotból indul, a súlypont csak függőlegesen mozog és nem lépnek fel ingaszerű mozgások.
(iv) A jojó „alvását" zárja ki ez a feltétel.
(v) Ez egyszerűsítés az első közelítés céljából. A valóságban a jojó a súrlódás következtében energiájának nagy részét elveszti. Az energia pótlására a játékosnak kell a fonalat megfelelően fel és le mozgatni.

*Koordináták:*

$x$: a súlypont függőleges távolsága a felső fordulóponttól,
$\varphi$: a forgás pillanatnyi szöge a legalsó helyzethez viszonyítva az óramutató járásával ellentétes irányban mérve.

*Kinematika.* A mozgásnak három fázisát különböztetjük meg: esés, fordulás és emelkedés (9. ábra).




*9. ábra*

Végtelen hosszú fonal feltételezése esetén a jojó csak akkor emelkedik és esik, ha az egy függőleges sík mentén mozog. A forduláskor ugyanis a fonal az egyik oldalról a másikra kerül, mert a súlypont csak függőlegesen mozog.

*Az esés és emelkedés dinamikája.* A $\dot{\varphi}=d\varphi/dt$ szögsebesség a mozgás három fázisa alatt az energia megmaradásból és a forgás kinematikai feltételéből határozható meg. Az esés és az emelkedés esetében egyszerű a helyzet, ugyanis

$$\dot{\varphi}=\sqrt{\frac{2mgx}{J+mr^2}}$$

$r$: az orsó sugara,
$m$: a tömeg,
$J$: a tehetetlenségi nyomaték,
$g$: a nehézségi gyorsulás.

*A fordulás dinamikája.* A fordulási fázisban $\dot{\varphi}$-nak zárt formában megkapható kifejezése nagyon bonyolult. Az energia és az impulzusnyomaték megmaradásából

levezethető, hogy

$$\dot{\varphi}|_{\varphi=-\pi/2} = \sqrt{\frac{2mgl}{J}}\left(1+\frac{mr^2}{J}\right)^{-1/2} \leqq \dot{\varphi} \leqq \sqrt{\frac{2mgl}{J}}\left(1+\frac{r}{l}\right)^{1/2} = \dot{\varphi}|_{\varphi=0},$$

ahol $l$ a fonal hossza.
Ha két, az orsó $r$ sugarának kicsinységét bizonyos értelemben kifejező, egyszerűsítő feltevést alkalmazunk, nevezetesen, hogy

$$\frac{mr^2}{J} \ll 1 \quad \text{és} \quad \frac{r}{l} \ll 1,$$

akkor láthatjuk, hogy $\dot{\varphi}$ megközelítőleg állandó és egyenlő $\sqrt{\frac{2mgl}{J}}$-vel a fordulat alatt. Ezt a közelítést felhasználva kapjuk (l. [3]), hogy

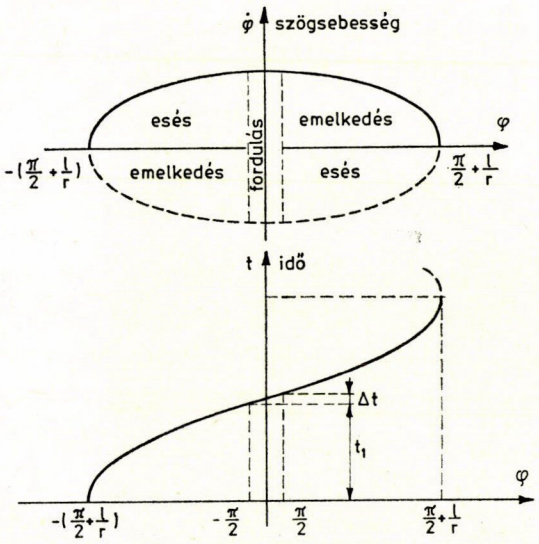
$$\dot{\varphi} = \begin{cases} \sqrt{\frac{2mgx}{J}} & \text{eséskor vagy emelkedéskor } (0 \leqq x \leqq l), \\ \sqrt{\frac{2mgl}{J}} & \text{fordulatkor } (l \leqq x \leqq l+r). \end{cases}$$

Ezt az egyenletet felhasználva nyerjük, hogy

$$t = \sqrt{\frac{J}{mr^2}}\sqrt{\frac{2x}{g}} \qquad \text{az idő az esési fázisban,}$$

$$\Delta t = \sqrt{\frac{J}{2mgl}}\,\pi \qquad \text{a fordulat ideje.}$$

A fenti összefüggéseket a 10. ábra két grafikonján szemléltettük.




*10. ábra*

*Példa* (nagy jojó)

$l=100$ cm
$m=200$ g
$J=104$ g cm$^2$
$r=2$ cm

$t_1 \approx 1{,}6$ sec az esés ideje,

$\Delta t \approx 0{,}05$ sec a fordulat ideje.

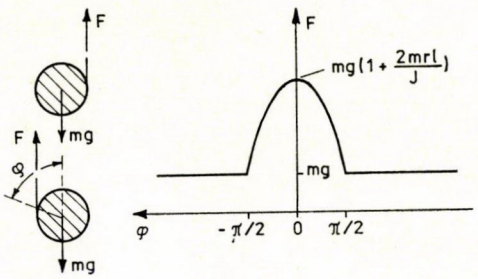
A fordulat időtartama csak mintegy 3%-a az esés idejének. A fenti, egyszerűsítő feltételek kielégítően teljesülnek, ugyanis

$$\frac{mr^2}{J} = 0{,}08 \ll 1 \quad \text{és} \quad \frac{r}{l} = 0{,}02 \ll 1.$$

A fordulat olyan gyors, hogy egy közelítő elméletben ütközésként lehet felfogni.

*A fonálban ébredő erő* (l. a 11. ábrát). A jojó gyorsulása az esés és az emelkedés alatt állandó, így a fonálban ébredő erő is az. Az impulzusnyomaték vizsgálatából kapjuk, hogy

$$F = \frac{mg}{1+\frac{mr^2}{J}} \approx mg \quad \left(\frac{mr^2}{J} \ll 1\right).$$




*11. ábra*

A $\dot{\varphi}$ szögsebesség a fordulás alatti ismert értékéből kaphatjuk $\left(\frac{mr^2}{J} \ll 1, \frac{r}{l} \ll 1\right)$ [3], hogy

$$F = mg\left(1+2\frac{mrl}{J}\cos\varphi\right) \quad |\varphi| \leqq \frac{\pi}{2}.$$

*Példa* (a fenti nagy jojó)

$$F_{\max} = mg\left(1+2\frac{mrl}{J}\right) \approx 9mg.$$

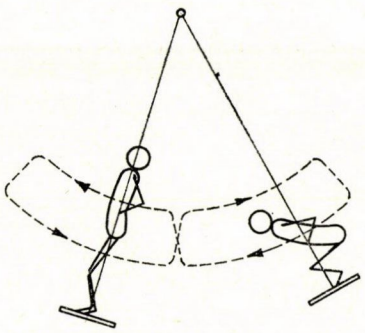
Tehát a maximális erő kilencszerese az orsó súlyának.

*Megjegyzés.* A jojó elméleti tárgyalásában a következő lépés az lenne, hogy különböző zavaró hatással bíró erőket (a súrlódás a fonal és a test között, a fonal

nem rugalmas megnyúlása a fordulás során a legnagyobb terheléskor, a légellenállás) tekintetbe vennénk és vizsgálnánk a játékos tevékenységét, aki a fonal felső végének szabályos mozgatásával energiával látja el a rendszert és a mozgást közel periodikussá teszi.

### 3. Rezgő játékok

A legtöbb rezgő játék valamely öngerjesztő és ezért nemlineáris oszcilláció felhasználásával működik. A legismertebb példa a gyermekek hintája (12. ábra), melynél paraméteres gerjesztés van.



