Alkalmazott matematikai lapok

1985/1-2

A MAGYAR TUDOMÁNYOS AKADÉMIA
MATEMATIKAI ÉS FIZIKAI TUDOMÁNYOK
OSZTÁLYÁNAK KÖZLEMÉNYEI

11.

KÖTET

AKADÉMIAI KIADÓ, BUDAPEST

ALKALMAZOTT MATEMATIKAI LAPOK

A MAGYAR TUDOMÁNYOS AKADÉMIA MATEMATIKAI ÉS FIZIKAI TUDOMÁNYOK OSZTÁLYÁNAK KÖZLEMÉNYEI

XI. 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 **11** (1985) 1—90

NAGYMÉRETŰ, RITKA, SZIMMETRIKUS MÁTRIXOK HATÉKONY SZÁMÍTÓGÉPES KEZELÉSE

ARANY ILONA

Budapest

Jelen dolgozatban a sávszélesség/profil-redukció problémáját elemezzük. Alkalmas terminológia bevezetése után értékeljük a *Gibbs—Poole—Stockmeyer-*, illetve a *George—Liu-eljárásokat*. Részletesen foglalkozunk a sávszélesség/profil-redukció algoritmusában a szintstruktúra kialakítási és számozási fázisok problémáival. Ennek eredményeképpen 17 algoritmust fogalmazunk meg ritka, szimmetrikus mátrixok sávszélesség/profil-redukciójára. Ezek közül 6 eljárás nem csupán eredményében, hanem gépidő szükségletében is jónak minősül a fenti jól ismert algoritmusokkal való összehasonlításban.

Bevezetés

Számos alkalmazási területen, különösen a műszaki gyakorlatban felvetődő problémák (parciális differenciálegyenletek numerikus megoldása, statikai számítások, stb.) gyakran eredményeznek nagyméretű, ritka, szimmetrikus mátrixot. A számítógépes hatékonyság növelése érdekében ekkor speciális mátrix-kezelési technikák alkalmazása szükséges.

Tekintsük például az

$$\mathbf{A}\mathbf{x} = \mathbf{b} \tag{1}$$

lineáris algebrai egyenletrendszert, ahol $\mathbf{A}$ nagyméretű, szimmetrikus, sok zérus-elemet tartalmazó pozitív definit mátrix. Ekkor (1) direkt módszerrel való számítógépes megoldásakor a faktorizáció során fellépő feltöltődés *(fill-in)* folytán a mátrix nem-zérus elemeinek száma erősen megnőhet, ezáltal megnő a tárigény és a végrehajtandó műveletek száma. Vagyis a megoldás gazdaságtalanná válik.

A számítógépes hatékonyság növelése, illetve igen nagy (több 10 000-es) méret esetén a központi tárban való kezelhetőség érdekében célszerű volna olyan $\mathbf{P}^0$ permutációs mátrixot meghatározni, hogy

$$\mathbf{A}^{\mathbf{P}^0} = \mathbf{P}^0\mathbf{A}\mathbf{P}^{0T}$$

zérus/nem-zérus szerkezete a választott megoldási módszer szempontjából optimális, azaz

$\mathbf{A}^{\mathbf{P}^0}$ olyan szerkezetű, hogy az elimináció során fellépő feltöltődés minimális; vagy

$\mathbf{A}^{\mathbf{P}^0}$ sáv-mátrix, minimális sávszélességgel.

Mindkét esetben azonban $\mathbf{P}^0$ meghatározása *NP-teljes probléma* (P. Z. CHINN és munkatársai [24]; M. R. GAREY és munkatársai [34]; CH. H. PAPADIMITRIU [73], [74]; J. A. GEORGE [50], [51]), ezért a gyakorlatban meg kell elégednünk $\mathbf{P}^0$ valami-

lyen fajta közelítésével, mely csupán a feltöltődés, illetve sávszélesség redukcióját eredményezi. E közelítéseket különböző heurisztikus algoritmusokkal állítják elő a gyakorlatban.

A dolgozatban ritka, szimmetrikus mátrixok sávszélesség/profil redukciójával foglalkozunk és 17 algoritmust közlünk. Ezekből 6 eljárás nem csupán eredményében, hanem gépidő-szükségletében is kedvező képet mutat a J. A. GEORGE és J. W-H.LIU [53], valamint N. E. GIBBS, W. G. POOLE és P. K. STOCKMEYER [55] ismert algoritmusaival való összehasonlításban.

A ritka mátrixok szerkezeti sajátságainak felhasználása iránti igény elsőként a nagyméretű lineáris programozási feladatok kapcsán fogalmazódott meg (G. B. DANTZIG, 1952, [29]). A műszaki alkalmazások, különösen a véges elemek módszerének gyors elterjedése sürgető kényszerként hatott a számítógépes megoldás hatékonyságát növelő eljárások kidolgozására. Számos munka jelent meg az elimináció során fellépő feltöltődés, az elimináció gráfelméleti szimulációja problémakörében. (H. M. MARKOWITZ [68]; D. J. ROSE [76], J. R. BUNCH [20], [21]; J. A. GEORGE [39], [47], [53]; R. P. TEWARSON [83]; S. C. EISENSTAT [31].)

Egyrészt adott feladat-típus jellegzetességein alapuló eljárásokat hoztak létre (B. M. IRONS [57]; C. A. FELIPPA [32]). Másrészt feladat-típustól független eljárások jelentek meg, melyekkel a feltöltődés jelentősen csökkenthető a mátrix elemeinek szintjén (H. M. MARKOWITZ [68]; D. J. ROSE [76]), illetve teljesen megszüntethető a mátrix bizonyos blokkjaiban (J. A. GEORGE és J. W-H. LIU [36], [37], [38], [40], [43], [44], [45], [46], [52], [53]).

Megindult a megfelelő software kidolgozása. Új programnyelvek születtek (GRAAL [75], LASCALA [84]) s program-rendszerek készültek ritka mátrixú rendszerek kezelésére (*Yale Sparse Matrix Package (Yale University)*; ill. SIRSM [85]). 1978-ra befejeződött az első, univerzálisnak nevezhető program-rendszer, a SPARSPAK [53] kidolgozása (J. A. GEORGE és J. W-H. LIU, *University of Waterloo*). Így lehetővé vált a szimmetrikus vagy csupán zérus/nem-zérus szerkezetében szimmetrikus, ritka mátrixú lineáris egyenletrendszerek hatékony direkt megoldásainak alkalmazása.

Az ESZR-ben 1975-ben elkészült a ritka, szimmetrikus mátrixok sávszélesség-redukcióját végző MÁTRIXOK program-rendszer [6].

A sávszélesség-redukció problémájára elsőként 1965-ben G. G. ALWAY és D. W. MARTIN [3], majd 1968-ban R. ROSEN [77] fogalmazták meg eljárásaikat, melyek azonban nagy munkaigényük miatt a gyakorlatban nem nyertek alkalmazást. Rövid időn belül számos új eljárás fogalmazódott meg [4], [7], [23], [27], [53], [55], [59], [81], [82], melyek közül [27], [53], [55] a gyakorlatban is hatékonynak bizonyultak.

Az alábbiakban az első, gyakorlatban is alkalmazható eljárás megszületésétől napjainkig közölt főbb eredményeket összegezzük.

1969-ben E. H. CUTHILL és J. MCKEE [27] fogalmazták meg az első, ma már klasszikusnak nevezhető eljárásukat. A szerzők irányítatlan gráf kifeszítő fájának alkalmas számozásával érték el a redukált sávszélességet.

1972-ben W. F. SMYTH ILO *(International Labour Office)* szakértővel és Szóda Lajos kollegámmal közösen új sávszélesség-redukciós eljárást közöltünk [4]. Bevezettük a gráfon értelmezett szintstruktúra fogalmát, s algoritmusunkat szintstruktúra kialakítása és szintstruktúra számozási fázisok együtteseként fogalmaztuk

meg. E munkánkra a nemzetközi szakirodalomban 16 dolgozatban és 3 könyvben találtunk hivatkozást.

1976-ban N. E. GIBBS, W. G. POOLE és P. K. STOCKMEYER közölték új algoritmusukat [55], melyet szintén szintstruktúra kialakítás és szintstruktúra számozás együtteseként írtak le. A maximális excentricitás közelítésére itt fogalmazták meg „pszeudo-átmérő"-t meghatározó algoritmusukat, de a pszeudo-átmérő fogalmát nem definiálták.

1976-ban J. W-H. LIU és A. H. SHERMAN bebizonyították [66], hogy a *Cuthill—McKee számozás* megfordításával a profil-érték nem nőhet, sok esetben viszont radikális csökkenést eredményez a számozás megfordítása.

1976-ban W. F. SMYTH ILO szakértővel közösen új számozás koncepcióját közöltük [81]. E munkánkra két irodalmi hivatkozást találtunk. 1978-ra elkészült a módszer, s annak számítógépes kidolgozása, mely profil-értékben kedvező csökkenést eredményezett.

1979-ben J. A. GEORGE és J. W-H. LIU három módosítási stratégiát közöltek [49] a *Gibbs—Poole—Stockmeyer pszeudo-átmérőt meghatározó eljárás* műveletigényének csökkentésére. Algoritmusaikat „pszeudo-perifériális pont"-ot előállító eljárásoknak nevezték, de a pszeudo-perifériális pontot ők sem definiálták.

1981-ben J. A. GEORGE és J. W-H. LIU tollából jelent meg az első monográfia [53], melyben a szerzők részletesen tárgyalják a ritka mátrixú lineáris egyenletrendszerek hatékony megoldására szolgáló eljárásaikat és elemzik a vonatkozó, SPARSPAK program-rendszer *(University of Waterloo)* egyes ágait. A SPARSPAK-beli sávszélesség/profil-redukciós eljárásuk szintstruktúra kialakítására a [49]-ben közölt stratégiájuk egyikét alkalmazták, s a szintstruktúra számozását a *fordított Cuthill—McKee számozással* végezték.

1982-ben J. K. PACHL *(University of Waterloo)* munkájában [72] jelenik meg a pszeudo-perifériális pontok definíciója.

1982-ben P. Z. CHINN, J. CHVÁTÁLOVA, A. K. DEWDNEY és N. E. GIBBS [24] megállapítása szerint a szakirodalomban a sávszélesség/profil-redukció tárgyalása nélkülözi a tétel/bizonyítás-jellegű leírást; kellő formalizmus híján az egyes eljárások informálisan nyertek megfogalmazást.

A dolgozatban a sávszélesség/profil-redukció problémájával foglalkozunk. Egzakt fogalmak bevezetésével felépítjük azt a terminológiát, mely lehetővé teszi, hogy e tisztán heurisztikusként kezelt problémát pontosabban, bizonyos részeit precíz matematikai megfogalmazásban tárgyaljuk. A kialakított terminológiában, mely esetileg eltérhet a gráfelméletben ismerttől, számos gráfelméleti összefüggést fogalmazunk meg. Esetenként, ismert tételekre olyan konstruktív bizonyításokat közlünk, melyek a sávszélesség/profil-redukció algoritmusában hatékony rész-eljárásokként alkalmazhatók. Ezáltal egyrészt lehetővé válik GIBBS—POOLE—STOCKMEYER és GEORGE—LIU heurisztikus algoritmusaiban a szintstruktúra kialakító fázis értékelése. Másrészt, mód nyílik további hatékony algoritmusok megfogalmazására.

Eredményeinket az alábbiak szerint tárgyaljuk.

Az 1. fejezetben bevezetjük a gráfon értelmezett általános és gyökérrel rendelkező szintstruktúra fogalmát és a szintstruktúrával kompatibilis számozást, s megfogalmazzuk a sávszélesség/profil-redukció feladatát. Rámutatunk, hogy a sávszélesség csökkentésének egyik szükséges feltétele egy közel-minimális szélességű szintstruktúra meghatározása.

A 2. fejezetben megmutatjuk, hogyan írható fel a gráfon értelmezett távolság

függvény adott gyökérrel rendelkező szintstruktúrában [14], illetve szintstruktúra rendszerben. Definiáljuk a gráf kvázi-perifériális, pszeudo-perifériális és szemi-pszeudo-perifériális pontjait. Megmutatjuk, hogy a pszeudo-perifériális pontok PACHL szerinti definíciója, mely azonos jelen tárgyalás szemi-pszeudo-perifériális pontjának definíciójával, excentricitás szempontjából gyengébb fogalom, mint jelen értelmezés szerinti pszeudo-perifériális pont. Ezért a továbbiakban a már bevezetett terminológiát követjük.

Rámutatunk a pszeudo-perifériális pontok néhány fontos tulajdonságára.

A 3. fejezetben bebizonyítjuk, hogy GIBBS—POOLE—STOCKMEYER „pszeudo-átmérő"-t meghatározó eljárása [55] pszeudo-perifériális végpontot eredményez [15].

Megmutatjuk, hogy a GEORGE—LIU által kidolgozott módosítások [49] mindegyike szemi-pszeudo-perifériális pontot eredményez. Következésképpen, a SPARSPAK megfelelő rutinja maximális excentricitás közelítéseként minimális excentricitású pontot is eredményezhet.

Hatékony módszert közlünk [16] pszeudo-perifériális pontok meghatározására.

A 4. fejezetben a perifériális pontok gráfbeli elhelyezkedését vizsgálva, fontos tételt bizonyítunk, melynek következményeképpen tetszőleges pontból kiindulva megkonstruálható a pontok olyan részhalmaza, amely tartalmaz perifériális pontot. Ennek alapján módszert (P) közlünk perifériális pontok meghatározására [18]. A műveletigény csökkentésére különböző módosított verziókat fogalmazunk meg. E verziók egyike heurisztikus eljárás ugyan [17], de alkalmazásakor a vizsgált esetek döntő többségében perifériális pontot eredményezett, műveletigénye kedvezően alacsony.

A szintstruktúra rendszerben a súlyok és excentricitások kapcsolatát vizsgálva, heurisztikus eljárást (P2) közlünk, mely a vizsgált esetek jelentős részében perifériális pontot eredményezett.

Az 5. fejezetben a perifériális pontot előállító eljárásaink felhasználásával három algoritmust dolgozunk ki, melyekkel kedvezően kis szélességű, gyökérrel rendelkező szintstruktúra állítható elő.

Bevezetjük a gráf minimális tagozódási halmazának fogalmát, s egy speciális tipusának egzisztenciájára olyan konstruktív bizonyítást közlünk, mely eljárást ad (CS) ilyen speciális minimális tagozódási halmazok előállítására.

Bemutatjuk az LS eljárást, mely tetszőleges tagozódási halmazból speciális szerkezetű általános szintstruktúrát generál. Mindezek alapján hét eljárást fogalmazunk meg a közel-minimális szélességű szintstruktúra előállítására, melyekből három gyökérrel rendelkező, négy speciális szerkezetű általános szintstruktúrát eredményez.

Eljárásaink — összehasonlítva GIBBS—POOLE—STOCKMEYER [55] és GEORGE—LIU [53] vonatkozó algoritmusaival — kedvező eredményeket adnak és műveletigényük is elfogadhatóan alacsony.

A 6. fejezetben ismertetjük [4]-ben alkalmazott számozási eljárásunkat (N1), mely speciális szerkezetű általános szintstruktúrák számozását végzi.

Általános szerkezetű általános szintstruktúra számozási problémáját elemezve, elméleti indoklást adunk a [81]-ben közölt számozási koncepciónk helyességére. Bemutatjuk a [9]-beli számozási eljárásunkat (NN) is, mely mind gyökérrel rendelkező, mind általános szintstruktúrák számozására alkalmas, azonban igen nagy gépidővonzata miatt a gyakorlatban nem használható.

A 7. fejezetben új sávszélesség/profil-redukciós eljárásokat fogalmazunk meg.

A bemutatott 7 szintstruktúra kialakító és két számozó eljárásunkat kiegészítjük a GEORGE—LIU sávszélesség/profil-redukciós algoritmusának szintstruktúra kialakító és számozó eljárásával. A fenti algoritmusok ésszerű kombinációjaként 17 sávszélesség/profil-redukciós eljárást fogalmazunk meg. Ezek közül 6 nem csupán eredményében, hanem gépidő szükségletében is jónak bizonyul a GEORGE—LIU [53] és GIBBS—POOLE—STOCKMEYER [55] módszereivel való összehasonlításban.

A 8. fejezetben röviden ismertetjük program-rendszerünket, mely pozitív definit, szimmetrikus, ritka mátrixú lineáris egyenletrendszert old meg profil-szimmetrikus faktorizációval. A leghatékonyabbnak ítélt 6 sávszélesség/profil-redukciós eljárásunk mindegyike aktivizálható, s az egyenletrendszer megoldását SPARSPAK-beli rutinok végzik.

A 9. fejezetben felvázoljuk eredményeink hasznosítási lehetőségeit.

1. A sávszélesség redukciójának alapfeladata

A mátrixok ritkasági tulajdonságának jellemzésére egységes meghatározás nem ismeretes.

Számos alkalmazási területen, bizonyos feladat-osztályokban tipikusan olyan szimmetrikus mátrixok állnak elő, melyek igen sok zérus-elemet tartalmaznak.

A véges elemes és véges differenciás közelítésekből származó mátrixok specialitását jellemezve R. P. TEWARSON [83] a mátrixot ritkának nevezi, ha rendjétől függetlenül az egysorban levő nem-zérus elemeinek száma nem haladja meg a 10-et.

F. L. ALVARADO [2] egy mátrix-családot ritkának nevez, ha az egyes mátrixok méretének növelésével a nem-zérus elemek száma nem négyzetesen növekszik.

Gyakran a 10—20%-os telítettségű mátrixot nevezik ritkának.

Megjegyezzük azonban, hogy pl. 30—40%-os telítettség esetén is feltétlenül célszerű elkerülni a zérus-elemek tárolását és velük való művelet-végzést, azaz ritka mátrixként kezelni.

1.1. Jelölések, elemi ismert összefüggések

Legyen $\mathbf{A}$ $N \times N$-es szimmetrikus mátrix.

Az $\mathbf{A}$ mátrix *sávszélességét* $b(\mathbf{A}) = \max_{a_{ij} \neq 0} |i-j|$ szerint értelmezzük, s a mátrix *sávján* a

$$\text{Band}\,(\mathbf{A}) = \{\{i, j\} | 0 < i-j \leqq b(\mathbf{A})\}$$

halmazt értjük.

Vezessük be az $f_i(\mathbf{A}) = \min\{j | a_{ij} \neq 0\}$ jelölést. A mátrix *profilját*

$$\text{Pr}(\mathbf{A}) = \{\{i, j\} | f_i(\mathbf{A}) \leqq j < i\}$$

szerint értelmezzük.

Nyilvánvalóan, $\text{Pr}(\mathbf{A}) \subseteq \text{Band}\,(\mathbf{A})$ teljesül, ahonnan $|\text{Pr}(\mathbf{A})| \leqq |\text{Band}(\mathbf{A})|$ következik, ahol $|\cdot|$ a halmaz elemeinek a számát jelöli.

A ritka mátrixok tárolására az alábbi két mód a leginkább használatos:

— sáv-szimmetrikus tárolás (R. S. MARTIN és J. H. WILKINSON [69]);

— profil tárolási séma (A. JENNINGS [58]),

melyek rendre $\text{Band}\,(\mathbf{A}) \cup \{\text{diagonális elemek}\}$, illetve $\text{Pr}\,(\mathbf{A}) \cup \{\text{diagonális elemek}\}$

elemeket tárolják. Vagyis a szükséges tárigény csökkentésének egyik hatékony eszköze a sávszélesség, illetve a |Pr (A)| profil-érték redukálása.

A ritka mátrixok és belőlük származó irányítatlan gráfok ismert összefüggései (1. sz. melléklet) szerint tetszőleges szimmetrikus A mátrix *sávszélességének minimalizálásához* meg kell határoznunk a G_A gráf olyan n_0 számozását, melyre

$$b(G_A^{n_0}) = \min_n b(G_A^n).$$

Mivel n_0 meghatározása NP-teljes probléma, ezért célunk a továbbiakban a sávszélesség, illetve profil-érték redukciója, azaz G_A olyan n számozásának meghatározása, melyre $b(G_A^n)$, illetve $\Pr(G_A^n)$ kedvezően kicsiny érték.

1.2. A szintstruktúra és tulajdonságai

Legyen $G=(X, E)$ irányítatlan, összefüggő, hurok és többszörös él nélküli nem teljes gráf: X a pontok, E az élek halmaza. (Ezt a továbbiakban feltételezzük.)

A $G=(X, E)$ gráfban legyen $Y \subset X$ tetszőleges, valódi részhalmaz. A gráf Y által definiált *metszetgráfját* $G(Y)=(Y, E(Y))$ szerint értelmezzük, ahol Y a pontok halmaza

$$E(Y) = \{(y_i, y_j) | y_i, y_j \in Y,\ (y_i, y_j) \in E\}$$

az élek halmaza.

A $G=(X, E)$ gráfban a $Z \subset X$ valódi részhalmazt szétválasztó vagy *tagozódási halmaznak* nevezik, ha $G(X \setminus Z)$ metszetgráf nem összefüggő. Az $x, y \in X \setminus Z$ pontokat Z *elválasztja*, ha x és y $G(X \setminus Z)$ különböző komponenseihez tartoznak.

Minden összefüggő, nem teljes gráfnak van tagozódási halmaza.

A $G=(X, E)$ gráfban tetszőleges $x,\ y \in X$ pontok *távolságán* a köztük létesíthető legrövidebb útban levő élek számát értjük és $d(x, y)$-ként jelöljük.

A gráfon értelmezett távolság kielégíti a háromszög-egyenlőtlenséget.

Tetszőleges $x \in X$ pont *szomszéd halmazát*

$$N(x) = \{y | y \in X;\ d(x, y) = 1\}$$

szerint értelmezzük. Hasonlóan értelmezhető tetszőleges $Y \subset X$ valódi részhalmaz szomszéd halmaza, mely

$$N(Y) = \{y | y \in X \setminus Y;\ d(Y, y) = 1\}$$

szerint írható fel, ahol $d(Y, y) = \min_{z \in Y} d(z, y)$.

Tetszőleges $x \in X$ pont *fokszámán* a benne összefutó élek számát értjük és deg (x)-ként jelöljük.

Tetszőleges $x \in X$ pont *excentricitását*

$$l(x) = \max_{y \in X} d(x, y) \tag{1.2.1}$$

szerint értelmezzük.

Megjegyezzük, hogy (1.2.1)-et C. BERGE [19] definiálta, mint az x pont „kapcsolódási számát" *(associated number)*. Jelen tárgyalásban J. A. GEORGE terminológiáját követjük, aki (1.2.1)-et excentricitásnak nevezte.

1.2.1. Definíció. (I. ARANY, W. F. SMYTH and L. SZÓDA [4]). Tekintsük a $G=(X, E)$ pontjainak

$$\{L_0, L_1, \ldots, L_k\} \tag{1.2.2}$$

partícióját tehát

$$L_i \subset X, \quad i = 0, 1, \ldots, k,$$

$$L_i \cap L_j = 0, \quad i, j = 0, 1, \ldots, k; \ (i \neq j),$$

$$\bigcup_{i=0}^{k} L_i = X.$$

Ha

$$N(L_0) \subseteq L_1,$$

$$N(L_k) \subseteq L_{k-1},$$

$$N(L_i) \subseteq L_{i-1} \cup L_{i+1}, \quad i = 1, 2, \ldots, k-1$$

teljesül, akkor (1.2.2)-t a gráf *általános szinstruktúrájának* nevezzük és GLS-ként* jelöljük.

L_0, L_k szélső
L_i $1 \leqq i \leqq k-1$ közbülső } szintek.

k a *szintstruktúra hossza.*

$|L_i|$-t az *i-edik szint szélességének* nevezzük, míg a *szintstruktúra szélességét*

$$W(\mathrm{GLS}) = \max_{0 \leqq i \leqq k} |L_i|$$

szerint definiáljuk.

1.2.1. TÉTEL. Bármely összefüggő, nem teljes $G=(X, E)$ gráfnak van általános szintstruktúrája.

Bizonyítás. Mivel a gráf nem teljes, így létezik $S \subset X$ tagozódási halmaza. Könnyű belátni, hogy

$$\{L_0 = X_1, \quad L_1 = S, \quad L_2 = X_2\}$$

általános szintstruktúra. ∎

GLS tulajdonságai

— minden közbülső szint a gráf tagozódási halmaza,

— tetszőleges $x \in L_i \ (0 \leqq i \leqq k)$ pont esetén minden $y \in N(x)$ pont olyan L_j szinthez tartozik, hogy $|i-j| \leqq 1$ teljesül. Másszóval, GLS azt biztosítja, hogy bármely $(x, y) \in E$ élt kifeszítő x, $y \in X$ csúcsok azonos vagy szomszédos szinthez tartoznak.

1.2.2. Definíció [4]. Tetszőleges $G=(X, E)$ gráfban legyen $Y \subset X$ tetszőleges részhalmaz. A gráf pontjainak

$$\{L_0(Y), L_1(Y), \ldots, L_k(Y)\}$$

* GLS jelölés az angol *"general level structure"* rövidítéséből származik.

partícióját, ahol

$$L_0(Y) \equiv Y,\ L_1(Y) = N(Y)$$

$$L_i(Y) = N(L_{i-1}(Y)) \setminus L_{i-2}(Y) \quad (i = 2, 3, \ldots, k)$$

$$L_{k+1}(Y) = \emptyset$$

teljesülnek, *Y-gyökerű szintstruktúrának* nevezzük és RLS (Y)-ként* jelöljük.

k a szintstruktúra *hossza,* míg a *szintstruktúra szélességét*

$$W(\mathrm{RLS}(Y)) = \max_{0 \leqq i \leqq k} |L_i(Y)|$$

szerint értelmezzük.

Nyilvánvalóan értelmes a definíció $Y \equiv \{x\}$ $(x \in X)$ esetén is. Ekkor az egyszerűség kedvéért RLS (x) jelölést alkalmazzuk.

Könnyű belátni, hogy a $G=(X, E)$ gráfban tetszőleges $Y \subset X$ esetén létezik RLS (Y).

Tetszőleges $x \in X$ pontból generált RLS (x) *tulajdonságai*

— RLS (x) egyben GLS is,

— bármely $z \in L_i(x)$ pontra $d(z, x)=i$ $(i=0, 1, \ldots, k)$, vagyis $L_i(x)$ azonos az x pontból képezett kifeszítő fában az x-től i távolságban levő pontok halmazával,

— a szintstruktúra hossza egyenlő a gyökér excentricitásával, azaz $k=l(x)$. Az $L_{l(x)}(x)$ szintet az x pont *excentricitási szintjének* nevezzük és $L_{ec}(x)$-ként jelöljük.

1.3. Szintstruktúrával kompatibilis számozás

1.3.1. Definíció [5]. A $G=(X, E)$ gráfon legyen $\mathrm{GLS}=\{L_0, L_1, \ldots, L_k\}$ tetszőleges szintstruktúra. A gráf n számozását GLS-sel *kompatibilis számozásnak* nevezzük, ha

$x \in L_i$, $y \in L_j$ pontokra $i<j$ teljesüléséből $n(x)<n(y)$ következik.

1.3.1. TÉTEL. Legyen adva a $G=(X, E)$ gráf tetszőleges GLS-je. A gráf n számozása akkor és csak akkor kompatibilis GLS-sel, ha bármely $y \in L_i$ $(i>0)$ esetén

$$W_{i-1}+1 \leqq n(y) \leqq W_i \tag{1.3.1}$$

teljesül, ahol

$$W_i = \sum_{j=0}^{i} |L_j| \quad (i = 0, 1, \ldots, k).$$

Bizonyítás. Tegyük fel, hogy az n számozás kompatibilis GLS-sel. Ekkor 1.3.1. definícióból közvetlenül következik (1.3.1) teljesülése.

Fordítva, tegyük fel, hogy az n számozásra (1.3.1) teljesül. Ekkor bármely $z \in L_{i-1}$ esetén is

$$W_{i-2}+1 \leqq n(z) \leqq W_{i-1}$$

igaz, ami (1.3.1) értelmében $n(z)<n(y)$ teljesülését eredményezi minden $z \in L_{i-1}$ és $y \in L_i$ pontokra. Vagyis, az n számozás kompatibilis GLS-sel. ∎

* RLS (Y) jelölés az angol *"rooted level structure"* rövidítéséből származik.

KÖVETKEZMÉNYEK

1.3.1. Valamely szintstruktúrával kompatibilis számozás úgy építhető fel, hogy a számozást szintenként, azok indexeinek monoton növekvő sorrendjében végezzük úgy, hogy az egyes csúcsszámok hozzárendelése az $I=\{1, 2, \ldots, |X|\}$ halmaz elemeinek monoton növekvő sorrendjében történik.

1.3.2. Tetszőleges szintstruktúra bármely kompatibilis n számozásának hatására a számozott gráf összefüggési mátrixa blokk-tridiagonális szerkezetű lesz, rendre $|L_0|, |L_1|, \ldots, |L_k|$ méretű diagonális blokkokkal; (2. sz. melléklet).

Megjegyzések

1.3.1. A kompatibilis számozás fogalma korábban nem szerepelt a nemzetközi szakirodalomban. E fogalom azonban hasznosnak bizonyult. Ugyanis J. A. GEORGE hatékony rendezési algoritmusai (*quotient tree method* [53], *one way dissection* [46], [52], [53], *nested dissection* [37], [40], [44], [45], [53]), melyek a mátrixon eliminációszempontjából kedvező blokkolásokat eredményeznek, úgy fogalmazhatók meg, mint adott szintstruktúra szintjeinek bizonyos permutációival nyert halmazrendszerek kompatibilis számozásai.

1.4. A sávszélesség redukciójának egyik szükséges feltétele

Mint azt [5]-ben megmutattuk, az alábbi állítás igaz.

1.4.1. TÉTEL. Legyen GLS tetszőleges szintstruktúra. Ekkor minden GLS-sel kompatibilis n számozásra

$$W(\mathrm{GLS}) \leqq b(G^n) \leqq 2W(\mathrm{GLS}) - 1 \tag{1.4.1}$$

érvényes, ahol $b(G^n)$ a számozott gráf sávszélességét jelöli.

Megjegyezzük, hogy a kifeszítő fa számozásának eredményét elemezve, elsőként E. H. CUTHILL és J. MCKEE tesz említést (1.4.1)-típusú összefüggés felismeréséről.

KÖVETKEZMÉNYEK

1.4.1. Tetszőleges RLS(x) szintstruktúra szélessége meghatározza a kompatibilis számozásaival nyerhető sávszélességek pontos alsó és felső korlátját. Következésképpen, a minimális sávszélesség közelítésének egyik fontos szükséges feltétele olyan RLS(x_0) meghatározása, melyre

$$W(\mathrm{RLS}(x_0)) = \min_{x \in X} W(\mathrm{RLS}(x)) \tag{1.4.2}$$

teljesül.

Megjegyzések

1.4.1. (1.4.2)-t teljesítő RLS(x) meghatározása $|X|$ számú szintstruktúra generálását igényli, ami műveletigényes feladat. Ezért az egyes szerzők (N. E. GIBBS, W. G. POOLE és P. K. STOCKMEYER [55], J. A. GEORGE és J. W-H. LIU [53], W. F. SMYTH és I. ARANY [81]) olyan szintstruktúra hatékony előállítására törekedtek, melynek szélessége kedvezően kicsi érték, s ezt a leghosszabb illetve közel leghosszabb RLS(x)-ek körében vélték megtalálni. (A gyakorlati tapasztalatok mindeddig megerősítették e közelítési mód helyességét.)

1.5. A sávszélesség-redukció feladata

Fentiek alapján a sávszélesség-redukció algoritmusa a következőképpen fogalmazható meg:

1. lépés (Szintstruktúra kialakítása)
Kedvezően kis szélességű (W) szintstruktúra generálása.

2. lépés (Szintstruktúra számozása)
A szintstruktúrával kompatibilis olyan számozás felépítése, mely W-hez közeli sávszélességet eredményez.

Megjegyezzük, hogy az eljárások algoritmikus megfogalmazásában J. A. GEORGE és J. W-H. LIU [53]-ban alkalmazott jelölésrendszerét követjük.

2. A gráf néhány nevezetes pontjának vizsgálata

A továbbiakban feltételezzük a gráf összefüggőségét. A későbbiekben bemutatásra kerülő eljárások ugyanis úgy dolgoznak a gráfon, mint összefüggő komponenseinek unicióján ([53], 44. old., 65. old.).

Az $x \in X$ pontot *perifériális pontnak* nevezzük, ha

$$l(x) = \max_{p \in X} l(p);$$

az $l(x)$-érték a gráf *átmérője*, melyet diam (G)-ként jelölünk.

Az 1.4.1. megjegyzés értelmében minimálishoz közeli szélességű szintstruktúra meghatározásához tehát egy perifériális pont ismerete szükséges. Ennek megtalálása azonban ismét $|X|$ számú RLS (x) képzését igényli, amely túl munkaigényes feladat. Ismeretes, hogy W. F. SMYTH és W. M. L. BENZI 1974-ben közölték heurisztikus algoritmusukat [80] a gráf perifériális pontjainak meghatározására. Ez azonban szintén túl munkaigényes eljárásnak bizonyult.

Különböző heurisztikus algoritmusok születtek az átmérő, azaz a maximális excentricitás közelítésére.

1976-ban GIBBS—POOLE—STOCKMEYER közölték új eljárásukat [55], melyet „pszeudo-átmérő"-t meghatározó algoritmusnak neveztek. Nem definiálták azonban a pszeudo-átmérő fogalmát. Módszerükre a továbbiakban GPS—PS eljárásként hivatkozunk, s ezzel részletesen a 3.1. fejezetben foglalkozunk.

J. A. GEORGE és J. W-H. LIU a GPS—PS eljárás hatékonyságának növelésére 1979-ben három módosítási stratégiát közöltek [53]. Eljárásaikat „pszeudo-perifériális pont"-ot meghatározó algoritmusoknak nevezték, a pszeudo-perifériális pontot azonban ők sem definiálták. Leghatékonyabbnak ítélt eljárásukkal, melyre a továbbiakban GL—SPS-ként hivatkozunk, a 3.2. fejezetben foglalkozunk.

2.1. A gráfon értelmezett távolság függvény és a szintstruktúra kapcsolata

2.1.1. Definíció ([14]). A $G=(X, E)$ gráfban legyenek $x, y \in X$ tetszőleges pontok. Azon pontok halmazát, melynek minden eleme hozzátartozik az x és y közti legrövidebb utak egyikéhez, az x és y pontok *reverzibilis halmazának* nevezzük és $R(x, y)$-ként jelöljük.

E fogalom segítségével a következőképpen fogalmazható meg a gráfelméletből jólismert állítás.

2.1.1. TÉTEL. Tetszőleges $x, y, z \in X$ pontokra

$$d(x, y) = d(x, z) + d(y, z) \tag{2.1.1}$$

akkor és csak akkor teljesül, ha $z \in R(x, y)$.

A következőkben rámutatunk, hogyan írható fel a távolság adott szintstruktúrában.

2.1.2. TÉTEL ([14]). Legyen RLS (x) tetszőleges $x \in X$ pontban generált szintstruktúra, s legyenek

$$y \in L_i(x) \quad 0 \leqq i \leqq l(x)$$

$$z \in L_j(x) \quad 0 \leqq j \leqq l(x)$$

tetszőleges pontok. Ekkor

$$i + j - 2k_x(y, z) \leqq d(y, z) \tag{2.1.2}$$

teljesül, ahol

$$k_x(y,z) = d(x, R(y, z)). \tag{2.1.3}$$

Bizonyítás. (2.1.3) szerint létezik olyan

$$t \in R(x, y), \tag{2.1.4}$$

melyre

$$d(x, t) = k_x(y, z) \tag{2.1.5}$$

teljesül. (2.1.4) miatt a 2.1.1. tétel értelmében

$$d(y, z) = d(y, t) + d(t, z) \tag{2.1.6}$$

következik. Másrészt, a háromszög egyenlőtlenség szerint

$$d(x, y) \leqq d(x, t) + d(t, y),$$

$$d(x, z) \leqq d(x, t) + d(t, z)$$

következik, ahonnan

$$d(x, y) + d(x, z) \leqq 2d(x, t) + d(t, y) + d(t, z)$$

adódik. Innen (2.1.5) és (2.1.6) miatt

$$d(x, y) + d(x, z) - 2k_x(y, z) \leqq d(y, z) \tag{2.1.7}$$

következik, mely (2.1.2) teljesülését jelenti. ∎

KÖVETKEZMÉNYEK

2.1.1. (2.1.2) és a háromszög egyenlőtlenség miatt

$$i + j - 2k_x(y, z) \leqq d(y, z) \leqq i + j$$

következik, miszerint van olyan $W_x(y, z) \geqq 0$ érték, amelyre

$$d(y, z) = i + j - W_x(y, z) \tag{2.1.8}$$

teljesül, ahol $0 \leqq W_x(y, z) \leqq 2k_x(y, z)$.

2.1.2. (2.1.3)-ból következik, hogy $k_x(y, z)=0$ akkor és csak akkor teljesül, ha $x \in R(y, z)$. Hasonlóan, $W_x(y, z)=0$ akkor és csak akkor áll fenn, ha $x \in R(y, z)$.

2.1.2. Definíció. Legyen $x \in X$ tetszőleges pont, melynek legyen $y \in L_{ec}(x)$ tetszőleges excentricitási pontja. Az RLS (x) és RLS (y) szintstruktúrák együttesét *szintstruktúra rendszer*nek nevezzük és $\langle x, y\rangle$-ként jelöljük. Ekkor bármely $z \in X$ pont $\langle x, y\rangle$-beli helyzete a

$$z_x = d(z, x); \quad z_y = d(z, y)$$

értékek együttesével jellemezhető, melyeket a *z pont pszeudo-koordinátáinak* nevezünk. A pszeudo-koordináták összegét a pont $\langle x, y\rangle$*-beli súlyának* nevezzük és $r_{xy}(z)$-ként jelöljük.

Megjegyzések

2.1.1. $\langle x, y\rangle$ jelölésben x és y rendezett pontpárt jelöl.

2.1.2. Adott $z \in X$ pont z_x, z_y értékei nem egyértelműen határozzák meg a z-t.

2.1.3. Tetszőleges $x \in X$ pontban generált $\langle x, y\rangle$ nem egyértelműen meghatározott.

2.1.3. TÉTEL. Legyen $\langle x, y\rangle$ tetszőleges szintstruktúra rendszer és legyenek p, $q \in X$ tetszőleges pontok. Ekkor

$$d(p, q) = \frac{p_x+p_y+q_x+q_y-W(p, q)}{2} \tag{2.1.9}$$

teljesül, ahol $p=(p_x, p_y)$; $q=(q_x, q_y)$ és

$$W(p, q) = W_x(p, q)+W_y(p, q).$$

Bizonyítás. Írjuk fel $d(p, q)$-t rendre az RLS (x), illetve RLS (y) szintstruktúrákban. Így

$$d(p, q) = p_x+q_x-W_x(p, q),$$

$$d(p, q) = p_y+q_y-W_y(p, q)$$

következik, ahonnan

$$2d(p, q) = p_x+p_y+q_x+q_y-(W_x(p, q)+W_y(p, q))$$

adódik. Ez pedig (2.1.9) teljesülését jelenti. ∎

2.2. Perifériális pontok és tulajdonságaik

A perifériális pontok definíciójából adódóan érvényesek az alábbi tulajdonságok:

2.2.1 Bármely összefüggő, véges gráfnak van legalább két perifériális pontja.

2.2.2. Ha $x \in X$ perifériális pont, akkor bármely $y \in L_{ec}(x)$ szintén perifériális pont. Ekkor x és y egymás perifériális megfelelői, s együtt perifériális párt alkotnak.

2.2.3. Az $x \in X$ perifériális pont ismeretében $|L_{ec}(x)|+1$ számú diam (G) hosszúságú szintstruktúra képezhető.

2.2.4. Ha $x \in X$ perifériális pont, akkor $X \setminus (\{x\} \cup L_{ec}(x))$ tartalmazhat további perifériális pontokat.

2.2.5. Legyen $x \in X$ perifériális pont, melynek legyen $y \in L_{ec}(x)$ tetszőleges perifériális megfelelője. Ekkor

$$l(x) = \max_{p \in L_{ec}(x)} l(p); \; l(y) = \max_{q \in L_{ec}(y)} l(q), \tag{2.2.1}$$

$$l(y) = l(x) \tag{2.2.2}$$

tulajdonságok triviálisan teljesülnek.

2.2.1. TÉTEL. Legyen $x \in X$ tetszőleges pont és legyen $y \in L_{ec}(x)$ tetszőlegesen választott pont. Ekkor

$$l(x) \leqq l(y) \leqq 2l(x) \tag{2.2.3}$$

teljesül.

Bizonyítás. (2.2.3) első egyenlőtlensége triviálisan teljesül.
Legyen $z \in L_{ec}(y)$ tetszőleges pont. Ekkor

$$d(y, z) \leqq d(x, y) + d(x, z)$$

igaz $\forall x \in X$ pontra, de $d(x, z) \leqq d(x, y) = l(x)$, így

$$l(y) \leqq 2l(x)$$

következik. ∎

KÖVETKEZMÉNYEK

2.2.1. Tetszőleges $x \in X$ pontra

$$\left\lfloor \frac{\operatorname{diam}(G)+1}{2} \right\rfloor \leqq l(x) \leqq \operatorname{diam}(G)$$

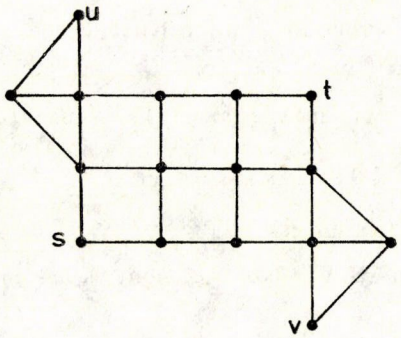
teljesül, ahol $\lfloor \; \rfloor$ az egész osztást jelöli. Más szóval, diam (G) egyértelműen meghatározza a gráfban lehetséges minimális excentricitás értékét, melyet *abszolút minimális excentricitásnak* nevezünk.

2.3. Pszeudo-perifériális pontok és tulajdonságaik

A 2.2.5. tulajdonsággal rámutattunk, hogy ha $x \in X$ perifériális pont, akkor (2.2.1) és (2.2.2) triviálisan teljesülnek. A gráfnak azonban lehetnek olyan pontjai, melyek kielégítik (2.2.1)-et és (2.2.2)-t, de nem perifériális pontok. Ilyen pontpárt szemléltetünk a 2.3.1. ábrán, ahol u és v perifériális párt alkotnak és diam $(G)=7$. $L_{ec}(s)=\{t\}$ és $l(s)=5$, míg $L_{ec}(t)=\{s\}$ és $l(t)=5$.

Tetszőleges $x \in X$ pont $y \in L_{ec}(x)$ excentricitási pontját az x *szélső pontjának* nevezzük, ha $l(y) = \max_{p \in L_{ec}(x)} l(p)$.

2.3.1. Definíció ([14]). Az $x \in X$ pontot *pszeudo-perifériális pontnak* nevezzük, ha tetszőleges $y \in L_{ec}(x)$ szélső pontjára $l(y)=l(x)$ teljesül. Ekkor x és y pszeudo-perifériális párt alkotnak, s a közös excentricitás a megfelelő *pszeudo-átmérő*, melynek x a *kezdő-*, y a *végpontja.*

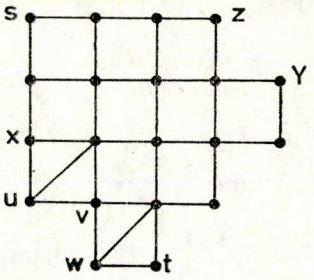


2.3.1. ábra

KÖVETKEZMÉNYEK

2.3.1. Minden perifériális pont egyben pszeudo-perifériális pont is.

2.3.2. Ha $x \in X$ pszeudo-perifériális pont, $(l(x) < \operatorname{diam}(G))$, akkor $y \in L_{ec}(x)$ pszeudo-perifériális végpont nem szükségszerűen pszeudo-perifériális pont (2.3.2. ábra).



2.3.2. ábra

Itt x pszeudo-perifériális pont és $l(x)=5$. Ugyanis $L_{ec}(x)=\{y, z\}$ és $l(y)=5$ és $l(z)=5$. Azonban $L_{ec}(y)=\{u, x, v, w, s, t\}$ és könnyű belátni, hogy s és t perifériális párt alkotnak és $\operatorname{diam}(G)=6$.

A bemutatott feladatban $L_{ec}(y)$ két pontot is tartalmaz, melyek excentricitása nagyobb, mint $l(y)$.

A pszeudo-perifériális pont fogalom tehát nem-szimmetrikus. Szimmetrikus megfelelőjét a következőképpen adjuk meg.

2.3.2. Definíció ([14]). Az $x \in X$ pontot *kvázi-perifériális pontnak* nevezzük, ha x pszeudo-perifériális pont és minden $y \in L_{ec}(x)$ szintén pszeudo-perifériális pont. Ekkor

$$x \quad \text{és minden } y \in L_{ec}(x);$$

tetszőleges $z \in L_{ec}(x)$ és minden $s \in L_{ec}(z)$ egymásnak megfelelő pszeudo-perifériális pontpárt alkotnak, melyeket *kvázi-perifériális pontpároknak* nevezünk; $l(x)$ a megfelelő *kvázi-átmérő*.

A 2.3.1. ábrán s és t kvázi-perifériális pontpár.

KÖVETKEZMÉNYEK

2.3.3. Minden perifériális pont kvázi-perifériális pont.

2.3.4. Minden kvázi-perifériális pont pszeudo-perifériális pont.

2.3.3. Definíció. Az $x \in X$ pontot *szemi-pszeudo-perifériális* pontnak nevezzük, ha van olyan $y \in L_{ec}(x)$, melyre $l(y)=l(x)$ teljesül. Ekkor x és y *szemi-pszeudo-perifériális* pontpárt alkotnak és $l(x)$ a megfelelő *szemi-pszeudo-átmérő.*

KÖVETKEZMÉNYEK

2.3.5. Minden perifériális, kvázi-, illetve pszeudo-perifériális pont egyben szemi-pszeudo-perifériális pont is.

2.3.6. A szemi-pszeudo-perifériális pont fogalom szimmetrikus.

A nemzetközi szakirodalomban 1982-ben J. K. PACHL *(University of Waterloo)* [72] munkájában jelenik meg a pszeudo-perifériális pontok alábbi definíciója:

2.3.4. Definíció (J. K. PACHL, [72]). Az $x \in X$ pontot *pszeudo-perifériális pontnak* nevezzük, ha létezik olyan $y \in L_{ec}(x)$, amelyre $l(y)=l(x)$ teljesül.

A pszeudo-perifériális pontok PACHL szerinti definíciója megegyezik jelen tárgyalás szemi-pszeudo-perifériális pontjának definíciójával.

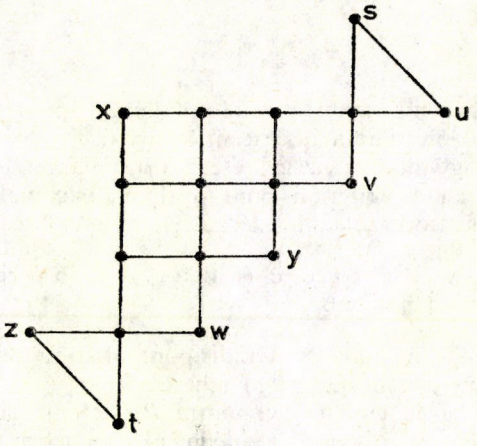
2.3.1. TÉTEL. Tetszőleges $x \in X$ szemi-pszeudo-perifériális pont excentricitására

$$l(x) \geqq \left\lfloor \frac{\operatorname{diam}(G)+1}{2} \right\rfloor$$

érvényes.

Bizonyítás. Tekintsük a 2.3.3. ábrán bemutatott gráfot, melyben s, t perifériális pontpárt jelöl és $\operatorname{diam}(G)=8$.
Ugyanakkor x pontra $l(x)=4$ és

$$L_{ec}(x) = \{s, u, v, y, w, t, z\},$$



2.3.3. ábra

továbbá $l(y)=4$, vagyis x és y szemi-pszeudo-perifériális pontpárt alkotnak és

$$l(x) = \left\lfloor \frac{\operatorname{diam}(G)+1}{2} \right\rfloor$$

teljesül. ∎

KÖVETKEZMÉNYEK

2.3.7. A szemi-pszeudo-perifériális pont lehet abszolút minimális excentricitású pont.

2.3.2. TÉTEL. Legyen $z\in X$, melyre $l(z)<\operatorname{diam}(G)$, a 2.3.1. definíció szerinti pszeudo-perifériális pont. Ekkor

$$l(z) > \left\lfloor \frac{\operatorname{diam}(G)+1}{2} \right\rfloor$$

teljesül.

Bizonyítás. Feltétel szerint z pszeudo-perifériális pont, és legyen $y\in L_{ec}(z)$ tetszőleges pont. Legyen $s, t\in X$ tetszőleges perifériális pár. Ekkor $s\notin L_{ec}(z)$ és $t\notin L_{ec}(z)$, vagyis

$$d(z, s) < d(z, y);\ d(z, t) < d(z, y)$$

teljesül, de ekkor

$$d(z, s)+1 \leqq d(z, y)$$

is igaz. Mindebből

$$d(z, s)+1+d(z, t) < 2d(z, y)$$

illetve

$$d(s, t) \leqq d(z, s)+d(z, t) < 2d(z, y)-1$$

adódik. Ez viszont

$$\left\lfloor \frac{\operatorname{diam}(G)+1}{2} \right\rfloor = \left\lfloor \frac{d(s, t)+1}{2} \right\rfloor < d(z, y) = l(z)$$

teljesülését jelenti. ∎

KÖVETKEZMÉNYEK

2.3.8. A 2.3.1. definíció szerinti pszeudo-perifériális pont excentricitása határozottan nagyobb az abszolút minimális excentricitásnál.

Megjegyezzük, hogy minden vizsgált esetben $l(z)$ diam (G)-hez közeli érték volt.

2.3.9. Mivel a pszeudo-perifériális pont fogalom a maximális excentricitású pont bizonyos közelítését hivatott szolgálni, ezért 2.3.1. definíciót fogadjuk el.