*12. ábra*

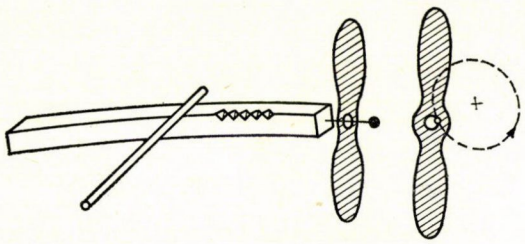
A hinta legegyszerűbb modellje egy változó hosszúságú matematikai inga. A madzag hirtelen megnyúlása a forduló pontokban és összehúzódása abban a pillanatban, amikor a zsineg a függőleges síkon halad keresztül, reprezentálja a gyermek irányító mozgását [4, 5]. Mint a 12. ábra mutatja, a gyermek a végpontokban leguggol és középen feláll.

A következő rezgő játékot egy egyszerű lineáris kényszerrezgésen alapuló modell segítségével lehet elmagyarázni:

*A mágikus szélmalom* (rovátkált bot (13. ábra))

Ezt a játékot az eliptikusan polarizált rezgés bemutatása végett találták ki mintegy 40 évvel ezelőtt [6].

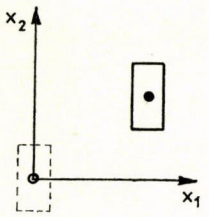
Első pillanatban csodának tűnik, hogy a légcsavar forog, amikor a botot megfelelően dörzsöljük a rovátkák mentén. Mindenekelőtt azt a mechanizmust kell



*13. ábra*

megértenünk, ami a légcsavart mozgatja. A lyuk átmérője egy kicsit nagyobb, mint a szög vastagsága. Ha a szög forog, akkor a propellert annak tehetetlensége és a surlódási erő következtében a gerjesztő mozgás irányába hajtja.

De miért mozog a bot szabad vége egy ellipszis vagy kör pályán, valahányszor a rovátkákat dörzsöljük. A játék akkor is működik, ha azt a kezünkben tartjuk vagy egy asztalhoz rögzítjük. Így szükségképpen egy rugalmas rúd görbevonalú gerjesztett rezgésének feltételezéséhez jutunk. A rezgés, amit szabad szemmel megfigyelhetünk, ha a bot vékony, erősen csillapított. Reménytelennek tűnik egy rögzített, rugalmas rúd gerjesztett rezgésének tárgyalása még a csillapítatlan esetben is. Avégett, hogy a szerkezet fő tulajdonságait megértsük, egy kettő szabadságfokú, egyszerű modellt fogunk tanulmányozni (a bot két egymásra merőleges irányban tud hajlékonyan rezegni (14. ábra)). Megadjuk a harmonikus gerjesztő erő által létrehozott




*14. ábra*

rezgés két, egymástól független egyenletét az $x_1$ és $x_2$ irányában, $\Lambda$ rezgésszám esetén:

$$m_1\ddot{x}_1+c_1\dot{x}_1+k_1x_1 = a_1\cos\Lambda t,$$

$$m_2\ddot{x}_2+c_2\dot{x}_2+k_2x_2 = a_2\cos\Lambda t,$$

ahol

$m_i$: az ekvivalens „rezgő" tömeg;
$k_i$: a rugalmas erő együtthatója;
$c_i$: a csillapítási együttható.

Világos, hogy a dörzsölés a frekvenciák egy egész spektrumát hozza létre, de a rezonancia miatt ezek közül csak meghatározottak erősödhetnek fel.

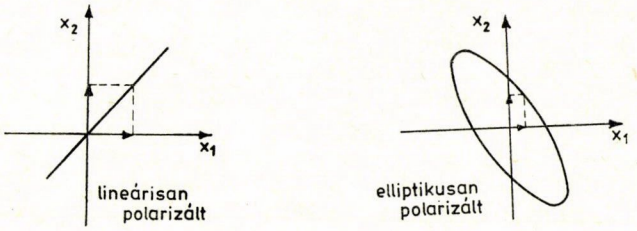
Elegendően hosszú idő múlva, amikor a tranziens tagok (a homogén egyenletek megoldása) exponenciálisan lecsengenek, csak a harmonikus rezgés marad meg $\Lambda$ gerjesztő frekvenciával:

$$x_i = A_i\cos(\Lambda t-\alpha_i) \quad (i = 1, 2)$$

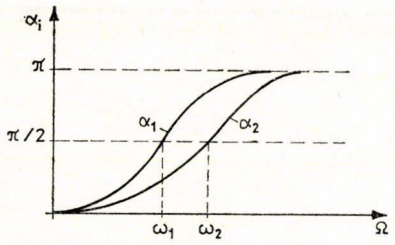
$$A_i = a_i((\omega_i^2-\Lambda^2)^2+(c_i\Lambda/m_i)^2)^{-1/2} \quad \text{(amplitudó)}$$

$$\alpha_i = \tan^{-1}\left(\frac{c_i\Lambda}{m_i(\omega_i^2-\Lambda^2)}\right) \quad \text{(fázis)}$$

$\left(\omega_i=\sqrt{\frac{k_i}{m_i}}\right.$ a csillapítatlan rezgés saját frekvenciái$\left.\right)$. Akkor kapunk nagy amplitudót, ha $\Lambda$ a két természetes frekvencia valamelyikéhez közel esik. Ha a két, egymásra merőleges $x_1$ és $x_2$ rezgés fázisban vannak egymással, akkor a bot vége lineáris mozgást végez, amely nem tudja mozgásba hozni a légcsavart. Csak ha a




*15. ábra*



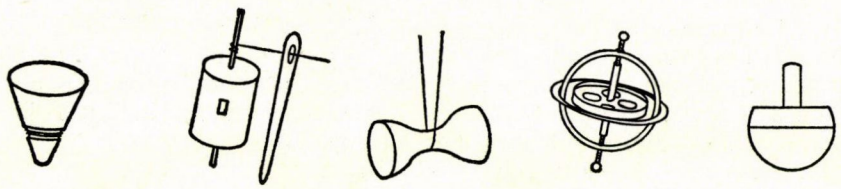

*16. ábra*

merőleges rezgések fáziskülönbséget mutatnak, akkor jár be a szög egy elliptikus utat (15. ábra).

Egy közönséges rezonálás esetén a fáziseltolódásokat mutatja a 16. ábra, melyből nyilvánvaló, hogy a két természetes frekvencia nem nagyon különbözhet egymástól. Téglalap keresztmetszetű botok jobban megfelelnek, mint kör vagy négyzet keresztmetszetűek. Ez magyarázza meg, hogy a négyzet keresztmetszetű szélmalmok nem működnek jól, kivéve ha a szimmetriát valamilyen módon elrontjuk, pl. az ujjunkkal az egyik oldalról megfeszítjük a botot. Természetesen a két merőleges oldalról különböző módon gerjeszthetjük a mozgást.

## 4. Pörgettyűk

Egy, a mechanikai játékokról szóló előadás igencsak szegényesen „peregne le", ha nem említenénk meg a pörgettyűket. A pörgettyűk mozgása ugyancsak misztikus. Nekem 40 darab különféle pörgettyűm van a legkülönbözőbb méretekben. A legfontosabb típusok az ostorral hajtott és a hegyes pörgettyű, a diaboló, a giroszkóp és a billenős pörgettyű (17. ábra). Rendkívül nehéz, ha nem lehetetlen, visszavezetni valamennyit a történelmi gyökereihez. Pörgettyűkkel már az *ókori Kínában* is játszottak. Az i. e. VIII. sz.-ból származó görög pörgettyűk találhatók a *British Múzeumban* és tudjuk, hogy a görög fiúk már az ókori időkben hajtották ostorukkal a pörgettyűjüket. Európában a középkortól játszottak velük és hosszú kultúrális hagyományokkal rendelkeznek pl. *Japánban*, ahol a pörgettyű készítése igen magas fokra fejlődött. Míg a nyugati államokban az emberek csak öt- vagy hatfélét ismernek, addig távolkeleten mintegy száz fajtáját különböztetik meg. Egy szerfölött



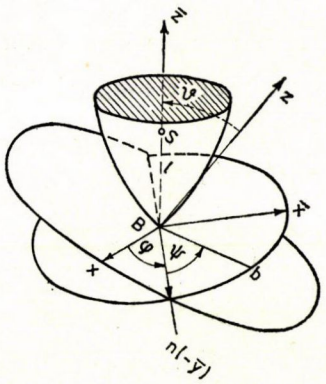

*17. ábra*

nevezetes játékot űznek *Malaysiában*, ahol 9 hüvelyk (22,86 cm — ford.) átmérőjű és 12 font (kb. 6 kg — ford.) súlyú, diszkoszhoz hasonló pörgettyűket a gyakorlott játékosok olyan sokáig forgatnak, ameddig csak lehet. A csúcs 1 óra 47 perc.