2.3.3. TÉTEL. Az $x\in X$ akkor és csak akkor pszeudo-perifériális pont, ha bármely $y\in L_{ec}(x)$ pontra $l(x)=l(y)$.

Bizonyítás. Ha $x\in X$ pszeudo-perifériális pont, akkor a definícióból következik, hogy minden $y\in L_{ec}(x)$ esetén $l(x)=l(y)$ teljesül.

Tegyük fel, hogy bármely $y\in L_{ec}(x)$ pontra $l(x)=l(y)$ igaz, de akkor ez x bármely szélső pontjára is igaz, vagyis 2.3.1. definíció értelmében x pszeudo-perifériális pont. ∎

KÖVETKEZMÉNYEK

2.3.10. Ha $x \in X$ nem pszeudo-perifériális pont, akkor van olyan $y \in L_{ec}(x)$, melyre $l(y) > l(x)$.

Megjegyzések

2.3.1. Tetszőleges $x \in X$ pszeudo-perifériális pont ismeretében $|L_{ec}(x)|+1$ számú, $l(x)$ hosszúságú gyökérrel rendelkező szintstruktúra képezhető.

2.3.4. TÉTEL ([14]). Legyen $x \in X$ tetszőleges pont, melynek jelölje $y \in L_{ec}(x)$ tetszőleges szélső pontját. Legyen

(2.3.1) $$z \in L_{ec}(y)$$

tetszőleges pont. Ha

(2.3.2) $$x \in R(y, z),$$

akkor z pszeudo-perifériális pont; z és y pszeudo-perifériális párt alkotnak.

Bizonyítás. Annak igazolásához, hogy a tétel feltételei mellett z pszeudo-perifériális pont, elegendő megmutatnunk, hogy bármely $u \in L_{ec}(z)$ pontra

(2.3.3) $$l(u) = l(z).$$

Annak igazolására, hogy z és y pszeudo-perifériális pontpárt alkotnak, elegendő

(2.3.4) $$l(z) = l(y)$$

igazolása. (2.3.1) miatt

(2.3.5) $$l(z) \geqq l(y)$$

következik. Megmutatjuk, hogy (2.3.5)-ben egyenlőség áll fenn.

Indirekt, tegyük fel, hogy

(2.3.6) $$l(z) > l(y)$$

azaz tetszőleges $u \in L_{ec}(z)$ pontra

(2.3.7) $$d(z, u) > d(z, y)$$

teljesül. Ekkor (2.3.2) miatt

(2.3.8) $$d(y, z) = d(x, y) + d(x, z)$$

következik, a háromszög egyenlőtlenségből viszont

(2.3.9) $$d(z, u) \leqq d(x, z) + d(x, u)$$

adódik, melyek (2.3.7) szerint

(2.3.10) $$d(x, u) > d(x, y) = l(x)$$

teljesítését eredményezik. (2.3.10) viszont ellentmond feltevésünknek, miszerint $y \in L_{ec}(x)$. Következésképpen (2.3.5)-ben egyenlőség van, azaz (2.3.4) teljesül.

Most (2.3.3)-at igazoljuk. Tetszőleges $u \in L_{ec}(z)$-re

(2.3.11) $$l(u) \geqq l(z) = l(y)$$

teljesül.

Könnyű belátni, hogy (2.3.4) teljesüléséből bármely $u \in L_{ec}(z)$ pontra (2.3.8) és (2.3.9) felhasználásával

$$d(u, x) = d(x, y) = l(x)$$

adódik, ami azt jelenti, hogy $L_{ec}(z) \subseteq L_{ec}(x)$. Vagyis $u \in L_{ec}(x)$ következik. Mivel $y \in L_{ec}(x)$ az x szélső pontja volt, így (2.3.11)-ben egyenlőségnek kell fennállnia, ami (2.3.3) teljesülését jelenti. ∎

3. Pszeudo-perifériális pontok meghatározási módjainak vizsgálata

Jelen fejezetben elemezzük GIBBS—POOLE—STOCKMEYER [55] és GEORGE—LIU [53] nagy excentricitást közelítő eljárásait, s algoritmust közlünk pszeudo-perifériális pontok előállítására.

3.1. A GPS—PS-módszer pszeudo-perifériális végpontot eredményez

A GPS—PS eljárás a következőképpen fogalmazható meg:

1. lépés (Kezdő pont választása)
Legyen $x \in X$ tetszőleges minimális fokszámú pont.

2. lépés (Szintstruktúra generálás)
Képezzük $\mathrm{RLS}(x) = \{L_0(x), L_1(x), \ldots, L_{ec}(x)\}$-et.

3. lépés (Rendezés)
Rendezzük $L_{ec}(x)$ pontjait fokszámaik növekvő sorrendjében, mellyel $L_{ec}(x) = \{y_1, \ldots, y_k\}$ sorrend adódik

$$i \leftarrow 0.$$

4. lépés (Excentricitási vizsgálat)

(4.1) $$i \leftarrow i+1$$

ha $l(y_i) > l(x)$, akkor $x \leftarrow y_i$, go to 3. lépés;

egyébként ha $i < k$, go to (4.1),

ha $i = k$, akkor go to 5. lépés.

5. lépés (Exit)
y_k az eredményül nyert pont, melynek excentricitása $l(x)$.

Megjegyzések

3.1.1. A módszer könnyen áttekinthető eljárás, mely a gyakorlati tapasztalatok szerint [55] sok esetben eredményez perifériális pontot.

3.1.2. A 4. lépésből látható, hogy az eljárás akkor tér rá a következő iterációs lépésre, ha $L_{ec}(x)$-ben $l(x)$-nél nagyobb excentricitású pontot talál.

3.1.1. TÉTEL ([15]). Tetszőleges kezdő pontból származtatott GPS—PS eljárás véges sok lépésben pszeudo-perifériális végpontot eredményez.

Bizonyítás. Mivel a gráf véges, így pontjai excentricitásainak szigorúan monoton növekvő sorozata is véges, legfeljebb $\left\lfloor \frac{\mathrm{diam}(G)}{2} \right\rfloor$ elemű halmaz.

Megmutatjuk, hogy az eredményül nyert pont pszeudo-perifériális végpont.

Az eljárás akkor ér véget, ha az aktuális x gyökérpont valamennyi $y_i \in L_{ec}(x)$ excentricitási pontjára $l(y_i)=l(x)$ teljesül. Ekkor a 2.3.3. tétel értelmében x pszeudo-perifériális pont.

Az eljárás eredményként x utolsóként vizsgált excentricitási pontját szolgáltatja, mely szükségszerűen pszeudo-perifériális végpont. ∎

Megjegyzések

3.1.3. Ha a gráf nem tartalmaz diam (G)-nél kisebb excentricitású kvázi-, illetve pszeudo-perifériális pontokat, akkor az eljárás szükségszerűen perifériális pontot eredményez. Ez elméletileg indokolja azt a fontos tapasztalati tényt, hogy az algoritmus sok esetben perifériális pontot eredményez [55].

3.1.4. Ha a gráf a perifériális pontokon kívül tartalmaz kvázi- vagy pszeudo-perifériális pontokat, akkor a GPS—PS által szolgáltatott excentricitás értéke függ a kezdő pont megválasztásától.

3.1.2. TÉTEL. Adott kezdőpontból származtatott GPS—PS eljárás eredményéül nyert excentricitás értéke nem egyértelműen meghatározott.

Bizonyítás. Legyen x az aktuális gyökérpont és legyenek $y_1, y_2 \in L_{ec}(x)$ minimális fokszámú excentricitási pontok. Mivel az eljárás nem tesz különbséget az azonos fokszámú pontok között, így sorrendiségük nem meghatározott. Ekkor ha

$$l(x) < l(y_1) < l(y_2)$$

teljesül és y_1 pszeudo-perifériális pont, akkor az y_1 és y_2 pontok sorrendjétől függően $l(y_1)$, illetve ennél nagyobb $l(y_2)$ excentricitás lesz az eredmény. ∎

Ilyen esetre mutatunk példát a 3. sz. mellékletben.

3.2. A GL—SPS-módszer szemi-pszeudo-perifériális pontot eredményez

J. A. GEORGE és J. W-H. LIU a GPS—PS eljárás műveletigényét úgy kívánták csökkenteni, hogy az egyes iterációs lépésekben az excentricitási szintnek csupán bizonyos, [49]-ben részletezett részhalmazaira korlátozva végezték az excentricitási vizsgálatokat. Jelen tárgyalásban a szerzők által leghatékonyabbnak ítélt verzióval [53] foglalkozunk, mely az alábbiak szerint fogalmazható meg.

1. lépés (Kezdőpont választás)
Legyen $x \in X$ tetszőleges pont.
2. lépés (Szintstruktúra képzése)
Képezzük $\mathrm{RLS}(x) = \{L_0(x), L_1(x), \ldots, L_{ec}(x)\}$-et.
3. lépés (Excentricitási pontok szelekciója)
Határozzuk meg $y \in L_{ec}(x)$ pontot, melynek fokszáma minimális.
4. lépés (Excentricitási vizsgálat)
Ha $l(y)=l(x)$, akkor go to 5. lépés;
egyébként $x \leftarrow y$;
go to 3. lépés.
5. lépés (Exit)
y az eredményül nyert pont.

Megjegyzések

3.2.1. Az eljárás tetszőleges kezdő pontból indul s az excentricitási szintet minden lépésben egyetlen, minimális fokszámú pontjára szűkíti le. Ezáltal a műveletszámban jelentős csökkenés tapasztalható a GPS—PS eljáráshoz képest. Az eljárás gépidőigénye még igen nagy méretű feladatokban is alacsony marad.

3.2.2. Az eljárás a SPARSPAK különböző ágainak fundamentális komponenseként nyert alkalmazást.

3.2.1. TÉTEL. Tetszőleges kezdőpontból származtatott GL—SPS eljárás véges sok lépésben szemi-pszeudo-perifériális pontot eredményez.

Bizonyítás. A 3.1.1. tétel bizonyításához hasonlóan könnyű belátni, hogy az eljárás véges sok lépésben szolgáltatja az eredményt.

Az eljárás akkor ér véget, ha az aktuális x gyökérpont minimális fokszámú $y \in L_{ec}(x)$ excentricitási pontjára $l(y)=l(x)$ teljesül. Ez azt jelenti, hogy x és y szemi-pszeudo-perifériális pontpár, vagyis y szemi-pszeudo-perifériális pont. ∎

Megjegyzések

3.2.3. Ha a gráf nem tartalmaz diam(G)-nél kisebb excentricitású kvázi-, pszeudo- vagy szemi-pszeudo-perifériális pontokat, akkor az eljárás szükségszerűen perifériális pontot eredményez.

3.2.4. Ha a gráf tartalmaz a perifériális pontokon kívül is szemi-pszeudo-perifériális pontot, akkor az eredményként adódó excentricitás függ a kezdőpont választásától.

3.2.2. TÉTEL. Adott kezdőpontból származtatott GL—SPS által nyert excentricitás értéke nem egyértelműen meghatározott.

Bizonyítás. Könnyen megadható olyan gráf, melyben az aktuális excentricitási szint két pontja minimális fokszámú, de excentricitásuk eltér, s egyik pont az aktuális gyökér szemi-pszeudo-perifériális megfelelője. Ekkor az eredmény-excentricitás attól függ, hogy melyik minimális fokszámú pontra esett a választás. ∎

A 4. sz. mellékletben ilyen esetet szemléltetünk.

KÖVETKEZMÉNYEK

3.2.1. A pszeudo-perifériális pontok PACHL szerinti értelmezése (2.3.4. definíció) egzakt módon írja le a GL—SPS eljárással nyert pontokat. Feltehetőleg PACHL nem ismerte fel a GL—SPS és GPS—PS eljárások révén adódó pont-típusok különbözőségét. Ezzel indokolható a „pszeudo-perifériális" pont elnevezés.

3.2.2. Mivel GL—SPS szemi-pszeudo-perifériális pontot eredményez, így az — a 2.3.7. következmény értelmében — a maximális excentricitás közelítéseként abszolút minimális excentricitást is eredményezhet.

3.2.3. GEORGE—LIU [49]-beli stratégiáinak alkalmazásával jelentős műveletszám csökkenés érhető el, de ennek fejében erős eredménybeli romlás állhat elő.

3.2.4. A fentiek alapján indokoltnak látszik, hogy a SPARSPAK megfelelő rutinja (FNROOT) esetileg gyengébb eredményeket is szolgáltat, mint a GPS—PS eljárás.

3.3. A pszeudo-perifériális pontok meghatározásának újabb közelítése

A 2.3.4. tételben megfogalmaztuk, hogy tetszőleges $x \in X$ pontból kiindulva és megfelelően választott y és z pontok esetén $x \in R(y, z)$ elégséges feltétel ahhoz, hogy z és y pszeudo-perifériális párt alkosson. E feltétel viszont $W_x(y, z) = 0$ teljesülését jelenti.

3.3.1. TÉTEL. Legyen $x \in X$ tetszőleges pont, melynek jelölje $y \in L_{ec}(x)$ tetszőleges szélső pontját. Ekkor minden

(3.3.1) $$z \in L_{ec}(y)$$

pont esetén

(3.3.2) $$W_x(y, z) = \min_{p \in L_{ec}(x)} W_x(z, p)$$

teljesül, ahol $0 \leqq W_x(y, z) \leqq 2 \cdot d(x, R(y, z))$.

Bizonyítás. Mivel y az x szélső pontja, így (3.3.1) miatt minden $p \in L_{ec}(x)$ pontban

$$d(z, y) \geqq d(z, p)$$

következik. Innen

$$d(z, x) + d(x, y) - W_x(z, y) \geqq d(z, x) + d(x, p) - W_x(z, p)$$

adódik, ahol $d(x, y) = d(x, p) = l(x)$, vagyis

$$W_x(z, p) \geqq W_x(z, y)$$

teljesül minden $p \in L_{ec}(x)$ esetén, ami (3.3.2) teljesülését jelenti. ∎

KÖVETKEZMÉNYEK

3.3.1. A tétel feltételei mellett (3.3.1) helyett válasszunk olyan $t \in L_{ec}(y)$-t, melyre

(3.3.3) $$d(t, x) = \min_{z \in L_{ec}(y)} d(z, x)$$

teljesül. Ekkor

(3.3.4) $$W_x(t, y) = \min_{z \in L_{ec}(y)} W_x(z, y).$$

Bizonyítás. Legyen $z \in L_{ec}(y)$ tetszőleges pont és $t \in L_{ec}(y)$-re feltétel szerint (3.3.3) teljesül. Ekkor $d(t, y) = d(z, y) = l(y)$ érvényes, melyből

$$d(t, x) + d(x, y) - W_x(t, y) = d(z, x) + d(x, y) - W_x(z, y)$$

illetve

$$W_x(z, y) - W_x(t, y) = d(z, x) - d(t, x)$$

következik, melyből (3.3.3) miatt

$$W_x(z, y) \geqq W_x(t, y)$$

adódik minden $z \in L_{ec}(y)$ pontra, ami pontosan (3.3.4) teljesülését jelenti. ∎

Megjegyzések

3.3.1. (3.3.2) és (3.3.4) együttesen

(3.3.5) $$W_x(t, y) = \min_{z \in L_{ec}(y)} \left(\min_{p \in L_{ec}(x)} W_x(z, p) \right)$$

teljesülését jelenti. Vagyis feltételezhető, hogy $W_x(t, y)$ kicsiny érték.

3.3.2. Számos vizsgált esetben $W_x(t, y)=0$ adódott, ami 2.3.4. tétel teljesülését eredményezi, miszerint ekkor t és y pszeudo-perifériális párt alkotnak.

Mindez egyben módszert is ad pszeudo-perifériális pontok meghatározására, mely a következőképpen fogalmazható meg [16].

1. lépés (Kezdőpont választása)
Legyen $x \in X$ tetszőleges pont.
2. lépés (Szintstruktúra képzése)
Képezzük RLS(x)-et.
3. lépés (Szélső pont meghatározása)
Határozzuk meg $y \in L_{ec}(x)$, melyre $l(y)=\max_{p \in L_{ec}(x)} l(p)$.
4. lépés (Excentricitási vizsgálat)
Ha $l(y)=l(x)$, akkor go to 5. lépés;
egyébként $x \leftarrow y$;
go to 3. lépés.
5. lépés (Exit)
x és y az előállított pszeudo-perifériális pár.

Megjegyzések

3.3.3. Könnyű belátni, hogy az eljárás, melyre PS-ként hivatkozunk, pszeudo-perifériális párt eredményez. 3.3.2 megjegyzéssel összhangban megállapítható, hogy a vizsgált esetek jelentős részében a kezdőpont szélső pontja perifériális végpontnak adódott.

3.3.4. Az eljárás úgy is tekinthető, mint a GPS—PS elméletileg megalapozott módosítása, származtatása azonban független attól.

3.3.5. Érvényesek a 3.1.3. és 3.1.4. megjegyzések.

3.3.6. A vizsgált esetek mindegyikében azt tapasztaltuk, hogy adott kezdőpont esetén az eljárás által szolgáltatott excentricitás értéke egyértelműen meghatározott.

3.3.7. Az eljárás műveletigénye erősen függ a gráf szerkezeti adottságaitól. Az iterációs lépések száma többnyire lényegesen kisebb, de a generálandó szintstruktúrák száma elérheti, sőt meg is haladhatja a GPS—PS eljárás megfelelő paramétereit.

4. Perifériális pontok meghatározása

A gráf átmérőjének meghatározásához $|X|$ számú RLS (x) generálása szükséges.

Jelen fejezetben különböző eljárásokat fogalmazunk meg, melyekkel előállítható X olyan, viszonylag kis elemszámú részhalmaza, mely tartalmaz perifériális pontot, illetve olyan nagy excentricitású pontot, mely a vizsgált esetek döntő többségében perifériális pont.

4.1. Elméleti meggondolások

4.1.1. Definíció. Tetszőleges x, $y \in X$ pontok esetén

$$M(x, y)=\left\{z \mid z \in R(x, y);\ d(z, x)=\left\lfloor \frac{d(x, y)}{2} \right\rfloor\right\}$$

halmazt az x és y pontok *középső halmazának* nevezzük.

4.1.1. TÉTEL. Legyen $x \in X$ tetszőleges pont, melyre $l(x) < \operatorname{diam}(G)$, és legyen $y \in L_{ec}(x)$ tetszőleges pont. Legyen $a \in M(x, y)$ tetszőlegesen választott pont. Ekkor bármely s, $t \in X$ pontokra, melyekre

$$d(s, t) > d(x. y) \tag{4.1.1}$$

fennáll, s és t közül legalább az egyik eleme $D(x, y, a) \subset X$ halmaznak, ahol

$$D(x, y, a) = \{z | z \in X;\ d(z, a) > d(a, x)\}. \tag{4.1.2}$$

Bizonyítás. Tekintsük s, $t \in X$ pontokat, melyek feltétel szerint kielégítik (4.1.1)-et. A háromszög egyenlőtlenségből

$$d(s, t) \leqq d(a, s) + d(a, t) \leqq 2 \cdot \max\left(d(a, s), d(a, t)\right) \tag{4.1.3}$$

következik.

Másrészt, $a \in M(x, y)$ feltételből

$$d(x, y) = d(a, x) + d(a, y) \geqq 2 \cdot d(a, x) \tag{4.1.4}$$

adódik. Most (4.1.3) és (4.1.4) felhasználásával (4.1.1)-ből

$$\max\left(d(a, s),\ d(a, t)\right) > d(a, x)$$

következik; vagyis s és t közül az a-tól távolabbi pont eleme a (4.1.2)-ben definiált halmaznak. ∎

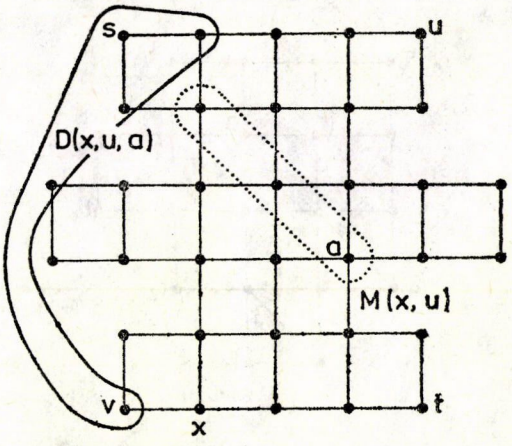
KÖVETKEZMÉNYEK

4.1.1. Minden nem-perifériális $x \in X$ esetén $D(x, y, a) \neq \{\emptyset\}$ és $D(x, y, a)$ tartalmazza a gráf valamennyi perifériális párjának legalább az egyik végpontját.

Megjegyezzük, hogy számos vizsgált esetben $D(x, y, a) \subset X$ kedvezően kis elemszámú halmaz, mint azt a 4.1.1. ábrán bemutatott példa is szemlélteti.

$\left.\begin{matrix} s, t \\ u, v \end{matrix}\right\}$ perifériális párok, $\operatorname{diam}(G) = 9$, $l(x) = 8$; $L_{ec}(x) = \{u\}$.

$s, v \in D(x, u, a)$, $|X| = 34$, $|D(x, u, a)| = 5$.

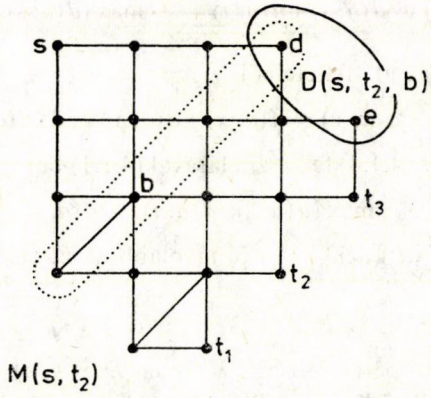


4.1.1. ábra

4.1.2. Ha $x \in X$ perifériális pont, akkor $D(x, y, a)$ nem tartalmaz szükségszerűen perifériális pontot (lásd a 4.1.2. ábrát) és előfordulhat $D(x, y, a) = 0$ speciális eset is (lásd a 4.1.3. ábrát).

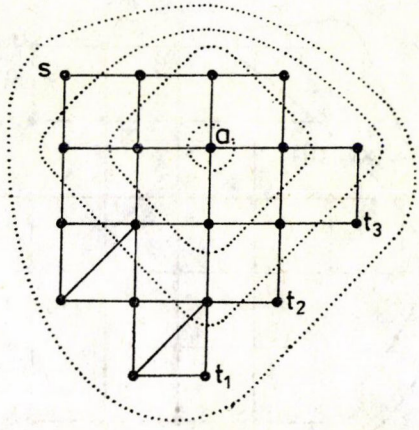
s, t_1
s, t_2 } perifériális párok
s, t_3
diam $(G) = 6$.

Itt $b \in M(s, t_2)$ esetén $D(s, t_2, b) = \{d, e\}$, vagyis e halmaz nem tartalmaz perifériális pontot.



4.1.2. ábra

$a \in M(s, t_2)$ választáskor $D(s, t_2, a) = 0$ adódik.



4.1.3. ábra

4.2. Eljárások perifériális pontok meghatározására

Fentiek alapján megállapítható, hogy tetszőleges $x \in X$ esetén minden olyan $z \in H \equiv \{x\} \cup D(x, y, a)$ pont, melyre

$$l(z) = \max_{p \in H} l(p)$$

teljesül, perifériális pont.

Mindez módszert ad perifériális pontok meghatározására, mely a következőképpen fogalmazható meg.

1. lépés (Kezdőpont választása)
Legyen $x \in X$ tetszőleges pont.

2. lépés ($M(x, y)$ meghatározása)
Tetszőleges $y \in L_{ec}(x)$ esetén képezzük $M(x, y)$-t.

3. lépés (Középső pont választása)
Legyen $a \in M(x, y)$ tetszőlegesen választott pont.

4. lépés (D halmaz meghatározása)
Határozzuk meg $D(x, y, a)$-t.

5. lépés (Perifériális pontok gyűjtése)
Határozzuk meg $\{x\} \cup D(x, y, a)$ maximális excentricitású IP számú pontjait, s tároljuk őket PNOD tömbben.

6. lépés (Exit)
PNOD (i) $(i = 1, 2, \ldots, \mathrm{IP})$ perifériális pontok.

Megjegyzések

4.2.1. A fenti algoritmus, melyre a továbbiakban P eljárásként hivatkozunk, legalább egy perifériális pontot eredményez.

4.2.2. Legyen $s \in X$ perifériális pont, melyet P eljárással határoztunk meg. Generáljuk RLS (s)-t, így $|L_{ec}(s)|$ számú újabb perifériális pontot nyerünk. Ha P eljárás kezdőpontja nem perifériális pont, akkor a fenti módon a gráf valamennyi perifériális párja előállítható.

4.2.3. Az eljárás során $|D(x, y, a)| + 3$ szintstruktúra generálása szükséges. Vagyis P algoritmus annál hatékonyabb, minél kevesebb pontot tartalmaz $D(x, y, a)$.

4.2.4. Ha egy nagyméretű gráfban $\mathrm{diam}\,(G) \leqq 6$, akkor $|D(x, y, a)|$ nagy, $|X|$-hez közeli érték is lehet.

Bizonyítás. Ha $\mathrm{diam}\,(G) = 6$, akkor a 2.2.1. következmény értelmében minden $z \in X$ pontra

$$3 \leqq l(z) \leqq 6$$

teljesül. Következésképpen, $d(a, x) \leqq 3$ igaz. Másrészt, RLS (a)-ban

$|L_0(a)| = 1$,
$|L_1(a)| \leqq m$, ahol m a gráfban a maximális fokszámot jelöli,
$|L_2(a)| \leqq (m-1) \cdot m$

érvényes, vagyis RLS (a) első három szintjében legfeljebb $m^2 + 1$ pont van. Ha a mátrix kielégíti a ritka mátrixok TEWARSON szerinti definícióját [83], akkor $m \leqq 10$ feltételezhető. Vagyis

$$D(x, y, a) = \sum_{j=d(a,x)+1}^{l(a)} L_j(a)$$

halmaz az X jelentős részét tartalmazhatja. ∎

Megjegyezzük, hogy a gyakorlatban a nagyméretű feladatok mindegyikében $\operatorname{diam}(G) \geqq 7$ teljesülését tapasztaltuk. Véletlen gráfok esetén viszont $\operatorname{diam}(G) \leqq 6$ gyakran előfordulhat (5. sz. melléklet).

Vizsgáljuk meg, hogy hogyan növelhető a P algoritmus hatékonysága.

4.2.1. TÉTEL. Tetszőleges $x \in X$, $y \in L_{ec}(x)$ és $a \in M(x, y)$ esetén tekintsük RLS (a)-t és jelölje $h(x, a)$ a $D(x, y, a)$ halmazban fekvő szintjeinek a számát. Ekkor

$$0 \leqq h(x, a) \leqq \frac{3 \cdot \operatorname{diam}(G)}{4} \tag{4.2.1}$$

teljesül.

Bizonyítás. Nyilvánvalóan,

$$h(x, a) = l(a) - d(a, x) \tag{4.2.2}$$

érvényes.

(4.2.1) első relációja triviálisan teljesül.

Tetszőleges $x \in X$ esetén $l(x) \geqq \operatorname{diam}(G)/2$ teljesüléséből $d(a, x) \geqq \operatorname{diam}(G)/4$ következik. Másrészt $l(a) \leqq \operatorname{diam}(G)$, így (4.2.2) alapján

$$h(x, a) \leqq \operatorname{diam}(G) - \frac{\operatorname{diam}(G)}{4} = \frac{3 \cdot \operatorname{diam}(G)}{4}$$

teljesülése következik. ∎

Megjegyzések

4.2.5. Előfordulhat, hogy $D(x, y, a)$ tartalmazza RLS (a) szintjeinek 3/4 részét.

4.2.6. Számos vizsgált esetben azt tapasztaltuk, hogy nagyobb $h(x, a)$ érték $|D(x, y, a)|$ növekedését eredményezte.

Könnyű belátni, hogy $h(x, a)$ kedvezően kis érték, ha $a \in M(x, y)$ pontra

$$l(a) = \min_{x, y \in C} \left(\min_{b \in M(x, y)} l(b) \right) \tag{4.2.3}$$

teljesül, ahol C a perifériális párok halmazát jelöli.

Célunk (4.2.3) közelítésének olyan előállítása, hogy a P eljárás ily módon nyert módosítása kisebb műveletigényű legyen, mint az eredeti P.

A. Vegyük észre hogy ha $x \in X$ viszonylag nagy excentricitású, akkor (4.2.2) alapján a $h(x, y)$ várhatóan csökken. Ezért módosítsuk a P eljárás 1. lépését a következőképpen.

1′ lépés (Kezdőpont választása)

Tetszőleges $z \in X$ pontnak határozzuk meg tetszőleges $x \in L_{ec}(z)$ szélső pontját.

Az így előálló P′ eljárás műveletigénye $|L_{ec}(z)| + |D(x, y, a)| + 3$ (szintstruktúrában). Ebből látható, hogy ugyanazon z kezdőpont esetén

(i) ha z excentricitása kicsi, akkor P′ gyorsabbá válik P-nél;

(ii) ha viszont z nagy excentricitású és excentricitási szintje sok pontot tartalmaz, akkor P′ gazdaságtalanná válhat.

B. Most azt vesszük figyelembe, hogy adott kezdőpont esetén $h(x, a)$ akkor is csökken, ha $l(a)$ kis érték.

Módosítsuk ezért a P eljárás 3. lépését a következőképpen:

3′ lépés (Középső pont választás)

Legyen $a \in M(x, y)$ minimális excentricitású pont.

Az így előálló P'' eljárás műveletigénye $|M(x,y)|+|D(x,y,a)|+2$, mely láthatóan a gráf szerkezeti adottságaitól függő érték.

C. Alkalmazzuk most egyidejűleg a bemutatott két módosítást, vagyis a P eljárás 1. és 3. lépése helyett alkalmazzuk 1′. és 3′. lépéseket. Az így nyert P''' eljárás műveletigénye $|L_{ec}(z)|+|M(x,y)|+|D(x,y,a)|+2$. Látható, hogy az esetenként kellemetlenül megnőhet, de előállhat az igen kedvező $|D(x,y,a)|=0$ eset is.

Megjegyzések

4.2.7. A P eljárás és annak P′, P″ és P‴ módosításainak hatékonysága erősen függ a gráf szerkezeti adottságaitól, és a gráf átmérőjétől.

4.3. Heurisztikus algoritmus perifériális pontok meghatározására

A 4.1.1. tétel következményei szerint a pszeudo-perifériális pontokat előállító eljárásokban az $a \in M(x, y)$ kezdőpont választás kedvezőnek ígérkezik.

Módosítsuk pszeudo-perifériális pontot meghatározó PS eljárásunk (3.3. fejezet) 1. lépését az alábbiak szerint:

1′ lépés (Kezdőpont választása)

Tetszőleges $z \in X$ és $y \in L_{ec}(z)$ pontok esetén legyen $x \in M(y, z)$ tetszőleges pont.

Az így nyert eljárásra PS1-ként hivatkozunk.

Megjegyzések

4.3.1. PS1 eljárás pszeudo-perifériális végpontot eredményez.

4.3.2. Minden vizsgált esetben PS1 az első lépésben szolgáltatta az eredményt. Ez azt jelenti, hogy valamennyi vizsgált esetben PS1 az $L_{ec}(a)$ halmazban találta meg a pszeudo-perifériális végpontot.

E tapasztalatok alapján fogalmazzuk meg az alábbi sejtésünket.

4.3.1. Sejtés ([17]). Legyen $x \in X$ tetszőleges pont és legyenek $y \in L_{ec}(x)$ és $a \in M(x, y)$ tetszőlegesen választott pontok. Ekkor a-nak bármely $s \in L_{ec}(a)$ szélső pontja pszeudo-perifériális végpont.

A fenti állítást nem sikerült igazolni, de nem sikerült ellenpéldát konstruálni, mely cáfolná az állítás helyességét.

KÖVETKEZMÉNYEK

4.3.1. Módosítsuk a PS1 eljárást úgy, hogy az excentricitási vizsgálatokat csupán $L_{ec}(a) \cup \{x\}$ halmazra szorítkozva végezzük. Az így nyert eljárásra P1 algoritmusként hivatkozunk.

Tetszőleges $x \in X$, $y \in L_{ec}(x)$ és $a \in M(x, y)$ pontok esetén ha $D(x, y, a) \neq \{\emptyset\}$, akkor $L_{ec}(a) \subset D(x, y, a)$. A 4.1.1. tétel alapján ésszerű feltételezésnek tűnik, hogy a nagy excentricitású pontok $L_{ec}(a)$-ban helyezkednek el. Ezért módosítsuk a P eljárás 4. lépését úgy, hogy az excentricitási vizsgálatokat csupán $L_{ec}(a)$ pontjaira szorítkozva végezzük. Láthatóan, az így előálló eljárás pontosan a fenti P1 algoritmust eredményezi, mely tehát a következőképpen fogalmazható meg:

1. lépés (Kezdeti beállítás)
Tetszőleges $x \in X$, $y \in L_{ec}(x)$ pontokhoz legyen $a \in M(x, y)$ tetszőlegesen választott pont.

2. lépés (Szintstruktúra képzése)
Generáljuk RLS $(a) = \{L_0(a), L_1(a), \ldots, L_{ec}(a)\}$-t.

3. lépés (Szélső pont meghatározása)
Határozzuk meg $s \in L_{ec}(a)$ pontot, melyre

$$l(s) = \max_{p \in L_{ec}(a)} l(p).$$

4. lépés („Perifériális" pont kiválasztása)
Ha $l(x) > l(s)$, akkor $s \leftarrow x$;

5. lépés (Exit)
s az eredmény pont.

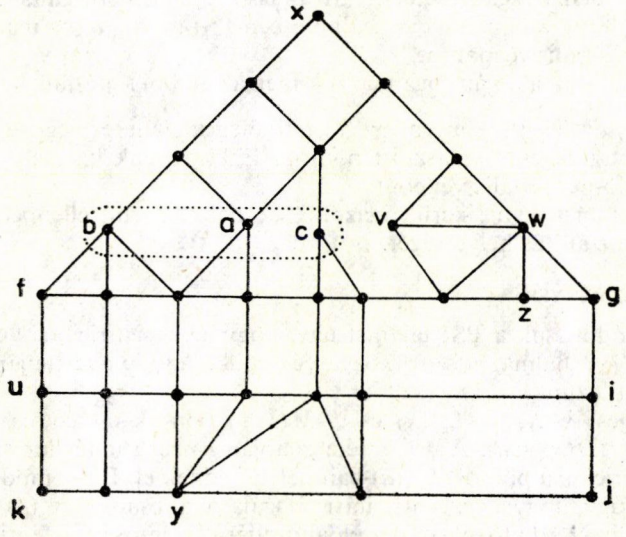
Megjegyzések

4.3.3. A P1 eljárás, mint a PS1 algoritmus módosítása — a 4.3.1. sejtéssel összhangban — minden vizsgált esetben pszeudo-perifériális végpontot eredményez.

4.3.4. A P eljárásban a $D(x, y, a)$ halmaznak $L_{ec}(a)$ halmazra való szűkítése önkényes, így a P1 algoritmus által nyert pont nem szükségszerűen perifériális pont, azaz P1 heurisztikus algoritmus perifériális pontok meghatározására.

4.3.5. Igen nagyszámú feladaton nyert tapasztalatok szerint megállapítható, hogy a P1 eljárással nyert pszeudo-perifériális végpont a vizsgált esetek döntő többségében perifériális pont. Ilyen esetet szemléltetünk a 4.3.1. ábrán.

$(f, g,)$ $(u, v,)$ (u, w), (u, z) perifériális párok
diam $(G) = 8$.
$l(x) = 6$.



4.3.1. ábra

Könnyen ellenőrizhető, hogy $L_{ec}(a) = \{z, g, i, j, k\}$,

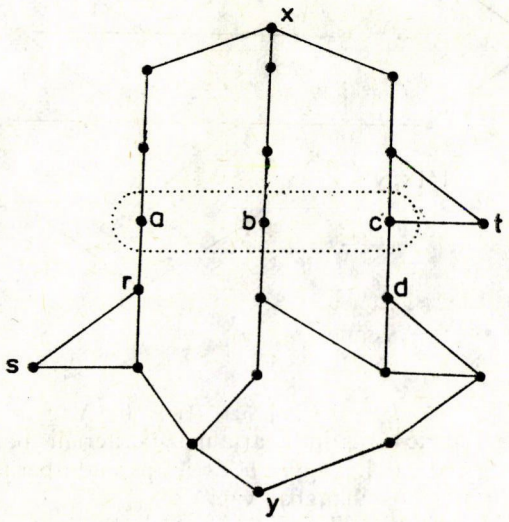
$$L_{ec}(b) = \{g\},$$

$$L_{ec}(c) = \{u\},$$

Vagyis $M(x, y)$ minden pontjának excentricitási szintje tartalmaz legalább egy perifériális pontot.

Hosszú ideig minden vizsgált esetben hasonló eredményre jutottunk.

1983-ban azonban W. F. SMYTH* professzor ellenpéldát konstruált (4.3.2. ábra), melyben van olyan $a \in M(x, y)$, hogy $L_{ec}(a) \cup \{x\}$ nem tartalmaz perifériális pontot.
s, t perifériális pár;
$\text{diam}(G) = 8$.
$l(x) = 7$
$y \in L_{ec}(x)$



4.3.2. ábra

Könnyen ellenőrizhető, hogy

$$L_{ec}(a) = \{d\},\ l(d) = 7;$$

$$L_{ec}(b) = \{a\},\ l(a) = 7;$$

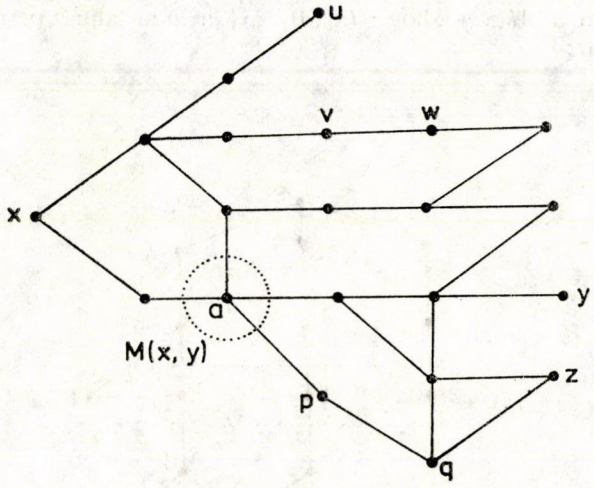
$$L_{ec}(c) = \{s, r\};$$

ami azt jelenti, hogy $M(x, y)$ két pontja is olyan, hogy excentricitási szintjében nincs perifériális pont. Azonban $L_{ec}(d) = \{a\}$, így a és d kvázi-perifériális pontpárt alkotnak, ami 4.3.1. sejtésünkkel összhangban van.

* *McMaster University, Unit for Computer Science, Hamilton, Ontario, Canada.*

1984-ben sikerült a bemutatottnál erősebb ellenpéldát konstruálnunk (4.3.3. ábra), melyben $M(x, y)$ egyetlen pontjának excentricitási szintje sem tartalmaz perifériális pontot.

u, y
u, z } perifériális párok, $l(x) = 5$
v, z
diam $(G) = 7$.



4.3.3. ábra

Most $M(x, y) = \{a\}$ és $L_{ec}(a) = \{w\}$ és $l(w) = 6$. Vagyis $M(x, y)$-ban nincs olyan pont, melynek excentricitási szintje tartalmaz perifériális pontot. Ugyanakkor $L_{ec}(w) = \{p, q, z\}$ és $L_{ec}(p) = \{w\}$, vagyis p és w pszeudo-perifériális párt alkotnak, ami sejtésünk állításával összhangban van.

Megjegyezzük, hogy mindkét ellenpéldában a P1 által szolgáltatott excentricitás nem kisebb (diam $(G) - 1$)-nél.

Megjegyzések

4.3.6. A gyakorlati esetek döntő többségében a P1 eljárás perifériális pontot eredményez, míg műveletigénye $|L_{ec}(a)| + 3$, ahol igen gyakran

$$|L_{ec}(a)| \ll |D(x, y, a)|$$

teljesül.

4.3.7. Ha nagyméretű gráfok esetén diam $(G) \leqq 6$, akkor P1 hatékonysága lecsökkenhet (4.2.4. megjegyzés).

Eljárásaink értékeléséhez szükséges számítógépes eredményeinket az 5. sz. mellékletben közöljük.

Vizsgálataink teszt-anyagát azonos (500) pontszámú, növekvő élsűrűségű

(1500, 2000, 2500 élű) véletlen gráfok képezik, s minden sűrűség esetén 3-3 feladatot tekintünk. A véletlen gráfokat a következőképpen képeztük:

> minden pontban azonos számú élt képezünk, de az élek végpontjait véletlenszám generátorral állítottuk elő. Így az élek 50%-a azonos fokszámot biztosít, a másik 50% véletlen változásokat idéz elő a fokszámokban is.

A táblázatokban az egyes gráfok mellett feltüntetjük a gráf átmérőjét (d).

Az egyes eljárásoknál feltüntetjük a szintstruktúrában mért műveletszámot (Op), a végrehajtáshoz szükséges, másodpercben mért gépidőt (t), az azonos sűrűségű gráfokon nyert átlagos műveletszámot (Op_a), illetve a gyökérpontok gráfbeli, átlagos százalékát $\left(\frac{Op_a}{5}\right)$.

A számítógépes futtatásainkat IBM 3031 gépen, OS operációs rendszer alatt végeztük.

A közölt gépidők CPU-időket jelölnek, melyek regisztrálását a LISKA TIBOR által kidolgozott CPUTIM rutin [65] tette lehetővé.

A gépidő-méréssel kapcsolatos néhány észrevételünket a 6. sz. mellékletben csatoljuk, mely szerint a közölt gépidőkre 0,001484 hibakorlát tekinthető érvényesnek. Számítógépes eredményeink alapján megállapítható, hogy

— P1 eljárás minden esetben perifériális pontot szolgáltatott.

— A $G=(500, 1500)$ gráfok mindegyikében az átmérő nagyobb (6), mint a többi gráfban (5). Részben ezzel magyarázható, hogy eljárásaink viszonylag kedvező műveletszám-csökkenést is tudnak előidézni: P‴ durván 30%-os, P′ csaknem ilyen mértékű átlagos műveletszám-csökkenést eredményez, míg P″ növeli az eredeti műveletszámot.

P1 eljárás igen gyors, 94%-os javulást eredményez.

— A $G=(500, 2000)$ gráfokon a P1 eljáráson kívül csupán P″ tud csekély javulást eredményezni. A I. és II. gráfokon P‴ különösen nagy romlást idéz elő.

P1 durván 58%-os gépidő-csökkenéssel dolgozik.

— A $G=(500, 2500)$-as sorozat gráfjain P″ durván P-vel azonos hatékonyságú, bár a III. gráfban jobbnak bizonyul. P′ és P‴ eljárások viszont többnyire nagyobb műveletszámot igényelnek, mint P.

P1 most is kiugróan jónak bizonyul, átlagosan durván 70%-os műveletszám-csökkenést eredményez a P-vel való összehasonlításban.

Összefoglalva megállapítható, hogy eljárásaink hatékonysága erősen függ a gráf szerkezeti adottságaitól, s mindenek előtt a gráf átmérőjétől.

P, P′, P″, P‴ eljárásaink durván azonos hatékonyságúaknak tekinthetők igen kis átmérő esetén is. Az átmérő növekedésével eljárásaink hatékonysági növekedése várható ($G=(500, 1500)$).

P1 eljárás e szélsőségesnek tekinthető kis átmérők esetén is igen kedvező műveletszám-csökkenést eredményezett a vizsgált esetek mindegyikében.

Perifériális pontok konkrét meghatározásakor, ha gráf átmérője nem becsülhető, — azaz a fenti szélsőséges esetek is előállhatnak — akkor P eljárás alkalmazását javasoljuk.

Ha az alkalmazási feladat természetéből adódóan elegendő csupán a gráf nagy, átmérőhöz közeli excentricitású pontjának meghatározása, akkor egyértelműen P1 eljárás alkalmazását javasoljuk.

Megjegyzések

4.3.8. 1984-ben W. F. SMYTH professzor kidolgozta a P1 eljárásunk továbbfejlesztését [82], melyben $L_{ec}(a)$ tetszőleges minimális fokszámú pontját fogadja el eredménynek. E módosítással jelentős műveletszám-csökkenést ért el, de az előálló excentricitás természetesen kisebb is lehet, mint P1 esetén.

4.4. Perifériális pontok meghatározásának más közelítése

A 2.1.2. definícióval bevezettük tetszőleges $x \in X$ pontban generált $\langle x, y \rangle$ szintstruktúra rendszert, s tetszőleges $z \in X$ pont $\langle x, y \rangle$-beli súlyát

$$r_{xy}(z) = z_x + z_y$$

szerint értelmeztük, ahol $z_x = d(z, x)$, $z_y = d(z, y)$ a z pont pszeudo-koordinátái.

Megjegyzések

4.4.1. Tetszőleges $\langle x, y \rangle$ egyértelműen definiálja a benne fellépő súlyok pontos alsó (k) és felső korlátját (K) és

$$k = l(x); \quad K \leqq l(x) + l(y)$$

érvényesül.

Az alábbiakban az $\langle x, y \rangle$-beli súlyok és excentricitások kapcsolatát vizsgáljuk.

4.4.1. TÉTEL. Tegyük fel, hogy $\langle x, y \rangle$-ban $k < K$. Legyen $s \in X$ tetszőleges pont, melyre $r_{xy}(s) < K$ s legyen $t \in L_{ec}(s)$ tetszőleges pont. Ha $z \in L_{ec}(t)$ pontra

$$W(z, t) > W(s, t) \tag{4.4.1}$$

teljesül, akkor

$$r_{xy}(z) > r_{xy}(s) \tag{4.4.2}$$

is fennáll.

Bizonyítás. A feltétel szerint tetszőleges $s \in X$ esetén $t \in L_{ec}(s)$, vagyis $l(t) \geqq l(s)$; amely szerint bármely $z \in L_{ec}(t)$ pontban

$$d(z, t) \geqq d(s, t) \tag{4.4.3}$$

következik.

Írjuk fel a (4.4.3)-beli távolságokat a 2.1.3. tétel alkalmazásával, így

$$2d(z, t) = r_{xy}(z) + r_{xy}(t) - W(z, t),$$

$$2d(s, t) = r_{xy}(s) + r_{xy}(t) - W(s, t)$$

adódik. Ezeket a (4.4.3)-ba helyettesítve

$$r_{xy}(z) - W(z, t) \geqq r_{xy}(s) - W(s, t),$$

azaz

$$r_{xy}(z) - r_{xy}(s) \geqq W(z, t) - W(s, t)$$

következik. Ekkor a (4.4.1) feltétel a (4.4.2) teljesülését vonja maga után, s ezt akartuk megmutatni. ∎

Megjegyzések

4.4.2. A 4.4.1. tétel állítása szerint bármely nem maximális súlyú s ponthoz tetszőleges $t \in L_{ec}(s)$ pont excentricitási szintjében a (4.4.1) feltételnek eleget tevő $v \in L_{ec}(t)$ pontok súlyai határozottan nagyobbak az s súlyánál.

4.4.2. TÉTEL. Tegyük fel, hogy $\langle x, y\rangle$-ban $k<K$. Legyen $s \in X$ tetszőleges pont, melyre $r_{xy}(s)<K$ és $l(s)<\text{diam}\,(G)$ teljesülnek. Legyen $t \in L_{ec}(s)$ tetszőleges pont, továbbá $z \in L_{ec}(t)$ olyan, hogy

(4.4.3) $$r_{xy}(z) > r_{xy}(s).$$

Ekkor a

(4.4.4) $$W(z, t) \leqq W(s, t)$$

egyenlőtlenségből

(4.4.5) $$d(z, t) > d(s, t)$$

következik.

Bizonyítás. Feltevésünk szerint tetszőleges $s \in X$ és $t \in L_{ec}(s)$ esetén $z \in L_{ec}(t)$ kielégíti (4.4.3)-at.
A 2.1.3. tétel alkalmazásával

$$r_{xy}(z) = 2d(z, t) - r_{xy}(t) + W(z, t)$$

$$r_{xy}(s) = 2d(s, t) - r_{xy}(t) + W(s, t)$$

érvényes. Ezeket (4.4.3)-ba helyettesítve

$$2d(z, t) + W(z, t) > 2d(s, t) + W(s, t),$$

vagy átrendezés után

$$2(d(z, t) - d(s, t)) > W(s, t) - W(z, t)$$

adódik. Ebből a (4.4.4) miatt (4.4.5) teljesülése következik, s ezt akartuk megmutatni. ∎

Megjegyzések

4.4.3. A 4.4.2. tétel állítása szerint bármely, nem maximális súlyú, nem-perifériális s pont esetén a nála nagyobb súlyú pontok halmazában a (4.4.4)-t kielégítő pontok szükségszerűen nagyobb excentricitásúak mint s.

4.4.4. A 4.4.1. és 4.4.2. tételek azt mutatják, hogy bizonyos esetekben az excentricitás növekedésével a pont súlya nőhet, illetve a súlyok növekedésével nagyobb excentricitás adódik.

4.4.5. A vizsgált esetek jelentős részében azt tapasztaltuk, hogy a maximális súlyú pontok H halmazában a maximális excentricitású pont perifériális pont. Ez a tapasztalati felismerés lehetőséget ad nagy excentricitású pontok meghatározására. A vonatkozó eljárás a következőképpen fogalmazható meg:

1. lépés (Kezdőpont választása)
Legyen $x \in X$ tetszőleges pont.

2. lépés ($\langle x, y\rangle$ generálása)
Tetszőleges $y \in L_{ec}(x)$ esetén képezzük RLS (x) és RLS (y) szintstruktúrákat, s regisztráljuk a pontbeli súlyokat.

3. lépés (H meghatározása)
Állítsuk elő H-ban a maximális súlyú pontok halmazát.

4. lépés (Excentricitási vizsgálat)
Határozzuk meg $s \in H$ pontot, melyre

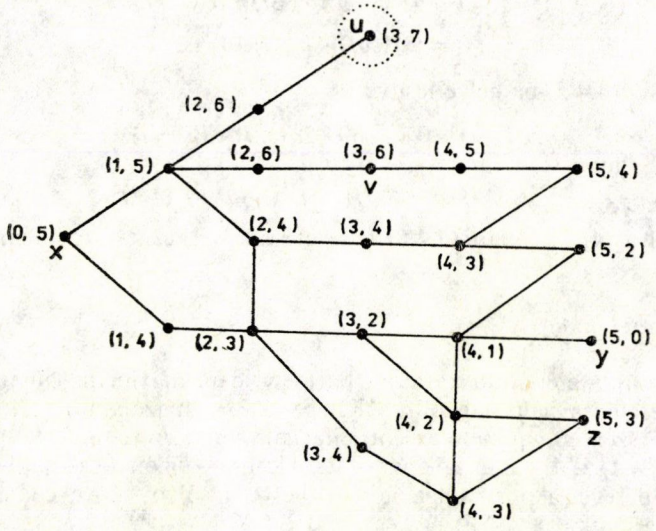
$$l(s) = \max_{p \in H} l(p)$$

teljesül.

5. lépés (Exit)
s nagy excentricitású pont.

Megjegyzések

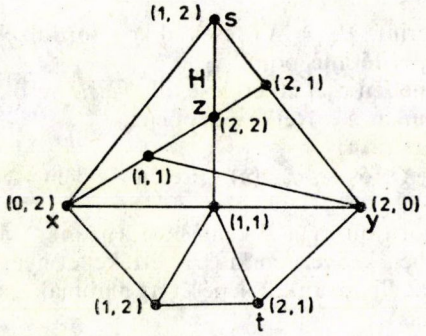
4.4.6. A fenti algoritmus, melyre P2-ként hivatkozunk, a vizsgált esetek jelentős részében perifériális pontot eredményez, melynek szemléltetésére a 4.4.1. ábrán mutatunk példát; (a pontok mellett feltüntetjük pszeudo-koordinátáikat).
Tekintsük $\langle x, y \rangle$-t.
u, y
u, z } perifériális párok.
v, z
$H \equiv \{u\}$ és u perifériális pont.



4.4.1. ábra

A 4.4.2. ábrán viszont arra mutatunk példát, hogy P2 eljárás nem eredményez perifériális pontot. (A példát KÖRNYEI IMRE* konstruálta 1983-ban).
Tekintsük $\langle x, y \rangle$-t.
$H \equiv \{z\}$ és $l(z) = 2$.
Könnyű belátni, hogy az s, t az egyetlen perifériális pár és diam $(G) = 3$.

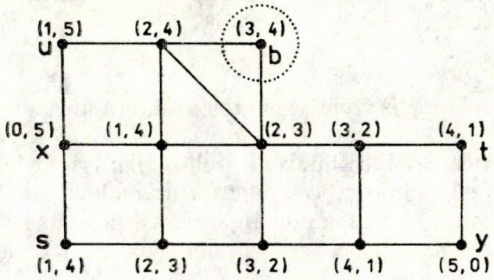
* ELTE Számítóközpont, Budapest.



4.4.2. ábra

1983-ban egyszerű síkbeli gráfok körében is sikerült ellenpéldát találnunk, melyet a 4.4.3. ábrán közlünk.

x, y
s, t } perifériális párok
u, y
diam $(G)=5$
$H=\{b\}$.

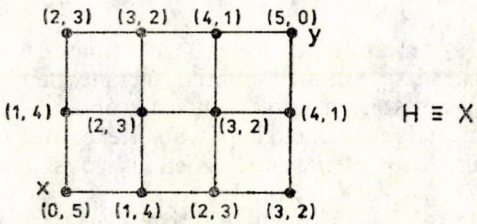


4.4.3. ábra

Megjegyzések

4.4.7. Az eljárás műveletigénye $|H|$-tól függ. $|H|$ általában kedvezően kicsi érték, de előfordulhat $|H|\equiv|X|$ szélsőséges eset is (4.4.4. ábra).

Tekintsük $\langle x, y\rangle$-t.
$H\equiv X$



4.4.4. ábra

Tapasztalataink szerint $|H|=|X|$ eset akkor fordul elő, ha a szintstruktúra rendszer x kezdőpontja perifériális pont.

4.4.8. A fenti kellemetlen eset elkerülése, illetve $|H|$ csökkentése érdekében módosítsuk P2 algoritmusunkat a következőképpen:

1′ lépés (Kezdőpont választása)

Tetszőleges $u\in X$ és $v\in L_{ec}(u)$ esetén legyen $x\in M(u,v)$ tetszőlegesen választott pont.

Az így előálló algoritmusra P2′-ként hivatkozunk. Megjegyezzük, hogy P2′ alkalmazásával sok esetben kedvezően kis $|H|$ értékeket nyertünk.

Módosítsuk most P2 eljárásunk 2. lépését az alábbiak szerint:

2′ lépés ($\langle x,y\rangle$ generálása)

Legyen $u\leftarrow x$; Legyen $x\in L_{l(u)-1}(u)$ tetszőleges pont. Képezzük $\langle x,y_i\rangle y_i\in L_{ec}(x)$ szintstruktúra rendszereket s válasszuk ki azt, amelyre $|H|$ minimális.

Ha $|H|\leqq 3$, akkor elfogadjuk kedvezően kis értéknek, nem vizsgálunk további $|H|$-értékeket.

Az így előálló módosításra P2″-ként hivatkozunk.

Megjegyezzük, hogy P2′ a kisebb munkaigényű, P2″ viszont sok esetben kisebb $|H|$-értéket eredményez.

Vonatkozó számítógépes eredményeinket a 7. sz. mellékletben közöljük, ahol P2″ eljárást alkalmaztuk.

5. Közel-minimális szélességű szintstruktúra meghatározása

A perifériális pontot előállító, illetve közelítő eljárásaink lehetővé teszik a gráf leghosszabb, illetve közel-leghosszabb szintstruktúráinak előállítását.

Könnyű belátni, hogy a gráf minimális szélességű gyökérrel rendelkező szintstruktúrája nem szükségszerűen perifériális pont gyökerű RLS (x).

5.1. A probléma közelítésének történeti áttekintése

1972-ben [4]-ben speciális szerkezetű általános szintstruktúrát állítottunk elő, melyben a maximális fokszámú pontok szintstruktúra-szélességet növelő hatását kívántuk ellensúlyozni. Ezt az 5.3. fejezetben ismertetjük.

1976-ban GIBBS és munkatársai [55] a GPS—PS eljárással meghatározott x,y pszeudo-perifériális párból kiindulva általános szintstruktúrát konstruáltak, az alábbi elvek szerint:

— Képezték $R(x,y)$-t.

— A $G(X\setminus R(x,y))$ metszetgráf összefüggő komponenseit RLS (x), illetve RLS (y) vonatkozásában vizsgálva, azok pontjait úgy illesztették $R(x,y)$-hoz, illetve ennek, mint x-ből és y-ból származtatott szintstruktúráknak a szintjeihez, hogy a bővítés során a szint-szélességek a lehető legkisebb értéken maradjanak.

Eljárásuk bonyolult, de sok esetben kedvezően kis szélességű általános szintstruktúrát eredményez.

1978-ban saját-fejlesztésű heurisztikus eljárásunkkal (lásd 8. sz. melléklet) előállítottuk a gráf két, átmérőhöz közeli excentricitású pontját, s ezekből kiindulva

[4]-ben közölt algoritmusunkkal speciális szerkezetű általános szintstruktúrát állítottunk elő.

1981-ben GEORGE-LIU [53] a GL—SPS eljárásukkal nyert szemi-pszeudo-perifériális pár között létesíthető szintstruktúrát módosítás nélkül elfogadták.

Jelen fejezetben azt vizsgáljuk, hogy a maximálist megközelítő excentricitású pontok ismeretében hogyan állítható elő közel-minimális szélességű gyökérrel rendelkező, illetve általános szintstruktúra.

5.2. Speciális szerkezetű általános szintstruktúrák előállítása

5.2.1. Definíció. A $G=(X, E)$ gráf $D \subset X$ tagozódási halmazát *minimális tagozódási halmaznak* nevezzük, ha D egyetlen valódi részhalmaza sem tagozódási halmaz.

5.2.1. TÉTEL. Tetszőleges nem teljes $G=(X, E)$ gráfban legyen $x, y \in X$ olyan pontpár, melyre $(x, y) \notin E$. Ekkor van a gráfnak olyan D minimális tagozódási halmaza, mely az x és y pontokat elválasztja.

Bizonyítás. A viszonylag egyszerűen bizonyítható állítás igazolására konstruktív bizonyítást adunk, azaz tetszőleges $x, y \in X$; $(x, y) \notin E$ pontpárból kiindulva előállítjuk azon D minimális tagozódási halmazt, mely x és y pontokat elválasztja.

Mivel G nem teljes, így van olyan $x, y \in X$, hogy $(x, y) \notin E$. Generáljuk rendre RLS (x) és RLS (y) szintstruktúrák azonos indexű szintjeit, s képezzük az aktuálisan előállított szintek közös részeinek egyesítését. Könnyű belátni, hogy az így előálló halmaz a gráf minimális tagozódási halmaza, mely x-et és y-t elválasztja.

A konstrukció a következőképpen fogalmazható meg.

1. lépés (Kezdeti beállítások)

Tetszőleges $x, y \in X$; $(x, y) \notin E$ pontokat tekintve

$$D \leftarrow \{\emptyset\}$$

$$J_0 \leftarrow \{x\}, \quad M_0 \leftarrow \{y\},$$

$$X_1 \leftarrow \{x\}, \quad X_2 \leftarrow \{y\}.$$

$$i \leftarrow 0.$$

2. lépés (RLS (x) aktuális szintjének vizsgálata)

$$i \leftarrow i+1$$

$$L_{i-1}(x) \leftarrow L_{i-1}(x) \setminus D,$$

$$J_i \leftarrow L_i(x) \setminus D$$

Ha $|J_i|=\{\emptyset\}$ és $|M_{i-1}|=\{\emptyset\}$, akkor go to 4. lépés.
Ha $|J_i|=\{\emptyset\}$ és $|M_{i-1}| \neq \{\emptyset\}$, akkor go to 3. lépés.

Ha $|J_i| \neq \{\emptyset\}$; akkor

$$D \to D \cup J_i \cap M_{i-1},$$

$$X_1 \leftarrow (X_1 \cup J_i) \setminus (D \cup X_2),$$

go to 3. lépés.

3. lépés (RLS (y) aktuális szintjének vizsgálata)

$$L_{i-1}(y) \leftarrow L_{i-1}(y) \setminus D,$$

$$M_i \leftarrow L_i(y) \setminus D$$

Ha $|M_i| = \{\emptyset\}$ és $|J_i| = \{\emptyset\}$, akkor go to 4. lépés.
Ha $|M_i| = \{\emptyset\}$ és $|J_i| \neq \{\emptyset\}$, akkor go to 2. lépés.
Ha $|M_i| \neq \{\emptyset\}$, akkor

$$D \leftarrow D \cup M_i \cap J_i,$$

$$X_2 \leftarrow (X_2 \cup M_i) \setminus (D \cup X_1),$$

go to 2. lépés.

4. lépés (Exit)

$$X_1 \leftarrow X_1 \setminus D; \quad X_2 \leftarrow X_2 \setminus D$$

D a gráf minimális tagozódási halmaza, mely az x és y pontokat elválasztja; $G(X_1) \cup G(X_2)$ nem összefüggő gráf.

Megmutatjuk, hogy $G(X_1) \cup G(X_2)$ nem összefüggő, s az előállított D a gráf minimális tagozódási halmaza, mely az x és y pontokat elválasztja.

Indirekt, tegyük fel, hogy $G(X_1) \cup G(X_2)$ összefüggő gráf, vagyis bármely két pontja, így $x \in X_1$ és $y \in X_2$ összeköthető úttal. Legyen $(x_1 = x, x_2, \ldots, x_k = y)$ x és y közötti legrövidebb út, $x_i \in X_1 \cup X_2$ $(i = 1, 2, \ldots, k)$. Ekkor van $x_{j-1} \in X_1$ $x_j \in X_2$ pontok között él valamely j indexre. A konstrukció miatt viszont ekkor

$$D \leftarrow D \cup \{x_j\}, \quad \text{illetve} \quad D \leftarrow D \cup \{x_{j-1}\}$$

lépések egyike kerül végrehajtásra. Következésképpen, a fenti legrövidebb útban szükségszerűen van D-beli pont. Másszóval, egyetlen j indexre sem teljesülhet az, hogy $x_{j-1} \in X_1$, $x_j \in X_2$ élt alkot $G(X_1) \cup G(X_2)$-ben. Vagyis D az eredeti gráf x és y pontokat szétválasztó tagozódási halmaza.

A konstrukcióból triviálisan következik, hogy D a gráf x és y pontokat elválasztó minimális tagozódási halmaza. ∎

Megjegyzések

5.2.1. A fenti eljárás az x és y pontokat elválasztó D minimális tagozódási halmazt eredményezi. Megjegyezzük, hogy D nem szükségszerűen minimális tagozódási halmaza a gráfnak; létezhet ugyanis olyan $z \in D$ pont, hogy $\{z\}$ a gráf tago-

zódási halmaza, de $\{z\}$ nem választja el az x és y pontokat. A fenti konstrukció nem feltétlenül érinti a gráf minden pontját, azaz előfordulhat $X_3 = X \setminus (X_1 \cup D \cup X_2) \neq \{\emptyset\}$ eset is. Az esetek döntő többségében azonban $X_3 = \{\emptyset\}$.

Ha $|X_3| \neq 0$, akkor

$$N(X_3) \cap D \neq \{0\}, \quad \text{míg} \quad N(X_3) \cap X_1 = \{\emptyset\} \quad \text{és} \quad N(X_3) \cap X_2 = \{\emptyset\}$$

teljesülnek.

5.2.2. $G(X_3)$ nem szükségszerűen összefüggő, míg $G(X_1)$ és $G(X_2)$ összefüggő részgráfok.

X_3 pontjai X_1, illetve X_2 halmazokba sorolhatók a következőképpen.

Jelölje Z_i az X_3 i-edik összefüggő komponensében a pontok halmazát.

Ha van y, $z \in Z_i \cap N(D)$ pontpár, melyre $(y, z) \notin E$, akkor a konstrukciós eljárással Z_i pontjai X_1, illetve X_2 halmazokba sorolhatók.

Ha minden y, $z \in Z_i \cap N(D)$ párra $(y, z) \in E$ tejesül, akkor Z_i-t X_1 vagy X_2 halmazok egyikéhez soroljuk.

Így $X = X_1 \cup D \cup X_2$ felbontás adódik, ahol D az x és y pontokat elválasztó tagozódási halmaz, mely azonban már nem feltétlen minimális.

5.2.3. A fenti konstrukcióval előállított D halmaz függ a kezdő pontpártól, s azok sorrendjétől. Adott, rendezett pontpár esetén viszont D egyértelműen meghatározott.

5.2.4. A fenti konstrukciós eljárás az 5.2.2-beli kiegészítéssel együtt módszert ad tetszőleges, nem teljes, összefüggő gráf olyan D tagozódási halmazának előállítására, mely a gráf pontjainak $X_1 \cup D \cup X_2$ felbontását eredményezi úgy, hogy $x \in X_1$, $y \in X_2$. E módszerre a továbbiakban CS eljárásként* hivatkozunk.

5.2.2. TÉTEL. Bármely nem teljes $G = (X, E)$ gráf minden $x \in X$ pontjához, melyre

$$G(N(x)) \tag{5.2.1}$$

nem teljes részgráf, megadható a gráf olyan tagozódási halmaza, mely tartalmazza x-et.

Bizonyítás. Mivel (5.2.1) nem teljes gráf, így van olyan y, $z \in N(x)$, hogy $(y, z) \notin E$. Ekkor y, z párból a CS eljárással olyan D tagozódási halmaz áll elő, melyre $x \in D$ szükségszerűen teljesül. ∎

Megjegyzések

5.2.5. A gyakorlati feladatok jelentős részében $G(N(x))$ nem teljességére vonatkozó feltétel nem jelent megszorítást, hiszen vizsgálatainkban a ritka mátrixokat leíró ún. „él-ritka" gráfok alkotják.

A továbbiakban azt vizsgáljuk, hogyan generálható a gráf tetszőleges tagozódási halmazából általános szintstruktúra.

* CS az angol "*cut set*" rövidítésből származik.

5.2.3. TÉTEL. Legyen a $G=(X, E)$ nem teljes gráfban D tetszőleges tagozódási halmaz. Ha van olyan $x\in X$, melyre $d(x, D)>1$, akkor megadható a gráfnak olyan nem-triviális (azaz $K\geqq 3$) általános szintstruktúrája, melynek D közbülső szintje.

Bizonyítás. Legyen a D tagozódási halmaz által generált felbontás $X=X_1\cup D\cup X_2$. Ez az általánosság megszorítása nélkül feltehető, ugyanis ha D kettőnél több diszjunkt részgráfra bontja G-t, akkor 5.2.2. megjegyzés értelmében elérhető a fenti felbontás.

Képezzük rendre $D\cup X_1$, illetve $D\cup X_2$ halmazokon a megfelelő D-gyökerű szintstruktúrákat, melyek

$$\mathrm{RLS}_1(D) = \{L_0(D),\ L_1(D),\ \ldots,\ L_k(D)\} \quad D\cup X_1\text{-}n,$$

$$\mathrm{RLS}_2(D) = \{J_0(D),\ J_1(D),\ \ldots,\ J_s(D)\} \quad D\cup X_2\text{-}n$$

alakban írhatók fel. Könnyű belátni, hogy

$$\{J_s(D),\ \ldots,\ J_0(D) = D = L_0(D),\quad L_1(D),\ \ldots,\ L_k(D)\}$$

halmazok együttese kielégíti GLS tulajdonságait. A $d(x, D)>1$ feltételből viszont $s+k+1\geqq 3$ teljesül. ∎

Megjegyzések

5.2.6. A bizonyításban előállított GLS a D által egyértelműen meghatározott. E konstrukcióra LS eljárásként* hivatkozunk.

5.2.7. Legyen $x,\ y\in X$ tetszőleges perifériális pár. Az x, y pontpárból a CS eljárással olyan D minimális tagozódási halmaz nyerhető, hogy $M(y, x)\subseteqq D$ és a vizsgált esetek jelentős részében

$$|D| < |L_j(x)|$$

$$|D| < |L_j(y)|$$

tapasztalható, ahol $j=\left\lfloor\dfrac{\mathrm{diam}\,(G)}{2}\right\rfloor$.

5.3. A maximális fokszámú pont torzító hatásának ellensúlyozása

Legyen $x\in X$ tetszőleges pont, melyből képezzük RLS (x)-et. Legyen $L_j(x)$ olyan közbülső szint, amely tartalmazza az $y\in X$ maximális fokszámú pontot. A tapasztalatok szerint, ha

$$|N(y)\cap L_{j-1}(x)| \ll |N(y)\cap L_{j+1}(x)|,$$

akkor gyakran

$$|L_{j+1}(x)| \gg |L_j(x)| \sim |L_{j-1}(x)|$$

* LS az angol "*level structure*" rövidítéséből származik.

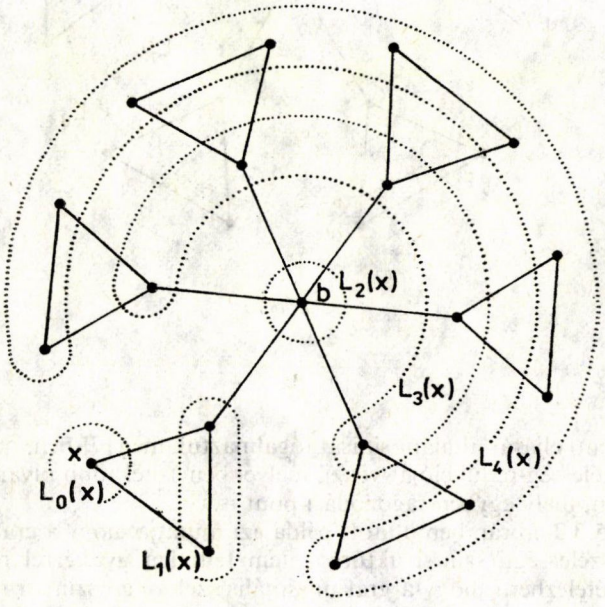
is érvényes. Másszóval, maximális fokszámú pont hatására a gyökérrel rendelkező szintstruktúra szélessége erősen megnőhet, mint azt az 5.3.1. ábrán szemléltetjük.
b maximális fokszámú pont

$$|L_1(x)| = 2$$

$$|L_2(x)| = 1$$

$$|L_3(x)| = 5$$

$$|L_4(x)| = 10$$



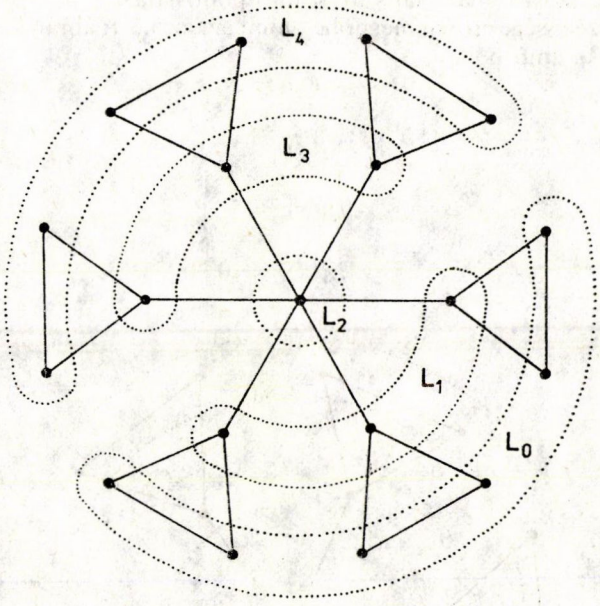
5.3.1. ábra

E probléma megoldását a következőképpen közelítettük [4]-ben.

Legyen $y \in X$ tetszőleges maximális fokszámú pont. Ekkor $G(N(y))$ nem teljessége esetén (s ez az 5.2.5. megjegyzés szerint feltételezhető) az 5.2.2. tétel felhasználásával CS eljárás révén előállítható a gráf olyan D tagozódási halmaza, mely tartalmazza y-t. Az LS eljárással D-ből egyértelműen előáll azon GLS, melynek szélessége az esetek jelentős részében kisebb, mint GLS szélső szintjeinek tetszőleges pontjából generált gyökérrel rendelkező szintstruktúráé.

Az 5.3.1. ábrán bemutatott gráf esetén a fenti eljárás által generált GLS-t az 5.3.2. ábrán szemléltetjük.

$$W(\mathrm{GLS}) = 6$$



5.3.2. ábra

Megjegyzések

5.3.1. A fenti eljárás általánosítását fogalmaztuk meg [7]-ben, mely olyan gráfok hatékony sávszélesség-redukcióját végzi, melyekben 1-nél több olyan maximális fokszámú pont van, mely egyben tagozódási pont is.

5.3.2. Az 5.3.2. ábrán bemutatott példa azt mutatja, hogy a gráf minimális vagy ahhoz közeli szélességű szintstruktúrája nem feltétlen gyökérrel rendelkező szintstruktúra. Feltételezhető, hogy a gráf minimális szélességű szintstruktúrájának meghatározása önmagában is NP-teljes probléma.

5.4. Eljárások közel-minimális szélességű szintstruktúra meghatározására

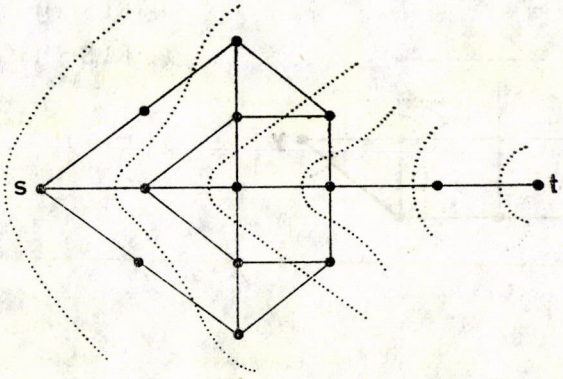
Amennyiben a minimálishoz közeli szélességű szintstruktúrát csupán a gyökérrel rendelkező szintstruktúrák körében keressük, akkor gyakorlati tapasztalatok alapján jó stratégiának bizonyul a perifériális pont gyökerű RLS (x) meghatározása. Előfordulhat azonban, hogy rövidebb RLS (x) kisebb szélességű, mint a hosszabb. Ilyen esetet szemléltetünk az 5.4.1. ábrán és 5.4.2. ábrákon.

KÖVETKEZMÉNYEK

5.4.1. A perifériális pont meghatározása során vizsgáljuk meg valamennyi előállított RLS (x) szélességét, s azt választjuk eredményül, melynek szélessége minimális.

s, t perifériális pár

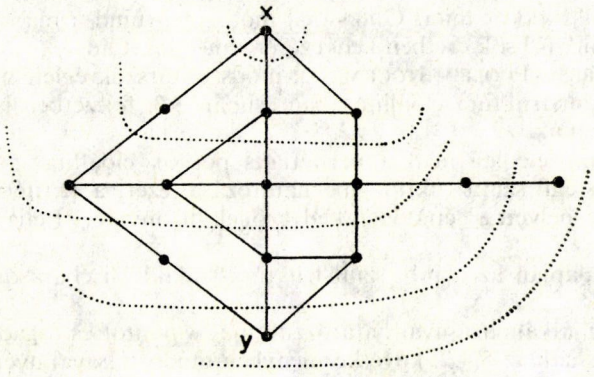
$$\operatorname{diam}(G) = 5$$
$$W(\mathrm{RLS}(t)) = 5$$
$$|L_4(t)| = 5$$



5.4.1. ábra

x, y szemi-pszeudo-perifériális pár;

$$l(x) = 4$$
$$W(\mathrm{RLS}(x)) = 4$$



5.4.2. ábra

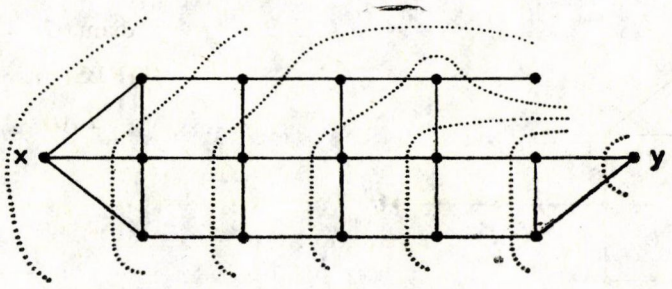
Megjegyezzük, hogy a számítógépes kidolgozásban regisztráljuk az aktuálisan minimálisnak minősülő szélességet (W) és tároljuk a megfelelő RLS (x)-et. Másrészt, ha az RLS-generálások során egy szint szélesebb, mint W, akkor nem képezzük a további szinteket, hanem rátérünk a következő RLS-generálásra.

Az eredmény javíthatósága szempontjából fontos az alábbi állítás:

Legyen $x, y \in X$ tetszőleges szemi-pszeudo-perifériális pár. Ekkor $W(\mathrm{RLS}(x))$ nem szükségszerűen egyenlő $W(\mathrm{RLS}(y))$ értékkel.

Az 5.4.3. ábrán feltüntetett gráf szemléletesen igazolja az állítást.

x, y szemi-pszeudo-perifériális pár egyben perifériális pár



$W(\mathrm{RLS}(x)) = 3$
$W(\mathrm{RLS}(y)) = 4$
$|L_4(y)| = 4$

5.4.3. ábra

KÖVETKEZMÉNYEK

5.4.2. Ha előállítottuk az s pontot melyre $W(\mathrm{RLS}(s))$ az eljárás során minimális a vizsgált szintstruktúrák vonatkozásában, akkor minden $t \in L_{ec}(s)$ pontból generáljuk RLS (t)-t, s azt fogadjuk el eredményül, melynek szélessége így adódik minimálisnak.

Megjegyezzük, hogy 5.4.1. következményben a nem befejezett szintstruktúrák hanyagolást is jelenthetnek, mivel előfordulhat, hogy egy nem befejezett RLS (x) valamely excentricitási pontjából kisebb szélességű szintstruktúra adódna.

Ha a közel-minimális szélességű szintstruktúrát az általános szintstruktúrák körében keressük, akkor mind GIBBS [55] módszere, mind a maximális fokszámon alapuló eljárásunk [4] sok esetben kedvező eredményeket ad.

Mivel a GIBBS—POOLE—STOCKMEYER-módszer társszükséglete magas (8. fejezet), így általános szintstruktúra előállításában csupán 5.3. fejezetben ismertetett eljárásunkra szorítkozunk.

Mivel célunk jelenleg nem a perifériális pontok előállítása, hanem a közelminimális szélességű szintstruktúra meghatározása, ezért a perifériális pontot előállító eljárásaink helyett a gépidőben kedvezőnek bizonyuló P1 eljárást fogjuk alkalmazni.

A fentiek alapján az alábbi szintstruktúra-kialakítási eljárások fogalmazhatók meg:

I. P1 módszer alkalmazásával határozzuk meg s pontot és fogadjuk el RLS (s)-t

II. A P1 eljárásnak az 5.4.2. következménybeli módosításával nyert szintstruktúrát fogadjuk el.

III. A II. eljárással határozzuk meg RLS (u)-t; az u kezdőpontból alkalmazzuk újra a II. eljárást. (Megjegyezzük ugyanis, hogy P1 eljárás ilyen ismételt alkalmazása — a gráf szerkezetétől függően — sok esetben újabb perifériális pontokat eredményezett.)

IV. A II. eljárással állítsuk elő RLS (x)-et.
Tetszőleges $y \in L_{ec}(x)$ pontot tekintve, x, y kezdőpárból állítsuk elő a CS eljárással a gráf azon D tagozódási halmazát, amely x és y-t elválasztja. Majd D-ből az LS eljárással generáljuk GLS-t.

V. Alkalmazzuk a IV. eljárást úgy, hogy a kezdeti RLS (x)-et III. eljárással állítjuk elő.

VI. A II. algoritmussal állítsuk elő RLS(x)-et. Legyen $L_j(x) \subset \mathrm{RLS}(x)$ olyan, melyre

$$|L_j(x)| = W(\mathrm{RLS}(x))$$

teljesül.

Határozzuk meg $y \in L_{j-1}(x)$ pontot, melyre

$$\deg(y) = \max_{p \in L_{j-1}(x)} \deg(p)$$

teljesül.

Határozzuk meg $u \in N(y) \cap L_{j-2}(x)$ és $v \in N(y) \cap L_j(x)$ pontpárt, melyből CS eljárással állítsuk elő e két pontot elválasztó D tagozódási halmazt. D-ből az LS eljárással generáljuk GLS-t.

(Megjegyezzük, hogy u és v fenti választásával azt szándékozzuk elérni, hogy az általuk meghatározott, D-ből származtatott GLS lehetőleg $R(x, y)$ „irányához" illeszkedjék.)

VII. Alkalmazzuk a VI. algoritmust, de a kezdeti RLS(x)-t a III. eljárással állítjuk elő.

Megjegyzések

5.4.1. Az I., II. és III. eljárások mindegyike gyökérrel rendelkező szintstruktúrát, míg IV., V., VI. és VII. általános szintstruktúrát állítanak elő.

5.4.2. Eljárásaink műveletigényére becslést nem sikerült adni.

5.4.3. Vonatkozó számítógépes eredményeinket a 7. sz. mellékletben foglaljuk össze, ahol a bevezetett hét eljárásunkon kívül közöljük a következő eljárások eredményeit is:

— GPS—PS
— GPS—LS, mely az általános szintstruktúra kialakítását végzi [55].
— GL—SPS
— NCC—LS, melyben saját fejlesztési heurisztikus eljárásunkkal (lásd a 8. sz. melléklet) meghatározott pontpárból CS és LS eljárásokkal GLS-t állítunk elő.
— P2″ (lásd a 4.4. fejezetet).

Az egyes eljárások mellett feltüntetjük az eredményezett szintstruktúra-szélességet (W) és a szükséges műveletek számát, szintstruktúrában mérve (Op).

Teszt-feladatainkat véletlen gráfok alkotják. Hat gráfot vizsgálunk, melyekből az első három csak sűrűségbeli növekedést demonstrál.

Az eredmények alapján megállapítható:

a) Az I. eljárás legtöbb esetben jobb eredményt ad, mint a GPS-eljárások, GL—SPS és NCC—LS, s műveletigénye kedvező.

b) A II. algoritmus az esetek döntő többségében jobb eredményeket szolgáltat, mint I. Műveletigénye is kedvezőnek mondható a befejezett szintstruktúrák kis száma miatt.

c) A III. eljárás a gráf szerkezeti adottságainak függvényében tud javítani II. eredményén. ($G=(300, 900)$, $G=(300, 1500)$). Műveletszáma természetszerűen nő II. eljáráshoz képest.

d) A IV. és V. eljárások eredménye erősen függ a gráf szerkezeti adottságaitól. $G=(400, 200)$ esetben jelentős javulást ad az I., II., III. eredményeihez képest, míg

$G=(600, 4800)$ esetekben nem tud javítani. A 300 pontú gráfokban rosszabb az eredménye mint az I., II. és III. eljárásoknak.

e) A közölt példákban V. csupán egyszer adott jobb eredményt, mint IV.

f) A VI. és VII. eljárások is bizonyos esetekken ($G=(300, 1500)$, $G=(400, 2000)$) jobb, míg $G=(300, 1200)$ esetén rosszabb eredményeket adnak, mint I., II., vagy III. A VII. eljárás csupán egyszer javít VI. eredményén.

g) A IV.—VII. eljárások műveletigénye természetszerűen nagyobb, mint az I.—III. algoritmusoké.

h) Műveletszám terén a GL—SPS a leghatékonyabb eljárás, de eredményében általában gyengébb, mint eljárásaink.

i) A GPS-eljárások is sok esetben gyengébb eredményeket adnak, mint algoritmusaink; gépidő-szükségletük erősen függ a gráf szerkezeti adottságaitól.

j) Az NCC—LS gyakran javít GL—SPS eredményén, s műveletigénye kedvezően alacsony.

k) A P″ sok esetben jobb eredményeket ad, mint GL—SPS vagy GPS-eljárások. Eredmény és műveletszám terén egyaránt az I. eljárással azonos hatékonyságúnak értékelhető.

6. A szintstruktúra számozása

Jelen fejezetben megmutatjuk, hogy a gráfban a minimális sávszélességét biztosító számozás közelítéseként reális célul csupán adott, minimálishoz közeli szélességű szintstruktúra közel-optimális számozásának előállítása tűzhető ki.

Ismertetjük [4]-beli számozási eljárásunkat.

Rámutatunk a [81], [9]-beli új számozási eljárásunk elveinek helyességére, s ismertetjük algoritmusunk lényegét.

6.1. A probléma közelítésének történeti áttekintése

1969-ben E. H. CUTHILL és J. MCKEE eljárásában [27] nyer megfogalmazást az első sávszélesség-redukciót eredményező számozás, mely az alábbiak szerint működik.

Legyen $x \in X$ tetszőleges, minimális fokszámú pont, melyhez rendeljük hozzá az 1-es csúcsszámot, azaz $n(x)=1$.

$i=1, 2, \ldots, l(x)$ esetén végezzük a következőket:

Tekintsük az x-től i távolságban levő pontok halmazát, mely jelen terminológiában $L_i(x)$.

Tekintsük az $L_{i-1}(x)$ pontjait a már hozzárendelt csúcsszámaik monoton növekvő sorrendjében, azaz

$$L_{i-1}(x) = \{z_1, z_2, \ldots, z_s\}, \quad \text{ahol} \quad s = |L_{i-1}(x)|,$$

és

$$n(z_j) < n(z_{j+1}), \quad j = 1, 2, \ldots, s-1.$$

Ekkor rendre

$$N(z_k) \cap L_i(x) \quad (k = 1, 2, \ldots, s)$$

halmaz pontjait fokszámaik növekvő sorrendjében látja el az eljárás a soronkövetkező csúcsszámokkal.

1972-ben a [4]-beli eljárásunkban általános szintstruktúrát generáltuk az 5.2. fejezetben tárgyalt módon, s ennek számozását a *Cuthill—McKee-féle számozás* ésszerű módosításával végeztük, melyet a 6.3. fejezetben ismertetünk.

1976-ban J. W-H. LIU és A. H. SHERMANN bebizonyították [66], hogy a *Cuthill—McKee-féle számozás* megfordításával, mely az $(1, |X|)$, $(2, |X|-1)$, $(3, |X|-2)$... csúcsszám-párokban a csúcsszámok cseréjét jelenti, a profil-érték nem nőhet; sok esetben viszont kedvező csökkenést eredményez a számozás megfordítása. (Megjegyezzük, hogy tapasztalati megfigyelései alapján CUTHILL [28] munkájában ezen összefüggés felismerésére utal.)

1976-ban N. E. GIBBS, W. G. POOLE és P. K. STOCKMEYER [55] általános szintstruktúrát állítottak elő, melynek számozását a *Cuthill—McKee-féle számozás* alábbi módosításával állították elő:

Legyen $\mathrm{GLS}=\{L_0, L_1, \ldots, L_k\}$.

L_0 számozása a pontok fokszámainak növekvő sorrendjében történik.

Az i-edik szint számozásakor $(1 \leqq i \leqq k)$ az $L_0, L_1, \ldots, L_{i-1}$ szinteket beszámozottaknak feltételezve először

$$H_i = N(L_{i-1}) \cap L_i \subset L_i$$

részhalmaz pontjait számozzák be a *Cuthill—McKee eljárás* alkalmazásával. A visszamaradó $L_i \setminus H_i$ halmaz pontjainak számozása a H_i-vel való élkapcsolataik figyelembevételével, [55]-ben részletezett bonyolult módon történik.

1976-ban W. F. SMYTH ILO szakértővel közösen új számozási stratégia elveit közöltük [81], amely alapján 1978-ra elkészült a módszer, s annak számítógépes kidolgozása [9]. Az eljárással 6.4. fejezetben foglalkozunk.

1981-ben J. A. GEORGE és J. W-H. LIU [55]-beli sávszélesség/profil-redukciós eljárásukban gyökérrel rendelkező szintstruktúrát képeztek, melynek számozását a *Cuthill—McKee-féle számozás* megfordításával állították elő.

6.2. A szintstruktúra számozásának problémái

Az 1.4. fejezetben rámutattunk, hogy bármely GLS szélessége meghatározza a tetszőleges kompatibilis számozásával nyerhető sávszélesség pontos alsó és felső korlátját, és

$$W(\mathrm{GLS}) \leqq b \leqq 2 \cdot W(\mathrm{GLS}) - 1 \tag{6.2.1}$$

érvényes.

Legyenek $x, y \in X$ olyan pontok, hogy

$$W(\mathrm{RLS}(x)) = W(\mathrm{RLS}(y))$$

teljesül. Ekkor a két szintstruktúrával kompatibilis, sávszélesség szerinti optimális számozás eltérő sávszélességet eredményezhet. A 6.2.1. ábrán szemléltetjük a fenti esetet.

Első esetben

$$W(\mathrm{RLS}(x)) = 5$$

$$b = 6$$

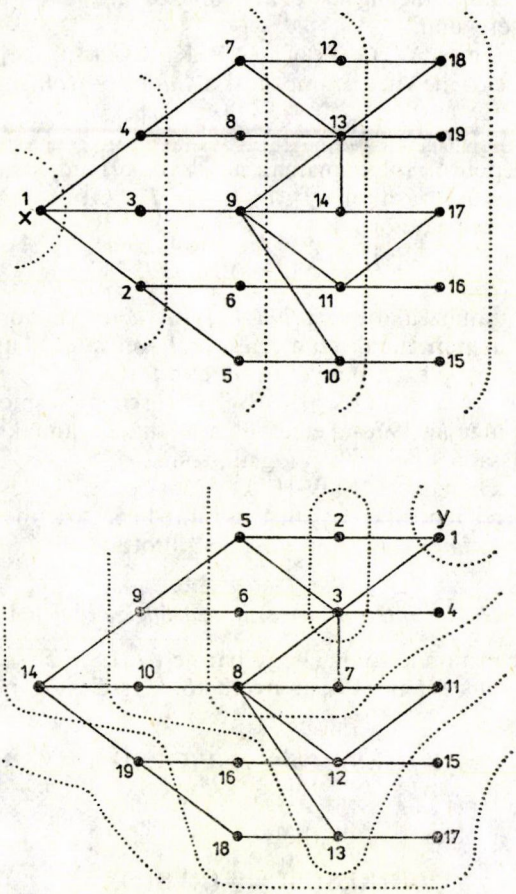
Sávszélességet kifeszítő élek: (7, 13) (11, 17)

Második esetben

$$W(\mathrm{RLS}(y)) = 5$$

$$b = 5$$

Sávszélességet kifeszítő élek: (3,8), (8,13), (9,14), (14,19), (13,18)



6.2.1. ábra

KÖVETKEZMÉNYEK

6.2.1. Tetszőleges minimális szélességű szintstruktúra sávszélesség szerinti optimális számozása nem szükségszerűen eredményezi a gráf minimális sávszélességét.

6.2.2. A gráf minimális sávszélességének közelítéseként reális célul csupán adott, minimálishoz közeli szélességű szintstruktúra sávszélesség szerinti közel-optimális számozásának meghatározása tűzhető ki.

A Cuthill—McKee-féle számozás tulajdonságai

— A kifeszítő fa kezdőponttól azonos távolságban levő pontjai a kezdőpont gyökerű szintstruktúra megfelelő szintjeit alkotják, így az eljárás e szintstruktúrával kompatibilis számozást épít fel.

— Az eljárás csupán gyökérrel rendelkező szintstruktúra számozására alkalmas.

— A *Cuthill—McKee számozás* az $L_i(x)$ szint ($i>0$) számozását csupán $L_{i-1}(x)$-beli élkapcsolatain és már felépített számozáson alapulva építi fel.

6.3. Speciális szerkezetű általános szintstruktúra számozása

A CS és LS eljárásokkal nyert GLS speciális szerkezetű, mivel az két halmazgyökerű szintstruktúra megfelelő illesztéséből áll elő. (5.2. fejezet.) E speciális adottságot felhasználjuk GLS számozásakor.

Legyen D a CS eljárással nyert tagozódási halmaz, s ekkor $X=X_1 \cup D \cup X_2$ felbontás előállítható.

Képezzük $D=D1 \cup D2$ diszjunkt felbontását úgy, hogy $D1$-be azon D-beli pontok kerüljenek, melyek több éllel kapcsolódnak X_1-hez, mint X_2 halmazhoz. Hasonlóan, a $D2$ halmazt az X_2-höz nagyobb számú éllel kötődő D-beli pontokból állítjuk elő.

Azon D-beli pontokat, melyek kapcsolódása egyenlő az X_1 és X_2 halmazokkal, ideiglenesen gyűjtjük T tömbben.

D valamennyi pontját megvizsgálva, ha $|T|=0$, akkor $D=D1 \cup D2$ felbontás előállt.

Ha $|T|\neq 0$, akkor elemeit egyenként a kisebb elemszámú halmazhoz soroljuk, hogy $D1$ és $D2$ elemszáma közeledjen egymáshoz.

Ezáltal $X=X_1 \cup D_1 \cup D_2 \cup X_2$ felbontást nyerjük. Képezzük rendre

RLS (D_1) szintstruktúrát a $D1 \cup X_1$
RLS (D_2) szintstruktúrát a $D2 \cup X_2$ halmazokon.

Lássuk el a két szintstruktúrát a *Cuthill—McKee számozás* alábbi módosításával:

— Rendezzük $D1$ elemeit X_1-beli élkapcsolódásaik monoton növekvő sorrendjébe, s ezen sorrendben lássuk el az 1, 2, ..., $|D1|$ csúcsszámokkal.

— RLS ($D1$) többi szintjeit lássuk el a *Cuthill—McKee számozással.* (Az utolsó ponthoz rendelt csúcsszám ($|D1|+|X_1|$).

— Fordítsuk meg az RLS ($D1$)-en generált számozást (6.1. fejezet).

— Rendezzük $D2$ elemeit X_2-beli élkapcsolódásaik szerinti monoton növekvő sorrendbe, s ezen sorrendben lássuk el a pontokat rendre $|D1|+|X_1|+1$, $|D1|+|X_1|+2$, ..., $|D1|+|X_1|+|D2|$ csúcsszámokkal.

— RLS ($D2$) többi szintjeit lássuk el a *Cuthill—McKee számozással.* (Az utolsó ponthoz rendelt csúcsszám $|X|$.)

Megjegyzések

6.3.1. Az eljárás, melyre N1-ként hivatkozunk, csupán ilyen speciális szerkezetű GLS-ek számozására alkalmas.

6.3.2. A vonatkozó számítógépes eredményeket a sávszélesség-redukciós eljárásaikon belüli alkalmazásban közöljük a 9—17. sz. mellékletekben.

6.4. Új számozási eljárásunk

Adott szintstruktúra sávszélesség-redukciót eredményező számozásának felépítésekor minden közbülső szint számozásánál az őt követő, számozatlan szinttel való élkapcsolatai is fontos szerepet töltenek be.

6.4.1. TÉTEL. Legyen GLS a $G=(X, E)$ tetszőleges általános szinstruktúrája. Legyen n tetszőleges, GLS-val kompatibilis számozás.

Az n számozás akkor és csak akkor eredményez b sávszélességet, ha minden $y\in L_i$ $(0<i<k)$ pontban

$$n(u)-b \leqq n(y) \leqq n(v)+b \tag{6.4.1}$$

teljesül, ahol

$$n(u)=\begin{cases}\max\limits_{q\in L_{i+1}\cap N(y)} n(q) & \text{ha} \quad L_{i+1}\cap N(y)\neq\{\emptyset\};\\ W_i & \text{ha} \quad L_{i+1}\cap N(y)=\{\emptyset\};\end{cases}$$

$$n(v)=\begin{cases}\min\limits_{p\in L_{i-1}\cap N(y)} n(p) & \text{ha} \quad L_{i-1}\cap N(y)\neq\{\emptyset\};\\ W_{i-1} & \text{ha} \quad L_{i-1}\cap N(y)=\{\emptyset\}.\end{cases}$$

Bizonyítás. Könnyű belátni, hogy az n kompatibilis számozásban bármely x, $y\in L_j$ $(0\leqq j\leqq k)$ pontokra

$$|n(x)-n(y)|<b \tag{6.4.2}$$

teljesül.

Megmutatjuk, hogy (6.4.1) teljesüléséből következik, hogy a sávszélesség b.

Legyen $u\in N(y)$, melyre

$$n(u)=\max_{q\in N(y)} n(q). \tag{6.4.3}$$

Ha $u\in L_{i+1}$, akkor (6.4.1)-ből $n(u)-b\leqq n(y)$ következik, ami

$$n(u)-n(y)\leqq b$$

teljesülését eredményezi. Ha $u\in L_i$, akkor (6.4.2) szerint $n(u)-n(y)<b$ adódik.

Most legyen $v\in N(y)$, melyre

$$n(v)=\min_{p\in N(y)} n(p). \tag{6.4.4}$$

Ha $v\in L_{i-1}$, akkor (6.4.1)-ből $n(y)-n(v)\leqq b$ adódik, melyből $v\in L_i$ esetén $n(y)-n(v)<b$ összefüggést nyerjük. Mindez azt igazolja, hogy az n számozással nyert sávszélesség b.

Most azt látjuk be, hogy ha az n kompatibilis számozással adódó sávszélesség b, akkor (6.4.1) szükségszerűen következik.

Tetszőleges $y\in L_i$ $(0<i<k)$ esetén legyen $u\in N(y)$, melyre (6.4.3) teljesül. Ha $u\in L_{i+1}$, akkor $n(u)-n(y)\leqq b$ érvényességéből $n(u)-b\leqq n(y)$ következik. $u\in L_i$ esetén viszont $n(u)-b<n(y)$ adódik. Mindez (6.4.1) első relációjának teljesülését jelenti.

Legyen most $v \in N(y)$, mely kielégíti (6.4.4)-et. Ekkor $n(y)-n(v) \leqq b$ teljesüléséből $n(y) \leqq b+n(v)$ adódik. E reláció $v \in L_i$ esetén szigorú egyenlőtlenségbe megy át. Vagyis (6.4.1) második relációja is teljesül. ∎

Megjegyezzük, hogy n kompatibilitása miatt (6.4.1) a következő alakban írható fel:

$$W_i+s-b \leqq n(y) \leqq n(v)+b, \tag{6.4.5}$$

ahol $0 \leqq s \leqq |L_{i+1}|$, mivel $n(u)=W_i+s$ érvényes.

KÖVETKEZMÉNYEK

6.4.1. Adott GLS-sel kompatibilis tetszőleges n számozás a közbülső szintek minden y pontjához hozzárendeli a (6.4.5)-ben definiált alsó, illetve felső korlátot, melyen belüli csúcsszámok bármelyikét rendelve az y-hoz, L_{i-1}, L_i és L_{i+1} vonatkozásában nem nő a sávszélesség.

6.4.2. (6.4.5) adott számozás esetén definiálja az alsó felső korlátokat. Ha a számozást úgy építjük fel, hogy minden közbülső szint minden pontjában (6.4.5)-típusú reláció figyelembevételével számozzuk be az illető pontot, akkor feltételezhető, hogy kedvezően kis sávszélesség adódik.

Megjegyzések

6.4.1. A *Cuthill—McKee számozás* az $L_i(x)$ közbülső szint számozását a már beszámozott $L_{i-1}(x)$ szinttel való élkapcsolatok és már meglevő számozás figyelembevételével végzi, s ekkor „ösztönösen" (6.4.5) jobb oldali relációját igyekszik kielégíteni. A bal oldali reláció figyelembevétele egyelőre nem is lehetséges, mivel $L_{i+1}(x)$ pontjai még számozatlanok.

Jelölje $a(y)$ az $y \in L_i$ pontnak az L_{i+1}-be nyúló éleinek számát, azaz

$$a(y) = |N(y) \cap L_{i+1}|. \tag{6.4.6}$$

Ekkor a (6.4.5)-beli s-re $0 \leqq a(y) \leqq s$ teljesül, melyet (6.4.2)-be helyettesítve, ott továbbra is marad az egyenlőtlenség. Így a (6.4.5)-ből

$$W_i+a(y)-b \leqq n(y) \leqq n(v)+b \tag{6.4.7}$$

adódik.