A játék célját szolgáló pörgettyűk kevés kivételtől eltekintve szimmetrikusak. A pörgettyűk legegyszerűbb osztálya, amivel foglalkozunk, a giroszkóp és a hegyes végű pörgettyű, mely utóbbinál az eszköznek a talajon való gurulásától és csúszásától eltekinthetünk. Ezek a játékpörgettyűk azonban nem működnek jól, ha a végük túl hegyes. A súlyos, egy fix ponttal rendelkező pörgettyűk elméletének legnagyobb része a XIX. századból eredő klasszikus anyag. Manapság a dinamika alapkurzusain adják elő. Mindazonáltal azonban mozgásuk elemzése egyáltalán nem könnyű. Térbeli kinematika és mozgó koordinátarendszerek feltétlenül szükségesek hozzá, azonban a tehetetlenség tenzorának fogalma elkerülhető, ha a tehetetlenség alkalmas főtengelyeit használjuk koordinátatengelyekként.

*Szimmetrikus pörgettyűk giroszkópikus mozgása*

Bevezetünk egy mozgó $\bar{x}, \bar{y}, \bar{z}$ koordinátarendszert, úgy, hogy a $\bar{z}$ tengely a pörgettyű szimmetriatengelye, melyen egyben az eszköz tömegközéppontja is található, és $\bar{y}$ merőleges $\bar{z}$-re a vízszintes síkban (18. ábra). Megjegyzendő, hogy $\bar{x}, \bar{y}$ és $\bar{z}$ a tehetetlenség főtengelyei, annak ellenére, hogy $\bar{x}$ és $\bar{y}$ *nincsenek* a testhez rögzítve, továbbá mindhárman a függőleges tengely körül forognak. A növekvő $\psi$ szöggel




*18. ábra*

haladó „$b$" egyenes a testhez van rögzítve. Szükségünk lesz továbbá egy $x, y, z$ inerciarendszerre, ahol $z$ a függőleges tengely, $x$ és $y$ (utóbbi nem látszik a 18. ábrán) a vízszintes síkot feszíti ki. Az $n$ egyenest (18. ábra) csúcsponti egyenesnek nevezik. A két koordinátarendszer közti kapcsolat a *három Euler-féle szöggel* írható le

$\varphi$ (precessziós szög) a $z$ tengely körül forog,
$\vartheta$ (nutációs szög) a $-\bar{y}$ tengely ($n$) körül forog,
$\psi$ (rotációs szög) a $\bar{z}$ tengely körül forog.

Az $\bar{x}, \bar{y}, \bar{z}$ koordináta rendszerben a mozgó vonatkoztatási rendszer $\vec{\Lambda}$ és a pörgettyű $\vec{\omega}$ forgási sebessége a következő:

$$\vec{\Lambda} = \begin{pmatrix} \Lambda_1 \\ \Lambda_2 \\ \Lambda_3 \end{pmatrix} = \begin{pmatrix} \dot{\varphi} \sin \vartheta \\ -\dot{\vartheta} \\ \dot{\varphi} \cos \vartheta \end{pmatrix} \quad \text{és} \quad \vec{\omega} = \begin{pmatrix} \omega_1 \\ \omega_2 \\ \omega_3 \end{pmatrix} = \begin{pmatrix} \dot{\varphi} \sin \vartheta \\ -\dot{\vartheta} \\ \dot{\varphi} \cos \vartheta + \dot{\psi} \end{pmatrix}.$$

Ha a szimmetriatengelyre, valamint a fixponton átmenő, merőleges tengelyekre vonatkozó tehetetlenségi nyomatékot rendre $C$-vel és $A$-val jelöljük, akkor az impulzusnyomaték

$$\vec{D} = \begin{pmatrix} A\omega_1 \\ A\omega_2 \\ C\omega_3 \end{pmatrix}.$$

Legyen $m$ a pörgettyű tömege és $l$ az $S$ súlypont távolsága a fixponttól (18. ábra). Ekkor a *mozgás Euler-féle egyenletei* a következő alakban írhatók [7, 8]:

$$\bar{x}: \quad A(\ddot{\varphi} \sin \vartheta + 2\dot{\varphi}\dot{\vartheta} \cos \vartheta) - C\omega_3 \dot{\vartheta} = 0$$

$$-\bar{y}: \quad A(\ddot{\vartheta} - \dot{\vartheta}^2 \sin \vartheta \cos \vartheta) + C\omega_3 \dot{\varphi} \sin \vartheta = mgl \sin \vartheta$$

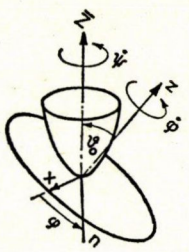
$$\bar{z}: \quad C\dot{\omega}_3 = 0$$

A harmadik egyenletből $\omega_3 =$ állandó. Egy fontos speciális mozgás az

*reguláris precesszió* ($\vartheta = \vartheta_0 =$ állandó, 19. ábra).

Az első egyenletből vagy $\vartheta_0 = 0$, vagy $\pi$, avagy $\dot{\varphi} =$ állandó. Ha $\vartheta_0 \neq 0, \pi$ akkor a második egyenlet $\dot{\varphi}$-re nézvést egy másodfokú algebrai egyenletté redukálódik:

$$(A \cos \vartheta_0 \dot{\varphi}^2 - C\omega_3 \dot{\varphi} + mgl) \sin \vartheta_0 = 0.$$




*19. ábra*

Megoldása a precesszió szögsebessége [8]:

$$\dot{\varphi} = \begin{cases} \dfrac{C\omega_3}{2A\cos\vartheta_0}\left(1 \mp \sqrt{1-\dfrac{4mglA\cos\vartheta_0}{C\omega_3^2}}\right) & \text{lassú precesszió} \quad \left(\vartheta_0 \neq \dfrac{\pi}{2}\right) \\ \dfrac{mgl}{C\omega_3} & \text{gyors} \quad \left(\vartheta_0 = \dfrac{\pi}{2}\right). \end{cases}$$

A lassú precesszió az, amit normális esetben megfigyelhetünk. Nyilvánvaló, hogy reguláris precesszió csak akkor lehetséges, ha az $\omega_3$ szögsebesség elegendően nagy (azaz nagy a $\dot{\psi}$ forgás)

$$|\omega_3| > \sqrt{\frac{4mglA\cos\vartheta_0}{C^2}}.$$

A pörgettyű akkor „gyors", ha $\dfrac{mgl}{C\omega_3^2} \ll 1$. Ekkor sorbafejtve az alábbi gyökös kifejezést kapjuk

$$\dot{\varphi} = \begin{cases} \dfrac{mgl}{C\omega_3}\left(1 - \dfrac{mglA\cos\vartheta_0}{4C^2\omega_3^2} \pm \ldots\right) & \text{lassú precesszió} \\ \dfrac{C\omega_3}{A\cos\vartheta_0}\left(1 - \dfrac{mglA\cos\vartheta_0}{C^2\omega_3^2} \pm \ldots\right) & \text{gyors precesszió} \end{cases}$$

Most $\omega_3 \to \infty$ határátmenet esetén azt kapjuk, hogy lassú a precesszió szögsebessége, $\omega_p = \dfrac{mgl}{C\omega_3}$, vagyis fordítottan arányos $\omega_3$-mal és független $\vartheta_0$-tól. Másrészt a gyors precesszió szögsebessége $\tilde{\omega}_p = \dfrac{C\omega_3}{A\cos\vartheta_0}$ asszimpatikusan függetlenné válik a gravitációtól.