A (6.4.7)-beli alsó határ a GLS, az n számozás (b), illetve a gráf szerkezete által definiált; a felső határt a már beszámozott pontokkal való kapcsolat írja le, mely szintén ismert.

6.4.2. Ha (6.4.7)-ben b az i-edik szint számozásának felépítésével nyert sávszélességet jelöli, akkor az $(i+1)$-edik szint pontjainak (6.4.7) szerinti számozásával nem nő a sávszélesség.

Ezen észrevételen alapul új számozási eljárásunk [81], [9] alapelve:
Legyen $GLS=\{L_0, L_1, \ldots, L_k\}$ tetszőleges szintstruktúra. Tegyük fel, hogy az $L_0, L_1, \ldots, L_{i-1}$ szinteket már beszámoztuk valamely, GLS-sel kompatibilis számozással, s az ennek eredményeképpen előálló (közbülső) sávszélességet jelölje b.

$L_i = \{y_1, y_2, \ldots, y_s\}$ minden y_j pontjához rendeljük hozzá

$$n \min(y_j) = \max(W_{i-1}+1,\ W_i + a(y_j)),$$

$$n \max(y_j) = \min(W_i,\ n(u)+b)$$

értékpárt, ahol

$$a(y_j) = |N(y_j) \cap L_{i+1}|,$$

$$n(u) = \begin{cases} \min\limits_{p \in L_{i-1} \cap N(y_j)} n(p) & \text{ha} \quad N(y_j) \cap L_{i-1} \neq \{\emptyset\}, \\ W_{i-1} & \text{ha} \quad N(y_j) \cap L_{i-1} = \{\emptyset\}. \end{cases}$$

Most (6.4.7) szerint bármely

$$n \min(y_j) \leqq m \leqq n \max(y_j)$$

feltételt kielégítő m csúcsszámot hozzárendelve y_j-hez, nem változik (nő) a sávszélesség. A b növekedése nélkül y_j-hez rendelhető csúcsszámok számát az y_j *számozási szabadsági fokának* nevezzük.

Rendezzük L_i pontjait

első rendben n min-értékeik

másod rendben számozási szabadsági fokuk

növekvő sorrendjében. Ezáltal elérjük, hogy a csúcsok a hozzárendelhető minimális csúcsszámok növekvő sorrendjében vannak úgy, hogy az azonos n min-értékű pontok közti sorrendiséget számozási szabadsági fokuk növekvő sorrendje szabja meg.

L_i pontjait az így előálló sorrendben lássuk el a soron következő csúcsszámokkal úgy, hogy egy csúcsszám hozzárendelésekor a vele azonos n min-értékű pontokban a hozzárendelhető értékek alsó határát 1-gyel növeljük, míg ugyanezen pontok számozási szabadsági foka 1-gyel csökken.

Megjegyezzük, ha valamely $y_r \in L_i$ pontban

$$n \min(y_r) > n \max(y_r)$$

teljesül, akkor a pont nem számozható be b növekedése nélkül. Ekkor $b \leftarrow b+1$ értékkel újra kell kezdenünk az aktuális szint számozását.

Eljárásunk algoritmikus leírását [9]-ben közöltük.

Megjegyzések

6.4.3. Számozási eljárásunk, melyre NN-ként hivatkozunk, mind általános, mind gyökérrel rendelkező szintstruktúrák sávszélesség/profil-redukciót eredményező számozását állítják elő.

6.4.4. Az eljárás szinte pontonként lokális optimumot keres, így sok esetben kedvező csökkenést tud előidézni az eredményként előálló sávszélességben.

6.4.5. Az eljárás rendkívül műveletigényes. Művelet számára becslést nem sikerült adni, de a számítógépes eredményeik gépidő-paramétere (9—17 sz. mellékletek) oly magas, hogy nagyméretű feladatoknál a gyakorlatban nem alkalmazható.

6.4.6. Megjegyezzük, hogy a számítógépes kidolgozásban a rendezéseket D. E. KNUTH [60] „radix-sorting" eljárásával végeztük, s ez csekély javulást eredményezett a gépidőkben, de a közölt eredmények e javított állapotot tükrözik.

6.4.7. LIU—SHERMAN eredménye [66], mely szerint a *Cuthill—McKee-féle számozás* (CN) megfordításával (RCN) a profil-érték nem nőhet, érvényes minden olyan

számozásra, melyhez tartozó összefüggési mátrix ún. *„monoton profil tulajdonságú"*, azaz

$$\text{minden } i \leqq j \text{ indexre } f_i(A) \leqq f_j(A)$$

teljesül.

NN eljárásunk nem eredményez monoton profil tulajdonságú összefüggési mátrixot, így megfordításával nyert RNN eljárással adódó profil-érték nem becsülhető. Ennek ellenére, a vizsgált esetek jelentős részében RNN kisebb profil-értéket szolgáltat, mint NN.

6.4.8. Tetszőleges $x \in X$ esetén tekintsük

$$\text{RLS}(x) = \{L_0(x), L_1(x), \ldots, L_{l(x)}(x)\}$$

szintstruktúrát. A

$$\text{GLS} = \{L_{l(x)}(x), \ldots, L_1(x), L_0(x)\} \tag{6.4.8}$$

szintstruktúrát az RLS (x) *megfordításának* nevezzük.

A vizsgált esetek mindegyikében a (6.4.8)-beli GLS NN számozása kisebb profil-értéket eredményezett, mint RNN. Az általa nyert eredmények számos esetben kiugróan nagyok az RLS (x) RNN számozásával adódó eredményekhez képest. Bizonyos esetekben viszont kiugróan nagy csökkenést eredményez nem csupán az RLS (x) RNN számozásához képest, hanem a közölt valamennyi sávszélesség/profil-redukciós eljárás eredményéhez viszonyítva is (9—17 sz. mellékletek).

E tapasztalati megfigyelésre eddig nem sikerült indoklást találnunk. További kutatásainkban e probléma vizsgálatát is tervezzük.

7. Új sávszélesség/profil-redukciós eljárások

A rendelkezésre álló szintstruktúra-kialakító és számozó eljárások lehetővé teszik új sávszélesség/profil-redukciós algoritmusok megfogalmazását.

Az I., II. és III. eljárások gyökérrel rendelkező szintstruktúrát eredményeznek, mely számozására RCN és RNN eljárásokkal egyaránt alkalmazhatók.

A IV., V., VI. és VII. eljárások speciális szerkezetű általános szintstruktúrát generálnak, melynek számozására RN1 és RNN számozási eljárásaink alkalmazhatók.

A tetszőleges gyökérrel rendelkező szintstruktúra megfordítása általános szintstruktúrát generál, mely NN eljárásunkkal számozható.

További szintstruktúra-kialakító eljárásként felhasználjuk a GL—SPS eljárást realizáló FNROOT rutint, mely a SPARSPAK sávszélesség/profil-redukciós GENRCM (*GEN*eral *R*everse *C*uthill *M*ethod) eljárásában a szintstruktúrát állítja elő.

A GPS—PS eljárás komponenseit nem alkalmazzuk, mert számítógépes kidolgozásuk tárigénye igen magas (8. fejezet).

Mindezek alapján az alábbi sávszélesség/profil-redukciós eljárások fogalmazhatók meg, ahol az eljárások nevei azonosak az őket realizáló rutinok neveivel. Az egyes eljárásokat úgy adjuk meg, hogy feltüntetjük egyes fázisait milyen algoritmusokkal állnak elő.

Eljárás neve	Szintstruktúra kialakítása	Szintstruktúra számozása
GENBRI	I.	RCN
PMM	I.	RNN
GENBRM—0	II.	RCN
GENBRM—1	III.	RCN
PPM—0	II.	RNN
PPM—1	III.	RNN
PPM—2	GLS*	NN
BBPRED—0	IV.	RNN
BBPRED—1	V.	RNN
BBPP—0	VI.	RNN
BBPP—1	VII.	RNN
BBPR—0	IV.	RN1
BBPR—1	V.	RN1
BBPP1—0	VI.	RN1
BBPP1—1	VII.	RN1
GGLAN	RLS**	RNN
GGLAA	GLS***	NN

* A III. eljárással képzett RLS (x) megfordítása.
** Az FNROOT rutinnal előállított RLS (x).
*** Az FNROOT rutinnal nyert RLS (x) megfordítása.

Megjegyzések

7.1. A GENBRI felel meg a SPARSPAK GENRCM eljárásának.

7.2. A IV., V., VI. és VII. eljárásokat használó algoritmusok számítógépes kidolgozását úgy végeztük, hogy ha a CS eljárásban a D tagozódási halmaz elemeinek száma meghaladja a megfelelő RLS (x) szélességét, akkor automatikusan RLS (x) minősül az előálló szintstruktúrának. Ezáltal BBPRED- és BBPP-eljárások átmehetnek PPM-eljárásokba. Hasonlóan BBPR- és BBPP1-eljárások GENBRM-ek által adott eredményt szolgáltathatják. Megjegyezzük, hogy N1 számozás csak speciális szerkezetű GLS számozására alkalmas, ezért, ha GLS helyett RLS (x) adódik, akkor értelemszerűen az RN1 számozást RCN eljárással kell cserélnünk.

Mindez azt a célt szolgálja, hogy az említett eljárások lehetőleg ne eredményezzenek a PPM-, illetve GENBRM-eljárásoknál gyengébb értékeket.

7.3. A II. és III. eljárásokat egyetlen (MMWLS) rutin realizálja, különböző paraméter-beállítással. A

(7.1)

GENBRM—0	GENBRM—1
PPM—0	PPM—1
BBPRED—0	BBPRED—1
BBPP—0	BBPP—1
BBPR—0	BBPR—1
BBPP1—0	BBPP1—1

eljárás-párok névbeli jelölése azt szimbolizálja, hogy a két hasonló nevű eljárásban RLS (x)-et ugyanazon rutin állítja elő, és az eltérő szintstruktúrákat az eltérő paraméterezéssel hívott MMWLS generálja.

7.1. TÁBLÁZAT

Sávszélesség/profil-redukciós eljárások értékelése a 9—17. sz. mellékletben csatolt számítógépes eredmények alapján

Tesztanyag :27 véletlen gráf.

	Legjobb esetek száma		*b*			*pr*		
	b	*pr*	jobb	azonos	rosszabb	jobb	azonos	rosszabb
GENRCM	2	1						
GPS	2	3	12	—	15	12	—	15
GENBRI	1	2	15	—	12	13	—	14
PMM	2	1	14	2	11	13	—	14
GENBRM—0	2	3	13	4	10	16	—	11
GENBRM—1	3	3	14	4	9	17	—	10
PPM—0	2	1	15	—	12	18	—	9
PPM—1	2	1	16	—	11	19	—	8
PPM—2	2	3	10	1	16	6	—	21
BBPRED—0	3	2	14	—	13	17	—	10
BBPRED—1	3	1	14	—	13	18	—	9
BBPP—0	2	2	14	—	13	17	—	10
BBPP—1	2	2	14	—	13	17	—	10
BBPR—0	6	7	15	2	10	20	—	7
BBPR—1	8	7	16	2	9	21	—	6
BBPP1—0	6	6	16	2	9	17	1	9
BBPP1—1	10	8	17	2	8	18	1	8
GGLAN	—	2	12	2	13	13	—	14
GGLAA	1	—	6	—	21	6	—	21

b — sávszélesség
pr — profil-érték

7.4. Mivel a gráf minimális sávszélessége és profilértéke nem ismeretes, így nem becsülhető, hogy adott eljárás eredménye mennyire közelíti az optimális megoldást.

7.5. Eljárásaink műveletigényére becslést nem sikerült adnunk. Ezért eljárásainkat a számítógépes eredmények alapján értékeljük.

Ugyanazon gráfok esetén fenti eljárásainkon túl a GEORGE—LIU (GENRCM) és GIBBS—POOLE—STOCKMEYER (GPS) *eljárásokat* is futtattuk, s ezáltal lehetővé válik az egyes algoritmusok összehasonlító értékelése.

Vizsgálataink teszt-anyagát véletlen gráfokból állítottuk össze az alábbiak szerint:

— három különböző (300, 500, 1000) pontszámú gráf-típust vizsgáltunk,
— mindhárom méret (pontszám) esetén három különböző él-sűrűséget tekintve, minden adott sűrűséggel 3-3 feladatot generáltunk.

Számítógépes futtatásainkat az alábbi 27 gráfon végeztük

300 pontú, 900 élű gráfok (9. sz. melléklet)
300 pontú, 1200 élű gráfok (10. sz. melléklet)
300 pontú, 1500 élű gráfok (11. sz. melléklet)

500 pontú, 1500 élű gráfok (12. sz. melléklet)
500 pontú, 2000 élű gráfok (13. sz. melléklet)
500 pontú, 2500 élű gráfok (14. sz. melléklet)

1000 pontú, 3000 élű gráfok (15. sz. melléklet)
1000 pontú, 4000 élű gráfok (16. sz. melléklet)
1000 pontú, 5000 élű gráfok (17. sz. melléklet)

Értékelésünkben vonatkozási alapul a GENRCM eljárás eredményeit tekintjük, mivel

— igen gyors eljárás;
— SPARSPAK részeként él.

Adott eljárás értékelését az alábbiak szerint végezzük:

— hány esetben adott jobb eredményt, mint GENRCM;
— hány esetben tudta a nyert értékek közül a legjobbat eredményezni;
— milyen a gépidő-szükséglete.

Eredmény-táblázatainkban az egyes eljárások mellett feltüntetjük a nyert sávszélességet (b), a profil-értéket (pr) s a végrehajtáshoz szükséges gépidőt (t), mely a vonatkozó CPU-időt jelenti másodpercekben mérve. Eredményeinkben a minimálisnak minősülő értékeket *-gal jelöltük.

Az első két értékelési szempont adatait a 7.1. táblázatban foglaljuk össze.

A sávszélesség redukciójának értékelése

— A GENRCM csupán két esetben tudta a legjobb értéket szolgáltatni.

— A GPS-eljárás 12 esetben tud a vonatkoztatási értékeknél jobbat adni, s ebből két esetben a relatív optimumot eredményezi.

— Összességében a legjobb eredményeket a BBPP1—0, BBPP1—1 eljárások adják; 16, illetve 17 esetben jobb b-t adnak, mint GENRCM, s ebből 6, illetve 10 esetben a legjobbnak minősülő értékeket adják.

— Majdnem azonos arányban (15, illetve 16 esetben) tudnak javítani GENRCM eredményén a PPM—0, PPM—1, BBPR—0, BBPR—1 eljárások; ebből 2-2, illetve 6-8 esetben adnak legkisebb b-t; vagyis a BBPR-eljárások a hatékonyabbak.

— A PPM—2 és GGLAA eljárások kivételével valamennyi algoritmusunk a vizsgált 27 esetből legalább 12 esetben jobb eredményt ad, mint GENRCM.

— A 7.2 megjegyzés értelmében

(7.2) BBPRED-, BBPP-, BBPR-, BBPP1- eljárások

a PPM-, illetve GENBRM-eljárások eredményein javítanak, s e javítások mértéke függ a gráf lokális szerkezeti adottságaitól.

A GENBRM- és PPM-eljárások durván az esetek 50%-ában adnak jobb eredményt, mint GENRCM. Amennyiben a (7.2)-beli eljárások nem tudtak volna javítani, akkor a vizsgált esetek csaknem 50%-ában e két eljárás szolgáltatta volna a legjobb eredményeket.

— A PPM—2 és GGLAA eljárások, melyek gyökérrel rendelkező szintstruktúra megfordításának számozását képezik, sok esetben kiugróan nagy sávszélességet és profil-értéket eredményeznek. Esetileg viszont (pl. $G=(300, 1200)$-as sorozat II., a $G=(1000, 5000)$-s sorozat II. feladatában) kiugróan kis sávszélességet eredményeznek.

A profil-érték redukciójának értékelése

— GENRCM egyetlen vizsgált esetben adott legjobbnak minősülő profil-értéket.

— GPS a profil-érték csökkentésében hatékonyabbnak bizonyul, mint sávszélesség-redukcióban, 3 esetben legjobbnak minősülő értéket eredményezett.

— Összességében legjobb eredményeket a BBPR—0, illetve BBPR—1 eljárások szolgáltatták, mivel 27 vizsgált esetből 20, illetve 21 alkalommal javítanak a vonatkozási értéken, s ebből 7-7 esetben legjobbnak minősülő eredményeket adtak.

— A közölt 17 eljárásunkból 12 algoritmus, 15 esetben kisebb profil-értéket szolgáltat, mint GENRCM.

— A (7.2)-beli eljárások értékelésére érvényes a korábbi észrevételünk, miszerint ezek a PPM- és GENBRM-eljárások eredményein javítanak — a gráf szerkezeti adottságaitól függően. Ez a két eljárás-típus a vizsgált esetek több mint 60%-ában jobb profil-értéket ad, mint GENRCM. A legjobbnak minősülő esetek száma viszonylag alacsony, de 7.2 megjegyzés értelmében ez a szám jelentősen nőhet.

A fentiek alapján megállapítható, hogy az eredmények értékelése során

BBPR-, BBPP1-, GENBRM- eljárások

tekinthetők a leghatékonyabbaknak.

Gépidő-igények elemzése

Bármely eljárás gyakorlati alkalmazhatóságát az általa nyert eredmények pontosságától függetlenül az dönti el, hogy gépidő-vonzata elfogadható korlátok alatt marad-e, vagy sem.

Eredmény-táblázatainkból világosan látszik, hogy az NN számozási eljárás kiugróan nagy gépidő-szükséglete miatt, az őt alkalmazó eljárások:

PMM

PPM-

BBPRED- eljárások

BBPP-

GGLAN-

GGLAA-

nagyméretű feladatokban nem alkalmazhatók.

További elemzésünkben csupán az RCN és RN1 számozási eljárásokkal operáló algoritmusaink vizsgálatára szorítkozunk. A 7.2. táblázatban az egyes eljárások mellett az azonos pontszámú gráfokon nyert gépideik átlagos értékét tüntettük fel.

7.2. TÁBLÁZAT

	CPU-idők átlagos értéke		
Gáf pontjainak száma	300	500	1000
GENRCM	0,17675	0,31039	0,69958
GPS	1,28199	1,34694	7,46595
GENBRM—0	0,58772	2,39421	2,71419
GENBRM—1	0,62769	2,41402	2,91414
BBPR—0	0,79188	2,87295	3,49981
BBPR—1	0,80387	3,17159	3,89352
BBPP1—0	0,87197	2,91712	5,55530
BBPP1—1	0,89621	2,94357	5,63958

— GENRCM a leggyorsabb eljárás; 1000-s méret esetén is átlagos gépidő-szükséglete 1 sec alatt marad.

— A GPS-eljárás gépidő-szükséglete erősen függ a gráf szerkezeti adottságaitól. Az 500-pontú gráfoknál alacsony, míg az 1000-pontú sorozatban indokolatlanul magas gépidőket eredményez.

— Eljárásaink, a felsorolási sorrendjüknek megfelelően monoton növekvő gépidő-szükségletet mutatnak, mely indokolt (algoritmusaik alapján). A gráf méretének növekedésével kedvező kis mértékben emelkedik a gépidő-felhasználás.

— Eljárásaink gépidő-szükséglete sokszorosa a GENRCM gépidő-igényének, de összevethető, sőt alatta marad a GPS eljárás gépidő-adatainak (1000-es méretű gráfok).

— A BBPR- és BBPP1-eljárások gépidő-vonzata természetszerűen erősen meghaladja a GENBRM-eljárásokét, de fenti tapasztalataink szerint számos esetben annál jobb eredményt is szolgáltatnak.

E megnövekedett gépidő-adatok még mindig kedvezőek a GPS-eljárás vonatkozó gépidő-szükségletéhez képest.

A fentiek alapján megállapítható, hogy

GENBRM-

BBPR- eljárásaink

BBPP1-

hatékonyabbnak értékelhetőek, mint GEORGE—LIU GENRCM, illetve GIBBS—POOLE—STOCKMEYER GPS-*eljárásai,* mivel a vizsgált esetek jelentős részében kedvezőbb sávszélességet és profil-értéket eredményeztek, míg gépidő-vonzatuk elfogadhatóan alacsony szinten marad.

Megjegyezzük, hogy vizsgált feladataink mindegyikében a gráf paramétereihez képest igen kicsi az átmérő (5, illetve 6). Ezzel magyarázható, hogy GENRCM igen kis műveletszámmal éri el az eredményt, míg GPS eljárásban és algoritmusainknál a vizsgált excentricitási pontok száma nagy (4.2.4. megjegyzés).

A teljesség kedvéért a 18. sz. mellékletben közöljük valamennyi, korábban kifejlesztett sávszélesség-redukciós eljárásunknak kisméretű, síkbeli gráfokon nyert eredményét. Itt

ROSEN R. ROSEN eljárása [77]
CUTHIL az eredeti *Cuthill—McKee eljárás* [27]
OURM maximális fokszámú ponton alapuló eljárásunk [4]
NCC RNN számozási algoritmusunkat elsőként alkalmazó eljárás (8. sz. melléklet)

eljárásokkal bővítjük ki a már vizsgált algoritmusok körét.

8. A számítógépes kidolgozás kérdései

Eljárásaink számítógépes kidolgozását FORTRAN nyelven készítettük, s futtatásainkat IBM 3031 gépen, OS operációs rendszer alatt végeztük.

A ritka mátrixok (gráfok) zérus/nem-zérus szerkezetének tárolását, kezelését az alábbi tömbök felhasználásával végeztük.

Legyen N a mátrix rendje, NZ a főátlón kívüli nem-zérus elemek száma.

ADJNCY a mátrix főátlón kívüli nem-zérus elemeinek oszlopindexeit regisztráló NZ hosszúságú tömb; a tárolás a sor-indexek növekvő sorrendjében, s egy soron belül az oszlop-indexek növekvő sorrendjében történik.

XADJ az ADJNCY vektor $(N+1)$ komponensű pointer tömbje, mellyel egyértelműen megadhatók, hogy a tárolt oszlop-indexek mely sorokra vonatkoznak.
XADJ $(I)=J$ azt jelenti, hogy a mátrix I-edik sorában az első nem-zérus elem oszlop-indexét az ADJNCY tömb J-edik komponenseként tároltuk. Vagyis, az I-edik sor nem-zérus elemeinek oszlop-indexei
ADJNCY (L) XADJ $(I) \leqq L \leqq$ XADJ $(I+1)-1$
szerint állnak elő.

MASK N-komponensű tömb, mely a gráf összefüggő komponenseinek regisztrálására szolgál. Kezdetben MASK $(I)=1$ $(I=1, 2, \ldots, N)$. Eljárásaink csupán MASK $(I)=1$ típusú pontokon dolgoznak. A már feldolgozott, összefüggő komponensek pontjait MASK $(I)=0$ szerint regisztrálva, valamennyi összefüggő részgráfon végrehajtódik a kívánt eljárás. (A nem összefüggő gráfok kezelhetőségének fontosságát a 21. sz. mellékletben közölt feladat szemlélteti.)

LS N-komponensű tömb, mely tetszőleges szintstruktúra pontjait tárolja a szint-indexek növekvő sorrendjében.

XLS Az LS vektor $(N+1)$ hosszúságú pointer tömbje, melynek I-edik komponense azt mutatja, hogy LS hányadik eleménél kezdődik az I-edik szint pontjainak felsorolása.

PERM N-komponensű tömbök, melyek az eljárásainkkal nyert számozást regisztrálják.

INVP PERM $(I)=J$ azt jelenti, hogy az új számozás az I csúcsszámot az eredeti számozásban J csúcsszámú ponthoz rendeli hozzá.
INVP $(I)=J$ azt jelenti, hogy az eredeti számozásban I csúcsszámmal rendelkező ponthoz az új számozás a J csúcsszámot rendeli hozzá. Nyilvánvalóan

$$\text{PERM}\,(\text{INVP}(I)) \equiv I$$

teljesül.
E tömböket permutációs tömböknek nevezzük, mivel egyértelműen definiálják számozás által generált permutációs mátrixot.

A fenti tömbök egész típusúak. Megjegyezzük, hogy konkrét mátrix tárolásakor további két valós vagy dupla pontos tömb szükséges, melyekben a fődiagonálisban levő (DIAG), illetve a fődiagonálison kívüli nem-zérus elemek (COEFF) numerikus értékeit tároljuk.

Eljárásaink tárigénye

A GENRCM eljárás a fenti tömböket (INVP kivételével) használja, így tárigénye 5N+NZ+2. (Az INVP tömb beállításáról az eljáráson kívül kell gondoskodni.)

Eljárásaink gondoskodnak INVP előállításáról is, így a fenti hat tömböt használják, s további négy N-komponensű tömböt igényelnek: így tárigényük

$$10N+NZ+3,$$

melyből 2N+1 nagyságú terület a legkeskenyebbnek minősülő szintstruktúra tárolására szolgál, így ez elkerülhető, de cserében újabb szintstruktúrát kell generálni.

Megjegyezzük, hogy GPS tárigénye 16N+NZ+2. 1982-ben J. G. LEWIS [61] munkája az eljárás igen hatékony verziójának kifejlesztéséről tudósít, melyben igen bonyolult program-szervezéssel a tárigényt sikerült 6N+NZ értékre csökkenteni.

Eljárásaink néhány szervezési kérdését a 7.2. és 7.3. megjegyzésekben ismertettük.

Eljárásainkat úgy építettük fel, hogy azok a SPARSPAK megfelelő rutinjával helyettesíthetők.

Megjegyezzük, hogy további program-szervezéssel elérhető a tárigény 7N+ +NZ+2 értékre csökkentése, melynek munkálatai elkezdődtek.

Ezáltal elkészült a SPARSPAK egyik ágának hazai kidolgozása, mellyel pozitív definit, ritka, szimmetrikus mátrixú lineáris egyenletrendszerek a profil-szimmetrikus faktorizációval [53] oldhatók meg.

A sávszélesség/profil-redukciós fázisban paraméterrel igényelhető

GENRCM-

GENBRM- eljárások

BBPR-

BBPPI-

bármelyikének végrehajtása.

Az egyenletrendszer megoldási fázisát a SPARSPAK megfelelő eljárásainak felhasználásával készítettük.

A program-rendszer működésének sematikus vázát a 19. sz. mellékletben szemléltetjük.

Mindezek alapján lehetővé válik a direkt módszerrel történő megoldás [53]-ban összegezett alábbi előnyeinek a felhasználása.

A ritka mátrixú lineáris egyenletrendszerek direkt megoldási módszereinek alkalmazásakor

— rendezés (sávszélesség/profil vagy fill-in csökkentés),
— tárolás szervezése,
— faktorizáció,
— megoldás

lépések egymásutánját kell végrehajtanunk, s ez további műveletszám-csökkentés lehetőségét hordozza magában.

a) Ha ugyanazon egyenletrendszert több jobb oldallal kell megoldanunk, akkor csak az első egyenletrendszer esetén kell végrehajtanunk a fenti négy lépést, a többi esetben elegendő csupán a megoldási fázist végrehajtani.

b) A PERM és INVP permutációs tömbök tárolásával elérhető, hogy

— zérus/nem-zérus szerkezetében azonos, numerikus értékekben eltérő mátrix esetén
— nem szimmetrikus, de egy, már korábban feldolgozottal azonos zérus/nem-zérus szerkezetű mátrixszal rendelkező egyenletrendszer esetén

} csupán a faktorizáció és megoldás fázisokat kell elvégezni.

A fenti lehetőségek felhasználásával további hatékonysági növekedés érhető el.

E mechanizmus sematikus működését szemléltetjük a 20. sz. mellékletben.

Eljárásaink alkalmazásával mutatkozó hatékonysági növekedést a 21. sz. mellékletben közölt statisztikai feladaton szemléltetjük. Demonstrációs példaként azért választottunk kisméretű feladatot, hogy szemléltetni tudjuk a merevségi mátrix szerkezetét, s a belőle származó gráfot. A példa rámutat a sávszélesség/profil-redukciós eljárások alkalmazásának folyamatára, s az alkalmazás révén nyert műveletszám-csökkenésre egyaránt.

9. Eredményeink hasznosíthatósága

— A perifériális pontokat meghatározó eljárásaink, illetve azok közel-minimális szélességű szintstruktúrát előállító módosításai közvetlen alkalmazásra találnak a J. A. GEORGE nagy hatékonyságot eredményező algoritmusaiban:

— *quotient tree method* [53];

— *one-way dissection ordering* [46], [52], [53];

— *nested dissection ordering* [37], [40], [44], [45], [53].

A fenti eljárások mindegyikének közös kiinduló lépése perifériális pont gyökerű, illetve közel-minimális szélességű RLS (x) előállítása. GEORGE ilyen céllal a sávszélesség/profil-redukciós eljárásukban használatos FNROOT rutint alkalmazta valamenynyi rendezési algoritmusának számítógépes kidolgozásában. Mint rámutattunk, az FNROOT rutin a maximális excentricitás közelítéseként abszolút minimális excentricitást is eredményezhet.

Eljárásaink alkalmazásával a fenti algoritmusokban lényeges eredménybeli javulás várható, csekély gépidő-növekedés mellett.

A számítógépes kidolgozásban eljárásainkat eleve úgy szerveztük, hogy azok közvetlenül alkalmazhatók a SPARSPAK-ban, annak megfelelő rutinjaival helyettesíthetők.

— J. A. GEORGE *nested dissection* algoritmusában a gráf olyan D tagozódási halmazának előállítása szükséges, mely „felezi" a gráfot, azaz két, durván azonos pontszámú részgráfra bontja. Az eljárás annál hatékonyabb, minél kevesebb pontot tartalmaz D [53]. GEORGE D-t egy szemi-pszeudo-perifériális pont gyökerű szintstruktúra közel-középső szintjéből állítja elő.

A közel-minimális szélességű szintstruktúrát előállító IV. eljárásunk közbülső lépéseként olyan D tagozódási halmazt állítunk elő, mely tartalmazza a kezdő pontpár középső halmazát (5.2.7. megjegyzés). Egyidejűleg $|D|$ igen gyakran kisebb, mint a megfelelő RLS (x)-ben a közel-középső szint szélessége.

A felvázolt módon tovább javítható a nested dissection algoritmus SPARSPAK-beli verziójának hatékonysága.

— A perifériális pontokat előállító eljárásaink alkalmazást nyerhetnek olyan operációkutatási feladatokban, melyekben a probléma irányítatlan gráffal szimulálható (pl. útvonalak), és közbülső lépésként az egymástól legtávolabbi pontok, vagy összes ilyen pontpár meghatározása szükséges.

— Sávszélesség-redukciós eljárásaink alkalmazása lehetővé teszi a gyári software, a standard programcsomagok teljesértékű felhasználását. Egyre több ugyanis a rendelkezésre álló, sáv-mátrix-szal operáló kész rutin, arra azonban nincs útmutatás, hogy hogyan állítható elő tetszőleges ritka, szimmetrikus mátrix kis sávszélességű sáv-mátrix alakban.

— Sávszélesség/profil-redukciós eljárásaink alkalmazásával a lineáris egyenletrendszerek iterációs [33], [78], [87], illetve kombinált [67], [85] eljárással való megoldásakor is könnyű kezelhetőség, s sok esetben csökkentett tárigény és műveletszám érhető el.

— Az IBM 3031 gépen élő, de ESz-gépekre is átvihető program-rendszerünk lehetővé teszi a nagyméretű lineáris egyenletrendszerek egy részének hatékony számítógépes kezelését és megoldását.

— Program-rendszerünk felhasználható olyan kész program-rendszerek hatékonyságának növelésére, melyekben a sávszélesség-redukció nem megoldott, így az egyenletrendszer megoldási fázisa nehézkes és rendkívül gépidőigényes; (pl. SAP IV.).

— Eljárásaink egy részének kisszámítógépekre történő kidolgozásával szélesíthető a kisgépeken megoldható feladatok köre.

IRODALOM

[1] AHO, A. V., HOPCROFT, J. E. and ULLMAN, J. D., *The Design and Analysis of Computer Algorithms*, (Addison—Wesley, 1974).

[2] ALVARADO, F. L., "Computational complexity of operations involving perfect elimination sparse matrices", *Int. J. of Comp. Math.* 6 (1977) 69—82.

[3] ALWAY, G. G. and MARTIN, D. W., "An algorithm for reducing the bandwidth of a matrix of symmetrical configuration", *Computer J.* 8 (1965) 264—272.

[4] ARANY, I., SMYTH, W. F. and SZÓDA, L., "An improved method for reducing the bandwidth for sparse symmetric matrices", *Information Processing 71*: Proc. of IFIP Congress, North-Holland Publ. Co., Amsterdam (1972) 1246—1250.

[5] ARANY, I. és SZÓDA, L., „Ritka, szimmetrikus mátrixok sávszélesség-redukciója", *Információ-Elektronika* 4 (1973) 273—282.

[6] ARANY, I., MÁTRIXOK, Országos Software Katalógus, 2.1277.00017 sz. dokumentáció, Budapest, (1977).

[7] ARANY, I., "Speciális típusú, ritka, szimmetrikus mátrixok sávszélességének és profil-értékének redukciója", *MÜM SZÁMTI Intézeti Tájékoztató*, 1. sz. Budapest, (1977) 101—115.

[8] ARANY, I., SMYTH, W. F. and SZÓDA, L., "Minimizing the bandwidth of sparse matrices", in *ANNALES, Sectio Computatorica*, tomus 1., ed. Eötvös Lorand University, Budapest, (1978) 129—151.

[9] ARANY, I., „Új számozási stratégia ritka, szimmetrikus mátrixok sávszélességének redukálására", *MÜM SZÁMTI Intézeti Tájékoztató*, 3. sz. Budapest, (1979) 18—28.

[10] ARANY, I., „Rúdszerkezet statikai vizsgálata mechanikai véges elemek módszerével", *MÜM SZÁMTI Intézeti Tájékoztató*, 4. sz. Budapest, (1980) 25—43.

[11] ARANY, I., "Notes on using quotient graphs in the elimination process", *Zeitschrift für Angewandte Mathematik und Mechanik* 63 (1983) T336—T337.

[12] ARANY, I., "How to find rooted level structure of near-minimum width", *Bulletins for Applied Mathematics*, No. XXIX. ed. Technical University of Budapest (1983) 75—97.

[13] ARANY, I., "A necessary condition for getting acceptable small bandwidth for sparse matrices", *Bulletins for Applied Mathematics*, No. XXX., ed. Technical University of Budapest, Budapest, (1983) 39—51.

[14] ARANY, I., "Distance — level structure — pseudo-peripheral nodes", *Models and Algorithms*, ed. Computing Centre, Eötvös Lorand University, Budapest, (1983) 5—32.

[15] ARANY, I., "The method of Gibbs—Poole—Stockmeyer is non-heuristic", in *ANNALES, Sectio Computatorica*, tomus 4., ed. Eötvös Lorand University, Budapest, (1983) 29—37.

[16] ARANY, I., "Another method for finding pseudo-peripheral nodes", in *ANNALES, Sectio Computatorica*, tomus 4., ed. Eötvös Lorand University, Budapest, (1983) 39—49.

[17] ARANY, I., "A heuristic method for finding peripheral nodes", W. P. MG/1. ed. Computer and Automation Institute, Hungarian Academy of Sciences, Budapest, (1984) 1—32.

[18] ARANY, I., "An efficient algorithm for finding peripheral nodes", *Proceedings of Colloquium on the Theory of Algorithms*, North-Holland Publ. Co., Amsterdam, to appear. (1985) 27—35.

[19] BERGE, C., *The Theory of Graphs and Its Applications*, (John Wiley and Sons Inc., New York, 1962).
[20] BUNCH, J. R., "Analysis of sparse elimination", *SIAM J. Numer. Anal.* **11** (1974) 847—873.
[21] BUNCH, J. R., "Block method for solving sparse linear systems", *Sparse Matrix Computations* (Acad. Press, New York, 1976) 39—58.
[22] CHEN, CH. K. E. and GARFINKEL, R. S., "The generalized diameter of a graph", *Networks* **12** (1982) 335—340.
[23] CHENG, K. Y., "Note on minimizing the bandwidth of sparse symmetric matrices", *Computing* **11** (1973) 27—30.
[24] CHINN, P. Z., CHVÁTALOVÁ, J., DEWDNEY, A. K. and GIBBS, N. E., "The bandwidth problem for graphs and matrices — a survey", *J. of Graph Theory* **6** (1982) 223—254.
[25] CHRISTOFIEDES, N., *Graph Theory — An Algorithmic Approach*, (Acad. Press, London, 1975).
[26] COLLINS, R. J., "Bandwidth reduction by automatic renumbering", *Int. J. for Num. Math. in Eng.* **6** (1973) 345—356.
[27] CUTHILL, E. H. and MCKEE, J., "Reducing the bandwidth of sparse symmetric matrices", *Proc. 24th National Conf. ACM* (1969) 157—172.
[28] CUTHILL, E. H., "Several strategies for reducing the bandwidth of matrices", *Sparse Matrices and Their Applications*, ed. D. J. Rose and R. A. Willoughby, SIAM Publications, New York, (1972) 157—166.
[29] DANTZIG, G. B. and ORCHARD—HAYS, W., Alternate algorithm for the revised simplex method using product form for the inverse, The RAND Corporation, Research memorandum, RM—1268 (1953).
[30] DUFF, I. S., "A survey of sparse matrix research", *Proc. of the IEEE* **65** (1977) 500—535.
[31] EINSENSTAT, S. C., SCHULTZ, M. H. and SHERMAN, A. H., "Software for sparse Gaussian elimination with limited core storage", *Sparse Matrix Proceedings* 1978, ed. I. S. Duff and G. W. Stewart, SIAM Publications, Philadelphia, (1979) 135—153.
[32] FELIPPA, C. A., "Solution of linear equations with skylinestored symmetric matrix", *Computers and Structures*, Pergamon Press, (1975) **5** 13—29.
[33] FORSYTHE, G. E. and MOLER, C. B., *Computer Solution of Linear Algebraic Systems* (Prentice—Hall Inc., Englewood Cliffs, New Jersey, 1967).
[34] GAREY, M. R., GRAHAM, R. L., JOHNSON, D. S. and KNUTH, D. E., "Complexity results for bandwidth minimization", *SIAM J. Appl. Math.* **34** (1978) 477—495.
[35] GEORGE, J. A., Computer Implementation of the Finite Element Method. Ph. D. Thesis, Techn. Rep. STAN — CS — 208., Standorfd University (1971).
[36] GEORGE, J. A., "Block elimination on finite element systems of equations", *Sparse Matrices and Their Applications*, ed. D. J. Rose and R. A. Willoughby, Plenum Press, New York (1972) 101—114.
[37] GEORGE, J. A., "Nested dissection of a regular finite element mesh", *SIAM J. Numer, Anal.* **10** (1973) 345—363.
[38] GEORGE, J. A. "On block elimination for sparse linear systems", *SIAM J. Numer, Anal.* **11** (1974) 585—603.
[39] GEORGE, J. A. and LIU, J. W-H., "A note on fill for sparse matrices", *SIAM J. Numer. Anal.* **12** (1975) 452—455.
[40] GEORGE, J. A. and LIU, J. W-H.,"An algorithm for automatic nested dissection and its applications to general finite element problems", *Proc. of the Sixth Manitoba Conference on Numerical Mathematics*, Winnipeg (1976) 59—95.
[41] GEORGE, J. A., "Numerical experiments using dissection methods to solve *nxn* grid problems", *SIAM J. Numer. Anal.* **14** (1977) 161—179.
[42] GEORGE, J. A. and MCINTYRE, D. R., "On the applications of the minimum degree algorithm to-finite element systems", *SIAM J. Numer. Anal.* **15** (1978) 90—121.
[43] GEORGE, J. A. and LIU, J. W-H., "Algorithms for matrix partitioning and the numerical solution of finite element systems", *SIAM J. Numer. Anal.*, **15** (1978) 297—327.
[44] GEORGE, J. A., POOLE, W. G. and VOIGT, R. G., "Incomplete nested dissection for solving *n* by *n* grid problems", *SIAM J. Numer. Anal.*, **15** (1978) 662—673.
[45] GEORGE, J. A. and LIU, J. W-H., "An automatic nested dissection algorithm for irregular finite element problems", *SIAM J. Numer. Anal.* **15** (1978) 1053—1069.
[46] GEORGE, J. A., "An automatic one-way dissection algorithm for irregular finite element problems", *Proc. 1977. Dundee Conf. on Numerical Analysis*, Lecture Notes No. 630., Springer-Verlag (1978) 76—89.

[47] GEORGE, J. A. and LIU J. W-H., "A quotient graph model for symmetric factorization", *Sparse Matrix Proceedings* 1978, I. S. Duff and G. W. Stewart, SIAM Publications, Philadelphia, (1979) 154—174.
[48] GEORGE, J. A. and LIU, J. W-H., "The desing of user interface for a sparse matrix package", *ACM Trans. on Math. Software* **5** (1979) 139—162.
[49] GEORGE, J. A. and LIU, J. W-H., "An implementation of a pseudo-peripheral node finder", *ACM Trans. on Math. Software* **5** (1979) 284—295.
[50] GEORGE, J. A., "Direct methods for the solution of large sparse system of linear equations, Part I", *SIAM News* **13** (June, 1980).
[51] GEORGE, J. A., "Direct methods for the solution of large sparse system of linear equations, Part II.". *SIAM News* **13** (August, 1980).
[52] GEORGE, J. A., "An automatic one-way dissection algorithm for irregular finite element problems", *SIAM J. Numer. Anal.* **17** (1980) 740—751.
[53] GEORGE, J. A. and LIU, J. W—H., *Computer Solution of Large Sparse Positive Definite Systems.* (Prentice Hall Inc., Englewood Cliffs, New Jersey, 1981).
[54] GERGELY, J., „Módszerek és programok ritka mátrixokra", *Alkalmazott Matematikai Lapok* **6** (1980) 407—442.
[55] GIBBS, N. E., POOLE, W. G. and STOCKMEYER, P. K., "An algorithm for reducing the bandwidth and profile of a sparse matrix", *SIAM J. Number. Anal.* **13** (1976) 236—250.
[56] GIBBS, N. E., POOLE, W. G. and STOCKMEYER, P. K., "A comparison of several bandwidth and profile reduction algorithms", *ACM Trans. on Math. Software* **2** (1976) 322—330.
[57] IRONS, B. M., "A frontal solution program for finite element analysis", *Int. J. for Num. Math. in Eng.* **2** (1970) 5—32.
[58] JENNINGS, A., "A compact storage scheme for the solution of symmetric linear simultaneous equations", *Computer J.* **9** (1966) 281—285.
[59] KING, I. P., "An automatic reordering scheme for simultaneous equations derived from net. work problems", *Int. J. for Num. Math. in Eng.* **2** (1970) 523—533.
[60] KNUTH, D. E., *The Art of Computer Programming, Vol. 3. Sorting and Searching* (Addison—Wesley Publ. Co., London, 1973).
[61] LEWIS, J. G., "Implementation of the Gibbs—Poole—Stockmeyer and Gibbs—King algorithms", *ACM Trans. on Math. Software* **8** (1982) 180—189.
[62] LEWIS, J. G., "Numerical experiments with SPARSPAK", *ACM SIGNUM Newsletter* **18** (July, 1983) 12—22.
[63] LIPTON, R. J., ROSE, D. J. and TARJAN, R. E., "Generalized nested dissection", Research Report No. STAN—CS—77—645, Stanford University, 1977.
[64] LIPTON, R. J., ROSE, D. J. and TARJAN, R. E., "Generalized nested dissection", *SIAM J. Numer. Anal.* **16** (1979) 346—358.
[65] LISKA, T., MTAFORT szubrutin könyvtár, *IBM 3031 Felhasználói Ismertetők*, 6. sz. MTA SZTAKI, Budapest, megjelenés alatt.
[66] LIU, J. W-H. and SHERMAN, A. H., "Comparative analysis of the Cuthill—McKee and the reverse Cuthill—McKee ordering algorithms for sparse matrices", *SIAM J. Numer. Anal.* **13** (1976) 197—213.
[67] MANTUEFFEL, T. A., The Shifted Incomplete Cholesky Factorization, Research Rep. No. Sand 78—8226, (Sandia Laboratories, Albuquerque, New Mexico and Livermore, 1978).
[68] MARKOWITZ, H. M., "The elimination form of inverse and its application to linear programming", *Management Science* **3** (1957) 255—269.
[69] MARTIN, R. S. and WILKINSIN, J. H., "Symmetric decomposition of positive definite band matrices", *Handbook for Automatic Computation, Vol. II.* ed. J. H. Wilkinson and C. Reinsch (Springer-Verlag, 1971).
[70] MAYEDA, W., *Alkalmazott gráfelmélet*, fordítás (Műszaki Könyvkiadó, Budapest, 1976).
[71] MUNKGAARD, N., New Factorization Codes for Sparse, Symmetric and Positive Definite Matrices, Rep. No. NI—78—08, Numerisk Institut of Technical University of Denmark, Lyngby 1978.
[72] PACHL, J. K., Finding Pseudoperipheral Nodes in Graphs, Research Report CS—82—20, Department of Computer Science, University of Waterloo, (June, 1982).
[73] PAPADIMITRIU, CH. H., "The NP-completeness of the bandwidth minimization problem", *Computing* **16** (1976) 263—270.
[74] PAPADIMITRIU, CH. H. and STEIGLITZ, K., *Combinatorial Optimization: Algorithms and Complexity*, (Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1982).

[75] RHEINBOLDT, W. C., BASILI, V. R. and MESZTÉNYI, C. K., "GRAAL — Graph Algorithmic Language", *Proc. of Sparse Matrices and Their Applications*, ed. D. J. Rose and R. A. Willoughby (Plenum Press, New York, 1972) 167—176.
[76] ROSE, D. J., "A graph-theoretic study of the numerical solution of sparse positive definite systems of linear equations", *Graph Theory and Computing*, (Acad. Press, New York, 1972) 183—217.
[77] ROSEN, R., "Matrix bandwidth minimization", *Proc. of 23rd Math. Conf. ACM, ACM Pubications P—68*, (Brandon/System Press, Princeton, New Jersey, 1968).
[78] Самарский, А. А., *Введение в теорию разностных схем*, (Изд. Москва, 1971).
[79] SMYTH, W. F. and RADACEANU, E., "Storage scheme for hierarchic structures", *The Computer J.* 17 (1973) 152—156.
[80] SMYTH, W. F. and BENZI, W. M. L., "An algorithm for finding the diameter of a graph", in *Proc. IFIP Congr. 74.*, (North-Holland Publ. Co., Amsterdam, 1974) 500—503.
[81] SMYTH, W. F. and ARANY, I., "Another algorithm for reducing bandwidth and profile of a sparse matrix", Proc. AFIPS 1976 NCC, AFIPS Press, Montvale, New Jersey, (1976) 987—994.
[82] SMYTH, W. F., "Algorithms for the reduction of matrix bandwidth and profile", *Journal on Computational and Applied Mathematics*, (1985) 551—561. to appear.
[83] TEWARSON, R. P., *Sparse Matrices* (Acad. Press., New York, 1973).
[84] WESTENBERG, A. W. and BERNA, T. J., "LASCALA — A language for large scale linear algebra", *Sparse Matrix Proceedings 1978*, ed. I. S. Duff and G. W. Stewart (SIAM Publications, Philadelphia, 1979) 90—106.
[85] ZLATEV, Z. and NIELSEN, H. B., "SIRSM, A package for the solution of sparse systems by iterative refinement", Rep. No. NI—77—13, Numerisk Institut of Technical University of Denmark, Lyngby, (1977).
[86] ZOMBORI, L. és KOLTAI, M., *Elektromágneses terek gépi analízise* (Műszaki Könyvkiadó, Budapest, 1979).
[87] YOUNG, D. M., *Iterative Solution of Large Linear Systems*, (Acad. Press., New York, 1971).

(Beérkezett: 1984. szeptember 20).

ARANY ILONA
MTA SZÁMÍTÁSTECHNIKAI ÉS AUTOMATIZÁLÁSI KUTATÓ INTÉZET
1014 BUDAPEST, ÚRI U. 49.

EFFICIENT TREATING OF LARGE SPARSE SYMMETRIC MATRICES

I. ARANY

In our present work we deal with the bandwidth/profile reduction problem. We introduce the terminology based on which we analyse the algorithms of *Gibbs—Poole—Stockmeyer* and *George—Liu*. On the other hand, we discuss both the level structure generation and the numbering phases of the bandwidth/profile reduction algorithm. As a result of the above discussion we describe several algorithms (by number 17) for reducing the bandwidth and profile of a sparse matrix and 6 of them seem to be more efficient than the above mentioned well-known algorithms.

FÜGGELÉK

JELÖLÉSEK

$G=(X, E)$	irányítatlan gráf (X pontjainak, E éleinek halmaza)
b	sávszélesség
$\mathrm{GLS}=\{L_0, L_1, \ldots, L_k\}$	általános szintstruktúra
$N(x)$	az $x\in X$ pont szomszéd halmaza
$\mathrm{RLS}(Y)=\{L_0(Y), L_1(Y), \ldots, L_k(Y)\}$ $Y\subset X$	gyökerű szintstruktúra
$l(x)$	$x\in X$ pont excentricitása
$L_{ec}(x)$	$x\in X$ pontból generált RLS (x) maximális indexű szintje
$d(x, y)$	$x,\ y\in X$ pontok távolsága
L_i	a szintstruktúra i-edik szintje
$\lvert L_i\rvert$	L_i szélessége
$W(\cdot)$	a $(\cdot)$ szintstruktúra szélessége
$W_x(y, z)$	$d(y, z)=d(y, x)+d(x, z)-W_x(y, z)$ által definiált mennyiség
W_i	a kompatibilis számozásban az i-edik szint számozása után az összes beszámozott pontok száma
$k_x(y, z)$	$d(y, z)\geqq d(x, y)+d(x, z)-2k_x(y, z)$ kifejezésből ismeretes mennyiség
$R(x, y)$	x és y pontok reverzibilis halmaza
$M(x, y)$	x és y pontok középső halmaza
$\langle x, y\rangle$	szintstruktúra rendszer
z_x, z_y	$z\in X$ pont $\langle x, y\rangle$-beli pszeudo-koordinátái
$W(p, q)$	$\langle x, y\rangle$-ban $p,\ q\in X$ pontok távolságaiban szereplő mennyiség

$$W(p, q) = W_x(p, q)+W_y(p, q)$$

diam (G)	a $G=(X, E)$ gráf átmérője
GPS—PS	*Gibbs—Poole—Stockmeyer pszeudo-perifériális pontot meghatározó eljárása*
GL—SPS	*George—Liu szemi-pszeudo-perifériális pontot meghatározó eljárása*
PS	saját fejlesztésű eljárás pszeudo-perifériális pontok meghatározására
PS1	PS eljárás speciális kezdőpont esetén
P-eljárás	perifériális pontot előállító algoritmus
P′-eljárás	P-eljárás 1. sz. módosítása
P″-eljárás	P-eljárás 2. sz. módosítása
P‴-eljárás	P-eljárás 3. sz. módosítása
P1-eljárás	heurisztikus algoritmus perifériális pontok meghatározására, mely P módosításaként is tekinthető
P2-eljárás	maximális súlyú pontok halmazára épülő heurisztikus algoritmus perifériális pontok meghatározására
P2′-eljárás	P2-eljárás 1. sz. módosítása
P2″-eljárás	P2-eljárás 2. sz. módosítása
CS-eljárás	tagozódási halmazt előállító algoritmus
LS-eljárás	tetszőleges tagozódási halmazból speciális szerkezetű GLS-t képező algoritmus
NCC—LS	saját fejlesztésű heurisztikus algoritmus max. excentricitás közelítésére
CN	*Cuthill—McKee számozási eljárás*
RCN	CN-számozás megfordítása

NI	speciális szerkezetű GLS-t számozó algoritmus
RNI	NI-számozás megfordítása
NN	általános szerkezetű GLS-t számozó algoritmus
RNN	NN-számozás megfordítása
GENRCM	SPARSPAK-ban használatos sávszélesség/profil-redukciós eljárás
GPS	*Gibbs—Poole—Stockemyer sávszélesség/profil-redukciós algoritmus*

GPS—LS-eljárás a GPS-eljárás GLS-t kialakító fázisa

1. MELLÉKLET

Ritka mátrixok és gráfok kapcsolata

Tetszőleges N-edrendű, szimmetrikus **A** mátrix $(a_{ii} \neq 0, i=1, 2, \ldots, N)$ zérus/nem-zérus szerkezetéhez egyértelműen hozzárendelhető $G_A=(X, E)$ irányítatlan, hurok és többszörös él nélküli gráf, ahol

$$X = \{x_1, x_2, \ldots, x_N\} \text{ pontok halmaza,}$$

$$E = \{(x_i, x_j) | a_{ij} \neq 0\} \text{ élek halmaza.}$$

Fordítva, bármely irányítatlan, hurok és többszörös él nélküli gráfnak egyértelműen megfeleltethető mátrixot, mely csupán zérus/nem-zérus szerkezetében meghatározott, a gráf *összefüggési mátrixának* nevezzük.

A $G=(X, E)$ gráf *számozásán*

$$n: X \to I = \{1, 2, \ldots, |X|\}$$

bijekciót értjük; s a számozással ellátott gráfot G^n-ként jelöljük. Az x_i $(1 \leqq i \leqq |X|)$ ponthoz rendelt $n(x_i)$ értéket a pont *csúcsszámának* nevezzük.

Az **A** mátrix a hozzárendelt G_A gráfon egyértelműen definiálja annak n_I számozását, ahol

$$n_I(x_i) = i \quad (1 \leqq i \leqq |X|)$$

érvényes, melyet a gráf kezdeti, vagy eredeti számozásának nevezünk.

Tekintsük a $G=(X, E)$ gráf tetszőleges n számozását; a G^n *gráf sávszélességét*

$$b(G^n) = \max_{(x_i, x_j) \in E} |n(x_i) - n(x_j)|$$

szerint értelmezzük, míg G^n profilján

$$\Pr(G^n) = \bigcup_{i=i}^{|x|} C(x_i)$$

halmazt értjük, ahol

$$C(x_i) = \{m | m \in I;\ \min_{(z, x_i) \in E} n(z) \leqq m < n(x_i)\}.$$

Könnyű belátni, hogy tetszőleges **A** mátrix esetén

$$b(A) = b(G_A^{n_I});\ |\Pr(A)| = |\Pr(G_A^{n_I})|$$

teljesülnek, továbbá tetszőleges G gráf bármely n számozásakor

$$b(G^n) = b(\overline{\mathbf{A}}); \quad |\mathrm{Pr}(G^n)| = |\mathrm{Pr}(\overline{\mathbf{A}})|,$$

ahol $\overline{\mathbf{A}}$ a G^n összefüggési mátrixa.

Tetszőleges **A** mátrixhoz rendelt G_A gráf bármely n számozása egyértelműen definiálja azon két (PERM, INVP) permutációs tömböt, melyek az $n_I \leftrightarrow n$ kapcsolatot írják le. PERM $(i)=j$ azt jelenti, hogy az n számozás az i csúcsszámot az n_I számozásban j csúcsszámmal rendelkező ponthoz, azaz x_j-hez rendeli hozzá. Az inverz transzformációban INVP $(i)=j$ azt jelenti, hogy az n_I számozásban i csúcsszámmal ellátott ponthoz, vagyis x_i-hez az n számozás a j csúcsszámot rendeli hozzá. Nyilvánvalóan

$$\mathrm{PERM}\,(\mathrm{INVP}\,(i)) \equiv i.$$

Tetszőleges **A** mátrix esetén G_A gráf bármely n számozása a PERM és INVP permutációs tömbök révén egyértelműen meghatároz egy **P** permutációs mátrixot, melyre

$$b(G_A^n) = b(\mathbf{PAP}^T)$$

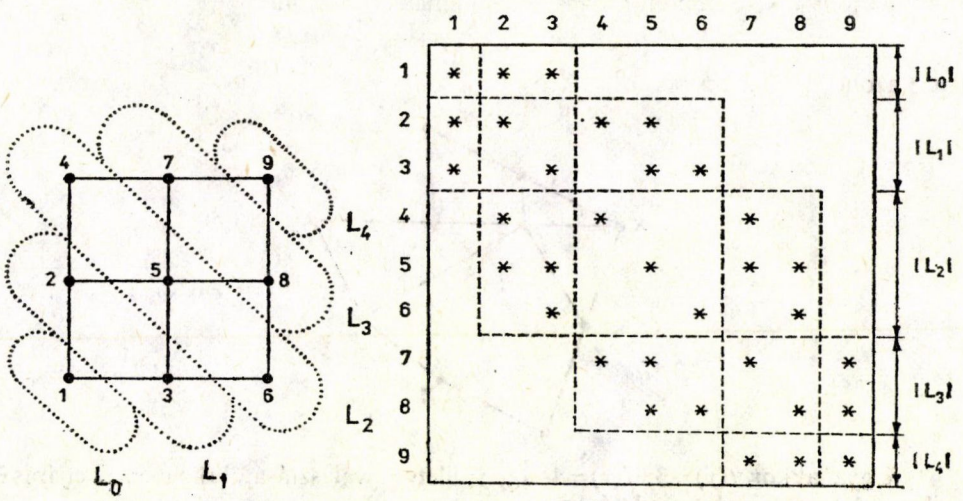
$$|\mathrm{Pr}(G_A^n)| = |\mathrm{Pr}(\mathbf{PAP}^T)|,$$

teljesülnek. **P** az egység-mátrix sorainak az n számozással definiált (PERM) permutációja révén előálló mátrix, melyet permutációs mátrixnak neveznek.

2. MELLÉKLET

GLS kompatibilis számozása az összefüggési mátrixban blokk-diagonális szerkezetet eredményez

$$\mathrm{GLS} = \{L_0, L_1, L_2, L_3, L_4\}$$

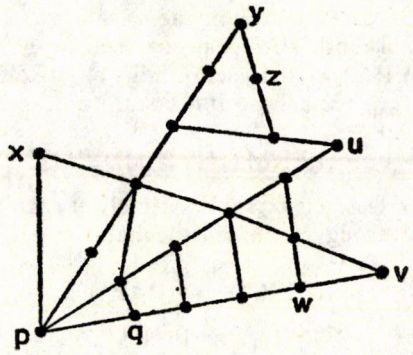


3. MELLÉKLET

A GPS—PS eljárással nyert excentricitás nem egyértelműen meghatározott

Legyen x a kezdőpont, $l(x)=4$;

$$L_{ec}(x) = \{y, z, u, v, w\},$$



y, z, u, v minimális fokszámú pontok.

$$l(y) = l(z) = l(v) = \text{diam}\,(G) = 6$$

$$l(u) = 5; \quad L_{ec}(u) = \{p, q\}; \quad l(p) = l(q) = 5;$$

vagyis u pszeudo-perifériális pont.

Ekkor az eljárás u, illetve v, y, z pontok közül választva, 5, illetve 6 excentricitást egyaránt eredményezhet.

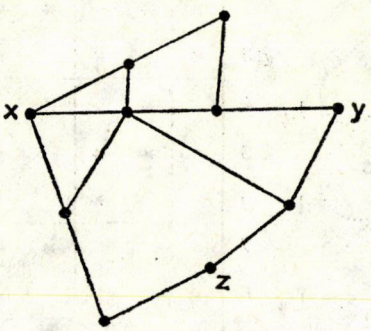
4. MELLÉKLET

A GL—SPS eljárással nyert excentricitás nem egyértelműen meghatározott

Legyen x a kezdőpont (mely nem minimális fokszámú)

$$l(x) = 3; \quad L_{ec}(x) = \{y, z\} \quad \text{és}$$

y, z azonos fokszámú pontok.



Ugyanakkor $l(y)=3$; $l(z)=4$. Így y, illetve z választásától függően az eljárás 3, illetve 4 excentricitást eredményez.

5. MELLÉKLET
Számítógépes eredmények
Perifériális pontok meghatározása

$G=(500, 1500)$

	I. $d=6$		II. $d=6$		III. $d=6$				
	Op	t	Op	t	Op	t	t_a	Op_a	$\frac{Op_a}{5}$
P	447	21,725905	458	22,244785	459	22,314108	22,094932	454,6̊	90,93
P′	317	15,666844	294	14,347027	335	16,354865	15,457507	315,3̊	63,06
P″	465	22,043375	469	22,313483	464	21,733223	22,030027	466,0	93,2
P‴	297	15,024319	194	14,809761	287	14,418746	14,750942	292,6̊	58,53
P1	30	1,468254	47	2,260129	13	0,634557	1,454313	30,0	6,00

$G=(500, 2000)$

	I. $d=5$		II. $d=5$		III. $d=5$				
	Op	t	Op	t	Op	t	t_a	Op_a	$\frac{Op_a}{5}$
P	380	26,149316	376	25,867076	425	29,237638	27,085010	393,6̊	78,73
P′	416	28,748950	414	28,706919	392	27,388144	26,398900	445,3̊	89,06
P″	432	24,511035	430	24,352337	434	24,374785	24,412719	432,0	86,53
P‴	430	26,478951	459	18,242518	437	26,941451	17,220973	442,0	88,4
P1	97	5,688201	174	9,870622	144	8,200102	7,919641	138,3̊	27,66

$G=(500, 2500)$

	I. $d=5$		II. $d=5$		III. $d=5$				
	Op	t	Op	t	Op	t	t_a	Op_a	$\frac{Op_a}{5}$
P	380	26,149316	376	25,867076	425	29,238638	27,085010	393,6	78,73
P′	416	28,748950	414	28,706919	392	27,388144	28,271337	407,3	81,46
P″	396	16,068378	390	25,789993	418	28,850773	26,903048	401,3	80,26
P‴	442	31,312518	414	29,979289	426	30,475825	30,589210	427,3	85,46
P1	51	3,501118	33	2,210520	29	1,977819	2,563152	37,6	7,53

$G=(A, B)$ véletlen gráf; A pontja B éle van.
P perifériális pontot előállító algoritmus (4.2. fejezet).
P′, P″, P‴ P eljárás módosításai (4.2. fejezet).
P1 perifériális pontot előállító heurisztikus algoritmus (4.3. fejezet).
Op műveletigény, melyet a generált szintstrukrúrák számával mérünk.
t CPU-idő (sec).
t_a CPU-idők átlaga.
Op_a átlagos műveletigény (Op-k átlaga).
d gráf átmérője.

6. MELLÉKLET

Gépidő-mérési tapasztalatok

A CPUTIM rutin [65] a két egymásutáni hívása között eltelt CPU-időt környezet-függetlenül regisztrálja egy egész típusú változóban. Egy CPUTIM-egység $26{,}04166 \cdot 10^{-6}$ másodpercnek felel meg.

Vizsgálataink során azt tapasztaltuk, hogy ugyanazon művelet-sorozat CPU-idejének mérésekor a CPUTIM eredményében ingadozás mutatkozik. Az ingadozás mértékének becslésére 5000, 10 000, illetve 15 000 szorzási-műveletet 20-szor ismétlő eljárások CPU-ideinek mérését 10 különböző alkalommal megismételtük. CMS-környezetben a CPUTIM eredményében az átlagtól való eltérések 128 egész számú többszöröseiként léptek fel, s a maximális eltérésre 896 adódott. OS-környezetben kisméretű, szabálytalan ingadozást tapasztaltunk, melynek maximális értéke 57 volt.

Fentiek alapján a CPUTIM által CMS-, illetve OS-környezetben regisztrált gépidőkre rendre 0,023̂, illetve 0,001484 (sec) hibakorlátok tekinthetők érvényeseknek.

7. MELLÉKLET

Számítógépes eredmények
Közel-minimális szélességű szintstruktúra meghatározása

G=	(300, 900)		(300, 1200)		(300, 1500)		(400, 2000)		(600, 3000)		(800, 4800)	
	W	*Op*	*W*	*Op*	*W*	*Op*	*W*	*Op*	*W*	*Op*	*W*	*Op*
GPS—PS	163	2	171	4	178	67	257	28	354	3	550	292
GPS—LS	148	*	140	*	178	*	257	*	306	*	560	*
GL—SPS	160	2	161	3	197	2	250	3	355	3	434	3
I.	151	14	157	71	177	15	228	24	347	137	573	53
II.	138	6(27)	132	4(74)	177	2(24)	227	6(179)	282	5(165)	437	5(296)
III.	132	8(40)	132	4(74)	170	5(28)	227	6(179)	282	5(165)	437	5(296)
IV.	157	8(29)	144	6(76)	189	4(26)	190	8(181)	282	7(167)	437	7(298)
V.	118	10(42)	144	6(76)	189	7(30)	190	8(181)	282	7(178)	437	7(298)
VI.	127	8(29)	182	6(76)	144	4(26)	190	8(181)	282	7(167)	437	7(298)
VII.	127	10(42)	182	6(76)	138	7(30)	190	8(181)	282	7(167)	434	7(298)
P2″	146	4(12)	140	8(12)	178	3(4)	251	4(7)	286	9(25)	462	3(52)
NCC—LS	162	3	181	3	177	3	197	3	294	3	390	3

W — szintstruktúra szélessége.
Op — generált szintstruktúrák száma;
$e(f)$ jelenti: f számú pontból indult RLS (x) generálás, de ebből csak e-számút fejezett be az eljárás.
GPS—LS — általános szintstruktúra előállítását végzi, melynek műveletszáma nem mérhető szintstruktúrában; (* jelölés).
NCC—LS — Saját-fejlesztésű heurisztikus eljárás (l. 8. melléklet)
P2″ — P2 eljárás módosítása (4.4. fejezet).

8. MELLÉKLET

Heurisztikus eljárás nagy excentricitású pontpár meghatározására

Legyen $x \in X$ tetszőleges pont, $G=(X, E)$;

— Képezzük $\mathrm{RLS}(x)=\{L_0(x), L_1(x), \ldots, L_{ec}(x)\}$-t.

— Határozzuk meg $y_i \in L_i(x)$ $(1 \leqq i \leqq l(x))$ pontokat, melyekre

$$|N(y_i) \cap L_i(x)| \leqq \min \{|N(y_i) \cap L_{i-1}(x)|, |N(y_i) \cap L_{i+1}(x)|\}$$

teljesül.

— $i=1, 2, \ldots, l(x)$ esetén a $G(\bigcup_{j=0}^{i} L_j(x))$ metszetgráfon generáljuk $\mathrm{RLS}(y_i)$-t, melynek hosszát h_i-ként jelöljük.

— Az $L=\max_{1 \leqq j \leqq l(x)} (l(x), h_j)$ az eredmény-excentricitás, mely y_s és $t \in L_{ec}(y_s)$ $(0 \leqq s \leqq l(x))$ pontpárt definiál.

Az eljárást s a számítógépes eredményeket *"A new numbering strategy for reducing the bandwidth for sparse symmetric matrices"* című előadásunkban ismertettük. (Fourth Symposium on Basic Problems of Numerical Mathematics, Plsen, 1978,)

9. MELLÉKLET

Sávszélesség/profil-redukciós eljárások
(véletlen gráfokon)
$G=(300, 900)$

	I.			II.			III.		
	b	*pr*	*t*	*b*	*pr*	*t*	*b*	*pr*	*t*
GENRCM	165	30128	0,14341	171	30215	0,14299	163*	30512	0,141832
GPS	169	30291	0,48379	181	30284	0,71874	165	30611	0,48739
GENBRI	163	30444	0,22794	164	30147	0,20005	174	29677*	0,47083
PMM	163	30309	15,39038	164	30115	13,83385	178	30082	21,67752
GENBRM—0	164	29972	0,23041	160	29910	0,36033	174	29677*	0,37458
GENBRM—1	164	29972	0,24955	160	29910	0,36033	174	29677*	0,37459
PPM—0	169	30075	5,00002	162	30027	8,80757	169	30353	11,03645
PPM—1	169	30075	5,00002	172	30027	8,81299	169	30353	11,05995
PPM—2	152*	30702	3,95872	176	31241	11,24465	172	30898	10,60168
BBPRED—0	171	31198	10,48747	162	30027	9,10171	169	30353	11,34538
BBPRED—1	171	31198	10,49989	162	30027	9,10172	169	30353	11,34539
BBPP—0	169	30075	5,18778	148	28422*	8,09525	169	30353	11,37754
BBPP—1	169	30075	5,18778	148	28422*	8,09525	169	30353	11,33754
BBPR—0	170	29909*	0,33961	160	29910	0,61895	174	29677*	0,63729
BBPR—1	170	29909*	0,33972	138*	29556	0,61896	174	29677*	0,63729
BBPP1—0	164	29972	0,39901	160	29910	0,58575	174	29677*	0,71903
BBPP1—1	174	29972	0,39902	138*	29556	0,58576	174	29677*	0,71904
GGLAN	166	30180	2,63677	169	30261	2,23327	171	30472	7,06538
GGLAA	169	30865	1,82934	179	30774	2,48124	168	30039	2,22624

b = sávszélesség
pr = profil
t = CPU-idő (sec)

10. MELLÉKLET

Sávszélesség/profil-redukciós eljárások
(véletlen gráfokon)
$G=(300, 1200)$

	I.			II.			III.		
	b	*pr*	*t*	*b*	*pr*	*t*	*b*	*pr*	*t*
GENRCM	190	32309	0,21205	290	32378	0,17299	190	32770	0,17893
GPS	180*	32251	0,55471	188	32574	0,53542	192	32350	0,54385
GENBRI	186	32587	0,79669	183	32398	1,76052	196	33468	1,36419
PMM	190	32759	18,43082	184	32298	19,76939	190	32913	18,67312
GENBRM—0	195	32563	0,59458	190	32950	0,83325	184	31938*	0,66273
GENBRM—1	195	32563	0,83325	190	32950	0,83326	184	31938*	0,66274
PPM—0	188	32334	12,67124	193	32199	15,85822	184	33451	21,51166
PPM—1	188	32334	12,67126	193	32199	15,89926	184	33451	21,51167
PPM—2	185	32016*	16,94293	191	33058	17,09127	183	32653	13,17955
BBPRED—0	188	32334	13,07424	193	32199	16,33140	191	32650	14,69965
BBPRED—1	188	32334	13,07426	193	32199	16,33145	191	32650	14,82764
BBPP—0	188	32334	13,02093	193	32199	16,32675	184	32212	11,33833
BBPP—1	188	32334	13,03185	193	32199	16,32676	184	32212	11,33834
BBPR—0	195	32563	0,77281	190	32950	0,11809	176*	32417	1,17856
BBPR—1	195	32563	0,77282	190	32950	0,11810	176*	32417	1,17857
BBPP1—0	195	32563	0,92591	190	32950	1,15713	184	31938*	1,01046
BBPP1—1	195	32563	0,93592	190	32950	1,15713	184	31938*	1,01047
GGLAN	190	32337	14,75299	187	32129**	3,47848	188	32921	12,61447
GGLAA	197	32242	8,18575	173*	32289	1,29817	177	32590	12,26924

b = sávszélesség
pr = profil
t = CPU-idő (sec)

11. MELLÉKLET

Sávszélesség/profil-redukciós eljárások
(véletlen gráfokon)
$G=(300, 1500)$

	I.			II.			III.		
	b	*pr*	*t*	*b*	*pr*	*t*	*b*	*pr*	*t*
GENRCM	212	34483	0,19963	208	34768	0,21112	204	34232	0,19895
GPS	213	34472	3,61875	207	34659	2,03182	206	34025*	2,56345
GENBRI	199	33730	0,42289	205	34162	0,42499	200*	34778	0,46382
PMM	200	34037	22,58059	207	34112	24,96929	212	34485	20,93416
GENBRM—0	209	34191	1,55262	208	34279	0,33809	207	34571	0,34296
GENBRM—1	209	34191	1,55264	208	34279	0,33809	207	34571	0,34298
PPM—0	209	34024	15,28947	213	34659	14,49715	203	34221	13,89731
PPM—1	209	34034	15,28949	213	34659	14,49718	203	34221	13,89731
PPMM—2	226	35297	6,07221	223	35772	9,15011	223	35572	10,64395
BBPRED—0	209	34024	22,09168	213	34659	15,53331	203	34221	15,67085
BBPRED—1	209	34024	22,09168	213	34659	15,53689	203	34221	15,67088
BBPP—0	220	35243	15,32611	202	34954	22,70187	203	34221	14,82119
BBPP—1	220	35243	15,32612	202	34954	22,70189	213	34221	14,82119
BBPR—0	209	34191	0,83039	208	34279	1,90622	207	34571	0,72505
BBPR—1	209	34191	0,83039	308	34279	1,90623	207	34571	0,72507
BBPP1—0	190	33363	1,72734	189*	33730*	0,53585	207	34571	0,78731
BBPP1—1	185*	32549*	1,93549	189*	33730*	0,53585	107	34571	0,78732
GGLAN	209	34541	1,40604	209	34937	1,83033	202	34429	2,28249
GGLAA	219	35534	2,16325	209	35310	1,83325	216	35407	2,65851

b = sávszélesség
pr = profil
t = CPU-idő (sec)

12. MELLÉKLET

Sávszélesség/profil-redukciós eljárások
(véletlen gráfokon)
$G = (500, 1500)$

	I.			II.			III.		
	b	*pr*	*t*	*b*	*pr*	*t*	*b*	*pr*	*t*
GENRCM	262	92835	0,28020	285	93140	0,27432	268	93217	0,27575
GPS	276	92934	0,89735	275	93174	1,34564	274	93311	0,85389
GENBRI	275	92661	1,50356	294	93353	2,27619	276	93090	0,75109
PMM	277	92757	34,90749	270	93163	22,87686	277	93398	34,86374
GENBRM—0	269	93095	0,88624	267	92647	1,33891	276	92568	0,53255
GENBRM—1	249	91895	0,88625	267	92647	1,33892	276	92568	0,53256
PPM—0	263	92703	11,72382	275	93509	23,45957	278	93032	35,50889
PPM—1	233	92703	11,83195	275	93509	23,45959	278	93032	35,50889
PPM—2	236	93156	3,23192	253	93868	23,68561	294	95004	29,46434
BBPRED—0	263	93464	23,25551	249	92327	31,16267	278	93042	25,74960
BBPRED—1	263	93464	23,25614	249	92327	31,16268	278	93032	25,92345
BBPP—0	273	92703	12,27210	275	93509	24,19015	287	92582	31,59436
BBPP—1	263	92703	12,27210	275	93509	24,19017	287	93282	21,59436
BBPR—0	228*	91891*	1,22645	217*	91081*	1,65534	276	92568	1,03450
BBPR—1	228*	91891*	1,22647	217*	91081*	1,65537	276	92568	1,03452
BBPP1—0	269	93095	1,18865	267	92647	1,63624	231*	92422*	0,86703
BBPP1—1	269	93095	1,18867	267	92647	1,63624	231*	92422*	0,86705
GGLAN	269	93413	5,37891	269	93574	17,38456	262	93215	2,84665
GGLAA	269	93821	4,32981	254	94198	21,01447	271	94556	4,53216

b = sávszélesség
pr = profil
t = CPU-idő (sec)

13. MELLÉKLET

Sávszélesség/profil-redukciós eljárások
(véletlen gráfokon)
$G=(500, 2000)$

	I.			II.			III.		
	b	*pr*	*t*	*b*	*pr*	*t*	*b*	*pr*	*t*
GENRCM	321	97303	0,27329	306*	96752	0,34716	308	97601	0,35278
GPS	315	95641	0,83453	307	96764*	1,98432	321	97847	2,04216
GENBRI	314	98171	5,53359	316	97218	9,78596	320	97607	8,11512
PMM	314	98125	74,08167	318	97801	95,31234	323	97493	85,02914
GENBRM—0	313	97907	2,22794	317	97331	3,49713	313	97097	3,06083
GENBRM—1	313	97907	2,22794	317	97331	3,49714	313	97097	3,06087
PPM—0	312	97513	54,58769	320	97517	64,14519	305*	96940	54,77282
PPM—1	312	97513	54,58769	320	97517	64,14519	305*	96940	54,77286
PPM—2	279	97541	27,70593	320	97528	64,14521	322	96082*	66,05073
BBPRED—0	312	97513	54,97923	320	97517	64,95881	305*	96940	55,87055
BBPRED—1	312	97513	54,98131	320	97517	64,96134	305*	96940	55,87134
BBPP—0	296	96812	54,80495	321	97517	65,04671	305*	96940	55,71713
BBPP—1	296	96812	54,81341	320	97517	56,04682	305*	96940	55,71713
BBPR—0	313	96906	2,77403	317	97331	4,01923	313	97097	4,02246
BBPR—1	313	97907	2,77502	317	97331	4,02001	313	97097	4,02302
BBPP1—0	273*	95411*	2,52534	317	97331	4,05736	313	97097	4,05640
BBPP1—1	273*	95411*	2,52544	317	97331	4,05801	313	97097	4,05641
GGLAN	321	96835	56,05871	314	96872	54,78615	330	97191	72,63541
GGLAA	353	97672	35,51672	314	97434	18,83056	342	98050	48,94532

b = sávszélesség
pr = profil
t = CPU-idő (sec)

14. MELLÉKLET

Sávszélesség/profil-redukciós eljárások
(véletlen gráfokon)
($G = 500, 2500$)

	I.			II.			III.		
	b	*pr*	*t*	*b*	*pr*	*t*	*b*	*pr*	*t*
GENRCM	351	101046	0,33645	331	99907	0,32416	343	100754	0,32942
GPS	356	101402	1,83953	307*	98764*	1,38352	321	100817	0,94246
GENBRI	339	100653	3,51278	351	101174	2,32260	336	100732	2,06294
PMM	338	100082	57,74076	354	104812	70,49940	324	99729*	67,08185
GENBRM—0	351	101046	4,69864	345	100541	4,26919	336	100732	1,03656
GENBRM—1	351	101046	4,69871	345	100541	4,26922	336	100732	1,03657
PPM—0	333	100840	53,47279	341	100674	49,34871	335	100423	40,54910
PPM—1	333	100840	54,47279	341	100674	49,34872	335	100423	40,54911
PPM—2	336	101178	82,10268	331	101213	80,60115	322	100957	69,00621
BBPRED—0	335	100849	79,60812	322	101045	73,93836	314*	100160	64,73485
BBPRED—1	335	100849	69,60815	322	101045	73,93911	314*	100160	64,73487
BBPP—0	333	100840	54,48771	341	100674	50,45096	335	100423	41,33694
BBPP—1	333	100840	54,48781	341	100674	51,01624	335	100423	41,33694
BBPR—0	313*	99745*	5,09436	308	98869	4,69377	315	100012	1,33526
BBPR—1	313*	99745*	5,09436	308	98869	4,69378	315	100012	1,33611
BBPP1—0	351	101046	5,34935	345	100541	4,91046	336	100732	1,66335
BBPP1—1	351	101046	5,34941	345	100541	4,91048	336	100732	1,66338
GGLAN	333	100840	48,94361	342	101098	34,56795	341	100360	32,16325
GGLAA	336	101078	77,55413	325	101846	66,56419	322	101687	61,76874

b = sávszélesség
pr = profil
t = CPU-idő (sec)

15. MELLÉKLET

Sávszélesség/profil-redukciós eljárások
(véletlen gráfokon)
$G=(1000, 3000)$

	I.			II.			III.		
	b	*pr*	*t*	*b*	*pr*	*t*	*b*	*pr*	*t*
GENRCM	544	423316	0,44218	536	422177	0,44234	557	421478*	0,44101
GPS	546	423514	5,34721	531	422004	7,39469	548	422124	9,34192
GENBRI	546	422438	0,72401	569	422088	0,58638	541	422072	24,95364
PMM	561	422862	166,54216	557	422202	166,90734	541	422022	159,23465
GENBRM—0	540	422109	1,42865	556	422880	0,92532	538	421864	10,16284
GENBRM—1	537	421707	2,49587	523*	421810	1,63874	538	421864	10,26284
PPM—0	539	421919	180,78536	541	422697	183,73411	530*	421907	190,56341
PMM—1	527	421873	165,36475	428	421924	180,01634	530*	421907	190,56351
PPM—2	555	421777	240,34061	554	422634	235,18347	553	422032	231,27643
BBPRED—0	539	421919	280,93541	538	421501*	182,30021	530*	421907	191,36874
BBPRED—1	527	421873	166,31819	528	421924	181,63974	530*	421907	191,36973
BBPP—0	539	421919	198,19056	541	422697	185,30401	530*	421907	193,53019
BBPP—1	453	420877	124,73815	528	421924	182,53190	530*	421907	195,63941
BBPR—0	540	422109	2,25391	556	422880	1,76874	538	421864	1,43561
BBPR—1	537	421707	3,33196	523*	421810	1,99083	538	421864	1,43564
BBPP1—0	540	422109	2,34539	556	422880	2,38392	538	421864	11,46532
BBPP1—1	433*	419866*	3,15674	523*	421810	1,83910	538	421864	11,46534
GGLAN	538	423273	195,37832	549	422107	106,73811	556	422078	206,30919
GGLAA	567	423055	211,83721	586	423051	59,59674	568	422184	73,53194

b = sávszélesség
pr = profil
t = CPU-idő (sec)

16. MELLÉKLET

Sávszélesség/profil-redukciós eljárások
(véletlen gráfokon)
$G=(1000, 4000)$

	I.			II.			III.		
	b	*pr*	*t*	*b*	*pr*	*t*	*b*	*pr*	*t*
GENCRM	625	428665	0,54521	618	427747	0,55337	626	427974	0,55124
GPS	628	428876	1,93451	622	428312	16,28465	622	428342	19,93472
GENBRI	609	426632*	0,97894	630	428141	7,07031	622	427681	2,88634
PMM	608*	427308	261,95465	638	428323	201,10833	619*	428033	470,63185
GENBRM—0	635	427961	0,89985	613*	427410*	4,29319	622	427681	2,43586
GENBRM—1	635	427961	0,89986	613*	427410*	4,29401	622	427681	2,43587
PPM—0	628	427778	364,14985	634	428136	405,27364	628	427087*	387,25364
PPM—1	628	427778	364,14991	634	428136	405,28012	628	427087*	387,26110
PPM—2	656	431091	320,77861	677	430264	332,73160	674	430264	333,16225
BBPRED—0	628	427778	360,80715	634	428136	407,36541	628	427087*	389,76874
BBPRED—1	628	427778	360,80718	634	428136	409,78162	628	427087*	390,12341
BBPP—0	638	427778	361,17072	634	428136	108,15341	628	427087*	290,12344
BBPP—1	638	427778	361,17072	634	428136	108,15341	628	427087*	290,12344
BBPR—0	635	427961	1,43952	613*	427410*	5,03252	622	427681	3,09132
BBPR—1	635	427961	1,43953	613*	427410*	5,83475	622	427681	3,09132
BBPP1—0	635	427961	2,09134	613*	427410*	5,65311	622	427681	3,64789
BBPP1—1	635	427961	2,93834	613*	427410*	5,68419	622	427681	3,65792
GGLAN	629	428656	9,59647	616	427619	26,83745	635	428674	26,21092
GGLAA	648	429229	18,36847	651	429987	26,93841	649	428311	20,38341

b = sávszélesség
pr = profil
t = CPU-idő (sec)

17. MELLÉKLET

Sávszélesség/profil-redukciós eljárások
(véletlen gráfokon)
$G = (1000, 5000)$

	I.			II.			III.		
	b	*pr*	*t*	*b*	*pr*	*t*	*b*	*pr*	*t*
GENRCM	683	434066	0,83116	682	433397	0,81653	689	433761	1,68324
GPS	681	434294	3,19384	684	433295	3,29735	671	433651	3,23465
GENBRI	669	432685	0,85101	696	433499	27,00016	690	433981	3,85195
PMM	671	433392	292,91365	685	433510	100,43538	692	433995	296,46871
GENBRM—0	687	432636*	1,45931	674	434144	1,33706	670*	433532	1,48374
GENBRM—1	687	432636*	1,45993	674	434144	1,33707	670*	433532	1,48574
PPM—0	673	433049	436,73519	686	432867	468,21871	688	433098	443,53641
PPM—1	673	433049	436,73519	686	432867	468,21881	688	433098	443,54501
PPM—2	709	434925	536,19874	624*	432005*	435,73819	694	433901	476,76819
BBPRED—0	673	433049	448,01982	686	432867	471,31015	688	433098	445,38412
BBPRED—1	673	433049	448,01984	686	432867	473,53192	688	433098	445,39421
BBPP—0	671	433827	465,19324	686	432867	469,29345	688	433098	445,39422
BBPP—1	671	433827	465,19410	686	432867	469,29346	688	433098	445,39424
BBPR—0	687	432636*	1,51637	674	434144	14,72835	670*	433432	1,74839
BBPR—1	687	432636*	1,51638	674	434144	14,72841	670*	433532	1,93843
BBPP1—0	622*	432899	1,59347	674	434144	15,61342	670*	433532	2,45902
BBPP1—1	622*	432899	1,60219	674	434144	15,91398	670*	433532	2,98341
GGLAN	690	432830	558,71315	674	432580	478,53915	699	432577*	391,73831
GGLAA	757	435524	373,53941	700	434955	397,70118	769	434480	155,91354

b = sávszélesség
pr = profil
t = CPU-idő (sec)

18. MELLÉKLET

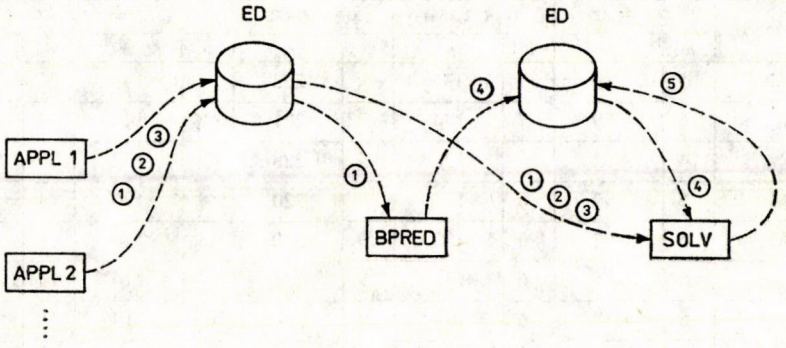
Sávszélesség/profil-redukciós eljárások
(kisméretű, síkbeli gráfokon)

	$G=(24,82)$		$G=(32,126)$		$G=(42,162)$		$G=(51,136)$		$G=(54,248)$	
	b	*pr*	*b*	*pr*	*b*	*pr*	*b*	*pr*	*b*	*pr*
GENRCM	5	89	7	135	9	223	8	145	21	425
GENBRI	5	89	7	134	9	221	8	145	22	422
PMM	4	82	6	133	9	228	9	192	15	359
GENBRM—0	5	93	6	133	9	220	8	170	14	320
GENBRM—1	5	93	6	133	9	220	8	170	14	320
PPM—0	5	85	6	136	9	231	10	214	14	310
PPM—1	5	83	6	134	9	231	10	214	14	310
PPM—2	5	93	6	135	8	209	8	174	13	324
BBPRED—0	4	82	6	134	7	216	7	204	11	299
BBPRED—1	4	82	6	134	7	216	7	204	11	299
BBPP—0	5	85	6	134	9	255	9	192	14	310
BBPP—1	5	85	6	134	8	250	9	192	14	310
BBPR—0	4	82	6	133	7	211	8	191	14	328
BBPR—1	4	82	6	133	7	211	8	191	14	328
BBPP1—0	5	89	6	133	9	272	8	170	14	320
BBPP1—1	5	89	6	133	9	272	8	170	14	320
GGLAN	5	93	7	145	9	227	8	176	15	371
GGLAA	5	86	7	134	9	216	8	146	19	392
CUTHIL	4	83	6	132	7	215	7	220	14	496
GPS	5	95	9	185	9	220	8	169	14	346
OURM	4	83	6	132	7	210	7	210	12	397
ROSEN	4	83	6	133	7	215	10	230	22	425
NCC	4	93	6	132	7	213	10	195	18	441

b = sávszélesség
pr = profil
t = CPU-idő (sec)

19. MELLÉKLET

A program-rendszer működésének sematikus vázlata



APPL 1

Alkalmazási feladatok
$\mathbf{Ax}=\mathbf{b}$ Egyenletrendszer összeállítása
Output: ① mátrix tömör alakja
② jobb oldal
③ segédinformációk

BPRED

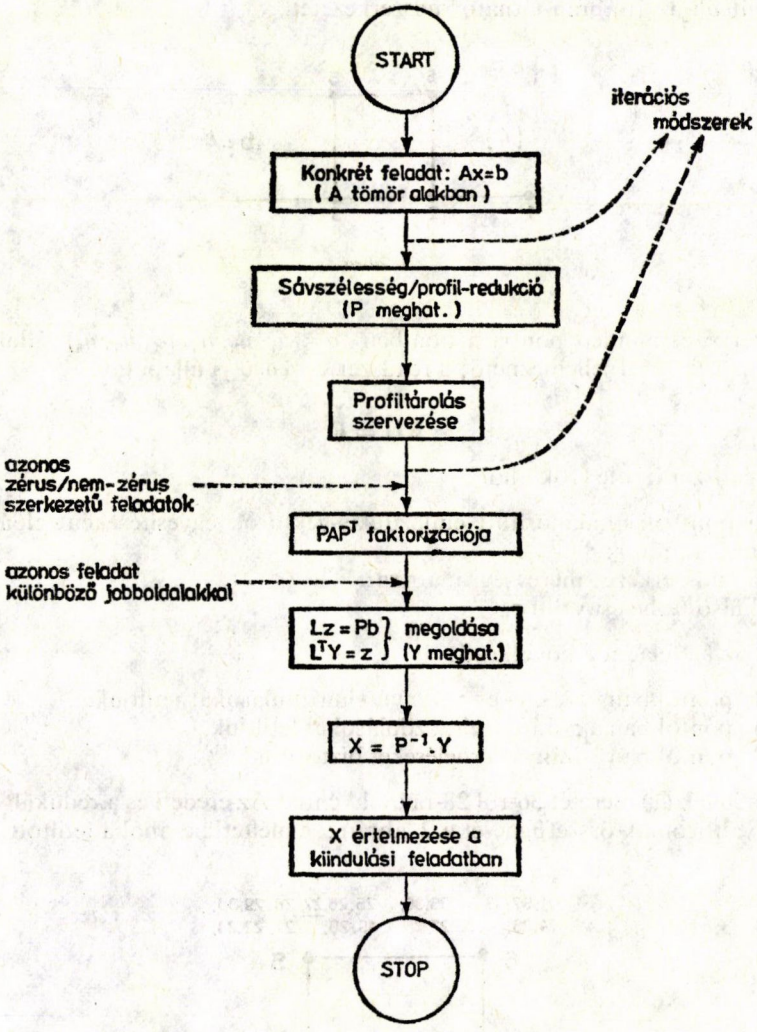
Sávszélesség/profil redukció
$\mathbf{P}$ permutációs mátrix meghatározása
Output: ④ permutációs tömbök (PERM, INUP)

SOLV

$(\mathbf{PAP}^T)\,\mathbf{y}=\mathbf{pb}$
egyenletrendszer megoldása
— $\mathbf{x}=\mathbf{p}^{-1}\mathbf{y}$ előállítása
— a segédinformációs révén $\mathbf{x}$-ből $\bar{\mathbf{x}}$ előállítása
Output: ⑤ az alkalmazási feladat $\bar{\mathbf{x}}$ megoldása.

20. MELLÉKLET

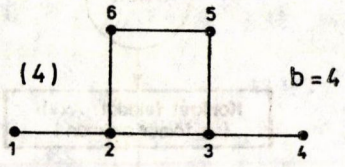
A programrendszer belépési pontjai



21. MELLÉKLET

Rúdszerkezet statikai számítása

Tekintsük az 1. ábrán látható rúdszerkezetet.



1. ábra

A szerkezet minden pontja a pontbeli $\mathbf{u}=(u_1, u_2, u_3, u_4, u_5, u_6)$ általánosított elmozdulás-vektorral jellemezhető; s rendszer egyensúlyi állapotát

$$\mathbf{KU} = \mathbf{F} \tag{1}$$

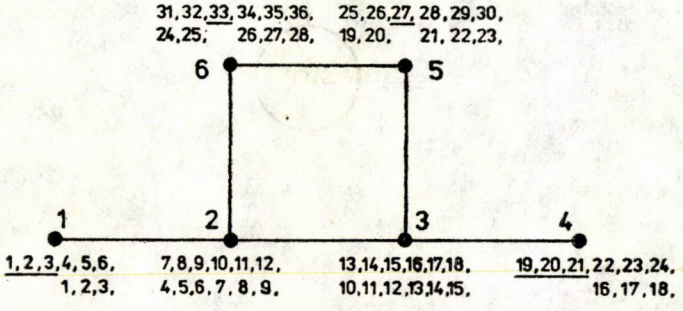
egyenletrendszer írja le [10], ahol

- **U** a pontbeli általánosított elmozdulás-vektorok egyesítéseként előálló ismeretlen vektor;
- **K** a rúdszerkezet merevségi mátrixa;
- **F** külső terhelési vektor.

Legyen a fizika feladat a következő:

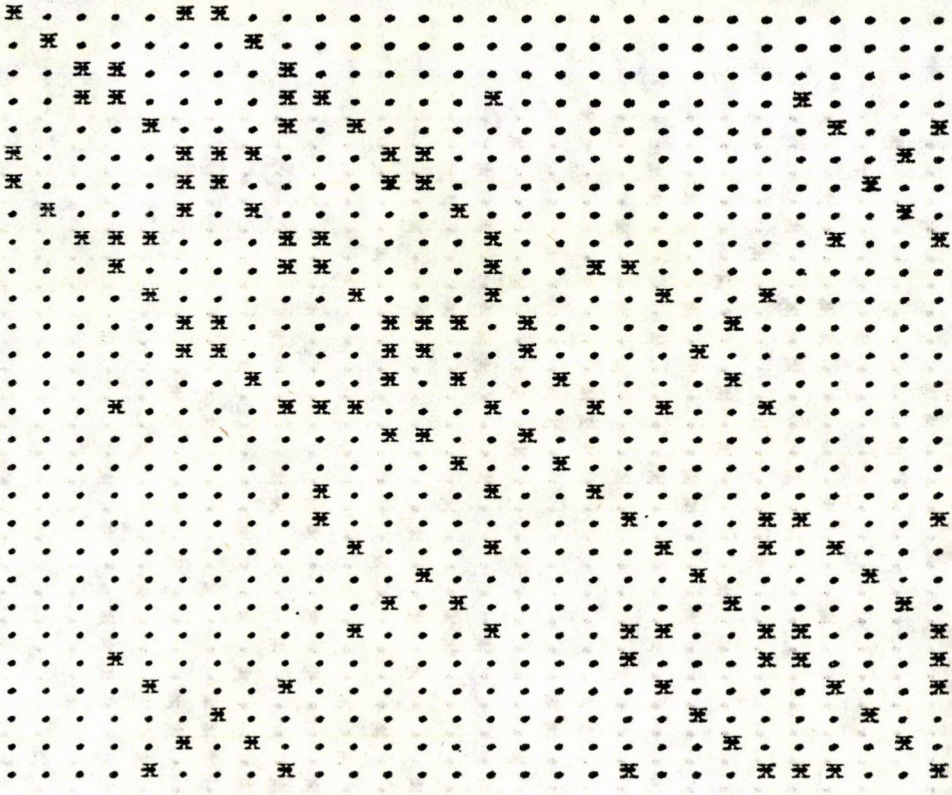
— 1 és 4 pontokban az x-, y- és z-irányú elmozdulásokat letiltjuk.
— 5 és 6 pontokban a z-irányú elmozdulásokat letiltjuk.
— 2 és 3 pontokban z-irányú terheléseket biztosítunk.

E feltételek (1) méretét 36-ról 28-ra csökkentik. Az eredeti és a redukált egyenletrendszer változóinak összefüggését a 2. ábrán szemléltetjük, ahol a letiltott



2. ábra

komponenseket aláhúzással jelöljük. A redukált rendszer merevségi mátrixának zérus/nem-zérus szerkezetét a 3. ábrán szemléltetjük, ahol $b=23$. Vagyis, azáltal, hogy egy ponthoz több paramétert rendelünk, a merevségi mátrix elemei „szétszóródnak".

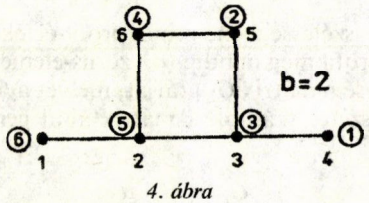


3. ábra

$$b = 23$$

$$pr = 232$$

Lássuk el a rúdszerkezetet sávszélesség-redukció szerinti optimális számozással, melyet a 4. ábrán a csúcsok mellett feltüntetett bekarikázott számokkal jelölünk.

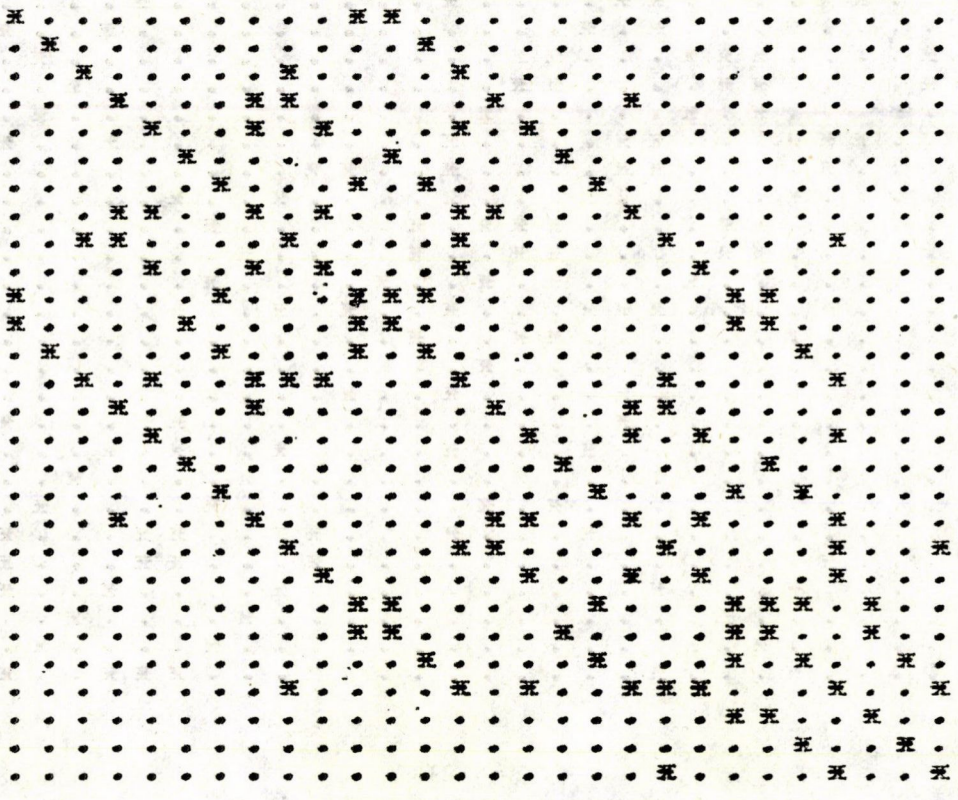


4. ábra

A megfelelő merevségi mátrix zérus/nem-zérus szerkezetét az 5. ábrán közöljük.

$$b = 16$$

$$pr = 204$$

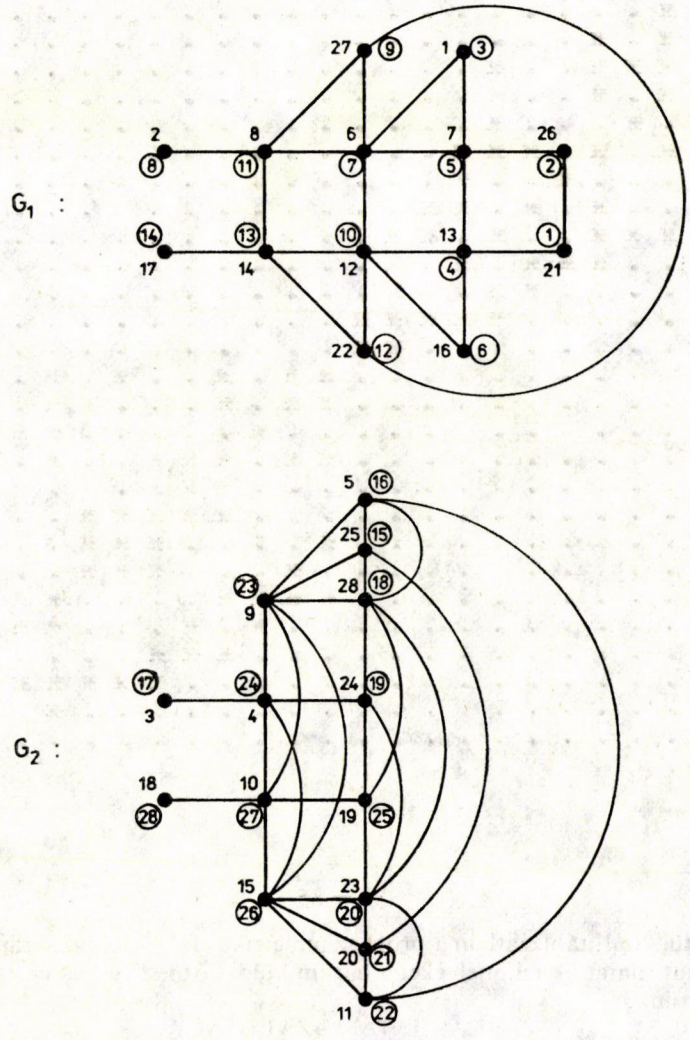


5. ábra

Láthatóan, mind sávszélességben, mind profil-értékben jelentős csökkenés tapasztalható, azonban a profil még mindig sok zérus-elemet tartalmaz.