*Nutáció.* Ha a lassú precesszió esetén a pörgettyü gyorsan forog $\vartheta_0$ deklimációval és megzavarják, akkor nutációba kezd $\vartheta_0$ körül. A $\dot{\varphi} - \omega_p$, $\vartheta - \vartheta_0$ eltérések az egyenletes precessziótól $\omega_3 \to \infty$ esetén az alábbi lineáris differenciál egyenletekkel írhatók le [9]:

$$\sin\vartheta_0\,\ddot{\varphi} - \omega_n\dot{\vartheta} = 0$$

$$\ddot{\vartheta} + \omega_n \sin\vartheta_0(\dot{\varphi} - \omega_p) = 0$$

$\omega_n = \dfrac{C\omega_3}{A}$: a nutáció szögsebessége

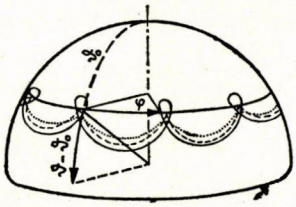
$\omega_p = \dfrac{mgl}{C\omega_3}$: a (lassú) precesszió szögsebessége

A $\vartheta = \vartheta_0 \neq 0$, $\varphi = 0$, $\dot{\vartheta} = 0$, $\dot{\varphi} = \omega_0$ ($t = 0$ esetén) kezdeti értékhez tartozó megoldás

$$\varphi(t) = \omega_p t - \frac{\omega_p - \omega_0}{\omega_n}\sin\omega_n t$$

$$\vartheta(t) = \vartheta_0 + \frac{\omega_p - \omega_0}{\omega_n}\sin\vartheta_0(1 - \cos\omega_n t).$$

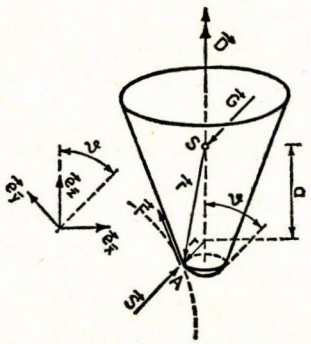
Ez a $\varphi - \vartheta$ síkon közönséges, rövidülő, ill. bővülő ciklois az $\omega_0 = 0$, $0 < \omega_0 < \omega_p$, ill. $\omega_0 < 0$ esetben (20. ábra).



*20. ábra*

*A pörgettyű felemelkedésének közelítő elmélete*

Általában a pörgettyűs játékoknak nincs rögzített pontjuk, hanem a támaszsíkon forognak és csúsznak, így ki vannak téve a súrlódási erő hatásának. A pörgettyű gördülése szempontjából a súrlódás szintén igen fontos. Alábbi fejtegetéseinket COULOMBnak egy érintkezési pontban, száraz körülmények között fellépő súrlódásra vonatkozó feltételezéseire alapozzuk ($\mu$ a súrlódási együttható). A pörgettyü végét egy $r$ sugarú gömbszeletre formázhatjuk (21. ábra). Az egyszerűség kedvéért tegyük



*21. ábra*

fel, hogy a pörgettyű olyan gyors, hogy az impulzusnyomaték-vektor csaknem a $\bar{z}$ (szimmetria) tengely irányába mutat, azaz $\vec{D} \approx C\omega_3 \vec{e}_{\bar{z}}$. A precesszió $\omega_p$ szögsebessége nagyon kicsi nagy $\omega_3$ esetén. Végezetül még azt tesszük fel, hogy a súlypont gyorsulása nagyon kicsi a gravitációhoz képest (ami szintén azt jelenti, hogy a pörgettyű nagyon gyors és a feltételezés jogossága a posteriori igazolható). Ezen felfeltételezések mellett az $A$ érintkezési pontnál ható $\vec{S}$ erő nagysága megegyezik, iránya ellentétes a $\vec{G}$ súlyéval:

$$\vec{S} \approx mg(\sin\vartheta \vec{e}_{\bar{x}} + \cos\vartheta \vec{e}_{\bar{z}}) = -\vec{G}.$$

A súrlódási erő irányát az érintkezési pontnál fellépő $-r\dot{\psi}\vec{e}_{\bar{y}}$ sebesség határozza meg és nagysága

$$\vec{F} \approx \mu m g \vec{e}_{\bar{y}} \quad (\text{ha } \omega_3 \approx \dot{\psi} > 0).$$

Az érintkezési ponttól a súlyponthoz mutató $\vec{r}$ vektorra

$$\vec{r} = -r\sin\vartheta \vec{e}_{\bar{x}} - (a + r\cos\vartheta)\vec{e}_{\bar{z}}$$

(21. ábra), és a súlypontra vonatkozó forgatónyomaték

$$\vec{M} = \vec{r} \times (\vec{F} + \vec{S}) = \begin{pmatrix} \mu mg(a + r\cos\vartheta) \\ -mga\sin\vartheta \\ -\mu mgr\sin\vartheta \end{pmatrix} \approx \begin{pmatrix} \mu mga \\ -mga\sin\vartheta \\ -\mu mgr\sin\vartheta \end{pmatrix},$$

ahol az $r \ll a$ összetevőt hanyagolhatjuk el az $\bar{x}$-komponensből. Ha csak azokat a tagokat tartjuk meg az impulzusnyomatéknál, amelyek befolyásolják $\omega_3 \to \infty$ esetén a határértékeket, akkor a következő egyenleteket kapjuk [9]

$$\bar{x}: \quad -C\omega_3\dot{\vartheta} = \mu mga$$

$$-\bar{y}: \quad C\omega_3\dot{\varphi}\sin\vartheta = mga\sin\vartheta$$

$$\bar{z}: \quad C\dot{\omega}_3 = -\mu mgr\sin\vartheta.$$

Ha $\vartheta \neq 0, \pi$ akkor azt kapjuk, hogy

$$\dot{\vartheta} = -\frac{\mu mga}{C\omega_3} = -\mu\omega_p \qquad \text{a pörgettyű emelkedésének szögsebessége}$$

$$\dot{\varphi} = \frac{mga}{C\omega_3} = \omega_p \qquad \text{a lassú precesszió szögsebessége}$$

$$\dot{\omega}_3 = -\mu\frac{mgr\sin\vartheta}{C} = -\mu\frac{r}{a}\omega_3\omega_p\sin\vartheta \qquad \text{szöggyorsulás (valójában lassulás)}$$

Így a súrlódási erő addig emeli a pörgettyű tengelyeit, amíg azok el nem érik a függőlegest. Világos, hogy ugyanakkor a súrlódás lelassítja a forgó mozgást, azaz $\dot{\psi}$-t egyre csökkenti. Ez azonban a pörgettyű emelkedéséhez képest lassú folyamat

$$\frac{\dot{\omega}_3}{\omega_3\dot{\vartheta}} = \frac{r\sin\vartheta}{a} \ll 1.$$

Máskülönben a pörgettyű emelkedése nem lenne megfigyelhető. Az emelkedés ideje $\vartheta = \vartheta_0$-ból indulva

$$t = \frac{C\omega_3}{\mu mga}\vartheta_0.$$

*A pörgettyű „alvása"*

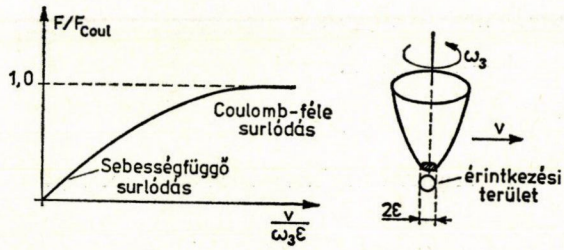
Miután a pörgettyű felemelkedett, függőleges helyzetben marad („alszik"). Alvás azonban csak akkor lehetséges, ha a forgás elegendően gyors. Habár a pörgettyű alvásának stabilitása nem függ explicit módon a súrlódási erő nagyságától, de mégis érzékeny a súrlódás természetére, azaz hogy *Coulomb-féle száraz súrlódás* vagy sebesség függő súrlódás lép-e fel.