Tekintsük a **K** merevségi mátrix G_k gráfját, melyet a 6. ábrán szemléltetünk, ahol a csúcsok melletti természetes számok a mátrix által generált, „eredeti" számozást jelölik.

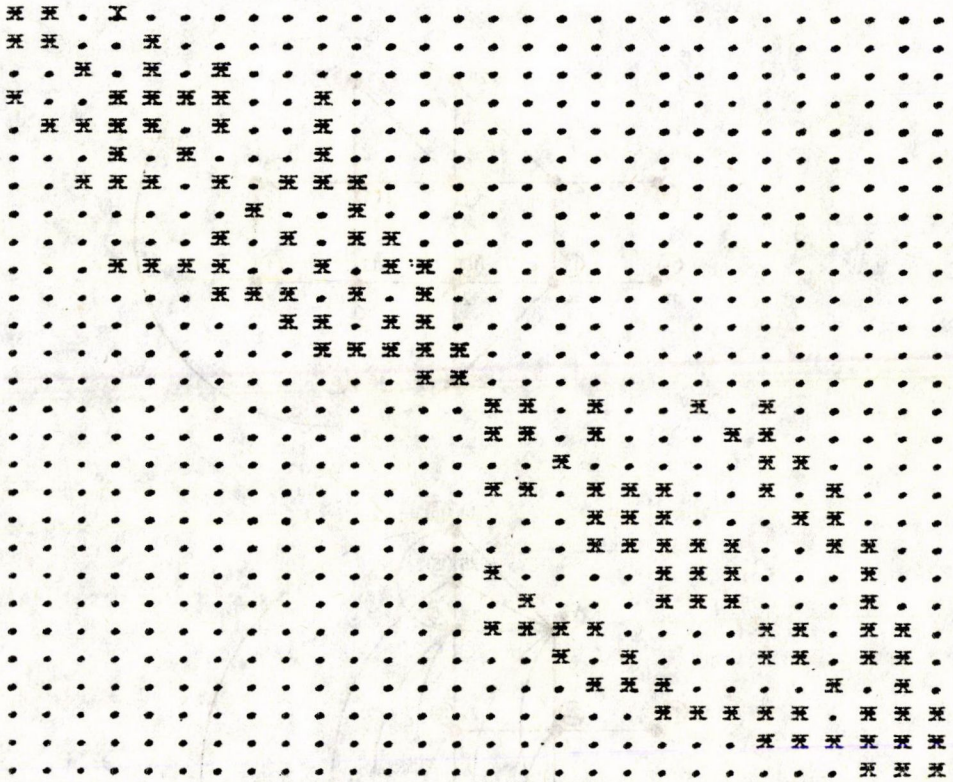
$$G_k = G_1 \cup G_2$$



6. ábra

Alkalmazva a $G_k = G_1 \cup G_2$ gráfra GENBRM—0 eljárásunkat, az így nyert „új" számozást a 6. ábrán bekarikázott számokkal tüntetjük fel. Az új számozással ellátott G_k gráf összefüggési mátrixának zérus/nem-zérus szerkezetét a 7. ábrán szemléltetjük. Világosan látható, a sávszélességben és profil-értékben mutatkozó radikális csökkenés.

A gráf számozásával a szerkezet különböző pontjainak paraméterei közti kapcsolatok is felhasználást nyernek, ezáltal sok esetben a bemutatotthoz hasonló, szembetűnő javulást mutatnak az eredmények.



7. ábra

$$b = 8$$

$$pr = 85$$

Végezetül, az 1. táblázatban a profil-szimmetrikus faktorizáció során adódó műveletszámokat tüntetjük fel, melyeket a három különböző sávszélesség és profil-érték mellett nyertünk.

1. TÁBLÁZAT

	b	pr	Műveletszám
Eredeti számozás	23	232	1971
Szerkezet optimális számozása	16	204	1574
G_k sávszélesség-csökkentő számozása	8	85	486

A sávszélesség/profil-redukcióval a műveletszámban közel 75%-os csökkenés tapasztalható. Következésképpen, a fenti utat követve, nagyméretű feladatok esetén jelentős gépidő-megtakarítás érhető el, s a műveletszám csökkenése egyben az aritmetikai pontosság javulását is eredményezi.

Alkalmazott Matematikai Lapok **11** (1985) 91—121

LINEÁRIS RENDSZEREK ÉS POLINOMMÁTRIXOK

G. VÁGÓ ZSUZSA

Budapest

Az irányítási rendszerek elméletének fontos fejezete a lineáris rendszerek vizsgálata az ún. frekvenciatartományban. A frekvenciatartományban értelmezzük az átviteli függvényt, amely lineáris rendszerek esetén racionális törtekből felépített mátrix. Ismertetjük a racionális törtek, ill. a polinommátrixok algebrai elméletének legfontosabb elemeit (legnagyobb közös osztó, jobb-bal törtek, *Smith—McMillan alak*, valódiság, pólus-zérus) és ezek kapcsolatát az időtartományban bevezetett fogalmakkal (irányíthatóság, megfigyelhetőség, realizáció). A dolgozatban bemutatott módszerek többdimenziós idősorok elemzésénél alkalmazásra kerültek.

1. Bevezetés

Dolgozatunkban állandó együtthatós, lineáris rendszerekkel fogunk foglalkozni. Ezeket az időtartományban a következőképpen írjuk le:

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t) \qquad \mathbf{x}(0) = \mathbf{x}_0, \tag{1.1}$$

$$\mathbf{y}(t) = \mathbf{C}\mathbf{x}(t), \tag{1.2}$$

ahol $\mathbf{x}(t) \in R^n$, $\mathbf{u}(t) \in R^m$, $\mathbf{y}(t) \in R^p$, $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ megfelelő alakú konstans mátrixok. $\mathbf{x}$ dimenzióját a rendszer dimenziójának nevezzük.

Vizsgálhatjuk a fenti rendszer *Laplace transzformáltját* is, ezt frekvenciatartománybeli leírásnak nevezzük. Ez a következő alakú ($\mathbf{x}_0 = 0$ esetben):

$$(s\mathbf{I} - \mathbf{A})\tilde{\mathbf{x}}(s) = \mathbf{B}\tilde{\mathbf{u}}(s) \tag{1.3}$$

$$\tilde{\mathbf{y}}(s) = \mathbf{C}\tilde{\mathbf{x}}(s) \tag{1.4}$$

ahol $\tilde{\mathbf{x}}$, $\tilde{\mathbf{u}}$, $\tilde{\mathbf{y}}$ rendre az $\mathbf{x}$, $\mathbf{u}$, $\mathbf{y}$ folyamatok *Laplace transzformáltjait* jelölik.

1.1. Definíció. A fenti lineáris rendszerhez tartozó rendszer-mátrix (system-mátrix) a következő blokk-mátrix:

$$\mathbf{P} = \begin{pmatrix} s\mathbf{I} - \mathbf{A} & -\mathbf{B} \\ \mathbf{C} & \mathbf{O} \end{pmatrix}. \tag{1.5}$$

1.2. Definíció. A fenti lineáris rendszerhez tartozó transfer-függvény (átmeneti függvény) a következő:

$$\mathbf{H}(s) = \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}. \tag{1.6}$$

A rendszer-mátrix a következő kapcsolat leírására szolgál:

$$\mathbf{P}\begin{pmatrix}\tilde{\mathbf{x}}(s)\\ \tilde{\mathbf{u}}(s)\end{pmatrix}=\begin{pmatrix}\mathbf{0}\\ \tilde{\mathbf{y}}(s)\end{pmatrix}.$$

A transfer függvény az input és output folyamatok *Laplace transzformáltjai* közötti kapcsolatot adja meg, nevezetesen könnyen ellenőrizhető, hogy

$$\tilde{\mathbf{y}}(s)=\mathbf{C}(s\mathbf{I}-\mathbf{A})^{-1}\mathbf{B}\tilde{\mathbf{u}}(s).$$

A lineáris rendszerekkel kapcsolatosan az irányíthatóság, megfigyelhetőség és minimális realizáció problémájával fogunk foglalkozni. Most először definiáljuk a fogalmakat, majd a kapcsolódó klasszikus tételeket mondjuk ki bizonyítás nélkül. Részletes tárgyalása a témának [2]-ben található.

1.3. Definíció. Az (1.1), (1.2) rendszert irányíthatónak nevezzük, ha tetszőleges ξ_0, $\xi_1\in R^n$ esetén létezik $0\leqq T<\infty$ és létezik $\mathbf{u}:[0,T]\to R^m$ szakaszonként folytonos irányítás, hogy $\mathbf{x}(0)=\xi_0$ és $\mathbf{x}(T)=\xi_1$.

1.1. TÉTEL. Az (1.1), (1.2) rendszer pontosan akkor irányítható, ha a $\mathscr{C}=(\mathbf{B}, \mathbf{AB}, \ldots, \mathbf{A}^{n-1}\mathbf{B})$ mátrix teljes rangú.

Megjegyzés. Az (1.1), (1.2) rendszer irányíthatósága helyett sokszor az (**A**, **B**) pár irányíthatóságáról beszélünk, hiszen ez a két mátrix határozza meg a rendszer irányíthatóságát.

1.4. Definíció. Az (1.1), (1.2) rendszert megfigyelhetőnek nevezzük, ha tetszőleges $T>0$ esetén az $\mathbf{u}(t)$, $\mathbf{y}(t)$, $t\in[0,T]$ input-output folyamatok ismeretében $\mathbf{x}(0)$ és így a teljes $\mathbf{x}(t)$ $t\in[0,T]$ trajektória meghatározható.

1.2. TÉTEL. Az (1.1), (1.2) rendszer pontosan akkor megfigyelhető, ha az

$$\mathscr{O}=\begin{pmatrix}\mathbf{C}\\ \mathbf{CA}\\ \mathbf{CA}^{n-1}\end{pmatrix}$$

mátrix teljes rangú.

Megjegyzés. Az (1.1), (1.2) rendszer megfigyelhetősége helyett a (**C**, **A**) pár megfigyelhetőségéről beszélhetünk hisz ez a két mátrix határozza meg a rendszer megfigyelhetőségét.

1.5. Definíció. Tetszőleges $\mathbf{H}(s)$ transfer-függvény esetén az (1.1), (1.2) rendszer ennek realizációja, ha $\mathbf{H}(s)=\mathbf{C}(s\mathbf{I}-\mathbf{A})^{-1}\mathbf{B}$. A rendszer minimális realizáció, ha az összes lehetséges realizáció között az állapottér minimális dimenziójú.

Megjegyzés. Minimális realizáció alatt sokszor a megfelelő (**A**, **B**, **C**) hármast értjük.

1.3. TÉTEL. Az (1.1), (1.2) rendszer pontosan akkor minimális realizációja a $\mathbf{H}(s)$ transfer-függvénynek, ha irányítható és megfigyelhető.

1.1. LEMMA. Legyen (**A**, **B**, **C**) minimális realizáció. Ekkor tetszőleges **T** R^n-beli bázistranszformáció esetén a transzformáció után kapott $(\mathbf{T}^{-1}\mathbf{AT}, \mathbf{T}^{-1}\mathbf{B}, \mathbf{CT})$ realizáció is minimális lesz.

A lemma állítása magától értetődő. Érdekes a fordított állítás:

1.4. TÉTEL. Legyenek $(\mathbf{A}_1, \mathbf{B}_1, \mathbf{C}_1)$ és $(\mathbf{A}_2, \mathbf{B}_2, \mathbf{C}_2)$ ugyanannak a $\mathbf{H}(s)$ transfer függvénynek a minimális realizációi. Ekkor létezik olyan $\mathbf{T}$ nemszinguláris mátrix, hogy $\mathbf{A}_1=\mathbf{T}^{-1}\mathbf{A}_2\mathbf{T}$, $\mathbf{B}_1=\mathbf{T}^{-1}\mathbf{B}_2$, $\mathbf{C}_1=\mathbf{C}_2\mathbf{T}$.

2. Polinommátrixok. A Smith-alak, invariáns polinomok

Azt mondjuk, hogy az $\mathbf{A}(s)$ $m\times r$-es mátrix polinommátrix, ha elemei polinomok.

2.1. Definíció. Az $\mathbf{A}(s)$ $m\times m$-es polinommátrix unimoduláris, ha létezik az inverze és ez szintén polinommátrix. Ez nyilván azt jelenti, hogy $|\mathbf{A}(s)|$ konstans nem nulla polinom, hiszen tetszőleges mátrix inverzét megkaphatjuk, ha az adjungált mátrix elemeit a determinánssal elosztjuk.

2.2. Definíció. Legyenek $\mathbf{A}(s)$ és $\mathbf{B}(s)$ $m\times$r-es polinommátrixok. $\mathbf{A}(s)$ és $\mathbf{B}(s)$ ekvivalensek, ha léteznek olyan $\mathbf{U}_1(s)$ és $\mathbf{U}_2(s)$ $m\times m$-es, illetve $r\times r$-es unimoduláris polinommátrixok, melyekre $\mathbf{A}(s)=\mathbf{U}_1(s)\mathbf{B}(s)\mathbf{U}_2(s)$.

2.3. Definíció. Legyen adott egy polinom-mátrix. Elemi sor-transzformáción a következő transzformációkat értjük:

— i-edik és j-edik sor felcserélése,

— i-edik sorhoz a j-edik sor $b(s)$-szeresének hozzáadása, ahol $b(s)$ tetszőleges polinom,

— i-edik sor szorzása tetszőleges $c\neq 0$ skalárral.

Hasonlóan értelmezzük az elemi oszloptranszformációkat.

2.1. LEMMA. Tetszőleges $\mathbf{A}(s)$ $m\times r$-es polinommátrix esetén az elemi sor-(oszlop-)transzformációk $m\times m$-es ($r\times r$-es) unimoduláris mátrixszal való bal-(jobb-) szorzásnak felelnek meg, a mátrix determinánsát skalárszorzó erejéig változatlanul hagyják.

Bizonyítás. Könnyen látható, hogy az elemi sor transzformációk az alábbi mátrixokkal való balszorzást jelentenek:

$$\mathbf{S}_1 = \begin{pmatrix} 1 & & & & \\ & 0 & 1 & & \\ & 1 & 0 & & \\ & & & 1 & \\ & & & & \ddots \end{pmatrix} \begin{matrix} \\ \cdot i \\ \cdot j \\ \\ \\ \end{matrix} \qquad \mathbf{S}_2 = \begin{pmatrix} 1 & & & & \\ & 1 & & 0 & \\ & & b(s) & & \\ & 0 & & 1 & \\ & & & & \ddots \end{pmatrix} \begin{matrix} \\ \\ \cdot i \\ \\ \\ \end{matrix} \qquad i \neq j$$

$$\begin{matrix} & & i & & j & & \\ & & & & & & \\ & & & & & & \\ & & & & & & \\ & & & & & & \end{matrix}$$

$$\mathbf{S}_3 = \begin{pmatrix} 1 & & & & \\ & 1 & & 0 & \\ & & c & & \\ & 0 & & 1 & \\ & & & & \ddots \end{pmatrix} \begin{matrix} \\ \\ \cdot i \\ \\ \\ \end{matrix}.$$

Az S_i mátrixok mindegyike unimoduláris (determinánsa skalár).

2.4. Definíció. Az $A(s)$ $m\times r$-es polinommátrix kanonikusan diagonális, ha $a_{ii}(s)=a_i(s)$ 1 főegyütthatójú polinom, $i=1,\ldots,p$, $a_i(s)|a_{i+1}(s)$ $i=1,\ldots,p-1$, az összes többi eleme 0. (l. GANTMACHER [1].)

2.1. TÉTEL. Tetszőleges $A(s)$ $m\times r$-es polinommátrix ekvivalens egy kanonikusan diagonális polinommátrixszal.

Bizonyítás. Elemi sor- és oszloptranszformációkkal a mátrixot kanonikusan diagonális alakra hozzuk. A konstrukció során a mátrixot végig $A(s)$-sel jelöljük, elemeit $a_{ij}(s)$-sel. A konstrukció lépései a következők:

1. Sor- és oszlopcserékkel elérjük, hogy $a_{11}(s)$ egy minimális fokú nem nulla polinom legyen.
2. Az első sor és az első oszlop elemeit maradékosan elosztjuk $a_{11}(s)$-sel:
$$a_{1k} = q_{1k}a_{11}+r_{1k} \quad k = 2, \ldots, r$$
$$a_{i1} = q_{i1}a_{11}+r_{i1} \quad i = 2, \ldots, m.$$
Ha létezik $r_{1k}\neq 0$, akkor az első oszlop q_{1k}-szorosát vonjuk le a k-adik oszlopból. Ha létezik $r_{i1}\neq 0$, akkor az első sor q_{i1}-szeresét levonjuk az i-edik sorból. Ezután újra 1. következik.
3. Ha $r_{1k}=r_{i1}=0, k=2,\ldots,r$; $i=2,\ldots,m$, akkor az első oszlop q_{1k}-szeresét levonva a k-adik oszlopból $(k=1,\ldots,r)$ és az első sor q_{i1}-szeresét levonva az i-edik sorból $(i=1,\ldots,m)$ a következő alakú mátrixhoz jutunk:
$$\begin{pmatrix} a_{11}(s) & 0 & \ldots & 0 \\ 0 & a_{22}(s) & \ldots & a_{2r}(s) \\ & \vdots & & \\ 0 & a_{m2}(s) & \ldots & a_{mr}(s) \end{pmatrix}.$$
4. Ha van olyan $a_{ij}(s)$, $i,j>1$, melyre $a_{11}(s)\nmid a_{ij}(s)$, akkor az i-edik sort hozzáadjuk az elsőhöz, és újra végigcsináljuk az 1, 2, 3. lépéseket. Mivel így minden lépésben $a_{11}(s)$ foka csökken, véges sok lépésben a fenti alakú mátrixot kapjuk, ahol $a_{11}(s)|a_{ij}(s)$ minden $i,j>1$ párra.
5. Ha létezik olyan $a_{ij}(s)$, $i,j>1$, mely nem nulla, akkor az $\overline{A}((s)=(a_{ij}(s))i, j\geq 2$ mátrixszal csináljuk végig az 1, 2, 3, 4. lépéseket.

Végül éppen a kívánt alakú mátrixhoz jutunk.

Megjegyzés. A tétel értelmében tehát minden polinommátrixhoz hozzárendelhetünk egy kanonikusan diagonális alakú mátrixot. Be fogjuk látni, hogy az egyértelmű, és ezt az egyértelműen létező kanonikusan diagonális mátrixot az eredeti mátrix *Smith-formájának* fogjuk nevezni.

2.5. Definíció. Legyen $A(s)$ tetszőleges $m\times r$-es, majdnem minden s-re p rangú polinommátrix. $k\leq p$ esetén jelölje $d_k(s)$ a mátrix összes $k\times k$-s nem nulla aldeterminánsának legnagyobb közös osztóját. Legyen $d_0(s)=1$. Az $A(s)$ mátrix invariáns polinomjainak nevezzük a következő polinomokat:
$$i_1(s) = d_p(s)/d_{p-1}(s), \ldots$$
$$i_p(s) = d_1(s)/d_0(s).$$

2.2. TÉTEL. Ekvivalens mátrixok invariáns polinomjai megegyeznek (konstans szorzó erejéig).

A bizonyításhoz szükségünk lesz a következő lemmákra.

2.1. LEMMA. Legyenek $\mathbf{A}(s)$, $\mathbf{B}(s)$, $\mathbf{C}(s)$ $m\times r$-es, $r\times n$-es illetve $m\times n$-es polinommátrixok, melyekre $\mathbf{C}(s)=\mathbf{A}(s)\mathbf{B}(s)$. Ekkor $k\leqq\min\{m, n, r\}$ esetén $\mathbf{A}(s)$ (ill. $\mathbf{B}(s)$) $k\times k$-s aldeterminánsainak legnagyobb közös osztója $\mathbf{C}(s)$ $k\times k$-s aldeterminánsainak legnagyobb közös osztójának osztója.

2.2. LEMMA. *(Binet—Cauchy-tétel):* Tetszőleges $\mathbf{A}(s)$ mátrix esetén jelölje $|\mathbf{A}|^n_{i,j}$ az $i=(i_1, \ldots, i_n)$ és $j=(j_1, \ldots, j_n)$ sor- illetve oszlopindexek segítségével kijelölt $n\times n$-es aldeterminánst. Ekkor $\mathbf{A}(s)$ $m\times r$-es, $\mathbf{B}(s)$ $r\times p$-s mátrixok esetén

$$|\mathbf{A}\cdot\mathbf{B}|^r_{i,j}=\sum_k |\mathbf{A}|^r_{i,k}|\mathbf{B}|^r_{k,j},$$

ahol a jobb oldalon az összegzés az összes szóbajöhető $k=(k_1, \ldots, k_r)$ indexsorozatra megy.

A lemma bizonyítására nézve l. GANTMACHER [1], a 2.1. lemma a 2.2. lemma közvetlen következménye. 2.2. tétel viszont közvetlen következménye a 2.1. lemmának, hiszen ha $\mathbf{A}(s)=\mathbf{U}(s)\mathbf{B}(s)$ és $\mathbf{B}(s)=\mathbf{V}(s)\mathbf{A}(s)$, akkor tetszőleges k esetén $\mathbf{A}(s)$ $k\times k$-s aldeterminánsainak legnagyobb közös osztója megegyezik $\mathbf{B}(s)$ $k\times k$-s aldeterminánsainak legnagyobb közös osztójával. Nyilván ugyanez igaz unimoduláris mátrixszal való jobb szorzás esetén is.

2.1. KÖVETKEZMÉNY. Tetszőleges mátrix kanonikusan diagonális alakja egyértelműen meghatározott.

Bizonyítás. Mivel az invariáns polinomok egyértelműek, így a 2.2. tétel alapján az állítás nyilvánvaló.

2.6. Definíció. Tetszőleges polinom-mátrix esetén a vele ekvivalens, egyértelműen létező kanonikusan diagonális mátrixot a mátrix *Smith-alakjának* nevezzük.

2.7. Definíció. Egy polinommátrixot szinguláris nak mondunk, ha determinánsa azonosan 0.

2.3. TÉTEL. Legyen $\mathbf{A}(s)$ $m\times m$-es mátrix. $\mathbf{A}(s)$ pontosan akkor szinguláris, ha létezik olyan $\mathbf{p}(s)$ m-dimenziós polinomvektor, melyre $\mathbf{A}(s)\mathbf{p}(s)=\mathbf{0}$.

Bizonyítás. A feltétel elegendősége triviális. Tegyük fel most, hogy $|\mathbf{A}(s)|=0$. Az előző tétel alapján léteznek olyan $\mathbf{U}(s)$, $\mathbf{V}(s)$ unimoduláris mátrixok, melyekre

$$\mathbf{A}(s)=\mathbf{U}(s)\mathbf{S}(s)\mathbf{V}(s),$$

ahol $\mathbf{S}(s)$ kanonikusan diagonális, fődiagonálisában a $\lambda_1(s), \ldots, \lambda_r(s)$ invariáns polinomok állnak és $r<m$. Legyen $\mathbf{q}(s)=(0, \ldots, 0, a(s))^T$, ahol $a(s)$ tetszőleges nem nulla polinom. Nyilván $\mathbf{S}(s)\mathbf{q}(s)=0$. Ekkor $\mathbf{V}(s)$ unimodularitása miatt $\mathbf{p}(s)=\mathbf{V}^{-1}(s)\mathbf{q}(s)\neq\mathbf{0}$. Így

$$\mathbf{A}(s)\mathbf{p}(s)=\mathbf{U}(s)\mathbf{S}(s)\mathbf{V}(s)\mathbf{V}^{-1}(s)\mathbf{q}(s)=\mathbf{0}.$$

3. Polinommátrixok legnagyobb közös osztója

3.1. Definíció. Legyenek $\mathbf{A}(s)$, $\mathbf{B}(s)$ $m\times r$-es, ill. $r\times n$-es polinommátrixok, és legyen $\mathbf{A}(s)\cdot\mathbf{B}(s)=\mathbf{C}(s)$. Ekkor azt mondjuk, hogy $\mathbf{A}(s)$ bal osztója, $\mathbf{B}(s)$ jobb osztója $\mathbf{C}(s)$-nek; $\mathbf{C}(s)$ jobb többszöröse $\mathbf{A}(s)$-nek és bal többszöröse $\mathbf{B}(s)$-nek. Ez alapján értelmezhető mátrixok közös bal osztója és közös jobb osztója. A legnagyobb közös bal osztók, illetve legnagyobb közös jobb osztók definiálásánál korlátoznunk kell a szóbajöhető mátrixok dimenzióját.

3.2. Definíció. Legyenek $\mathbf{N}(s)$ és $\mathbf{D}(s)$ $m\times r$-es, illetve $m\times n$-es polinommátrixok. A $\mathbf{W}(s)$ $m\times m$-es négyzetes polinommátrix legnagyobb közös bal osztója a mátrixoknak, ha közös bal osztója $\mathbf{N}(s)$-nek és $\mathbf{D}(s)$-nek és tetszőleges $\mathbf{V}(s)$ $m\times p$-s közös bal osztó esetén létezik $\mathbf{U}(s)$ $p\times m$-es polinommátrix, hogy $\mathbf{W}(s)=\mathbf{V}(s)\mathbf{U}(s)$.

A legnagyobb közös jobb osztó teljesen hasonlóan értelmezhető $n\times m$-es és $r\times m$-es mátrixok esetén.

A következő fejezetekben és általában az alkalmazások során legnagyobb közös osztó meghatározásakor az egyik mátrix (pl. $\mathbf{D}(s)$) négyzetes alakú. Az általános eset semmivel sem nehezebb ennél, így a könnyebb áttekinthetőség kedvéért mi csak az előbbi esettel foglalkozunk.

A bal osztókra vonatkozó állítások minden esetben könnyen átfogalmazhatók jobb osztókra vonatkozó állításokra.

3.1. TÉTEL. Legyenek $\mathbf{N}(s)$ és $\mathbf{D}(s)$ $m\times r$-es, ill. $m\times m$-es polinommátrixok. Ekkor létezik $\mathbf{W}(s)$ legnagyobb közös bal osztójuk, továbbá léteznek $\mathbf{X}(s)$ és $\mathbf{Y}(s)$ polinommátrixok, úgy, hogy $\mathbf{D}(s)\mathbf{X}(s)+\mathbf{N}(s)\mathbf{Y}(s)=\mathbf{W}(s)$.

Bizonyítás. Először belátjuk, hogy a $(\mathbf{D}(s)\mathbf{N}(s))$ mátrix elemi oszloptranszformációkkal $(\mathbf{W}(s)\mathbf{0})$ alakra hozható, ahol $\mathbf{W}(s)$ $m\times m$-es polinommátrix.

Tekintsük az első sort. Oszlopcserékkel elérhető, hogy a legkisebb fokú nem 0 tag kerüljön az (1, 1) helyre. Ezzel az elemmel maradékosan elosztva a sor többi elemét, majd az első oszlop megfelelő polinomszorosait kivonva a többiből az első sor második elemétől kezdve mindegyik elem kisebb fokú lesz, mint az első elem, vagy 0 lesz. Ezután újra kiválasztva a legkisebb fokú nem 0 tagot, újra végrehajtjuk a fenti oszloptranszformációkat mindaddig, míg az első sorban az első elemen kívül az összes többi elem 0-vá válik.

Ezután a k-adik ($k=2, 3, \ldots$) sort tekintve, a mátrix k-adiktól kezdődő oszlopait felhasználva addig csináljuk a fenti transzformációt, míg a k-adik sorban az első k elemet kivéve az összes többi 0-vá válik. Így szerkesztettünk egy olyan

$$\mathbf{U}(s)=\begin{pmatrix}\mathbf{U}_{11}(s) & \mathbf{U}_{12}(s)\\ \mathbf{U}_{21}(s) & \mathbf{U}_{22}(s)\end{pmatrix}$$

unimoduláris mátrixot, melyre

$$(\mathbf{D}(s)\mathbf{N}(s))\begin{pmatrix}\mathbf{U}_{11}(s) & \mathbf{U}_{12}(s)\\ \mathbf{U}_{21}(s) & \mathbf{U}_{22}(s)\end{pmatrix}=(\mathbf{W}(s)\mathbf{0}).$$

Tehát speciálisan

$$\mathbf{W}(s)=\mathbf{D}(s)\mathbf{U}_{11}(s)+\mathbf{N}(s)\mathbf{U}_{21}(s).$$

Ebből következik, hogy ha $\mathbf{V}(s)$ közös bal osztó, azaz $\mathbf{D}(s)=\mathbf{V}(s)\mathbf{D}_0(s)$, $\mathbf{N}(s)=\mathbf{V}(s)\mathbf{N}_0(s)$, akkor $\mathbf{W}=\mathbf{V}\mathbf{D}_0\mathbf{U}_{11}+\mathbf{V}\mathbf{N}_0\mathbf{U}_{21}=\mathbf{V}(\mathbf{D}_0\mathbf{U}_{11}+\mathbf{N}_0\mathbf{U}_{21})$, azaz minden közös bal osztónak a $\mathbf{W}(s)$ mátrix jobb többszöröse.

Másrészt megszorozva a fenti egyenlőséget az $\mathbf{U}(s)$ unimoduláris mátrix

$$\mathbf{U}^{-1}(s)=\begin{pmatrix}\mathbf{V}_{11}(s) & \mathbf{V}_{12}(s)\\ \mathbf{V}_{21}(s) & \mathbf{V}_{22}(s)\end{pmatrix} \quad \text{inverzével:}$$

$$(\mathbf{D}\ \ \mathbf{N})=(\mathbf{W}\ \ \mathbf{0})\begin{pmatrix}\mathbf{V}_{11} & \mathbf{V}_{12}\\ \mathbf{V}_{21} & \mathbf{V}_{22}\end{pmatrix}$$

azaz

$$\mathbf{D}(s)=\mathbf{W}(s)\mathbf{V}_{11}(s),$$

$$\mathbf{N}(s)=\mathbf{W}(s)\mathbf{V}_{12}(s).$$

Ebből látható, hogy $\mathbf{W}(s)$ közös bal osztója $\mathbf{D}(s)$-nek és $\mathbf{N}(s)$-nek. Tehát az $\mathbf{X}(s)=\mathbf{U}_{11}(s)$ és $\mathbf{Y}(s)=\mathbf{U}_{21}(s)$ választással a legnagyobb közös bal osztó kívánt előállítását kapjuk.

3.2. TÉTEL. Ha van nem szinguláris legnagyobb közös bal osztó, akkor mindegyik legnagyobb közös bal osztó nem szinguláris és ezek csak unimoduláris jobb osztóban különbözhetnek egymástól.

Bizonyítás. Legyen $\mathbf{W}_1(s)$ nem szinguláris, $\mathbf{W}_2(s)$ tetszőleges legnagyobb közös bal osztó. Ekkor a legnagyobb közös bal osztó tulajdonsága miatt $\mathbf{W}_1(s)$

$$\mathbf{W}_1(s)=\mathbf{W}_2(s)\mathbf{U}_1(s) \quad \text{és}$$

$$\mathbf{W}_2(s)=\mathbf{W}_1(s)\mathbf{U}_2(s)$$

azaz

$$\mathbf{W}_1(s)=\mathbf{W}_2(s)\mathbf{U}_1(s)=\mathbf{W}_1(s)\mathbf{U}_2(s)\mathbf{U}_1(s)$$

ahonnan $\mathbf{U}_1(s)\mathbf{U}_2(s)=\mathbf{I}$ és $\mathbf{U}_1(s)$, $\mathbf{U}_2(s)$ unimodularitása következik.

Megjegyzés. Nem szinguláris legnagyobb közös bal osztó létezése $(\mathbf{D}(s)\mathbf{N}(s))$ teljes rangúságával ekvivalens, hiszen az állításbeli $(\mathbf{D}(s)\mathbf{N}(s))$ és $\mathbf{W}(s)$ rangja ugyanaz. A továbbiakban feltesszük, hogy $\mathbf{D}(s)$ teljes rangú.

Definíció. Az $\mathbf{N}(s)$ és $\mathbf{D}(s)$ polinommátrixok bal relatív prímek, ha csak unimoduláris közös bal osztójuk van.

3.1. KÖVETKEZMÉNY. Ha $\mathbf{D}(s)$ és $\mathbf{N}(s)$ bal relatív prímek, akkor léteznek olyan $\tilde{\mathbf{X}}(s)$ és $\tilde{\mathbf{Y}}(s)$ polinommátrixok, hogy

$$\mathbf{D}(s)\tilde{\mathbf{X}}(s)+\mathbf{N}(s)\tilde{\mathbf{Y}}(s)=\mathbf{I}.$$

Bizonyítás. A 3.1. tétel alapján léteznek olyan $\mathbf{X}(s)$ és $\mathbf{Y}(s)$ polinommátrixok, hogy

$$\mathbf{D}(s)\mathbf{X}(s)+\mathbf{N}(s)\mathbf{Y}(s)=\mathbf{W}(s),$$

ahol $\mathbf{W}(s)$ unimoduláris mátrix. Ekkor

$$\tilde{\mathbf{X}}(s) = \mathbf{X}(s)\mathbf{W}^{-1}(s), \quad \tilde{\mathbf{Y}}(s) = \mathbf{Y}(s)\mathbf{W}^{-1}(s)$$

választással igaz az állítás.

3.3. TÉTEL. Legyen $\mathbf{D}(s)$ nem szinguláris $m\times m$-es, $\mathbf{N}(s)$ $m\times r$-es polinommátrix. $\mathbf{N}(s)$ és $\mathbf{D}(s)$ pontosan akkor bal relatív prímek, ha $(\mathbf{N}(s)\,\mathbf{D}(s))$ teljes rangú minden s-re.

Bizonyítás. Tegyük fel először, hogy a mátrixok bal relatív prímek. A 3.1. tétel alapján az $(\mathbf{N}(s)\mathbf{D}(s))$ mátrix oszloptranszformációkkal $(\mathbf{W}(s)\mathbf{0})$ alakra hozható. Ez utóbbi mátrix minden s-re teljes rangú, hiszen $\mathbf{W}(s)$ unimoduláris.

Fordítva, ha a polinommátrixok nem relatív prímek, akkor $\mathbf{W}(s)$ determinánsa legalább elsőfokú polinom. Ha ennek egy gyöke s_0, akkor $\mathbf{W}(s_0)$ szinguláris, így $(\mathbf{D}(s_0)\mathbf{N}(s_0))$ nem teljes rangú.

4. Irányíthatósági-, megfigyelhetőségi kritériumok a frekvenciatartományban

Tekintsük a következő lineáris rendszert:

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}+\mathbf{B}\mathbf{u} \quad \mathbf{x}\in R^n, \tag{4.1}$$

$$\mathbf{y} = \mathbf{C}\mathbf{x}. \tag{4.2}$$

Az 1.1. tétel szerint a rendszer irányíthatósága ekvivalens a $\mathscr{C}=(\mathbf{B}, \mathbf{AB}, \ldots, \mathbf{A}^{n-1}\mathbf{B})$ mátrix teljes rangúságával, míg az 1.2. tétel szerint a megfigyelhetőség az $\mathcal{O}=$ $=(\mathbf{C}^T\mathbf{C}^T\mathbf{A}^T\mathbf{C}^T\mathbf{A}^{T(n-1)})^T$ mátrix teljes rangúságával ekvivalens. Ezekre szeretnénk másfajta feltételeket kapni.

4.1. TÉTEL. Az $(\mathbf{A}, \mathbf{B})$ pár pontosan akkor irányítható, ha $\mathbf{A}$ tetszőleges bal oldali sajátvektora (amely komplex is lehet) nem eleme $\operatorname{Ker}\mathbf{B}^T$-nak, azaz

$$\mathbf{v}^T\mathbf{A} = \lambda\mathbf{v}^T, \quad \mathbf{v}^T\mathbf{B} = \mathbf{0} \qquad \mathbf{v} \neq \mathbf{0} \tag{4.3}$$

nem oldható meg.

Bizonyítás. Tegyük fel, hogy van megoldása a (4.3) feltételrendszernek. Ekkor $\mathbf{v}^T(\mathbf{B}, \mathbf{AB}, \ldots, \mathbf{A}^{n-1}\mathbf{B})=\mathbf{0}$, ami azt jelenti, hogy $(\mathbf{A}, \mathbf{B})$ nem irányítható.

Fordítva, tegyük fel, hogy (4.3) nem megoldható. Ekkor $\operatorname{Ker}\mathbf{B}^T$ nem tartalmazhatja $\mathbf{A}^T$ invariáns alterét, így tetszőleges $\mathbf{v}\in\operatorname{Ker}\mathbf{B}^T$, $\mathbf{v}\neq\mathbf{0}$ esetén létezik i, hogy $(\mathbf{A}^i)^T\mathbf{v}$ nem tartozik $\operatorname{Ker}\mathbf{B}^T$-ba. A *Cayley—Hamilton-tétel* alapján n-nél kisebb i is létezik ezzel a tulajdonsággal, tehát $\mathbf{v}\in\operatorname{Ker}\mathbf{B}^T$ $\mathbf{v}\neq\mathbf{0}$ esetén $\mathbf{v}^T(\mathbf{B}, \mathbf{AB}, \ldots, \mathbf{A}^{n-1}\mathbf{B})=$ $=\mathbf{v}^T\mathscr{C}\neq\mathbf{0}$. Ha viszont $\mathbf{v}\notin\operatorname{Ker}\mathbf{B}^T$, akkor már $\mathbf{v}^T\mathbf{B}\neq\mathbf{0}$. Tehát tetszőleges $\mathbf{v}\neq\mathbf{0}$ vektorra $\mathbf{v}^T\mathscr{C}\neq\mathbf{0}$.

4.2. TÉTEL. (Irányítható rendszerek direkt összege.) Legyenek

$$\dot{\mathbf{x}}_1 = \mathbf{A}_1\mathbf{x}_1+\mathbf{B}_1\mathbf{u}, \tag{4.4}$$

$$\dot{\mathbf{x}}_2 = \mathbf{A}_2\mathbf{x}_2+\mathbf{B}_2\mathbf{u}. \tag{4.5}$$

irányítható rendszerek. Tegyük fel, hogy $\mathbf{A}_1$-nek és $\mathbf{A}_2$-nek nincsen közös sajátértéke. Ekkor a

$$\begin{pmatrix}\dot{\mathbf{x}}_1\\ \dot{\mathbf{x}}_2\end{pmatrix}=\begin{pmatrix}\mathbf{A}_1 & 0\\ 0 & \mathbf{A}_2\end{pmatrix}\begin{pmatrix}\mathbf{x}_1\\ \mathbf{x}_2\end{pmatrix}+\begin{pmatrix}\mathbf{B}_1\\ \mathbf{B}_2\end{pmatrix}\mathbf{u} \tag{4.6}$$

rendszer is irányítható.

Bizonyítás. Legyen $(\mathbf{v}_1^T\,\mathbf{v}_2^T)\begin{pmatrix}\mathbf{A}_1 & 0\\ 0 & \mathbf{A}_2\end{pmatrix}=\lambda(\mathbf{v}_1^T\,\mathbf{v}_2^T)$. Mivel az $\mathbf{A}_i$ mátrixoknak nincsen közös sajátértéke, így $\mathbf{v}_1=\mathbf{0}$ vagy $\mathbf{v}_2=\mathbf{0}$. Legyen például $\mathbf{v}_1=\mathbf{0}$. Ekkor $\mathbf{v}_2^T$ az $\mathbf{A}_2$ sajátvektora és a (4.5) rendszer irányíthatósága miatt $(\mathbf{v}_1^T\,\mathbf{v}_2^T)\begin{pmatrix}\mathbf{B}_1\\ \mathbf{B}_2\end{pmatrix}=\mathbf{v}_2^T\mathbf{B}_2\neq\mathbf{0}$, tehát a (4.6) rendszerre a 4.1. tétel (4.3) feltételrendszere nem oldható meg.

4.3. TÉTEL. Az $(\mathbf{A}, \mathbf{B})$ pár pontosan akkor irányítható, ha $((s\mathbf{I}-\mathbf{A}), \mathbf{B})$ teljes rangú minden s-re.

Bizonyítás. A tétel állítása a 4.1. tétel alapján nyilvánvaló, hiszen rögzített λ esetén $((\lambda\mathbf{I}-\mathbf{A}), \mathbf{B})$ pontosan akkor teljes rangú, ha nem létezik olyan $\mathbf{v}\neq\mathbf{0}$, melyre $\mathbf{v}^T((\lambda\mathbf{I}-\mathbf{A}), \mathbf{B})=\mathbf{0}$, azaz melyre $\lambda\mathbf{v}^T=\mathbf{v}^T\mathbf{A}$ és $\mathbf{v}^T\mathbf{B}=\mathbf{0}$.

A megfigyelhetőségre vonatkozó összefüggések a fentiek mintájára könnyen megfogalmazhatók a dualitás alapján. A dualitás szerint az $(\mathbf{A}, \mathbf{B})$ pár irányíthatósága ekvivalens a $(\mathbf{B}^T, \mathbf{A}^T)$ pár megfigyelhetőségével. Így a fenti állítások megfelelői.

4.4. TÉTEL. $(\mathbf{C}, \mathbf{A})$ pontosan akkor megfigyelhető, ha teljesül az alábbi feltétel:

$$\mathbf{A}\mathbf{v}=\lambda\mathbf{v},\quad \mathbf{C}\mathbf{v}=\mathbf{0},\quad \mathbf{v}\neq\mathbf{0} \tag{4.7}$$

nem oldható meg.

4.5. TÉTEL. $(\mathbf{C}, \mathbf{A})$ pontosan akkor megfigyelhető, ha $\begin{pmatrix}\mathbf{C}\\ s\mathbf{I}-\mathbf{A}\end{pmatrix}$ teljes rangú minden s-re.

5. Jobb és bal oldali mátrixtörtek

5.1. Definíció. Legyen $\mathbf{H}(s)$ tetszőleges racionális tört elemekből álló mátrix. A $\mathbf{H}(s)=\mathbf{D}(s)^{-1}\mathbf{N}(s)$ felbontás bal oldali mátrixtört-felbontás, ha $\mathbf{D}(s)$ $m\times m$-es, $\mathbf{N}(s)$ $m\times r$-es polinommátrixok és $\mathbf{D}(s)$ nem szinguláris. A bal oldali mátrixtört-felbontás irreducibilis, ha a felbontásban szereplő polinommátrixok bal relatív prímek.

Hasonlóan definiáljuk a jobb oldali mátrixtört-felbontást is. A fenti definícióban szereplő $\mathbf{H}(s)=\mathbf{D}^{-1}(s)\mathbf{N}(s)$ bal oldali mátrixtörtet szeretnénk jobb oldali mátrixtört alakban felírni. A 3. fejezetben, a legnagyobb közös bal osztó meghatározásakor láttuk, hogy létezik olyan

$$\mathbf{U}(s)=\begin{pmatrix}\mathbf{U}_{11}(s) & \mathbf{U}_{12}(s)\\ \mathbf{U}_{21}(s) & \mathbf{U}_{22}(s)\end{pmatrix} \tag{5.1}$$

unimoduláris polinommátrix, melyre

$$(\mathbf{D}(s)\mathbf{N}(s))\mathbf{U}(s)=(\mathbf{W}(s)\mathbf{0}),$$

ahol $\mathbf{W}(s)$ a $\mathbf{D}(s)$ és $\mathbf{N}(s)$ mátrixok legnagyobb közös bal osztója.

5.1. TÉTEL. A fenti feltételekkel és jelölésekkel $\mathbf{U}_{22}(s)$ nem szinguláris és

$$\mathbf{H}(s) = \mathbf{D}^{-1}(s)\mathbf{N}(s) = -\mathbf{U}_{12}(s)\mathbf{U}_{22}(s)^{-1}, \tag{5.2}$$

ahol az $\mathbf{U}_{12}\mathbf{U}_{22}^{-1}$ felírás irreducibilis mátrixtört-felbontás. Továbbá, ha $\mathbf{D}(s)$ és $\mathbf{N}(s)$ bal relatív prímek, akkor deg $|\mathbf{D}(s)|$=deg $|\mathbf{U}_{22}(s)|$.

Bizonyítás. Ha $\mathbf{U}_{22}(s)$ nem szinguláris, akkor az (5.2) egyenlőség nyilván teljesül, hisz (5.1) alapján

$$\mathbf{D}(s)\mathbf{U}_{12}(s)+\mathbf{N}(s)\mathbf{U}_{22}(s) = \mathbf{0}. \tag{5.3}$$

Tegyük fel, hogy $\mathbf{U}_{22}(s)$ szinguláris. Láttuk, (2.3. tétel), hogy ez ekvivalens azzal, hogy létezik $p^{(s)}\neq 0$ polinomvektor, melyre $\mathbf{U}_{22}(s)\mathbf{p}(s)\equiv\mathbf{0}$. (5.3)-at beszorozva a $\mathbf{p}(s)$ vektorral

$$\mathbf{D}(s)\mathbf{U}_{12}(s)\mathbf{p}(s) = \mathbf{0} \tag{5.4}$$

eredményre jutunk, ahonnan $\mathbf{U}_{12}(s)\mathbf{p}(s)=\mathbf{0}$ következik, hiszen $\mathbf{D}(s)$ nem szinguláris. Ezután az $\mathbf{U}(s)$ unimoduláris mátrixot megszorozva a $\tilde{\mathbf{p}}(s)=\begin{pmatrix}0\\ \mathbf{p}(s)\end{pmatrix}$ vektorral, $\mathbf{U}(s)\tilde{\mathbf{p}}(s)=\mathbf{0}$-hoz jutunk, ami ellentmond $\mathbf{U}(s)$ unimodularitásának.

Tekintsük most az $\mathbf{U}(s)$ unimoduláris mátrix inverzét, legyen ez $\mathbf{V}(s)$. Megfelelő alakú blokkmátrixokra való felbontása legyen a következő:

$$\mathbf{V}(s) = \begin{pmatrix}\mathbf{V}_{11}(s) & \mathbf{V}_{12}(s)\\ \mathbf{V}_{21}(s) & \mathbf{V}_{22}(s)\end{pmatrix}.$$

Ekkor a következő felírásból azonnal látszik, hogy $\mathbf{U}_{12}(s)$ és $\mathbf{U}_{22}(s)$ jobb relatív prímek:

$$\mathbf{V}(s)\begin{pmatrix}\mathbf{U}_{12}(s)\\ \mathbf{U}_{22}(s)\end{pmatrix} = \begin{pmatrix}\mathbf{0}\\ \mathbf{I}\end{pmatrix}. \tag{5.5}$$

Hátra van még a fokszámokra vonatkozó összefüggés igazolása. Tegyük fel, hogy $\mathbf{D}(s)$ és $\mathbf{N}(s)$ bal relatív prímek. Szorozzuk meg (5.1)-et $\mathbf{V}(s)$-sel jobbról:

$$(\mathbf{W}(s)\,\mathbf{0})\begin{pmatrix}\mathbf{V}_{11}(s) & \mathbf{V}_{12}(s)\\ \mathbf{V}_{21}(s) & \mathbf{V}_{22}(s)\end{pmatrix} = (\mathbf{D}(s)\,\mathbf{N}(s)). \tag{5.6}$$

Ebből azonnal adódik, hogy

$$\mathbf{W}(s)\mathbf{V}_{11}(s) = \mathbf{D}(s). \tag{5.7}$$

$\mathbf{W}(s)$ unimodularitása miatt tehát

$$\deg|\mathbf{V}_{11}(s)| = \deg|\mathbf{D}(s)|. \tag{5.8}$$

Az $\mathbf{U}(s)$ mátrix determinánsát — ami konstans — a blokk-mátrixok determinánására vonatkozó összefüggés alapján a következőképpen számíthatjuk ki, (mivel $\mathbf{U}_{22}(s)$ nem szinguláris):

$$|\mathbf{U}(s)| = |\mathbf{U}_{22}(s)|\cdot|\mathbf{U}_{11}(s)-\mathbf{U}_{12}(s)\mathbf{U}_{22}^{-1}(s)\mathbf{U}_{21}(s)|.$$

Rövid számolással a $\mathbf{V}(s)\cdot\mathbf{U}(s)=\mathbf{I}$ felhasználásával azt kapjuk, hogy a jobb oldal második tagja éppen $|\mathbf{V}_{11}(s)|^{-1}$. Valóban (az s argumentumot most elhagyjuk):

$$\mathbf{V}_{11}(\mathbf{U}_{11}-\mathbf{U}_{12}\mathbf{U}_{22}^{-1}\mathbf{U}_{21}) = \mathbf{V}_{11}\mathbf{U}_{11}-\mathbf{V}_{11}\mathbf{U}_{12}\mathbf{U}_{22}^{-1}\mathbf{U}_{21} =$$

$$= (\mathbf{I}-\mathbf{V}_{12}\mathbf{U}_{21})-(-\mathbf{V}_{12}\mathbf{U}_{22})\mathbf{U}_{22}^{-1}\mathbf{U}_{21} = \mathbf{I}.$$

Innen

$$\mathbf{V}_{11} = (\mathbf{U}_{11}-\mathbf{U}_{12}\mathbf{U}_{22}^{-1}\mathbf{U}_{21})^{-1} \tag{5.9}$$

és

$$1 = |\mathbf{U}| = |\mathbf{U}_{22}|\,|\mathbf{U}_{11}-\mathbf{U}_{12}\mathbf{U}_{22}^{-1}\mathbf{U}_{21}| = |\mathbf{U}_{22}|\,|\mathbf{V}_{11}|^{-1}. \tag{5.10}$$

(5.8) és (5.10) alapján következik az állítás.

Megjegyzés. A tételben leírt mátrixtörtátírás lényegében egyértelmű is, ezt fogjuk most belátni.

5.2. TÉTEL. Legyen a $\mathbf{H}(s)$ $m\times r$-es racionális törtfüggvény két irreducibilis balfelbontása $\mathbf{H}(s)=\mathbf{D}_1^{-1}(s)\mathbf{N}_1(s)=\mathbf{D}_2^{-1}(s)\mathbf{N}_2(s)$. Ekkor létezik olyan $\mathbf{U}(s)$ unimoduláris mátrix, melyre

$$\mathbf{D}_1(s) = \mathbf{U}(s)\mathbf{D}_2(s), \quad \mathbf{N}_1(s) = \mathbf{U}(s)\mathbf{N}_2(s). \tag{5.11}$$

Bizonyítás. $\mathbf{H}(s)$ kétféle felírásából azonnal adódik, hogy

$$\mathbf{N}_1(s) = \mathbf{D}_1(s)\mathbf{D}_2^{-1}(s)\mathbf{N}_2(s).$$

Jelölje $\mathbf{U}(s)=\mathbf{D}_1(s)\mathbf{D}_2^{-1}(s)$. Belátjuk, hogy ez unimoduláris polinommátrix.

Mivel $\mathbf{N}_1(s)$ és $\mathbf{D}_1(s)$ bal relatív prímek, ezért léteznek olyan $\mathbf{X}(s)$ és $\mathbf{Y}(s)$ polinommátrixok, melyekre

$$\mathbf{N}_1(s)\mathbf{X}(s)+\mathbf{D}_1(s)\mathbf{Y}(s) = \mathbf{I}.$$

Ebből azt kaphatjuk, hogy

$$\mathbf{D}_1(s)\mathbf{D}_2^{-1}(s)\mathbf{N}_2(s)\mathbf{X}(s)+\mathbf{D}_1(s)\mathbf{D}_2^{-1}(s)\mathbf{D}_2(s)\mathbf{Y}(s) = \mathbf{I},$$

azaz

$$\mathbf{U}(s)\big(\mathbf{N}_2(s)\mathbf{X}(s)+\mathbf{D}_2(s)\mathbf{Y}(s)\big) = \mathbf{I}. \tag{5.12}$$

Ebből következik, hogy $\mathbf{U}^{-1}(s)=\mathbf{N}_2(s)\mathbf{X}(s)+\mathbf{D}_2(s)\mathbf{Y}(s)$, tehát $\mathbf{U}^{-1}(s)=$ $=\mathbf{D}_2(s)\mathbf{D}_1^{-1}(s)$ polinommátrix. Hasonlóan belátható (az indexek felcserélésével), hogy $\mathbf{U}(s)$ is polinommátrix, tehát $\mathbf{U}(s)$ unimoduláris.

5.1. KÖVETKEZMÉNY. Legyen $\mathbf{H}(s)=\mathbf{D}_1^{-1}(s)\mathbf{N}_1(s)$ irreducibilis törtfelbontás, $\mathbf{H}(s)=\mathbf{D}_2^{-1}(s)\mathbf{N}_2(s)$ tetszőleges törtfelbontás. Ekkor létezik $\mathbf{R}(s)$ nem szinguláris polinommátrix, melyre $\mathbf{D}_2(s)=\mathbf{R}(s)\mathbf{D}_1(s)$ és $\mathbf{N}_2(s)=\mathbf{R}(s)\mathbf{N}_1(s)$.

Bizonyítás. Jelölje $\mathbf{W}(s)\mathbf{D}_2(s)$ és $\mathbf{N}_2(s)$ legnagyobb közös bal osztóját; $\mathbf{D}_2(s)=$ $=\mathbf{W}(s)\mathbf{D}_0(s)$, $\mathbf{N}_2(s)=\mathbf{W}(s)\mathbf{N}_0(s)$. Ekkor $\mathbf{H}(s)=\mathbf{D}_0^{-1}(s)\mathbf{N}_0(s)$ nyilván irreducibilis törtfelbontás. Az 5.2. tétel alapján létezik $\mathbf{U}(s)$ unimoduláris mátrix, melyre $\mathbf{D}_0(s)=\mathbf{U}(s)\mathbf{D}_1(s)$ és $\mathbf{N}_0(s)=\mathbf{U}(s)\mathbf{N}_1(s)$. $\mathbf{R}(s)=\mathbf{W}(s)\cdot\mathbf{U}(s)$ választással nyilván igaz a következmény állítása.

6. Racionális tört mátrixok Smith—McMillan alakja Pólusok, zérusok értelmezése

Racionális törtmátrixok esetében is be lehet vezetni a *Smith-formához* hasonló fogalmat; elemi sor és oszloptranszformációkkal ezeket a mátrixokat is speciális diagonális alakúra hozhatjuk.

Tekintsük a $\mathbf{H}(s)$ $m\times r$-es racionális törtmátrixot. Tegyük fel, hogy a mátrix elemei irreducibilis törtek. Legyen $d(s)$ a mátrix-elemek nevezőinek legkisebb közös többszöröse. (Az egyértelműség kedvéért legyen $d(s)$ főegyütthatója 1). Ekkor $\mathbf{H}(s)=\mathbf{N}(s)/\mathbf{d}(s)$, ahol $\mathbf{N}(s)$ olyan polinommátrix, melynek elemei relatív prímek. Legyen $\mathbf{N}(s)$ *Smith-alakja*

$$\mathbf{\Lambda}(s) = \begin{pmatrix} \lambda_1(s) & & \\ & \ddots & \\ & & \lambda_p(s) \\ & & 0 \end{pmatrix}. \tag{6.1}$$

Ekkor léteznek $\mathbf{U}_1(s)$, $\mathbf{U}_2(s)$ $m\times m$-es, illetve $r\times r$-es unimoduláris mátrixok, melyekre $\mathbf{N}(s)=\mathbf{U}_1(s)\mathbf{\Lambda}(s)\mathbf{U}_2(s)$, azaz

$$\mathbf{H}(s) = \mathbf{U}_1(s)\mathbf{\Lambda}(s)/\mathbf{d}(s)\mathbf{U}_2(s),$$

6.1. Definíció. Legyen $\mathbf{M}(s)=\mathbf{\Lambda}(s)/\mathbf{d}(s)$, és egyszerűsítsük $\mathbf{M}(s)$ elemeit irreducibilis racionális törtekké:

$$\mathbf{M}(s) = \begin{pmatrix} \dfrac{\varepsilon_1(s)}{\psi_1(s)} & & \\ & \ddots & \\ & & \dfrac{\varepsilon_p(s)}{\psi_p(s)} \\ & & 0 \end{pmatrix}. \tag{6.2}$$

Ekkor $\mathbf{M}(s)$-t a $\mathbf{H}(s)$ mátrix *Smith—McMillan alakjának* nevezzük.

6.1. TÉTEL. A *Smith—McMillan alak* tulajdonságai:
(1) az *S—M alak* egyértelmű,
(2) $\varepsilon_i(s)|\varepsilon_{i+1}(s)$ $i=1, \ldots, p-1$,
(3) $\psi_{i+1}(s)|\psi_i(s)$ $i=1, \ldots, p-1$,
(4) $\psi_1(s)=\mathbf{d}(s)$.

Bizonyítás.
(1) Az egyértelműség következik a *Smith-forma* egyértelműségéből.
(2), (3) Következik a $\lambda_i(s)$-ekre vonatkozó $\lambda_i(s)|\lambda_{i+1}(s)$ tulajdonságból.
(4) $\psi_1(s)\neq\mathbf{d}(s)$ azt jelentené, hogy $\lambda_1(s)$-nek és $\mathbf{d}(s)$-nek van közös osztója. A *Smith-forma* bevezetésekor azonban láttuk, hogy egy mátrix első invariáns polinomja nem más, mint elemeinek legnagyobb közös osztója. A feltételek szerint $\mathbf{N}(s)$ esetében ez 1, tehát $\lambda_1(s)$ és $\mathbf{d}(s)$ legnagyobb közös osztója is 1.

Megjegyzés. Jelölje $\mathbf{E}(s)$ azt az $m\times r$-es polinommátrixot, melynek fődiagonálisának első p eleme $\varepsilon_1(s), \ldots, \varepsilon_p(s)$, az összes többi eleme 0. Jelölje $\mathbf{F}_r(s)$, (illetve

$\mathbf{F}_l(s)$) azt az $r\times r$-es (illetve $m\times m$-es) polinommátrixot, melynek fődiagonálisának első p eleme $\psi_1(s), \ldots, \psi_p(s)$, a többi diagonális elem 1, az összes többi eleme 0. Ekkor $\mathbf{H}(s)=\mathbf{E}(s)\cdot\mathbf{F}_r^{-1}(s)=\mathbf{F}_l^{-1}(s)\mathbf{E}(s)$ irreducibilis törtfelbontások. Az 5.2. tétel alapján ha adottak tetszőleges $\mathbf{H}(s)=\mathbf{N}_r(s)\mathbf{D}_r^{-1}(s)=\mathbf{D}_l^{-1}(s)\mathbf{N}_l(s)$ irreducibilis törtfelbontások, akkor léteznek $\mathbf{U}(s)$ és $\mathbf{V}(s)$ unimoduláris mátrixok úgy, hogy $\mathbf{N}_r(s)=\mathbf{E}(s)\mathbf{U}(s)$, $\mathbf{N}_l(s)=\mathbf{V}(s)\mathbf{E}(s)$, $\mathbf{D}_r(s)=\mathbf{F}_r(s)\mathbf{U}(s)$, $\mathbf{D}_l(s)=\mathbf{V}(s)\mathbf{F}_l(s)$. Így beláttuk a következő tényt:

6.1. KÖVETKEZMÉNY. Legyenek adottak $\mathbf{H}(s)=\mathbf{N}_r(s)\mathbf{D}_r^{-1}(s)=\mathbf{D}_l^{-1}(s)\mathbf{N}_l(s)$ tetszőleges irreducibilis törtfelbontások. Ekkor $\mathbf{N}_r(s)$ és $\mathbf{N}_l(s)$ *Smith-formája* megegyezik, továbbá $\mathbf{D}_r(s)$ és $\mathbf{D}_l(s)$ nem egység invariáns polinomjai is megegyeznek.

6.2. Definíció. A $\mathbf{H}(s)$ mátrix zérusai az *S—M alakban* szereplő törtek számlálóinak gyökei, $\mathbf{H}(s)$ pólusai pedig ugyanezen törtek nevezőinek gyökei.

Megjegyzés. A pólusok és zérusok nem alkotnak diszjunkt halmazokat, lehet valami pólus is, zérus is egyszerre, pl.

$$\mathbf{H}(s)=\begin{pmatrix}(s-1)^{-1} & 0\\ 0 & (s-1)\end{pmatrix}.$$

A pólus és zérus definíciójának megértését segíti a következő tétel:

6.2. TÉTEL. $\mathbf{H}(s)$ pólusai megegyeznek $|\mathbf{D}(s)|=0$ gyökeivel, ahol $\mathbf{D}(s)$ tetszőleges irreducibilis tört felbontás nevezője. Továbbá $\mathbf{H}(s)$ zérusai megegyeznek $\mathbf{N}(s)$ invariáns polinomjainak gyökeivel, ahol $\mathbf{N}(s)$ tetszőleges irreducibilis törtfelbontás számlálója.

Bizonyítás. A 6.1. következmény alapján azonnal adódik a zérus-pólus definíciójából.

A továbbiakban foglalkozunk a *Smith—McMillan alak* nem nulla részével, legyen $\tilde{\mathbf{M}}(s)=\operatorname{diag}\left\{\frac{\varepsilon_i(s)}{\psi_i(s)}: i=1,\ldots p\right\}$. Ez felírható a következő szorzat-alakban:

$$\mathbf{M}(s)=\prod_{\substack{\alpha\,\text{zérus}\\ \{\text{vagy pólus}\}}}\mathbf{M}_\alpha(s),\ \mathbf{M}_\alpha(s)=\operatorname{diag}\{(s-\alpha)^{\sigma_1(\alpha)}\ldots(s-\alpha)^{\sigma_p(\alpha)}\}.$$

6.3. Definíció. A $\{\sigma_1(\alpha)\ldots\sigma_p(\alpha)\}$ számokat az α-hoz tartozó *strukturális indexeknek* nevezzük. (A 6.1. tétel alapján $\sigma_1(\alpha)\leqq\sigma_2(\alpha)\leqq\ldots\leqq\sigma_p(\alpha)$.) Ha $\sigma_1(\alpha)<0$, akkor α pólus; rendje $\sigma_1(\alpha)$, foka $-\sum_{\sigma_i(\alpha)<0}\sigma_i(\alpha)$. Ha $\sigma_p(\alpha)>0$, akkor α zérus; rendje $\sigma_p(\alpha)$, foka $\sum_{\sigma_i(\alpha)>0}\sigma_i(\alpha)$.

6.4. Definíció. Legyen $g(s)$ tetszőleges racionális törtfüggvény, $g(s)=(s-\alpha)^\sigma p(s)q(s)^{-1}$, ahol $p(s)$ és $q(s)$ nem tartalmaznak $(s-\alpha)$ faktort, $|\alpha|<\infty$. A σ véges számot a g függvény α-beli kiértékelésének nevezzük, jele $v_\alpha(g)=\sigma$. $g\equiv O$ esetén $v_\alpha(g)=\infty$ minden α esetén. Tetszőleges $g(s)=p(s)q^{-1}(s)$ racionális törtfüggvény esetén, ha p és q relatív prímek, a függvény ∞-beli kiértékelésének nevezzük a következő számot: $v_\infty(g)=\deg q(s)-\deg p(s)$.

Definíció. Legyen $\mathbf{H}(s)$ tetszőleges $m\times r$-es racionális törtfüggvény. A $\mathbf{H}$ mátrix α-beli i-edik kiértékelésének nevezzük $1\leqq i\leqq p$ esetén $\mathbf{H}$ összes $i\times i$-s aldeterminánsának kiértékelésének minimumát. Jele $v_\alpha^{(i)}(\mathbf{H})$. Tömören tehát a definíció:

$$v_\alpha^{(i)}(\mathbf{H})=\min\{v_\alpha(|\mathbf{H}|^i):\ |\mathbf{H}|^i\ \ i\times i\text{-s aldeterminánsa}\}.$$

Megegyezés szerint legyen $v_\alpha^{(0)}(\mathbf{H})=0$.

$\mathbf{H}(s)$ kiértékeléseit vissza szeretnénk vezetni *Smith—McMillan alakjának* kiértékelésére. A 2. fejezetben kimondott *Binet—Cauchy-tétel* alapján 2.2. lemma véges α esetén:

$$v_\alpha^{(i)}(\mathbf{H})=v_\alpha^{(i)}(\mathbf{M})=v_\alpha^{(i)}(\mathbf{M}_\alpha).$$

Ez utóbbi számértéket könnyen meghatározhatjuk: $\mathbf{M}_\alpha$ definíciója alapján

$$v_\alpha^{(i)}(\mathbf{M}_\alpha)=\sum_1^i \sigma_j(\alpha).$$

Tehát az α-hoz tartozó strukturális indexeket a következő módon határozhatjuk meg:

$$\sigma_i(\alpha)=v_\alpha^{(i)}(\mathbf{H})-v_\alpha^{(i-1)}(\mathbf{H}),\quad i=1,\ldots,p.$$

Megjegyezzük, hogy a p-edik kiértékelésre:

$$v_\alpha^{(p)}(\mathbf{H})=\sum_{j=1}^{p}\sigma_j(\alpha)=\sum_{\sigma_j>0}\sigma_j(\alpha)-\Big(-\sum_{\sigma_j>0}\sigma_j(\alpha)\Big)$$

tehát ez nem más, mint az α-beli zérusok és pólusok számának (fokának) különbsége.

7. Proper és szigorúan proper racionális törtmátrixok

Kiindulásképpen tekintsünk egy általános állapottér-modellt:

$$\dot{\mathbf{x}}=\mathbf{A}\mathbf{x}+\mathbf{B}\mathbf{u}. \tag{7.1}$$

$$\mathbf{y}=\mathbf{C}\mathbf{x}+\mathbf{D}(d)\mathbf{u} \tag{7.2}$$

ahol $\mathbf{D}(d)$ a $d=\partial/\partial t$ differenciáloperátor polinomja. Ennek transzfer függvénye

$$\mathbf{H}(s)=\mathbf{C}(s\mathbf{I}-\mathbf{A})^{-1}\mathbf{B}+\mathbf{D}(s) \tag{7.3}$$

A (7.1), (7.2) rendszer fizikailag megvalósítható, ha az inputot nem kell deriválni, azaz $\mathbf{D}(d)$ konstans. Ez annak felel meg, hogy $s\to\infty$ esetén $\mathbf{H}(s)$-nek van véges határértéke. Ilyen típusú mátrixokkal fogunk most bővebben foglalkozni.

7.1. Definíció. A $\mathbf{H}(s)$ racionális törtmátrix proper, ha $\lim_{s\to\infty}\mathbf{H}(s)$ létezik és véges. $\mathbf{H}(s)$-et szigorúan propernek nevezzük, ha $\lim_{s\to\infty}\mathbf{H}(s)=\mathbf{O}$.

Skalár esetben egy függvény proper (ill. szigorúan proper), ha a tört nevezőjének foka nagyobb vagy egyenlő (illetve nagyobb) mint a számláló foka. Hasonló állítást szeretnénk megfogalmazni több dimenziós esetben is.

7.2. Definíció. Jelölje tetszőleges $\mathbf{D}(s)$ polinommátrix esetén $\deg_{c_i}\mathbf{D}(s)$ a mátrix i-edik oszlopának fokát, azaz az i-edik oszlopbeli polinomok fokának maximumát.

7.1. TÉTEL. Legyen $\mathbf{H}(s)$ $m\times r$ dimenziós szigorúan proper (proper) racionális törtmátrix, melyre $\mathbf{H}(s)=\mathbf{N}(s)\mathbf{D}(s)^{-1}$. Ekkor

$$\deg_{ci}\mathbf{N}(s) < \deg_{ci}\mathbf{D}(s) \tag{7.4}$$

(ill. $\deg_{ci}\mathbf{N}(s) \leqq \deg_{ci}\mathbf{D}(s)$) $\quad i = 1, \ldots, r.$

Bizonyítás. Mivel $\mathbf{N}(s)=\mathbf{H}(s)\mathbf{D}(s)$, ezért

$$n_{ij} = \sum_{k=1}^{r} h_{ik}d_{kj}.$$

$\mathbf{H}$ szigorú propersége miatt minden eleme valódi tört. Így

$$\deg(h_{ik}d_{kj}) < \deg d_{kj} \leqq \deg_{cj}\mathbf{D}(s).$$

Ebből

$$\deg n_{ij} = \deg\Big(\sum_{1}^{r} h_{ik}d_{kj}\Big) < \deg_{cj}\mathbf{D}(s).$$

Ez pedig igaz az $\mathbf{N}$ mátrix j-edik oszlopának minden elemére, így ezek közül a maximális fokúra is.

Az állítás megfordítása nem igaz. Legyen például

$$\mathbf{H}(s) = \begin{pmatrix} s & 1 \\ 1 & 1 \end{pmatrix}, \quad \mathbf{D}(s) = \begin{pmatrix} 1 & 2 \\ s & s \end{pmatrix}, \quad \mathbf{N}(s) = \begin{pmatrix} 2s & 3s \\ s+1 & s+2 \end{pmatrix}.$$

Könnyen látható, hogy $\mathbf{H}(s)=\mathbf{N}(s)\mathbf{D}(s)^{-1}$, $\mathbf{D}(s)$ oszlopainak foka egyenlő $\mathbf{N}(s)$ megfelelő oszlopainak a fokával, $\mathbf{H}(s)$ azonban nem proper. Ez azzal magyarázható, hogy $\mathbf{D}(s)$-ben „felesleges" deriválás van. Tekintsük az $\mathbf{U}(s)=\begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}$ unimoduláris mátrixot. Az $\mathbf{U}(s)$-sel való jobbszorzás azt jelenti, hogy az első oszlopot kivonjuk a másodikból.

Legyen $\mathbf{D}'(s)=\mathbf{D}(s)\mathbf{U}(s)=\begin{pmatrix} 1 & 1 \\ s & 0 \end{pmatrix}$,

$\mathbf{N}'(s)=\mathbf{N}(s)\mathbf{U}(s)=\begin{pmatrix} 2s & s \\ s+1 & 1 \end{pmatrix}$. Ekkor $\mathbf{H}(s)=\mathbf{N}'(s)\mathbf{D}'(s)^{-1}$ és itt $\mathbf{D}'$ és $\mathbf{N}'$ oszlopaira nem teljesül a 7.1. tétel feltétele.

Tehát a 7.1. tétel megfordításához előbb ki kell szűrni a „felesleges" deriválásokat $\mathbf{D}(s)$-ből.

7.3. Definíció. Tetszőleges $\mathbf{D}(s)$ polinommátrix esetén jelölje $\mathbf{D}_{hc}$ azt a konstans mátrixot, melynek i-edik oszlopa $\mathbf{D}(s)$ i-edik oszlopa $\deg_{ci}\mathbf{D}(s)$ fokú tagjainak főegyütthatóiból áll. A $\mathbf{D}(s)$ mátrixot *oszloppropernek* nevezzük, ha $\mathbf{D}_{hc}$ teljes rangú.

7.1. LEMMA. A $\mathbf{D}(s)$ $r\times r$-es négyzetes polinommátrix pontosan akkor oszlopproper, ha

$$\deg|\mathbf{D}(s)| = \sum_{j=1}^{r} \deg_{cj}\mathbf{D}(s). \tag{7.5}$$

Bizonyítás. A lemma rögtön következik az alábbi felírásból:

$$\mathbf{D}(s) = \mathbf{D}_{hc}\begin{pmatrix} s^{k_1} & 0 \\ & \ddots \\ 0 & s^{k_r} \end{pmatrix} + \mathbf{L}(s), \tag{7.6}$$

ahol $k_j = \deg_{cj} \mathbf{D}(s)$, $\mathbf{D}_{hc}$ a definícióban szereplő konstans mátrix, és $\deg_{cj} \mathbf{L}(s) < k_j$. Így $\det \mathbf{D}(s) = \det \mathbf{D}_{hc}\, s^{\Sigma k_j} +$ alacsonyabb fokú tagok.

Most már kimondhatjuk az előző állítás megfordítását.

7.2. TÉTEL. Legyen $\mathbf{H}(s) = \mathbf{N}(s)\mathbf{D}^{-1}(s)$, $\mathbf{D}(s)$ oszloppropert, és $\deg_{cj} \mathbf{D}(s) > \deg_{cj} \mathbf{N}(s)$, $j = 1, \ldots, r$. Ekkor $\mathbf{H}(s)$ szigorúan proper racionális törtmátrix. Ha $>$ helyett $\geqq$ áll, akkor $\mathbf{H}(s)$ proper.

Bizonyítás. Az inverzmátrixra vonatkozó összefüggés alapján: $\mathbf{H}(s) = \mathbf{N}(s)$ Adj $\mathbf{D}(s)/\det \mathbf{D}(s)$.

Elemenként megvizsgálva:

$$h_{ij}(s) = \frac{1}{\det \mathbf{D}(s)} \sum_{k=1}^{r} n_{ik}(s) D^{kj}(s)$$

ahol $\mathbf{D}^{kj}(s)$ az adjungált mátrix (k,j)-edik eleme, azaz a $\mathbf{D}(s)$ mátrix (j,k)-adik eleméhez tartozó aldeterminánsa. Emiatt a bal oldalon a Σ mögötti rész nem más mint annak a $\overline{\mathbf{D}}^{ij}$ mátrixnak a determinánsa, amit úgy kapunk, hogy $\mathbf{D}(s)$ j-edik sorát $\mathbf{N}(s)$ i-edik sorára cseréljük ki. Így

$$h_{ij}(s) = \frac{\det \overline{D}^{ij}(s)}{\det \mathbf{D}(s)}. \tag{7.7}$$

Mivel $\mathbf{N}(s)$ oszlopainak foka kisebb, mint $\mathbf{D}(s)$ megfelelő oszlopainak foka, így a sorcserénél kapott mátrix oszlopfokai nem nőttek. Tehát $\overline{\mathbf{D}}^{ij}(s)$ is felírható (7.6)-nak megfelelő alakban.

$$\overline{\mathbf{D}}^{ij}(s) = \overline{\mathbf{D}}^{ij}_{hc}\begin{pmatrix} s^{k_1} & 0 \\ & \ddots \\ 0 & s^{k_r} \end{pmatrix} + \mathbf{L}^{ij}(s). \tag{7.8}$$

A $\overline{\mathbf{D}}^{ij}(s)$ mátrix a j-edik sortól eltekintve azonos $\mathbf{D}(s)$-sel, így ez igaz $\overline{\mathbf{D}}^{ij}_{hc}$ és $\overline{\mathbf{D}}_{hc}$-re is. $\overline{\mathbf{D}}^{ij}_{hc}$ j-edik sora azonosan nulla, hisz $\overline{\mathbf{D}}^{ij}(s)$ j-edik sora $\mathbf{N}(s)$ egy sorával egyenlő; és a feltétel szerint $\mathbf{N}(s)$ oszlopainak a foka kisebb, mint $\mathbf{D}(s)$ megfelelő oszlopának foka. $\overline{\mathbf{D}}^{ij}_{hc}$ szingularitása miatt a 7.1. lemma alapján

$$\deg |\overline{\mathbf{D}}^{ij}(s)| < \sum_{l=1}^{r} k_l = \deg |\mathbf{D}(s)|.$$

Ez azt jelenti, hogy $h_{ij}(s)$ valódi tört, így $\mathbf{H}(s)$ szigorúan proper.

Ha $<$ helyett $\leqq$ áll, akkor elegendő a $\deg |\overline{\mathbf{D}}^{ij}| \leqq \sum_{l=1}^{r} k_l$ egyenlőtlenségre hivatkozni.

Ha tehát egy $\mathbf{H}(s)$ racionális-törtmátrix properségét szeretnénk megvizsgálni, akkor azt $\mathbf{H}(s) = \mathbf{N}(s)\mathbf{D}^{-1}(s)$ alakba kell átírni, ahol $\mathbf{D}(s)$ oszloppropert.

7.2. LEMMA. Tetszőleges $\mathbf{D}(s)$ $r\times r$-es mátrix oszloptranszformációkkal oszlopproperré transzformálható.

Bizonyítás. Legyen

$$\mathbf{D}(s) = \mathbf{D}_{\mathrm{hc}} \begin{pmatrix} s^{k_1} & & \\ & \ddots & \\ & & s^{k_r} \end{pmatrix} + \mathbf{L}_{\mathrm{hc}}$$

ahol most $k_1 \geqq k_2 \ldots \geqq k_r$ feltehető hiszen ez oszlopcserékkel elérhető. Tegyük fel, hogy $\mathbf{D}_{\mathrm{hc}}$ szinguláris. Ekkor például első oszlopa kifejezhető a többi lineáris kombinációjával: $\mathbf{D}_{\mathrm{hc}}^{(1)} = \sum_{j=2}^{r} \alpha_j \mathbf{D}_{\mathrm{hc}}^{(j)}$.

Hajtsuk most végre a $\mathbf{D}(s)$ mátrixon a következő oszloptranszformációkat: Az első oszlopból levonjuk a j-edik oszlop $\alpha_j s^{k_1-k_j}$-szeresét, $j=2, \ldots, r$. Ezek után az első oszlop fokszáma csökken. Mivel $\sum_{j=1}^{r} k_j \geqq \deg |\mathbf{D}(s)|$, így az eljárást ismételve véges sok lépésben oszloppropert mátrixhoz jutunk.

A lemma alapján tetszőleges $\mathbf{H}(s)=\mathbf{N}(s)\mathbf{D}^{-1}(s)$ mátrix-törtfelíráshoz tudunk konstruálni olyan $\mathbf{H}(s)=\mathbf{N}'(s)\mathbf{D}'^{-1}(s)$ felbontást, ahol $\mathbf{D}'(s)$ oszloppropert, így $\mathbf{H}(s)$ propersége ellenőrizhető. Még egy apróságot itt látnunk kell: ez a transzformáció bizonyos értelemben egyértelmű. Erről szól a következő lemma:

7.3. LEMMA. Legyenek $\mathbf{D}(s)$ és $\mathbf{D}'(s)$ $r\times r$-es oszloppropert mátrixok. Legyen $\mathbf{D}(s)=\mathbf{D}'(s)\mathbf{U}(s)$, ahol $\mathbf{U}(s)$ unimoduláris. Legyen $k_j=\deg_{\mathrm{cj}} \mathbf{D}(s)$, $k'_j=\deg_{\mathrm{cj}} \mathbf{D}'(s)$. Tegyük fel, hogy $k_1 \leqq \ldots \leqq k_r$, $k'_1 \leqq \ldots \leqq k'_r$. Ekkor $k_j=k'_j$, $j=1, \ldots, r$.

Bizonyítás. Először belátjuk az alábbi tulajdonságot:

7.4. LEMMA. (Predictable degree property.) Legyen $\mathbf{D}(s)$ oszloppropert polinommátrix, $\mathbf{q}(s)$ tetszőleges polinomvektor. Legyen $\mathbf{p}(s)=\mathbf{D}(s)\mathbf{q}(s)$. Ekkor

$$\deg \mathbf{p}(s) = \max_{q_j(s)\neq 0} \{\deg_{\mathrm{cj}} \mathbf{D}(s) + \deg q_j(s)\}.$$

(Emlékeztetünk arra, hogy polinomvektor foka az elemek fokainak maximumával egyenlő.)

Bizonyítás. Legyen $d = \max_{q_j(s)\neq 0} \{\deg_{\mathrm{cj}} \mathbf{D}(s) + \deg q_j(s)\}$. $\deg \mathbf{p}(s) \leqq d$ azonnal látható. Így $\mathbf{p}(s) = \sum_{j=0}^{r} \mathbf{p}^{(j)} s^j$, ahol $\mathbf{p}^{(j)}$ r-dimenziós vektor. Legyen α r-dimenziós vektor, melynek j-edik komponense $\mathbf{q}(s)$ j-edik komponensének $d-\deg_{\mathrm{cj}} \mathbf{D}(s)$ fokú tagjának főegyütthatója. d definíciója alapján $\alpha \neq 0$. Ekkor $\mathbf{p}^{(d)} = \mathbf{D}_{\mathrm{hc}} \cdot \alpha$, ami $\mathbf{D}_{\mathrm{hc}}$ teljes rangúsága miatt nem nulla.

Ezek után rátértünk a 7.3. lemma bizonyítására.

Jelölje $\mathbf{D}^{(i)}(s)$ és $\mathbf{U}^{(i)}(s)$ a $\mathbf{D}(s)$ és $\mathbf{U}(s)$ mátrixok i-edik oszlopát. Tegyük fel, hogy valamely j-re $k_j < k'_j$. A lemma alapján számítsuk ki $\mathbf{D}^{(i)}(s)$ fokát $i \leqq j$ esetén.

Mivel $\mathbf{D}^{(i)}(s)=\mathbf{D}(s)\mathbf{U}^{(i)}(s)$, így a lemma szerint

$$\deg \mathbf{D}^{(i)}(s) = \max_{r:\, \mathbf{U}_{ri}(s)\neq 0} \{\deg_{\mathrm{cr}} \mathbf{D}'(s) + \deg \mathbf{U}_{ri}(s)\}$$

azaz

$$k_i = \max_{r:\, \mathbf{U}_{ri}(s)\neq 0} \{k'_r + \deg \mathbf{U}_{ri}(s)\}.$$

$k_j<k'_j$ miatt $k_i<k'_r$ $r \geqq j \geqq i$ esetén, így a fenti azonosság csak $\mathbf{U}_{ri}(s)=\mathbf{O}$ $r \geqq j$ mellett állhat fenn. Ez igaz $i=1, 2, \ldots, j$ esetén, tehát azt kaptuk, hogy az $\mathbf{U}(s)$ mátrix első j oszlopában legfeljebb az első $j-1$ elem lehet nem nulla, ami ellentmond annak, hogy $\mathbf{U}(s)$ unimoduláris, így minden s-re teljes rangú.

8. Racionális mátrix-egyenletek speciális megoldásai

Legyenek $\mathbf{A}(s)$, $\mathbf{B}(s)$ $m\times r$-es, illetve $p\times r$-es racionális törtmátrixok. Tegyük fel, hogy $\mathbf{A}(s)$ teljes oszloprangú. A

$$\mathbf{H}(s)\cdot\mathbf{A}(s) = \mathbf{B}(s) \tag{8.1}$$

egyenlet megoldásait vizsgáljuk. Feltételt adunk arra, hogy létezzen olyan megoldás, amelynek nincsen előre megadott α-ban pólusa.

Legyen $\mathbf{F}(s)=\begin{pmatrix}\mathbf{B}(s)\\ \mathbf{A}(s)\end{pmatrix}$, rögzített véges szám. $\mathbf{F}(s)$ *Smith—McMillan alakját* felhasználva, a következő alakban írható:

$$\mathbf{F}(s) = \mathbf{U}(s)\mathbf{M}(s)\mathbf{V}(s) = \mathbf{U}(s)\overline{\mathbf{M}}(s)\mathbf{M}_\alpha(s)\mathbf{V}(s),$$

ahol

$$\mathbf{M}_\alpha(s) = \begin{pmatrix} (s-\alpha)^{\sigma_1(\alpha)} & & & 0 \\ & \ddots & & \\ & & (s-\alpha)^{\sigma_k(\alpha)} & \\ & & & 1 \\ & & & & \ddots \\ 0 & & & & & 1 \end{pmatrix}$$

és $\sigma_1(\alpha)\ldots\sigma_k(\alpha)$ jelölik az $\mathbf{F}(s)$ mátrix α-hoz tartozó strukturális indexeit. Jelölje

$$\overline{\mathbf{F}}(s) = \mathbf{U}(s)\overline{\mathbf{M}}(s) = \begin{pmatrix}\overline{\mathbf{B}}(s)\\ \overline{\mathbf{A}}(s)\end{pmatrix}$$

és

$$\mathbf{Q}(s) = \mathbf{M}_\alpha(s)\mathbf{V}(s).$$

Ekkor nyilván az teljesül, hogy $\overline{\mathbf{F}}(s)$-nek α-ban sem zérusa, sem pólusa nincsen. Megjegyezzük, hogy ekkor $\overline{\mathbf{A}}(s)$-nek és $\overline{\mathbf{B}}(s)$-nek sincsen pólusa α-ban, hiszen egy racionális törtfüggvény pólusai az irredubilis alakban felírt mátrixelemek nevezőinek legkisebb közös többszöröse. Mivel $\mathbf{Q}(s)$ nem szinguláris, így (8.1) átírható a következő alakba:

$$\mathbf{H}(s)\overline{\mathbf{A}}(s) = \overline{\mathbf{B}}(s).$$

8.1. LEMMA. Legyen $\mathbf{A}(s)$ $m\times r$-es, teljes oszloprangú racionális törtmátrix. Ekkor létezik $\mathbf{A}^L(s)$ balinverz, azaz olyan $r\times m$-es racionális törtmátrix, melyre $\mathbf{A}^L(s)\mathbf{A}(s)=\mathbf{I}_r$. Továbbá ha $\mathbf{A}(s)$-nek α-ban nincsen zérusa, akkor a fenti $\mathbf{A}^L(s)$-nek α-ban nincsen pólusa.

Bizonyítás. Először belátjuk, hogy a fenti feltételekkel $\mathbf{A}^T(s)\mathbf{A}(s)$ nem szinguláris. Ugyanis ellenkező esetben a 2.3. tétel alapján létezne olyan $\mathbf{x}(s)\neq 0$ polinomvektor, melyre

$$\mathbf{A}^T(s)\mathbf{A}(s)\mathbf{x}(s) = \mathbf{0}.$$

Ekkor nyilván

$$\mathbf{x}^T(s)\mathbf{A}^T(s)\mathbf{A}(s)\mathbf{x}(s) = \mathbf{0},$$

$$\|\mathbf{A}(s)\mathbf{x}(s)\|^2 = 0,$$

azaz $\mathbf{A}(s)\mathbf{x}(s)=\mathbf{0}$, amiből $\mathbf{x}(s)=\mathbf{0}$ következne, hiszen $\mathbf{A}(s)$ teljes oszloprangú. Ez ellentmondás, így $\mathbf{A}^T(s)\mathbf{A}(s)$ valóban nem szinguláris. Legyen

$$\mathbf{A}^L(s) \triangleq [\mathbf{A}^T(s)\mathbf{A}(s)]^{-1}\mathbf{A}^T(s).$$

Ez nyilván balinverze $\mathbf{A}(s)$-nek, azaz

$$\mathbf{A}^L(s)\mathbf{A}(s) = \mathbf{I}_r.$$

A lemma második részének bizonyításához írjuk fel $\mathbf{A}(s)$ *Smith—McMillan alakját*, legyen

$$\mathbf{A}(s) = \mathbf{U}_1(s)\mathbf{M}(s)\mathbf{U}_2(s).$$

Ekkor egyszerű számítások után azt kapjuk, hogy

$$\mathbf{A}^L(s) = [\mathbf{U}_2^T(s)\mathbf{M}^T(s)\mathbf{U}_1^T(s)\mathbf{U}_1(s)\mathbf{M}(s)\mathbf{U}_2(s)]^{-1}\times$$

$$\times\mathbf{U}_2^T(s)\mathbf{M}^T(s)\mathbf{U}_1^T(s) = \mathbf{U}_2(s)^{-1}\mathbf{M}^+(s)\mathbf{U}_1(s)^{-1},$$

ahol $\mathbf{M}^+(s)$ az $\mathbf{M}(s)$ mátrix általánosított inverze, azaz olyan $r\times m$-es racionális törtmátrix, melynek fődiagonálisában $\mathbf{M}(s)$ megfelelő elemeinek reciproka áll, illetve a többi eleme 0. Így $\mathbf{A}^L(s)$ *Smith—McMillan-alakját* kaptuk, amiből a lemma második része következik.

Visszatérve a feladatunkhoz a lemma alapján azt kapjuk, hogy ha $\overline{\mathbf{A}}(s)$-nek nincsen zérusa α-ban, akkor a $\mathbf{H}(s)=\overline{\mathbf{B}}(s)\overline{\mathbf{A}}^{-L}(s)$ megoldásnak nincsen pólusa α-ban. Bebizonyítjuk ennek megfordítását is; belátjuk, hogy ha $\mathbf{A}(s)$-nek van zérusa α-ban, akkor minden $\mathbf{H}(s)$ megoldásnak pólusa van α-ban.

Induljunk ki az

$$\overline{\mathbf{F}}(s) = \begin{pmatrix}\mathbf{H}(s)\\ \mathbf{I}\end{pmatrix}\overline{\mathbf{A}}(s)$$

felírásból. Mivel $\overline{\mathbf{F}}(s)$-nek α-ban sem pólusa, sem zérusa nincs, így $\overline{\mathbf{F}}(\alpha)$ véges elemű, teljes oszloprangú mátrix. Ha $\overline{\mathbf{A}}(\alpha)$ nem teljes rangú, akkor $\mathbf{H}(\alpha)$ nem lehet véges, hiszen véges mátrixot nem teljes rangú mátrixszal szorozva nem kaphatunk teljes rangú mátrixot. Beláttuk tehát, hogy:

8.1. TÉTEL. A (8.1) egyenletnek pontosan akkor van olyan $\mathbf{H}(s)$ megoldása, melynek nincsen pólusa α-ban, ha $\overline{\mathbf{A}}(s)$-nek nincsen zérusa α-ban.

Ezt a feltételt szeretnénk az $\mathbf{A}(s)$ és $\mathbf{B}(s)$ mátrixok segítségével megfogalmazni.

Mivel $\overline{\mathbf{A}}(s)$-nek α-ban nincsen pólusa, így

$$v_\alpha^{(r)}(\overline{\mathbf{A}}) = \overline{\mathbf{A}} \text{ zérusainak foka } \alpha\text{-ban.}$$

$\bar{\mathbf{A}}(s)$ definíciója alapján

$$\bar{\mathbf{A}}(s) = \mathbf{A}(s)\mathbf{Q}^{-1}(s),$$

ahol $\mathbf{Q}(s)$ $r\times r$-es nem szinguláris mátrix, így

$$\begin{aligned} v_\alpha^{(r)}(\bar{\mathbf{A}}) &= \min\{v_\alpha(|\bar{\mathbf{A}}|^r)\colon\ |\bar{\mathbf{A}}|^r \quad r\times r\text{-es aldeterminánsok}\} = \\ &= \min\{v_\alpha(|\mathbf{A}|^r|\mathbf{Q}^{-1}|)\colon\ |\mathbf{A}|^r \quad r\times r\text{-es aldeterminánsok}\} = \\ &= \min\{v_\alpha(|\mathbf{A}|^r)\colon\ |\mathbf{A}|^r \quad r\times r\text{-es aldeterminánsok}\} + v_\alpha^{(r)}(\mathbf{Q}^{-1}) = \\ &= v_\alpha^{(r)}(\mathbf{A}) - v_\alpha^{(r)}(\mathbf{Q}) = v_\alpha^{(r)}(\mathbf{A}) - v_\alpha^{(r)}(\mathbf{M}_\alpha) = v_\alpha^{(r)}(\mathbf{A}) - v_\alpha^{(r)}(\mathbf{F}), \end{aligned}$$

hiszen $\mathbf{Q}(s)=\mathbf{M}_\alpha(s)\mathbf{V}(s)$, ahol $\mathbf{M}_\alpha(s)$-ben éppen az $\mathbf{F}(s)$ mátrix α-hoz tartozó strukturális indexei vannak. Így igaz a következő:

8.2. TÉTEL. A (8.1) egyenletnek pontosan akkor létezik olyan megoldása, melynek nincsen pólusa α-ban, ha

$$v_\alpha^{(r)}(\mathbf{A}) = v_\alpha^{(r)} F.$$

Megjegyzés. A fenti gondolatmenetet véges α esetén végeztük el, azonban könnyen látható, hogy $s=\lambda^{-1}$ helyettesítéssel $\alpha=\infty$-re is ugyanezt az eredményt kapjuk.

8.1. KÖVETKEZMÉNY. A (8.1) egyenletnek pontosan akkor létezik proper megoldása, ha

$$v_\infty^{(r)}(\mathbf{A}) = v_\infty^{(r)}(\mathbf{F}).$$

8.2. KÖVETKEZMÉNY. A (8.1) egyenletnek pontosan akkor van stabil megoldása, ha

$$v_\alpha^{(r)}(\mathbf{A}) = v_\alpha^{(r)}(\mathbf{F})$$

teljesül minden olyan α-ra, melyre $\operatorname{Re}\alpha\geqq 0$.

Alkalmazzuk a 8.1. következményt abban az esetben, mikor $\mathbf{F}(s)$ polinommátrix. Feltehető, hogy oszloppproper, hiszen ez unimoduláris mátrixszal való jobb szorzással elérhető.

8.2. LEMMA. Tetszőleges $\mathbf{F}(s)$ $n\times r$-es oszloppproper mátrix esetén

$$v_\infty^{(r)}(\mathbf{F}) = -\sum_{i=1}^{r} \deg_{ci}\mathbf{F}.$$

Bizonyítás. A kiértékelések definíciója szerint

$$v_\infty^{(r)}(\mathbf{F}(s)) = \min\{v_\infty^{(r)}(|\mathbf{F}|^r)\colon\ |\mathbf{F}|^r \quad r\times r\text{-es aldeterminánsok}\}.$$

Tetszőleges $g(s)$ polinom esetén

$$v_\infty(g(s)) = -\deg g(s),$$

így

$$v_\infty^{(r)}(\mathbf{F}(s)) = -\max\{\deg(|\mathbf{F}|^r)\colon\ |\mathbf{F}|^r \quad r\times r\text{-es aldeterminánsok}\}.$$

$\mathbf{F}(s)$ oszloppropersége alapján a jobb oldal éppen az oszlopfokok összegének -1-szerese, így a lemmát beláttuk.

Így polinommátrixok esetén a következő eredményt kaptuk:

8.3. KÖVETKEZMÉNY. Ha $\mathbf{F}(s)$ oszloppproper, akkor a (8.1) egyenletnek pontosan akkor létezik proper megoldása, ha $\mathbf{A}(s)$ oszloppproper és $\deg_{ci}\mathbf{A}(s) \geqq \deg_{ci}\mathbf{B}(s)$.

Bizonyítás. A $v_\infty^{(r)}(\mathbf{A}) = v_\infty^{(r)}(\mathbf{F})$ feltétel és az előző lemma alapján nyilvánvaló.

9. Skalár differenciálegyenlet irányítható realizációja

Tekintsük a következő differenciálegyenletet:

$$A(d)y(t) = C(d)u(t) \tag{9.1}$$

ahol $\mathbf{A}(d)$ a $d=\partial/\partial t$ differenciáloperátornak n-ed fokú, míg $C(d)$ $n-1$-ed fokú polinomja.

$$A(d) = \sum_0^n a_i d^i, \quad a_n = 1, \tag{9.2}$$

$$C(d) = \sum_0^{n-1} c_i d^i. \tag{9.3}$$

Bevezetve a $\xi = \mathbf{A}^{-1}u$ segédváltozót, a (9.1) egyenlet a következőképpen írható:

$$A(d)\xi(t) = u(t), \tag{9.4}$$

$$y(t) = C(d)\xi(t). \tag{9.5}$$

Fejezzük ki az új változó legmagasabb-fokú deriváltját az alacsonyabb fokú deriváltak segítségével.

$$\xi^{(n)}(t) = -a_{n-1}\xi^{(n-1)}(t) - \ldots - a_0\xi(t) + u(t). \tag{9.6}$$

Természetesen adódik a következő állapotvektor bevezetése:

$$\mathbf{x}(t) = \left(\xi(t), \ldots, \xi^{(n-1)}(t)\right)^T. \tag{9.7}$$

Az állapottéregyenletek:

$$\dot{\mathbf{x}}(t) = \mathbf{A}_c\mathbf{x}(t) + \mathbf{B}_c u(t), \tag{9.8}$$

$$y(t) = \mathbf{C}_c\mathbf{x}(t) \tag{9.9}$$

ahol $\mathbf{A}_c$ az A polinom együtthatóiból alkotott ún. kísérő mátrix (companion-matrix)

$$\mathbf{A}_c = \begin{pmatrix} 0 & & 1 \\ & \ddots & \\ & 0 & 1 \\ -a_0 & \ldots & -a_{n-1} \end{pmatrix},$$

$\mathbf{B}_c = [0 \ldots 0\ 1]^T$, $\mathbf{C}_c = [c_0 \ldots c_{n-1}]$.

9.1. TÉTEL. Az $(\mathbf{A}_c, \mathbf{B}_c)$ pár irányítható.

Bizonyítás. A 4.3. tétel alapján azt kell igazolni, hogy $[s\mathbf{I}-\mathbf{A}_c\mathbf{B}_c]$ teljes rangú minden s-re. Valóban,

$$[s\mathbf{I}-\mathbf{A}_c\mathbf{B}_c] = \begin{pmatrix} s & -1 & & & 0 \\ & s & -1 & & \\ & & \ddots & -1 & 0 \\ & & & \ddots & \ddots \\ a_0 & a_1 & \dots & s+a_{n-1} & 1 \end{pmatrix}$$

az utolsó n oszlopból álló mátrix determinánsa s-től függetlenül $+1$ vagy -1.

9.2. TÉTEL. A $(\mathbf{C}_c, \mathbf{A}_c)$ pár pontosan akkor megfigyelhető, ha az $A(d)$ és $C(d)$ polinomok relatív prímek.

A tétel bizonyítása előtt három elemi lemmát igazolunk:

9.1. LEMMA. $\det(s\mathbf{I}-\mathbf{A}_c)=\mathbf{A}(s)$.

Bizonyítás.

$$s\mathbf{I}-\mathbf{A}_c = \begin{pmatrix} s & -1 & & \\ & s & -1 & \\ & & s & -1 \\ a_0 & & & a_{n-1}+s \end{pmatrix}.$$

Hajtsuk végre a mátrixon a következő elemi oszloptranszformációkat:

adjuk hozzá a j-edik oszlophoz a $j+1$-edik oszlop s-szeresét $j=n-1, n-2, \dots, 1$. A transzformáció után kapott mátrix:

$$\begin{pmatrix} 0 & -1 & & 0 \\ 0 & \ddots & \ddots & \\ & & 0 & -1 \\ \mathbf{A}(s) & x & x & x \end{pmatrix}$$

ahol $x, \dots, x$ számunkra érdektelen kifejezéseket jelölnek. Ennek a mátrixnak a determinánsa $A(s)$, és mivel az oszloptranszformációk a determinánst nem változtatják meg, így az eredeti mátrix determinánsa is $A(s)$.

9.2. LEMMA. Cseréljük ki $(s\mathbf{I}-\mathbf{A}_c)$ utolsó sorát a $\mathbf{C}_c=(c_0 \dots c_{n-1})$ vektorral. Az így kapott mátrix determinánsa $C(s)$.

Bizonyítás. Ugyanazokat az oszloptranszformációkat végrehajtva a mátrixon mint az előző bizonyításban tettük, a

$$\begin{pmatrix} 0 & -1 & & \\ & \ddots & \ddots & \\ & & 0 & -1 \\ C(s) & x & \dots & x \end{pmatrix}$$

mátrixhoz jutunk. Ebből a lemma következik.