*Súrlodás*

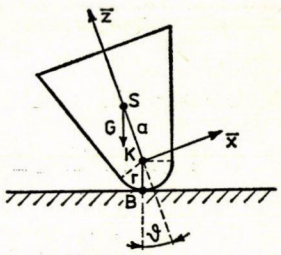
Csak mostanában mutatott rá CONTENSOU [10], hogy a pörgettyű fúró jellegű mozgása, amely hozzáadódik a haladáshoz, egy olyan súrlódási erőt kelt, amely erő arányos a haladó mozgás sebességével, ha a fúró mozgás a domináns. Ha a haladó

mozgás az uralkodó, akkor a súrlódás törvényei megközelítik a *Coulomb-féle száraz súrlódását* (22. ábra). Ez az összefüggés az érintkezési felület elemein fellépő súrlódási erők integrálásából ered.

*A pörgettyű alvásának stabilitása* (beleértve a billenős pörgettyűt is). A kérdést MAGNUS [7] vizsgálta éppen a *súrlódás Contensou-féle törvénye* alapján. A függőlegestől kicsit eltérő, gyorsan forgó pörgettyű esetében a sebességgel arányos súrlódás feltételezése megfelelő közelítés. Szimmetrikus pörgettyűket vizsgálunk, amelyek vége $r$ sugarú gömbsüveg (23. ábra), az $S$ súlypont és a gömb $K$ középpontjának távolsága $a$; $G$ a pörgettyű súlya. Mivel általában a testnek nincs fixpontja, ezért a




22. ábra



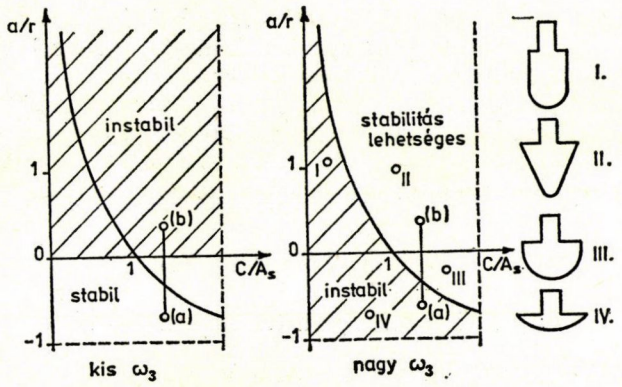

23. ábra

vizsgálatot a tömegközéppontra vonatkozó impulzus és impulzusnyomaték megmaradása alapján végezzük. A tömegeloszlás a súlypontra vonatkozó $C_s=C$ és $A_s$ tengelyirányú és oldalsó tehetetlenségi nyomaték formájában lép be. Az alvó mozgás a megfelelő egyenletek egy univerzális megoldása tetszőleges $\omega_3$ szögsebességre. MAGNUS levezetett egy linearizált egyenlőtlenséget, amely az „alvó" mozgás stabilitásának szükséges feltétele kis perturbáció esetén, és független a súrlódás hatásáról, eltekintve attól, hogy a hatás nem tűnhet el:

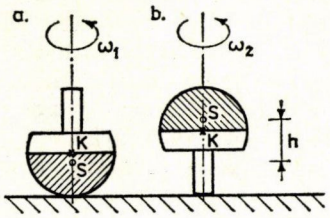
$$\frac{C}{A_s}\left(1+\frac{a}{r}\right)-1>\frac{a}{r}\left(1+\frac{a}{r}\right)^2\frac{rG}{A_s\omega_3^2}.$$

Valóban, ezen egyenlőtlenség szerint az alvó pörgettyű stabilitása független a súrlódási erő nagyságától (habár érzékeny a súrlódás természetére). A továbbiakban adottnak vesszük az $r$, $A_s$ és $G$ mennyiségeket és a $C/A_s-a/r$ síkon $\omega_3$ paraméterrel vizsgáljuk a stabilitás tartományát. (Világos, hogy az egyenlőtlenség $\omega_3$ négyzetét tartalmazza, hiszen a szimmetrikus pörgettyű stabilitása független a forgás irányától.) A definíciója szerint $a/r$ nagyobb, mint 1, és $C/A_s$-0-tól 2-ig változhat (a botszerűtől a lemezalakú pörgettyűig). A stabilitási feltétel teljes kifejtése megtalálható a 24. ábrán.

Egy közönséges pörgettyű — II — nyilván stabil nagy $\omega_3$ esetén. Ha a súrlódási erő már eléggé lelassította a pörgettyűt, akkor a nagy $\omega_3$-ra vonatkozó stabilitási diagram többé nem alkalmas a helyzet leírására. A pörgettyű bukdácsolni kezd, végül ledől.




*24. ábra*




*25. ábra*

*Billenős pörgettyű* (megforduló pörgettyű)

Ez egy valódi modern játék. Az első szabadalmat egy német nő, FRÄULEIN HELENE SPERL kapta rá 1891-ben *Münchenben,* habár a működés magyarázata a szabadalmi leírásban messze állt a valóságtól. A szokásos formája egy gömbszelet némileg túlméretezett nyéllel (25. ábra). Amikor nem forog, akkor a mechanikai tulajdonságai egy keljfeljancsihoz hasonlítanak. Ha a játékot elegendően nagy perdülettel mozgásba hozzuk egy közepesen síma asztalon, akkor megfordul és a nyelére áll, miközben forgási irányát megtartja. A megfordulás dinamikája meglehetősen bonyolult és még manapság is vizsgálják (l. [24, 25]). Azonban két egyszerű kérdés megválaszolható anélkül, hogy belemennék a játék bonyolult dinamikájának részleteibe:

i) miért pörög stabilan a nyelén, amikor a gömbtesten nem,

ii) a súrlódási erők okozzák-e a megfordulását.

i) Az „alvó" billenős pörgettyű stabilitása:

A játék a 25. ábrán látható normál (nyéllel fölfelé álló) pozícióban (a) és megfordítva (b). Keljfeljancsiként viselkedik, ha $\omega_3 = 0$. Az $S$ súlypontnak alacsonyabban kell lennie, mint a gömb $K$ középpontjának $(a/r < 0$ az (a) helyzetben). A stabilitási diagrammok alátámasztják a megfigyelést, hogy a billenős pörgettyű csak kis $\omega_3$ esetén forog stabilan a gömbi részén, míg ha $\omega_3$ elegendően nagy, akkor stabil a nyelén.

ii) Súrlódás nélkül a megfordulás nem lehetséges:

Az (a) helyzetből a (b)-be való forduláshoz a súrlódás alapvetően fontos, mert mint meg fogom mutatni, elvileg szükséges hozzá (megcáfolva ezzel egy ír fizikus, J. L. SYNGE egy korábbi állítását) [11]. Mint már korábban észrevettük, a játék normál, (a) helyzete statikailag stabil, ami azt jelenti, hogy az $S$ súlypont lejjebb van a gömb $K$ középpontjánál. Tehát az ellenkező helyzetben fordítva kell lennie. Így a potenciális energia növekszik, amikor a játék a nyelére áll, és ennek következtében a mozgási energiája csökken (mindenképpen van energia veszteség a súrlódás miatt):

$$T_2 = \frac{C}{2}\omega_2^2 < \frac{C}{2}\omega_1^2 = T_1.$$

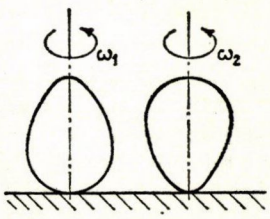
Az is következik, hogy az impulzusnyomaték szintén csökken:

$$D_2 = C\omega_2 < C\omega_1 = D_1.$$

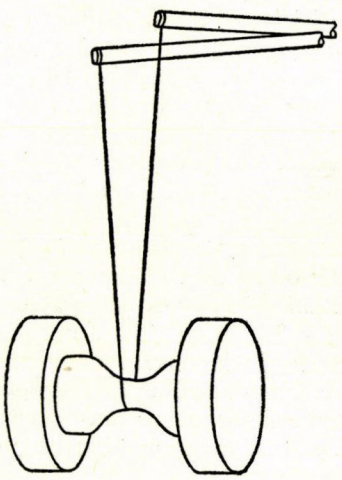
Ez azonban csak akkor lehetséges, ha a forgatónyomatékvektornak a megfordulási ideje alatt vett integrálja függőleges irányba mutat. Függőleges körüli forgatónyomaték azonban nem származhat függőleges erőkből (mint a gravitáció vagy vertikális támaszerő), hanem csak vízszintes erőtől (súrlódás). Tehát a súrlódási erő szükséges a pörgettyű megfordulásához. Egy nagyon sima asztalon ez a megfordulás igen hosszú ideig tart, ha egyáltalán megtörténik. Megjegyezzük, hogy a főtt keménytojás is egyfajta billenős pörgettyű (26. ábra).

A pörgettyűk tárgyalásának végéhez érve még két ilyen típusú játékhoz kívánok megjegyzéseket fűzni.

*Diaboló.* A diaboló (zsineges pörgettyű, „kínai ördög"), amelyet nagyszüleink oly lelkesen játszottak, a legegyszerűbb formájában két megfordított fém vagy műanyag kúpból áll (27. ábra). Két pálcához erősített zsineg segítségével pörgetik. A gyakorlott játékosok képesek magasra feldobni a levegőbe és ismét elkapni vagy egy másik játékosnak így átadni. Az „ördög" nevet gyakran azzal magyarázzák, hogy néhány




26. ábra



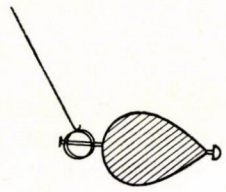
27. ábra

eredeti kínai darab magas fordulatszám esetén a beépített sípok miatt zümmögő hangot adott. A diaboló minden valószínűség szerint XV. századi kínai találmány és a XVIII. század végén hozta Európába az akkori brit nagykövet. A játék meghódította a szalonokat és hamarosan *Párizsban* is ismertté vált. 1812-ben annyira népszerű volt *Franciaországban*, hogy azt állítják, hogy az emberek több figyelmet szenteltek neki, mint *Napoleon* oroszországi hadjáratának [23]. Utcai mutatványosok még a jelen század elején is használtak nagyméretű diabolókat *Kínában*. Nyugaton az 1907-es „diaboló őrület" során éledt fel újra a játék.

A diaboló mozgását szabad repülés esetén könnyű megérteni mint nyomatékmentes pörgettyűt. Mozgása stabil, ha a $C_s/A_s$ hányados, vagyis a súlyponton keresztülhaladó szimmetria és oldalsó tengelyre vonatkozó nyomaték aránya nagyobb, mint 1 (rövid pörgettyű) vagy kisebb, mint 1 (hosszú pörgettyű).

Igen könnyű bemutatni olyan diabolót, amely nem működik, mert tehetetlensége gömbi szimmetriát mutat, azaz $C_s/A_s=1$ [12]. Egy ilyen diabolót zsineggel irányítva alig tapasztalunk forgást, az egyébként igen stabil szimmetriatengely körül, mert a zsineg a forgatónyomaték segítségével fejti ki hatását. Az említett eset bemutatása csak akkor lehet sikeres, ha a tehetetlenségi nyomatékok igen közel, néhány százalékos hibahatáron belül vannak egymáshoz viszonyítva [12].

*Török pörgettyű.* A török pörgettyű valójában egy igen egyszerű giroszkóp (28. ábra). Ez egy tojásalakú, fa testből áll, amelyben a szimmetriatengely helyén szabadon mozog egy szeg. A szeg egy gyűrűn halad keresztül. Ehhez az utóbbihoz erősítenek egy félméternyi zsineget. Először a zsineget a testre csavarják a csúcsnál kezdve, majd elengedik a pörgettyűt, miközben a zsineg szabad végét a játékos egyik kezével tartja. Amikor a zsineg letekeredik, akkor a pörgettyű már tekintélyes sebességgel forog és egy rövid ideig stabil precesszióval rendelkezik. A precesszió iránya az ellenkezőjére fordul, ha a csúcs a földet éri (miért?), mert a forgatónyomaték iránya megfordul. Sajnos a játék túl gyorsan felemelkedik, miután az érdes felületet elérte. Feltételezhetőleg ezt még javítani lehet, ha a végét még hegyesebbre készítjük.

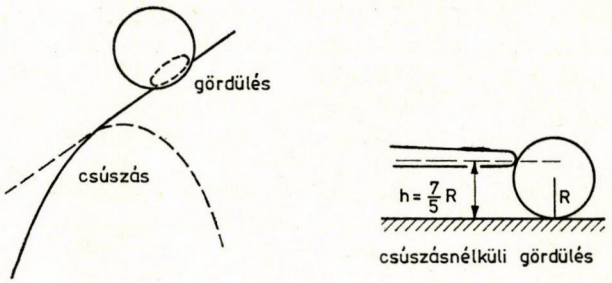


*28. ábra*

## 5. Néhány egyéb játék

Végül hadd említsek meg röviden néhány másik játékot, amelyek meglepő módon viselkednek és ezért furcsa megértenünk a mozgásukat irányító törvényeket, de ugyanakkor arra késztetnek minket, hogy nagyobb gyakorlatra tegyünk szert bennük.

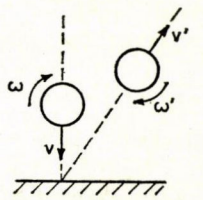
A billiárd már régóta megragadta több fizikus figyelmét is. GUSTAVE GASPARD CORIOLIS műve a *Théorie mathématique des effects du jeu de billiard,* 1835 [13] a billi-



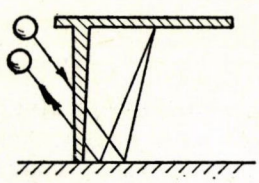

*29. ábra*

árddal kapcsolatban mind a mai napig az alapvető munka maradt. Az egyszerű jelenségek, amelyeket kevés matematika felhasználásával is tanulmányozni lehet, a golyó csúszó mozgása és gördülése a billiárdasztal érdes felületén (29. ábra). Az érintkezési pont egy parabolát ír le, ha a golyó csúszik, míg ha gördül, akkor egy egyenest. Tehát az elemzés ahhoz a meglepő eredményhez vezet, hogy a gördülés végső iránya kiolvasható a kezdeti feltételekből, nevezetesen nem más, mint az az irány, ami a golyó és az asztal érintkezési pontjától indulva halad az asztalnak ahhoz a pontjához, amelyre a dákó akkor mutat, amikor megüti a golyót. Azt a kérdést is könnyű megválaszolni, hogy a középponthoz viszonyítva milyen magasságban kell a dákónak meglöknie a golyót ahhoz, hogy guruljon és ne csússzon ($h=7R/5$, 1. a 29. ábrát). Gyakorlott játékosok képesek úgy megütni a golyót, hogy az visszafelé haladjon vagy valahol megálljon vagy egy másik, eredetileg álló golyóval ütközzön és úgy menjen tovább.

*Szuperlabda* trükkök azóta népszerűek, mióta ragadós felületű, igen rugalmas labdák olcsón kaphatók a játékboltokban. Ezek a labdák képesek rugalmas, rotációs ütközésre, mert felületük nem csúszik (30. ábra). Egy nemzedékkel korábban az ilyen tulajdonságú labdák csak elméleti feltételezésekben léteztek. Könnyen táncoltathatjuk őket a padlón, de a legnagyobb hatást keltő gyakorlat, amit meg lehet velük csinálni a „lövés az asztal alatt" (*"shot under the table"*), amelynek során a mozgás csaknem teljesen megfordul (31. ábra). Nagyon gyors labdák mozgása (amelyek esetében elhanyagolhatjuk a helyzeti energiát a nagy mozgási energia miatt) előre meghatározható az ütközési pontra vonatkoztatott impulzusnyomaték megmaradása és megfelelő, az ütközést leíró kinematikai hipoztézisek alapján [9, 14].

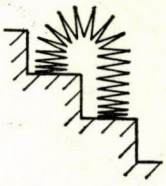



*30. ábra*
Rotációs ütközés

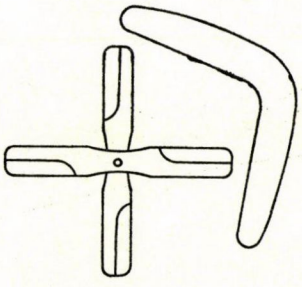


*31. ábra*

*A sétáló rugó* (32. ábra) egy egyszerű mechanikai játék, amelyet ismereteim szerint *Amerikában* találtak fel az 1940-es években. Nem más, mint egy súlyos, lágy rugó, amely képes arra, hogy (ritka) hullámként lemenjen a lépcsőn. Több, mint 25 évvel ezelőtt jelent meg mozgásának egy nem teljes elméleti vizsgálata [15]. Ez a mű a rugó hullámmozgását egy egyenes mentén vizsgálja és megbecsüli a sebességet. De nem tud felvilágosítással szolgálni arról, hogy mit fog csinálni a rugó azután, hogy egy fokot megtett. Valóban, a rugó sohasem indul állásból anélkül, hogy az egyik vége ne fordulna meg. Ez lesz ugyanis a rugó feje a következő lépcsőn és így halad tovább.



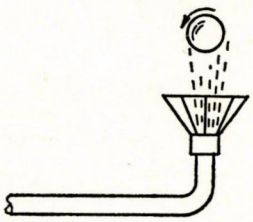
*32. ábra*



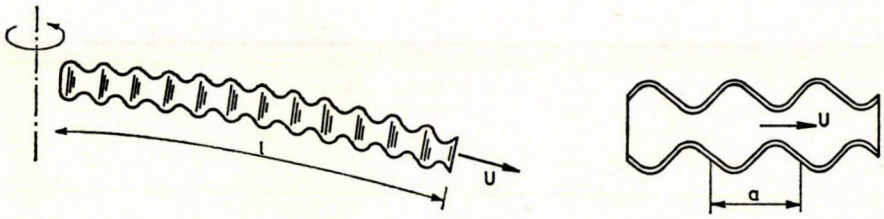
*33. ábra*
Visszatérő bumerángok

Az *aerodinamikai játékok,* mint a bumeráng vagy az egyszerűbb hajiga *(frisbee)*, magyarázata igen bonyolult (33. ábra). A légellenállás, az emelő erők és a stabilitást biztosító giroszkópikus erők rejtélyes kölcsönhatása folytán a bumeráng bonyolult, háromdimenziós úton tér vissza [16—19]. Meg kell jegyezni, hogy a széles körben elterjedt véleménnyel ellentétben a visszatérő bumerángot *Ausztrália* őslakói is csak játéknak használták, míg a nem visszatérőket, amelyeket csak kissé hajlítottak meg, vadászatra vagy messzehordó fegyverként. Igen egyszerű elemzést lehet adni arra az akadémikus, speciális esetre, amikor a bumeráng négy, szimmetrikusan elhelyezett lapátból áll, amikor a súly az egyéb fellépő erőkhöz képest elhanyagolható [9]. Ha megfelelően indítják, akkor körpályát tesz meg úgy, hogy a rotáció tengelye merőleges a lapátok síkjára és mindig a pálya középpontja felé mutat. A bumeráng visszatérése ebben az esetben könnyen származtatható az alsó és felső lapátokon működő különböző emelő erők által létrehozott forgatónyomaték hatására keletkezett precessziós mozgásból. A bumeráng általános mozgásának átfogó számítási módszerét adta egy monumentális, holland doktori értekezés [17], amelyben a szerző a súlypont mozgására átlagos egyenleteket vezet le (átlagos a forgás egy periódusa alatt). A pálya ezt követő számítását numerikusan kell végezni, ezért ez a munka csak csekély hozzájárulás a bumeráng dinamikájának megértéséhez.

*A lebegő labda* (34. ábra) gyakran szereplő feladat a hidrodinamikával foglalkozó előadásokon. Míg könnyű látni az impulzusmegmaradásból és a *nemviszkózus folyadékok áramlásának Bernoulli-féle törvényéből*, hogy az áramlás impulzusában a sugár elhajlása által okozott változás olyan erőt hoz létre, ami támasztja a labdát, addig nehéz megérteni a labda egyensúlyi helyzetének stabilitását. A kísérlet mindazonáltal sikeres, ha egy pingponglabdát a szánkkal vagy egy porszívó segítségével fújunk.



*34. ábra*




*35. ábra*

Hadd fejezzem be a dolgozatot egy olyan játékkal, melyet mintegy 10 évvel ezelőtt árultak az utcákon *Németországban* és még mos is használják, mint hangszert. Ez a „szuperbőgő" vagy barázdás kürt (35. ábra), amely kb. 1 m hosszú olcsó műanyagból készült hullámos falú cső. Átlagos átmérője 3 cm, a falon a fodrok hullámhossza 0,6 cm. Körbe forgatva hangok diszkrét spektrumát lehet vele előállítani. Valójában a légoszlop longitudinális rezgésének felhangjait kapjuk meg, mint a furulyában vagy az orgonában. A szuperbőgő hangot bocsájt ki, ha egy levegőfolyam keresztül halad rajta. Ha körbe forgatjuk, akkor ezt a légfolyamot a centrifugális szivattyú hatás kelti fel (35. ábra). Növekvő szögsebesség esetén a csőbeli légfolyam sebessége és a kibocsájtott hang magassága növekszik. Kevés kísérlet után is már látható, hogy valamelyik felhang megfelel az $U/a$ frekvenciának, amelyet az $U$ átlagos áramlási sebesség és a fal fodrainak hullámhossza határoz meg, l. pl. [20]. Ezt a „rezonancia hipotézist" egy újabbkeletű elméleti munka [21] igazolta, amely a kibocsájtott hangot annak az örvényeknek tulajdonítja, amelyek rezonanciafrekvenciával oszcillálva a hullámos fal mentén áramlanak vagy éppen a fodrok csapdájába esnek.

## IRODALOM

[1] JEARL WALKER: The Flying Circus of Physics, Wiley, New York, 1975.
[2] MARTIN GARDNER: Mathematischer Karneval, Ullstein, Frankfurt/M., 1975.
[3] HELMUT VOLZ: Einführung in die Theoretische Mechanik, Band 1, Akad. Verlagsges., Frankfurt/M., 1972.
[4] KURT MAGNUS: Schwingungen, B. G. Teubner, Stuttgart, 1961.
[5] STEVEN M. CURRY: How Children Swing, Amer. J. Phys. 44 (1976), 924—926.
[6] ROBERT W. LEONARD: An Interesting Demonstration of Two Linear Harmonic Vibrations to Produce a Singlę Elliptical Vibration. Amer. Phys. Teacher 5 (1937), 175—176.
[7] KURT MAGNUS: Kreisel — Theorie und Anwendungen, Springer, Berlin, 1971.
[8] JENS WITTENBURG: Dynamics of Systems of Rigid Bodies, B. G. Teubner, Stuttgart, 1977.
[9] V. BARGER & M. OLSSON: Classical Mechanics, a Modern Perspective, McGraw-Hill, New York, 1973.

[10] PIERRE CONTENSOU: Couplage entre frottement de glissement et frottement de pivotement dans la théorie de la toupie, IUTAM-Symposium Kreiselprobleme — Gyrodynamics, Celerina, 1962, Hsg. H. Ziegler, Springer, Berlin, 1963, 201—216.
[11] ANGELO R. DEL CAMPO: Tippe Top (Topsy Turnee Top) Cont'd, Amer. J. Phys 23 (1955), 544—545.
[12] HAROLD CRABTREE: Spinning Tops and Gyroscopic Motion, 3rd ed., Chelsea Publ. Co., New York, 1967.
[13] GUSTAVE GASPARD CORIOLIS: Théorie mathématique des effets du jeu de billard, Carilian-Goeury, Paris, 1835.
[14] H. H. MÜLLER & K. MAGNUS: Übungen zur Technischen Mechanik, Teubner Studienbücher, Band 23, B. G. Teubner Stuttgart, 1974.
[15] M. S. LONGUET-HIGGINS: On Slinky: The Dynamics of a Loose, Heavy Spring, Proc. Cambr. Phil. Soc. 50 (1954), 347—351.
[16] FELIX HESS: The Aerodynamics of Boomerangs, Sci. Amer., Nov. 1968, 124—136.
[17] FELIX HESS: Boomerangs, Aerodynamics and Motion, Proefschrift, Rijksuniversiteit te Groningen, 1975, 555pp+1 pair of polarisation glasses.
[18] JEARL WALKER: Boomerangs! How to make them and also how to fly them, Sci. Amer., March, 1979, 130—135.
[19] JEARL WALKER: More on boomerangs, including their connection with the dimpled golf ball, Sci. Amer., April 1979, 134—139.
[20] FRANK S. CRAWFORD: Singing Corrugated Pipes, Amer. J. Phys. 42 (1974), 278—288.
[21] PAUL H. TAYLOR: The Singing of Corrugated Pipes, Personal Communication by J. E. Ffowcs Williams (University of Cambridge), 1980.
[22] PAUL HILDEBRANDT: Das Spielzeug im Leben des Kindes, Berlin, 1904.
[23] FREDERIC V. GRUNFELD & EUGEN OKER: Spiele der Welt, Krüger-Verlag, Frankfurt/M., 1976.
[24] N. M. HUGENHOLTZ: On Tops Rising by Friction, Physica 18 (1952), 515—527.
[25] R. MERTENS & L. DE CORTE: An Exact Mathematical Solution of the Problem of Top Rising by Friction, ZAMM 58 (1978), T 116—T 118.

FORDÍTOTTA:

VÍZVÁRI BÉLA
MTA SZÁMÍTÁSTECHNIKAI ÉS AUTOMATIZÁLÁSI KUTATÓ INTÉZET
1132 BUDAPEST, VICTOR H. U. 18.

A kiadásért felelős az Akadémiai Kiadó és Nyomda főigazgatója
Műszaki szerkesztő: Sándor István
A kézirat nyomdába érkezett: 1984. augusztus 10. — Terjedelem: 22,05 (A/5 ív)
84-3301 — Szegedi Nyomda — F. v.: Dobó József igazgató

# ÚTMUTATÁS A SZERZŐKNEK

Az Alkalmazott Matematikai Lapok csak magyar nyelvű dolgozatokat közöl. A kéziratok gépelését olyan formában kérjük, hogy minden gépelt oldal 25, egyenként átlag 50 betűhelyes sort tartalmazzon. A közlésre szánt dolgozatokat három példányban kell beküldeni.

A kéziratok szerkezeti felépítésének a következő követelményeket kell kielégíteni. A fejlécnek tartalmaznia kell a dolgozat címét, a szerző teljes nevét, valamint annak a városnak a nevét, ahol a szerző dolgozik. A fejléc után egy, képletet nem tartalmazó, legfeljebb 200 szóból álló kivonatot kell minden esetben megadni. A dolgozatot címmel ellátott szakaszokra kell bontani, és az egyes szakaszokat arab sorszámmal kell ellátni. Az esetleges bevezetésnek mindig az első szakaszt kell alkotnia. Az irodalomjegyzék mindig az utolsó szakasz kell hogy legyen, és azt nem kell sorszámmal ellátni. Az irodalomjegyzék után, a kézirat befejezéseképpen fel kell tüntetni a szerző teljes nevét és a munkahelye (illetve lakása) pontos postai címét. A dolgozatban előforduló képleteket szakaszonként újrakezdődően, a képlet előtt két zárójel közé írt kettős számozással kell azonosítani. Természetesen nem szükséges minden képletet számozással ellátni. Az esetleges definíciókat és tételeket (segédtételeket és lemmákat) ugyancsak szakaszonként újrakezdődő, kettős számozással kell ellátni. Kérjük a szerzőket, hogy ezeket, valamint a tételek bizonyítását a szövegben kellő módon emeljék ki. Minden dolgozathoz csatolni kell egy angol, német, francia vagy orosz nyelvű, külön oldalra gépelt összefoglalót. Amennyiben lehetséges, kérjük a nyomtatás számára különösen nehézkes matematikai jelölések használatának az elkerülését.

A dolgozat ábráit és az esetleges lábjegyzeteket a dolgozat végén, különálló lapokon kérjük beküldeni. Mind az ábrákat, mind a lábjegyzeteket a dolgozat szakaszokra bontásától független, folytatólagos arab sorszámozással kell ellátni. Az ábrák elhelyezését a dolgozat megfelelő helyén, széljegyzetként feltüntetett, ábraazonosító sorszámokkal kell megadni. A lábjegyzetekre a dolgozaton belül az azonosító sorszám felső indexkénti használatával lehet hivatkozni.

Az irodalmi hivatkozások formája a következő. Minden hivatkozást fel kell sorolni a dolgoza végén található irodalomjegyzékben, a szerzők, illetve társszerzők esetén az első szerző neve szerint alfabetikus sorrendben úgy, hogy külön, de folytatólagos sorszámozású listát alkossanak a latin és a cirill betűs nevű szerzők műveire vonatkozó hivatkozások, és mindkét részben a megfelelő alfabetikus sorrend legyen kialakítva. A folyóiratban megjelent cikkekre [1], a könyvekre [5], a kötetben megjelent dolgozatokra [4], a disszertációkra [3] és a gépi program leírásokra [2] a következő minta szerint kell hivatkozni:


[1] Farkas, J., »Über die Theorie der einfachen Ungleichungen«, *Journal für die reine und angewandte Mathematik* **124** (1902) 1—27.

[2] Kéri, G., „DUALSIMP", rutin a CDC 3300-as gépekre (Magyar Tudományos Akadémia Számítástechnikai és Automatizálási Kutató Intézete, CDC 3300 felhasználói ismertetők 2. 1973. május) 19—20.

[3] Prékopa, A., „Sztohasztikus rendszerek optimalizálási problémáiról", doktori értekezés. Magyar Tudományos Akadémia, Budapest, 1970.

[4] Prabhu, N. U., "Recent research on the ruin problem of collective risk theory", in: *Inventory Control and Water Storage* Ed. A. Prékopa (János Bolyai Mathematical Society and North-Holland Publishing Company, Amsterdam—London, 1973) 221—228.

[5] Zoutendijk, G., *Methods of Feasible Directions* (Elsevier Publishing Company, Amsterdam and New York, 1960).


A dolgozatok szövegében az irodalmi hivatkozás számait szögletes zárójelben kell megadni mint például [5] vagy [4, 76—78]. A szerzők a dolgozatukról 100 darab különlenyomatot kapnak' ezek költsége — nyomott oldalanként 25 forint — a szerzői díjat terheli.

Ára: 50 Ft

ISSN 0133—3399

# TARTALOMJEGYZÉK



# INDEX

# ALKALMAZOTT MATEMATIKAI LAPOK

A MAGYAR TUDOMÁNYOS AKADÉMIA
MATEMATIKAI ÉS FIZIKAI
TUDOMÁNYOK OSZTÁLYÁNAK KÖZLEMÉNYEI



X. KÖTET

AKADÉMIAI KIADÓ, BUDAPEST
1984

# TARTALOMJEGYZÉK

# INDEX




84-3301 — Szegedi Nyomda — F. v.: Dobó József igazgató