9.3. LEMMA. Az $A(s)$ és $C(s)$ polinomok helyettesítési értékét megkaphatjuk a következőképpen:

$$(s_0\mathbf{I}-\mathbf{A}_c)\begin{pmatrix}1\\ s_0\\ \vdots\\ s_0^{n-1}\end{pmatrix}=\begin{pmatrix}s_0 & -1 & & \\ & s_0 & -1 & \\ & & s_0 & -1\\ a_0 & a_1 & & a_{n-1}+s_0\end{pmatrix}\begin{pmatrix}1\\ s_0\\ \vdots\\ s_0^{n-1}\end{pmatrix}=\begin{pmatrix}0\\ \vdots\\ 0\\ A(s_0)\end{pmatrix}$$

és

$$(c_0 c_1 \dots c_{n-1})\begin{pmatrix}1\\ s_0\\ \vdots\\ s_0^{n-1}\end{pmatrix}=C(s_0).$$

Bizonyítás. A lemma közvetlenül látható a szorzások elvégzése után.

A 9.2. tétel bizonyítása. A 4.5. tétel alapján $(\mathbf{C}_c, \mathbf{A}_c)$ pontosan akkor megfigyelhető, ha $\mathbf{R}(s)=\begin{pmatrix}s\mathbf{I}-\mathbf{A}_c\\ \mathbf{C}_c\end{pmatrix}$ teljes rangú minden s-re.

Tegyük fel először, hogy $\mathbf{R}(s)$ teljes rangú minden s-re. Ekkor tetszőleges s_0 esetén a 9.3. lemma alapján

$$\mathbf{R}(s_0)\begin{pmatrix}1\\ s_0\\ \vdots\\ s_0^{n-1}\end{pmatrix}=\begin{pmatrix}0\\ \vdots\\ 0\\ A(s_0)\\ C(s_0)\end{pmatrix}\neq 0.$$

Ez éppen azt jelenti, hogy a polinomok relatív prímek.

Tegyük fel most fordítva, hogy a polinomoknak nincs közös gyöke. Ekkor ha $A(s_0)\neq 0$, akkor $R(s_0)$ első n sorából álló mátrix determinánsa nem nulla, hiszen ez a 9.1. lemma alapján éppen $A(s_0)$. Ha $A(s_0)=0$, akkor $C(s_0)\neq 0$. Így $\mathbf{R}(s_0)$-ból az n-edik sor elhagyásával kapott mátrix determinánsa nem nulla, hiszen a 9.2. lemma alapján ez $C(s_0)$.

Megjegyzés. A 9.2. tétel nem meglepő, hiszen közös gyök esetén az n-ed fokú differenciálegyenlet redukálható lenne alacsonyabb fokú egyenletre, és így kisebb dimenziójú állapottér modellel lehetne a fenti módon realizálni.

10. Jobb oldali mátrixtörtfelbontás irányítható realizációja

Tekintsük a $\mathbf{H}(s)$ $r\times m$ dimenziós szigorúan proper átmeneti függvényt. A frekvenciatartományban ez az

$$\mathbf{y}(s)=\mathbf{H}(s)\mathbf{u}(s) \tag{10.1}$$

kapcsolatot jelöli az input- és output-folyamatok *Laplace transzformáltjai* között. Ennek szeretnénk irányítható állapottér realizációját felírni, ha $\mathbf{H}(s)=\mathbf{N}(s)\mathbf{D}^{-1}(s)$

jobb oldali törtfelbontása ismeretes. Vezessük be a $\xi(s)=\mathbf{D}^{-1}(s)\mathbf{u}(s)$ folyamatot. Így (10.1) átírható a következő alakba:

$$\mathbf{D}(s)\xi(s) = \mathbf{u}(s), \tag{10.2}$$

$$\mathbf{y}(s) = \mathbf{N}(s)\xi(s). \tag{10.3}$$

Skalárváltozók esetén ugyanígy kezdtünk hozzá az irányítható realizáció megkonstruálásához. Ott ezután az új változó legmagasabb rendű deriváltját kifejeztük alacsonyabb rendű deriváltak segítségével, és állapotterünk ezekből a deriváltakból állt. Itt is hasonló módon szeretnénk eljárni. Válasszuk le $\mathbf{D}(s)$-ból oszloponként a legmagasabb fokú tagokat. Legyen.

$$\mathbf{D}(s) = \mathbf{D}_{hc}\mathbf{S}(s)+\mathbf{D}_{lc}\psi(s) \tag{10.4}$$

ahol

$$\mathbf{S}(s) = \begin{pmatrix} s^{k_1} & & 0 \\ & \ddots & \\ 0 & & s^{k_m} \end{pmatrix}, \qquad \psi(s) = \begin{pmatrix} s_1^{k-1} & & & \\ \vdots & & & \\ 1 & & & \\ & s_2^{k-1} & & \\ & \vdots & & \\ & 1 & & \\ & & \ddots & \\ & & & s_n^{k-1} \\ & & & \vdots \\ & & & 1 \end{pmatrix} \tag{10.5}$$

továbbá $k_1, \ldots, k_m$ $\mathbf{D}(s)$ oszlopainak foka. Így

$$\mathbf{D}_{hc}\mathbf{S}(s)\xi(s) = -\mathbf{D}_{lc}\psi(s)\xi(s)+\mathbf{u}(s).$$

Látható, hogy $\mathbf{D}_{hc}$ regularitása szükséges az átosztáshoz. Emiatt a következőkben feltesszük, hogy $\mathbf{D}(s)$ oszlopproper mátrix. Ekkor

$$\mathbf{S}(s)\xi(s) = -\mathbf{D}_{hc}^{-1}\mathbf{D}_{lc}\psi(s)\xi(s)+\mathbf{D}_{hc}^{-1}\mathbf{u}(s).$$

Írjuk ki ezt az összefüggést koordinátánként:

$$\begin{pmatrix} s^{k_1}\xi_1(s) \\ \vdots \\ s^{k_m}\xi_m(s) \end{pmatrix} = -\mathbf{D}_{hc}^{-1}\mathbf{D}_{lc}\begin{pmatrix} s^{k_1-1}\xi_1(s) \\ \vdots \\ \xi_1(s) \\ \vdots \\ s^{k_n-1}\xi_m(s) \\ \vdots \\ \xi_m(s) \end{pmatrix} + \mathbf{D}_{hc}^{-1}\mathbf{u}(s). \tag{10.6}$$

Visszatérve időtartománybeli leírásba, vezessük be az

$$\mathbf{x}(t) = \begin{pmatrix} \xi_1^{(k_1-1)}(t) \\ \vdots \\ \xi_1(t) \\ \vdots \\ \xi_m^{(k_n-1)}(t) \\ \vdots \\ \xi_m(t) \end{pmatrix} \tag{10.7}$$

állapotvektort. (Itt ξ-t és *Laplace transzformáltját* az egyszerűség kedvéért ugyanígy jelöltük.)

A fenti (10.6) összefüggés állapottér felírásban a következőképpen néz ki:

$$\dot{\mathbf{x}}(t) = \mathbf{A}_c\mathbf{x}(t) + \mathbf{B}_c\mathbf{u}(t), \tag{10.8}$$

ahol

$$\mathbf{A}_c = \begin{pmatrix} x \dots x & x & \dots & x & 1 \\ 1 & & & 0 & \\ & 1\ 0 & & & \\ x \dots x & x \dots x & \dots & x & k_1+1 \\ & & 1 & & \\ & & & 1\ 0 & \\ & 0 & & x \dots x & k_1+\dots+k_{m-1}+1, \\ & & & 1\ 1\ 0 & \end{pmatrix} \tag{10.9}$$

$$\mathbf{B}_c = \begin{pmatrix} x \dots x & 1 \\ 0 & \\ x \dots x & k_1 \\ 0 & \\ \vdots & \vdots \\ x \dots x & k_1+\dots+k_{m-1}+1. \\ 0 & \end{pmatrix}$$

Az $\mathbf{A}_c$ mátrix 1, k_1+1-edik, $\dots k_1+\dots+k_{m-1}+1$-edik sorában rendre a $-\mathbf{D}_{hc}^{-1}\mathbf{D}_{lc}$ mátrix 1, 2, ... m-edik sorai szerepelnek. A diagonálisban álló blokkok főátló alatti elemei 1-esek, az összes többi elem 0. A $\mathbf{B}_c$ mátrix 1, k_1+1-edik, ..., $k_1+\dots+k_{m-1}+1$-edik sorában a $\mathbf{D}_{hc}^{-1}$ mátrix 1, 2, ... m-edik sora áll, az összes többi elem 0.

A megfigyelési egyenlet felírásához elegendő azt észrevenni, hogy $\mathbf{D}(s)$ oszlopproperségе miatt $\mathbf{N}(s)$ oszlopainak foka határozottan kisebb mint $\mathbf{D}(s)$ megfelelő oszlopainak foka. Így

$$\mathbf{N}(s) = \mathbf{N}_{lc}\boldsymbol{\psi}(s) \tag{10.10}$$

alakban írható, tehát az $\mathbf{x}(t) = \mathscr{L}^{-1}(\boldsymbol{\psi}(s)\boldsymbol{\xi}(s))$ választás miatt az időtartományban

a (10.3) megfigyelési egyenlet:

$$\mathbf{y}(t) = \mathbf{N}_{lc}\mathbf{x}(t). \tag{10.11}$$

Belátjuk, hogy ez a realizáció irányítható.

10.1. TÉTEL. Az $(\mathbf{A}_c, \mathbf{B}_c)$ pár irányítható.

Előkészítésképpen bebizonyítjuk a következő lemmát:

10.1. LEMMA. Legyen

$$\mathbf{A}_0 = \begin{pmatrix} 0 & 0 & & & & & & & \\ 1 & \ddots & \ddots & & & & & & \\ & \ddots & 1 & 0 & & & & & \\ & & & 0 & 0 & & & & \\ & & & 1 & \ddots & \ddots & & & \\ & & & & \ddots & 1 & 0 & & \\ & & & & & & 0 & 0 & \\ & & & & & & 1 & \ddots & \\ & & & & & & \ddots & 1 & 0 \end{pmatrix} \begin{matrix} \} k_1 \\ \\ \\ \} k_2 \\ \\ \\ \} k_m \\ \\ \end{matrix}$$

$$\underbrace{\qquad}_{k_1} \quad \underbrace{\qquad}_{k_2} \quad \underbrace{\qquad}_{k_m}$$

$$\mathbf{B}_0 = \begin{pmatrix} 1 & 0 \dots & 0 \\ & 0 & \\ 0 & 1 \; 0 \dots & 0 \\ & 0 & \\ & \vdots & \\ 0 & \dots \; 0 & 1 \\ & 0 & \end{pmatrix} \begin{matrix} \} k_1 \\ \\ \} k_2 \\ \\ \\ \} k_m \\ \end{matrix}$$

azaz $\mathbf{A}_0$ olyan blokkmátrix, melynek diagonális blokkjaiban a főátló alatti elemek 1, az összes többi elem 0; $\mathbf{B}_0$ olyan blokkmátrix melyben 1, k_1+1-edik, ..., $k_1+\dots$ $\dots+k_{m-1}+1$-edik sorának rendre első, második, ..., m-edik eleme 1, az összes többi 0. Ekkor $(\mathbf{A}_0, \mathbf{B}_0)$ irányítható.

Bizonyítás. Be kell látni, hogy $(s\mathbf{I}-\mathbf{A}_0 \; \mathbf{B}_0)$ teljes rangú minden s-re. Ha $\mathbf{A}_0$, $\mathbf{B}_0$ egy-egy blokkból állna csak, akkor

$$((s\mathbf{I}-\mathbf{A}_0)\,\mathbf{B}_0) = \begin{pmatrix} s & & & 1 \\ -1 & s & & 0 \\ & & & \vdots \\ & & -1 \; s & 0 \end{pmatrix}.$$

Ez a $k\times(k+1)$-es mátrix nyilván teljes rangú minden s-re, hiszen van olyan $k\times k$-s aldeterminánsa, ami 1. (Az utolsó előtti oszlop elhagyásával kapott mátrix). Több blokkból álló mátrixokra a bizonyítás teljesen hasonló.

A 10.1. tétel bizonyítása. Tetszőleges $\mathbf{X}$ $m\times k$-s mátrix esetén a $\mathbf{B}_0\mathbf{X}$ mátrix olyan $(\sum_{i=1}^{m} k_i)\times k$-s mátrix lesz, melynek $1, k_1+1$-edik, ..., $k_1+\dots+k_{m-1}+1$-edik sora $\mathbf{X}$ 1, 2, ..., m-edik sorával egyenlő, az összes többi eleme pedig 0. Így

$$\mathbf{A}_c = \mathbf{A}_0 - \mathbf{B}_0\mathbf{D}_{hc}^{-1}\mathbf{D}_{lc},$$

$$\mathbf{B}_c = \mathbf{B}_0\mathbf{D}_{hc}^{-1}.$$

Az irányíthatósághoz igazoljuk, hogy a (4.3) feltételrendszer nem oldható meg. Legyen $\mathbf{v}^T\mathbf{A}_c=\lambda\mathbf{v}^T$, $\mathbf{v}^T\mathbf{B}_c=\mathbf{0}$. Ekkor

$$\mathbf{v}^T(\mathbf{A}_0-\mathbf{B}_0\mathbf{D}_{hc}^{-1}\mathbf{D}_{lc}) = \lambda\mathbf{v}^T,$$

$$\mathbf{v}^T\mathbf{B}_0\mathbf{D}_{hc}^{-1} = \mathbf{0}.$$

Ebből következik, hogy

$$\mathbf{v}^T\mathbf{B}_0 = \mathbf{0},$$

$$\mathbf{v}^T\mathbf{A}_0 = \lambda\mathbf{v}^T,$$

amiből $(\mathbf{A}_0, \mathbf{B}_0)$ irányíthatósága miatt a $\mathbf{v}=\mathbf{0}$ eredményre jutunk, ami a 4.1. tétel alapján $(\mathbf{A}_c, \mathbf{B}_c)$ irányíthatóságát jelenti.

Megjegyzés. A fenti bizonyítás a következő általánosabb állítás bizonyítására is alkalmas: Ha (A, B) irányítható, akkor nem szinguláris **G** és tetszőleges **K** mellett (A−BGK, BG) is irányítható.

Most megvizsgáljuk, hogy az $(\mathbf{A}_c, \mathbf{B}_c, \mathbf{C}_c)$ realizáció milyen feltételek mellett lesz minimális, azaz mikor teljesül $(\mathbf{C}_c, \mathbf{A}_c)$ megfigyelhetősége is.

Induljunk ki a következő azonosságból:

$$\mathbf{N}(s)\,\mathbf{D}^{-1}(s) = \mathbf{C}_c(s\mathbf{I}-\mathbf{A}_c)^{-1}\mathbf{B}_c. \tag{10.12}$$

Felhasználva $\mathbf{C}_c$ definícióját:

$$\mathbf{N}_{lc}\psi(s)\mathbf{D}^{-1}(s) = \mathbf{N}_{lc}(s\mathbf{I}-\mathbf{A}_c)^{-1}\mathbf{B}_c. \tag{10.13}$$

Ez az egyenlőség minden $\mathbf{N}_{lc}$ mellett igaz, ($\mathbf{A}_c$, $\mathbf{B}_c$ függetlenek N-től), így

$$\psi(s)\mathbf{D}^{-1}(s) = (s\mathbf{I}-\mathbf{A}_c)^{-1}\mathbf{B}_c. \tag{10.14}$$

Ezek a felírások nyilván irreducibilisek: a jobb oldal az irányíthatóság miatt, a bal oldal pedig amiatt, hogy $\begin{pmatrix}\psi(s)\\ \mathbf{D}(s)\end{pmatrix}$ teljes rangú $\forall s$-re (hiszen már maga $\psi(s)$ is az). $\mathbf{A}_c$ sajátvektorai a következő módon jellemezhetők:

10.2. TÉTEL. Legyen $\mathbf{A}_c\mathbf{p}=\lambda\mathbf{p}$, $\mathbf{p}\neq\mathbf{0}$. Ekkor létezik olyan **q** vektor, melyre $\mathbf{D}(\lambda)\mathbf{q}=\mathbf{0}$ és $\psi(\lambda)\mathbf{q}=\mathbf{p}$.

Bizonyítás. Írjuk át (10.14)-et a következő alakba:

$$(s\mathbf{I}-\mathbf{A}_c)\psi(s) = \mathbf{B}_c\mathbf{D}(s). \tag{10.15}$$

A 6.1. következmény miatt $|s\mathbf{I}-\mathbf{A}|=|\mathbf{D}(s)|$. Legyen λ az **A** mátrix sajátértéke. Ekkor $|\mathbf{D}(\lambda)|=0$, így létezik olyan $\mathbf{q}\neq\mathbf{0}$ vektor, melyre $\mathbf{D}(\lambda)\mathbf{q}=\mathbf{0}$. Szorozzuk be (10.15) mindkét oldalát az $s=\lambda$ helyen a **q** vektorral. Ekkor

$$(\lambda\mathbf{I}-\mathbf{A}_c)\psi(\lambda)\mathbf{q} = \mathbf{B}_c\mathbf{D}(\lambda)\mathbf{q},$$

azaz $(\lambda\mathbf{I}-\mathbf{A}_c)\psi(\lambda)\mathbf{q}=\mathbf{0}$. Tehát a $\mathbf{p}=\psi(\lambda)\mathbf{q}$ az **A** mátrix λ-hoz tartozó sajátvektora.

Ezen előkészületek után beláthatjuk a realizáció megfigyelhetőségére vonatkozó tételt:

10.3. TÉTEL. $(\mathbf{C}_c, \mathbf{A}_c)$ pontosan akkor megfigyelhető, ha $\mathbf{N}(s)$ és $\mathbf{D}(s)$ jobb relatív prímek.

Bizonyítás. Tegyük fel, hogy $(\mathbf{C}_c, \mathbf{A}_c)$ nem megfigyelhető. A 4.4. tétel alapján ekkor létezik $\mathbf{p}$ sajátvektora $\mathbf{A}_c$-nek, melyre $\mathbf{C}_c\mathbf{p}=\mathbf{0}$. Legyen λ a hozzá tartozó sajátérték. Az előző tétel szerint $\mathbf{p}$ előáll $\mathbf{p}=\psi(\lambda)\mathbf{q}$ alakban, ahol $\mathbf{D}(\lambda)\mathbf{q}=\mathbf{0}$. Ekkor azonban $\mathbf{N}(\lambda)\mathbf{q}=\mathbf{N}_{lc}\psi(\lambda)\mathbf{q}=\mathbf{C}_c\mathbf{p}=\mathbf{0}$, így $\begin{pmatrix}\mathbf{N}(\lambda)\\ \mathbf{D}(\lambda)\end{pmatrix}\mathbf{q}=\mathbf{0}$, amiből az következik, hogy $\mathbf{N}(s)$, $\mathbf{D}(s)$ nem jobb relatív prímek.

Tegyük fel most, hogy $\mathbf{N}(s)$ és $\mathbf{D}(s)$ nem jobb relatív prímek. Ekkor $\mathbf{H}(s)=\overline{\mathbf{N}(s)}\,\overline{\mathbf{D}(s)}^{-1}$, ahol deg det $\overline{\mathbf{D}}(s)=\bar{n}<$deg det $\mathbf{D}(s)=n$. Az ismertetett konstrukcióval akkor konstruálható $\bar{n}$ dimenziós irányítható realizáció. Emiatt az n- dimenziós realizáció nem lehet megfigyelhető is, hiszen így ő minimális lenne.

11. Pólusok, zérusok értelmezése minimálrealizáció alapján

Tekintsünk egy $\mathbf{H}(s)$ szigorúan proper racionális törtmátrixot, és ennek $\mathbf{H}(s)=\mathbf{N}(s)\mathbf{D}^{-1}(s)$ jobb törtfelbontását. Legyen $(\mathbf{A}_c, \mathbf{B}_c, \mathbf{C}_c)$ ennek a 10. fejezetben megkonstruált irányítható realizációja. Ekkor láttuk, hogy

$$\psi(s)\mathbf{D}^{-1}(s)=(s\mathbf{I}-\mathbf{A}_c)^{-1}\mathbf{B}_c \tag{11.1}$$

irreducibilis törtfelbontások. ($\psi(s)$ definícióját l. (10.5.).)

11.1. LEMMA. A fenti jelöléseket felhasználva léteznek $\mathbf{X}(s)$, $\mathbf{Y}(s)$, $\tilde{\mathbf{X}}(s)$, $\tilde{\mathbf{Y}}(s)$ polinommátrixok, melyre

$$\begin{pmatrix}s\mathbf{I}-\mathbf{A}_c & \mathbf{B}_c\\ -\tilde{\mathbf{X}} & \tilde{\mathbf{Y}}\end{pmatrix}\begin{pmatrix}\mathbf{X} & -\psi\\ \mathbf{Y} & \mathbf{D}\end{pmatrix}=\begin{pmatrix}\mathbf{I} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}\end{pmatrix}. \tag{11.2}$$

Bizonyítás. Felhasználva azt, hogy $s\mathbf{I}-\mathbf{A}_c$ és $\mathbf{B}_c$ bal relatív prímek és $\psi(s)$, $\mathbf{D}(s)$ jobb relatív prímek, így léteznek $\mathbf{X}(s)$, $\mathbf{Y}(s)$, $\overline{\mathbf{X}}(s)$, $\overline{\mathbf{Y}}(s)$ polinommátrixok, melyekre

$$(s\mathbf{I}-\mathbf{A}_c)\mathbf{X}(s)+\mathbf{B}_c\mathbf{Y}(s)=\mathbf{I}$$

$$\overline{\mathbf{X}}(s)\psi(s)+\overline{\mathbf{Y}}(s)\,\mathbf{D}(s)=\mathbf{I}.$$

Azaz összefoglalva:

$$\begin{pmatrix}s\mathbf{I}-\mathbf{A}_c & \mathbf{B}_c\\ -\overline{\mathbf{X}(s)} & \overline{\mathbf{Y}(s)}\end{pmatrix}\begin{pmatrix}\mathbf{X}(s) & -\psi(s)\\ \mathbf{Y}(s) & \mathbf{D}(s)\end{pmatrix}=\begin{pmatrix}\mathbf{I} & \mathbf{0}\\ \mathbf{G}(s) & \mathbf{I}\end{pmatrix},$$

ahol $\mathbf{G}(s)$ valamilyen polinommátrix. Balról megszorozva mindkét oldalt az $\begin{pmatrix}\mathbf{I} & \mathbf{0}\\ -\mathbf{G}(s) & \mathbf{I}\end{pmatrix}$ mátrixszal, épp a fenti egyenlőséget kapjuk. (Ez utóbbi transzformáció annak felel meg, hogy a blokk mátrixok első blokk-sorának $\mathbf{G}(s)$-szeresét kivonjuk a második blokksorból.)

KÖVETKEZMÉNY. A lemmában szereplő blokk mátrixok unimodulárisak.

Megjegyzés. Tetszőleges $\mathbf{N}_1(s)\mathbf{D}_1^{-1}(s)=\mathbf{D}_2^{-1}(s)\mathbf{N}_2(s)$ irreducibilis törtfelbontások esetén megfogalmazható a 11.1. lemma analogonja.

11.1. TÉTEL. $(s\mathbf{I}-\mathbf{A}_c)$ és $\mathbf{D}(s)$ nem egység invariáns polinomja megegyeznek.

Bizonyítás. (11.1) miatt $s\mathbf{I}-\mathbf{A}_c$ és $\mathbf{D}(s)$ ugyanannak a racionális törtfüggvénynek irreducibilis bal-, illetve jobb törtfelbontásbeli nevezője, így a tétel a 6.1. következmény speciális esete.

11.2. TÉTEL. $\begin{pmatrix} s\mathbf{I}-\mathbf{A}_c & \mathbf{B}_c \\ \mathbf{C}_c & \mathbf{0} \end{pmatrix}$ és $\mathbf{N}(s)$ nem egység invariáns polinomjai megegyeznek.

Bizonyítás. Szorozzuk meg a rendszermátrixot (11.2) baloldalán szereplő második unimoduláris mátrixszal:

$$\begin{pmatrix} s\mathbf{I}-\mathbf{A}_c & \mathbf{B}_c \\ \mathbf{C}_c & \mathbf{0} \end{pmatrix}\begin{pmatrix} \mathbf{X}(s) & -\psi(s) \\ \mathbf{Y}(s) & \mathbf{D}(s) \end{pmatrix} = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{C}_c\mathbf{X}(s) & -\mathbf{C}_c\psi(s) \end{pmatrix}.$$

$\mathbf{C}_c$ definíciója alapján $\mathbf{C}_c\psi(s)=\mathbf{N}(s)$, így a jobb oldalon szereplő mátrix nem egység invariáns polinomjai megegyeznek $\mathbf{N}(s)$ nem egység invariáns polinomjaival. A bal oldal invariáns polinomjait az unimoduláris mátrixszal való szorzás nem befolyásolja, így ott éppen a rendszermátrix invariáns polinomjait kapjuk.

KÖVETKEZMÉNY. Legyen $\mathbf{H}(s)$ tetszőleges szigorúan proper racionális törtmátrix, $(\mathbf{A}, \mathbf{B}, \mathbf{C})$ ennek tetszőleges minimális realizációja. Ekkor

1. $\mathbf{H}(s)$ pólusai megegyeznek $(s\mathbf{I}-\mathbf{A})$ invariáns polinomjainak gyökeivel (azaz $\mathbf{A}$ sajátértékeivel).
2. $\mathbf{H}(s)$ zérusai megegyeznek $\begin{pmatrix} s\mathbf{I}-\mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{0} \end{pmatrix}$ invariáns polinomjainak gyökeivel.

12. Zérusok fizikai interpretációja (Transmission blocking)

Tekintsük a $\mathbf{H}(s)$ szigorúan proper racionális törtmátrix egy minimális realizációját:

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}+\mathbf{B}\mathbf{u} \quad \mathbf{x}(0) = \mathbf{x}_0, \tag{12.1}$$

$$\mathbf{y} = \mathbf{C}\mathbf{x}. \tag{12.2}$$

A 11. fejezetben láttuk, hogy a fenti törtmátrix zérusai ugyanazok, mint a következő system-mátrix invariáns plinomjainak gyökei.

$$\begin{pmatrix} s\mathbf{I}-\mathbf{A} & -\mathbf{B} \\ \mathbf{C} & \mathbf{0} \end{pmatrix}. \tag{12.3}$$

Legyen s_0 $\mathbf{H}(s)$ egy tetszőleges zérusa. Ekkor a fenti rendszer-mátrix nem lesz teljes rangú $s=s_0$-ban. Ezt az esetet vizsgálja a következő tétel.

12.1. TÉTEL. Tegyük fel, hogy létezik olyan s_0, $\mathbf{x}_0$, $\mathbf{u}_0$, melyekre

$$\begin{pmatrix} s_0\mathbf{I}-\mathbf{A} & -\mathbf{B} \\ \mathbf{C} & \mathbf{0} \end{pmatrix}\begin{pmatrix} \mathbf{x}_0 \\ \mathbf{u}_0 \end{pmatrix} = \mathbf{0}. \tag{12.4}$$

Ekkor $\mathbf{x}(0)=\mathbf{x}_0$, $\mathbf{u}(t)=\mathbf{u}_0 e^{s_0 t}$ esetén az output-folyamat azonosan nulla, azaz $\mathbf{y}(t)\equiv\mathbf{0}$.

Bizonyítás. Térjünk át frekvenciatartománybeli leírásra.
A folyamatok *Laplace transzformáltjait* jelölje rendre $\tilde{\mathbf{x}}(s)$, $\tilde{\mathbf{u}}(s)$, $\tilde{\mathbf{y}}(s)$. Ekkor

$$(s\mathbf{I}-\mathbf{A})\tilde{\mathbf{x}}(s) = \mathbf{x}_0+\mathbf{B}\tilde{\mathbf{u}}(s), \tag{12.5}$$

$$\tilde{\mathbf{y}}(s) = \mathbf{C}\tilde{\mathbf{x}}(s). \tag{12.6}$$

A tételben szereplő input folyamat *Laplace transzformáltját* ismerjük:

$$\tilde{\mathbf{u}}(s) = \mathbf{u}_0(s-s_0)^{-1}.$$

Ezt behelyettesítve:

$$(s\mathbf{I}-\mathbf{A})\tilde{\mathbf{x}}(s) = \mathbf{x}_0+\mathbf{B}\mathbf{u}_0(s-s_0)^{-1}. \tag{12.7}$$

Írjuk fel a tétel feltételeit koordinátánként:

$$(s_0\mathbf{I}-\mathbf{A})\mathbf{x}_0-\mathbf{B}\mathbf{u}_0 = \mathbf{0} \tag{12.8}$$

$$\mathbf{C}\mathbf{x}_0 = \mathbf{0}. \tag{12.9}$$

A (12.8) összefüggést (12.7)-be beírva, majd átalakítva

$$(s\mathbf{I}-\mathbf{A})\tilde{\mathbf{x}}(s) = \mathbf{x}_0+(s_0\mathbf{I}-\mathbf{A})\mathbf{x}_0(s-s_0)^{-1} =$$

$$= (s-s_0)^{-1}[(s-s_0)\mathbf{x}_0+(s_0\mathbf{I}-\mathbf{A})\mathbf{x}_0] = (s-s_0)^{-1}(s\mathbf{I}-\mathbf{A})\mathbf{x}_0.$$

Tehát azt kaptuk, hogy

$$\tilde{\mathbf{x}}(s) = (s-s_0)^{-1}\mathbf{x}_0.$$

Inverz Laplace transzformációt alkalmazva:

$$\mathbf{x}(t) = \mathbf{x}_0 e^{s_0 t}. \tag{12.10}$$

Az output-folyamat ez alapján:

$$\mathbf{y}(t) = \mathbf{C}\mathbf{x}(t) = \mathbf{C}\mathbf{x}_0 e^{s_0 t} = \mathbf{0}. \tag{12.11}$$

Megjegyzés. Az $\mathbf{x}(t)$-re vonatkozó zárt formula olyan mellékeredmény, amihez a tétel feltételeiből csak annyit használtunk fel, hogy $\mathbf{u}(t)=\mathbf{u}_0 e^{s_0 t}$, $\mathbf{x}(0)=\mathbf{x}_0$.

IRODALOM

[1] GANTMACHER, F. R., *Matrix Theory* (Nauka, Moszkva, 1966, oroszul).
[2] KAILATH, TH., *Linear Systems* (Prentice-Hall, Inc., 1980).
[3] KUCERA, V., *Discrete Linear Control. The Polynomial Equation Approach* (1979).
[4] LANCASTER, P., *Theory of Matrices* (Nauka, Moszkva, 1978, oroszul).
[5] MAC DUFFE, C. C, *The Theory of Matrices* (Chelsea, New York, 1956).
[6] ROSENBROCK, H. H., *State-space and Multivariable Theory* (Thomas Nelson and Sons Ltd., London, 1970).
[7] WOLOVICH, W. A., *Linear Multivariable Systems* (Applied Mathematical Sciences) (Springer Verlag, New York, 1974).

(Beérkezett: 1984. június 12.)

G. VÁGÓ ZSUZSA
CSEPEL MŰVEK SZÁMÍTÁSTECHNIKAI VÁLLALAT
1751 BUDAPEST, POSTAFIÓK 65.

LINEAR SYSTEMS AND POLINOM-MATRICES

G. VÁGÓ ZSUZSA

An important topic in the theory of control systems is the study of linear systems in frequency domain. The central notion is the transfer function of a linear system which is a matrix with rational elements. We give an introduction to relevant results of rational matrices (greatest common divisor, right-left matrix fraction, Smith—McMillan form, properness, pole-zero), and explain their connection with notions of time domain description (controllability, observability, realization). The methods of the paper can be applied in the analysis of multidimensional time series.

Alkalmazott Matematikai Lapok **11** (1985) 123–155

VÉGES ELEMEK MÓDSZERE LINEÁRIS, PARABOLIKUS TÍPUSÚ FELADATOK MEGOLDÁSÁRA

FARAGÓ ISTVÁN

Gödöllő

A dolgozatban áttekintést szeretnénk nyújtani a véges elemek módszerének időfüggő, parabolikus típusú feladatokra történő alkalmazásáról. Bevezetjük az ilyen feladatosztályra értelmezett általánosított megoldás fogalmát és a *Galjorkin-féle szemidiszkretizációs módszer* segítségével megadjuk annak numerikus megoldási módszereit. Az egyszerűbb feladatok tárgyalása mellett kitérünk a legáltalánosabb eset megvizsgálására is. Az elméleti jelentőségű hibabecslések mellett kitérünk a konkrét számítási algoritmusokra is.

1. Bevezetés

A [8] cikkben a lineáris, elliptikus típusú parciális differenciálegyenletek véges elemes megoldási módszereit tárgyaltuk. A műszaki-fizikai folyamatok egy jelentős része azonban nem modellezhető az ilyen típusú differenciálegyenletekkel, mivel a folyamat vagy

a) stacionárius, nemlineáris; vagy

b) instacionárius, lineáris; vagy

c) instacionárius, nemlineáris.

Ebben a cikkben a második esetet tárgyaljuk, azon belül is a parabolikus típusú feladatokat. Az ilyen típusú feladatok jellegzetessége az idő szerinti változó megjelenése, mégpedig oly módon, hogy a teljes feladat operátora nem szimmetrikus. Ezért ebben az esetben a *Ritz-módszer* a [8]-ban leírt módon nem alkalmazható. A *Galjorkin-módszer* alkalmazható a feladat megoldására, de ebben az esetben a megfelelő egzisztencia, unicitás, konvergencia és stabilitás vizsgálatok elvégzése szükséges. Megjegyezzük, hogy a *Galjorkin-módszer* alkalmazása egy olyan bázisfüggvényrendszer definiálását igényli, amely a teljesség mellett jó approximációs tulajdonsággal is rendelkezik [8]. Ez utóbbi tulajdonságba az is beletartozik, hogy ezekkel a bázis-függvényekkel a feladat megoldási tartományán jól tudjunk approximálni. Ugyanakkor vegyük észre, hogy ez csak a térváltozók szerinti tartományra jelent megszorítást; az idő szerinti változó mindig egy intervallumon változik, aminek az approximálása nem jelent problémát. Emellett, ha — az elliptikus feladatokhoz hasonlóan — a bázisfüggvényeket úgy választjuk meg, hogy azok valamennyi változótól (tehát tér- és idő szerintitől is) függjenek, akkor

a) a megoldandó lineáris algebrai egyenletrendszer mérete megnő;

b) a megoldási folyamat nem őrzi meg az idő szerinti folyamatos előrehaladás elvét;

c) az idő szerinti változóra nézve nem alkalmazhatunk explicit sémákat.

Ezeket a problémákat elkerülendő célszerű az ún. *szemidiszkrét Galjorkin-típusú eljárást* alkalmazni, amelynek lényege, hogy a térbeli változók szerint véges elemes,

az idő szerinti változó szerint pedig véges differencia közelítést alkalmazunk. Egyszerűbb típusú feladatokra ezen módszer alkalmazását már ismertettük [6], [7], [9]. Jelen cikkünk célja, hogy a módszer részletes ismertetésén túlmenően megadjuk azokat a feladatosztályokat, amelyekre a módszer alkalmazhatósága bizonyított, azaz az előzőekben felsorolt matematikai vizsgálatok elvégzése (amelyek általában igen bonyolultak) elhagyhatók.

Cikkünk 2. szakasza feladatunkra a variációs elvet fogalmazza meg; a 3. szakaszban a módszer egy-egy modell jellegű, egydimenziós feladatra történő alkalmazását vizsgáljuk meg. A 4. szakaszban az előző szakasz eredményeit általánosítjuk: a megoldandó feladat operátorának elliptikus részét kiterjesztjük tetszőleges, időtől független, szigorúan pozitív definit operátorra. A különböző normákban megadott hibabecslések mellett különböző véges differencia sémákat definiálunk az idő szerinti változó diszkretizációjára. Befejezésül az 5. szakaszban a 4. szakasz eredményeit általánosítjuk: megengedjük, hogy az elliptikus rész időfüggő legyen. Ebben a szakaszban az időrétegenként szükséges lineáris algebrai egyenletrendszer megoldására egy olyan iterációs eljárást javasolunk, amely a számítási munka leegyszerűsítése mellett a feladat elméleti pontosságát nem rontja el. A 4. és 5. szakaszban nemcsak a feladatosztályt, hanem a módszerek osztályát is kibővítjük: állításainkat az általános projekciós módszerekre mondjuk ki, de mindig utalunk a *szemidiszkrét Galjorkin módszerre* is.

2. A parabolikus feladatok variációs elve

A továbbiakban a lineáris, parabolikus típusú parciális differenciálegyenletek variációs elvét és azok általánosított megoldásainak fogalmát definiáljuk. Legyenek $u(t)$, $f(t)$: $[0, T] \to H(\Omega)$, $u_0 \in H(\Omega)$ és tekintsük a

$$D_t u + Lu = f(t) \quad 0 < t \leqq T, \tag{2.1}$$

$$u(0) = u_0 \tag{2.2}$$

absztrakt Cauchy-feladatot [6], [20], ahol $u_0, f(t)$ adottak.

Az 5. szakaszban olyan feladattípusokat vizsgálunk, amikor L időfüggő operátor, azaz $L = L(t)$ $(0 \leqq t \leqq T)$ és tetszőleges rögzített $t \in [0, T]$ érték esetén $L(t)$: $H(\Omega) \to H(\Omega)$ lineáris, szigorúan pozitív definit operátor. Ekkor, az előzőekhez hasonlóan, tekintsük a

$$D_t u + L(t)u = f(t) \quad 0 < t \leqq T, \tag{2.1a}$$

$$u(0) = u_0 \tag{2.2a}$$

absztrakt Cauchy feladatot.

Könnyen belátható, hogy a lineáris, másodrendű parabolikus típusú parciális differenciálegyenletek vegyes kitűzésű feladatai felírhatók ilyen alakban: a kiegészítő feltételek közül a kezdeti feltételt u_0 függvény jelenti és a Γ-n felvett peremfeltételeket az L operátor dom L értelmezési tartománya tartalmazza.

Példaként tekintsük a

$$D_t u - D_x(pD_x u) + qu = f(x, t) \quad 0 < x < \pi; \quad 0 < t \leqq T, \tag{2.3}$$

$$u(0, t) = D_x u(\pi, t) = 0 \quad 0 \leqq t \leqq T \tag{2.4}$$

$$u(x, 0) = u_0(x) \quad 0 \leqq x \leqq \pi \tag{2.5}$$

feladatot. Legyen $\Omega=(0, \pi)$; $H(\Omega)=H^0(\Omega)$ és jelölje

$$Lv = -D_x(pD_x v) + gv; \quad \text{dom } L = \{v \in C^2(0, \pi); \quad v(0) = D_x v(\pi) = 0\}$$

az $L: C^2(0, \pi) \to H^0(0, \pi)$ képező operátort. Ekkor megmutatható [8], hogy $p(x) \geqq$ $\geqq p_0 > 0$; $g(x) > 0 (x \in \Omega)$ esetén az L szigorúan pozitív definit operátor. Nyilvánvaló, hogy a fenti jelölések mellett (2.3)—(2.5) feladat felírható (2.1), (2.2) alakban.

Jelölje H_E azon $H(\Omega \times (0, T))$-beli elemek halmazát, amelyekre rögzített $v_0 \in V$ esetén az

$$\Phi(u, v_0) = ((D_t u, v_0)) + ((Lu, v_0)) \quad \forall u \in H_E \tag{2.6}$$

bilineáris funkcionál értelmes. A definícióból következik, hogy érvényes a következő tartalmazás:

$$\text{dom } L \times V \subset H_E \times V \subset H(\Omega \times (0, T)) \text{ sűrűn.}$$

2.1. Definíció. Azt a (2.2) kezdeti feltételt kiegészítő elemet, amelyre

$$\Phi(u^*, v) = ((f, v)) \quad \forall v \in V \tag{2.7}$$

a (2.1) (2.2) feladat általánosított megoldásának nevezzük.

A (2.7) kifejezés konkrét alakja a feladattól függő $H(\Omega \times (0, T))$ tér és a V sűrű altér megválasztásától függ. A feladatok döntő többségénél $H(\Omega \times (0, T))$ térnek a $H^0(\Omega \times (0, T))$ teret választjuk meg és ekkor a skaláris szorzat

$$((u, w)) = \int_0^T \int_\Omega uw \, dx \, dt;$$

azaz (2.7) felírható a következő alakban

$$\int_0^T [(D_t u, v) + (Lu, v) - (f, v)] \, dt = 0, \tag{2.8}$$

ahol V az $H^0(\Omega \times (0, T))$ tér valamilyen sűrű altere.

Mivel $\tilde{V}_0$ sűrű $H^0(\Omega \times (0, T))$-ben, ezért 2.1 definíció és (2.8) alapján az imént rögzített H és $V = \tilde{V}_0$ terekben definiáljuk az általánosított megoldás fogalmát.

2.2. Definíció. Azt az $u^* \in H_{L \times T}$ függvényt, amelyre

$$\int_0^T [-(u^*, D_t v) + [u^*, v]_L - (f, v)] \, dt = (u_0, v(\cdot, 0)) \quad \forall v \in \tilde{V}_0 \tag{2.9}$$

a (2.1) (2.2) feladat általánosított megoldásának nevezzük.

2.1. Megjegyzés. Vegyük észre, hogy (2.9) kifejezés (2.8)-ból az első tag parciális deriválásával; a $\tilde{V}_0$-beli függvények tulajdonságával és a H_L energetikai tér definíciójából közvetlenül nyerhető.

Térjünk vissza a (2.3)—(2.5) feladatra. Mint ismeretes [8], az L szigorúan pozitív definit operátorának energetikai tere:

$$H_L = \{v \in H^1(0, \pi);\quad v(0) = 0\},$$

$$[u, v]_L = \int_0^T (pD_x u \cdot D_x v + quv)\, dx \tag{2.10}$$

és így (2.9) felírható

$$\int_0^T \left\{ -\int_0^\pi (u \cdot D_t v + pD_x u \cdot D_x v + quv - fv)\, dx \right\} dt = \int_0^\pi u_0 \cdot v(x, 0)\, dx \tag{2.11}$$

alakban.

Így a 2.2 definíciót a (2.3)—(2.5) feladatra alkalmazva a következő érvényes: a feladat általánosított megoldásán azon $u^*(x, t)$ függvényt értjük, amelyre

a) $u^* \in H^0((0, \pi) \times (0, T))$,

b) $D_x u^* \in H^0((0, \pi) \times (0, T))$,

c) $u^*(0, t) = 0 \quad 0 < t \leqq T$,

d) minden olyan $v \in H^1(\Omega \times (0, T))$ függvényre, amelyre $v(0, t) = 0$ és $v(x, T) = 0$ $(0 \leqq t \leqq T, x \in \Omega)$ a (2.11) összefüggés teljesül.

A konkrét példán is látható, hogy a (2.9) módon definiált általánosított megoldás fogalma nagymértékben csökkenti a feladat megoldására tett simasági feltételeket és így lényegesen bővebb függvényosztályon kereshetjük a megoldást. Ugyanakkor (2.9), illetve (2.11) alakból látható, hogy ha valamennyi változó szerint (tehát tér- és idő szerint egyaránt) vezetjük be a diszkretizációhoz szükséges $\{\alpha_j(t)\varphi_i(x)\}$ alakú bázisfüggvényrendszert, akkor az idő szerinti változóra nézve is implicit sémát nyerünk. Ezért általában célszerű olyan általánosított megoldást definiálni, amelynél ugyan az általánosított megoldás t szerinti simaságát illetően a követelmény magasabb, de a kapott integrálösszefüggés numerikus realizálása lényegesen egyszerűbb.

Legyen:

$$\tilde{H}_{L \times T} = \{u \in H_{L \times T};\ D_t u \in H^0(\Omega \times (0, T))$$

$$\tilde{V}^* = \{v \in \tilde{V}_0,\ v(x, t) = \alpha(t)\varphi(x),\ D_t v \in H^0(\Omega \times (0, T))\}.$$

Tehát $\tilde{H}_{L \times T}$ azon $H_{L \times T}$-beli függvényekből áll, amelyek idő szerinti deriváltjai $H^0(\Omega \times (0, T))$-beliek, $\tilde{V}^*$ elemei pedig egy H_L-beli és egy $T = 0$ helyen nulla értéket felvevő $H^0(0, T)$-beli függvények szorzataként állnak elő.

Tegyük fel, hogy a (2.1), (2.2) feladatnak létezik 2.2 definíció szerinti $u^* \in H_{L \times T}$ általánosított megoldása és $D_t u^* \in H^0(\Omega \times (0, T))$. (Azaz $u^* \in \tilde{H}_{L \times T}$). Ekkor (2.9) felírható $\forall \varphi \in H_L$ esetén

$$\int_0^T \{(D_t u^*, \varphi) + [u^*, \varphi]_L - (f, \varphi)\}\alpha(t)\, dt + \alpha(0)(u^*(x, 0) - u_0, \varphi) = 0 \tag{2.13}$$

alakban. (Ugyanis $\tilde{V}^* \subset \tilde{V}_0$). Mivel $\alpha(t)$ az előzőekben említett megkötéseken túl

tetszőleges függvény, ezért (2.13) azt jelenti, hogy tetszőleges $t\in(0, T)$ rögzített értékre fennáll a

$$\begin{aligned}(D_t u^*, \varphi)(t)+[u^*, \varphi]_L(t) &= (f, \varphi)(t) \quad 0 < t \leqq T,\\ ((u^*(\cdot, 0)-u_0), \varphi) &= 0 \quad \forall \varphi\in H_L\end{aligned} \tag{2.14}$$

összefüggés. Ezek alapján vezessük be az általánosított megoldás egy újabb, a gyakorlatban használt definícióját.

2.3. Definíció. Azt az $u^*\in\tilde{H}_{L\times T}$ függvényt, amelyre tetszőleges $\varphi\in H_L$ esetén (2.14) fennáll, a (2.1), (2.2) feladat általánosított megoldásának nevezzük.

Tekintsük ismét a (2.3), (2.5) feladatot! Megmutatható [20], hogy ha $u_0\in H_L$, akkor a (2.9) feladatnak létezik egyetlen megoldása $H^1((0, \pi)\times(0, T))$-ben. Így u^* olyan függvény, hogy

a) tetszőleges $t\in[0, T]$ esetén a (2.10) által definiált H_L-beli;

b) $D_t u^*\in H^0((0, \pi)\times(0, T))$;

c) tetszőleges $t\in[0, T]$ és $\varphi\in H_L$ esetén

$$\int_0^\pi u^*(x, 0)\varphi(x)\,dx = \int_0^\pi u_0\varphi\,dx. \tag{2.15}$$

A továbbiakban a (2.14) típusú, 2.3 definíció szerinti általánosított megoldások meghatározásával foglalkozunk. A pontos megoldás meghatározása helyett — mivel az a gyakorlatban általában nem hajtható végre — a [8]-ban leírtakhoz hasonlóan a közelítő megoldást keressük, mégpedig oly módon, hogy a (2.14) feladatot szemidiszkrét feladatok sorozatára vezetjük vissza. Ezt a következő módon valósítjuk meg.

Legyenek $(E_n(t))\subset\tilde{H}_{L\times T}$ és $(V_n)\subset H_L$ $(n=1, 2.....)$ sűrű altérsorozatok, mégpedig olyanok, hogy (V_n) és az $x\in\Omega$ változóra nézve $(E_n(t))$ véges dimenziósak. Ekkor a (2.14) feladat megoldása helyett tekintsük a következő $u_n^*(x, t)\in E_n(t)$ $(n=1, 2.....)$ elemek meghatározását jelentő szemidiszkrét feladatok sorozatát:

$$\begin{aligned}&(D_t u_n^*, \varphi)(t)+[u_n^*, \varphi]_L(t) = (f, \varphi),\\ &(u_n^*(\cdot, 0), \varphi) = (u_0, \varphi) \qquad \forall \varphi\in V_n.\end{aligned} \tag{2.16}$$

Így (2.16) $n=1, 2...$ megoldásával az $(u_n^*(x, t))$ függvénysorozatot nyerjük. Megjegyezzük, hogy (2.16) a térváltozók szerinti *Galjorkin-típusú diszkretizáció* a gyakorlatban egy közönséges, elsőrendű differenciálegyenlet-rendszerre kitűzött *Cauchy-feladatot* jelent. Ennek megoldására egy újabb — most már az időváltozó szerinti — diszkretizációt alkalmazva nyerjük a tér- és időváltozó szerint egyaránt diszkretizált közelítő megoldást. Ismételten megjegyezzük, hogy az idő szerinti diszkretizációt véges differencia módszerrel hajtjuk végre.

A módszerek realizálása előtt fogalmazzuk meg a módszer alkalmazásánál felmerülő alapvető kérdéseket!

a) Létezik-e a (2.16) feladatnak $u_n^*(x, t)$ megoldása?

b) Tart-e az (u_n^*) sorozat u^* általánosított megoldáshoz és ha igen, akkor milyen becslés adható meg ezen konvergenciára?

c) Milyen diszkretizáció alkalmazható (2.14) közönséges differenciálegyenlet-rendszer megoldására és milyen globális becslés adható az így nyert megoldásnak az eredeti u^* megoldástól való eltérésére?

3. Egydimenziós feladat megoldása véges elemek módszerével

Először az előző szakaszban általánosságban leírt módszert ismertetjük a (2.1), (2.2) feladatra, valamint kitérünk a szakasz végén felvetett kérdésekre is.

Legyen $\{\varphi_i(x)\}\subset H_L$ teljes rendszer. Mi a továbbiakban ezen teljes rendszernek mindig valamelyik p-ed fokú spline-tér bázisfüggvényeit választjuk és ekkor legyen

$$\mathscr{L}[\varphi_1, \dots \varphi_n] = V_n, \tag{3.1}$$

$$\mathscr{L}(t)[\varphi_1 \dots, \varphi_n] = E_n(t)$$

ahol $\mathscr{L}$, ill. $\mathscr{L}(t)$ a $\varphi_1 \dots \varphi_n$ elemek állandó együtthatós ill. t-től függő együtthatós lineáris burka. Így egy tetszőleges $u_n^*(x, t)\in E_n(t)$ függvény

$$u_n^*(x, t) = \sum_{i=1}^{n} \alpha_i(t)\varphi_i(x) \tag{3.2}$$

alakban írható fel, ahol $\alpha_i(t)$ $(i=1, 2, \dots n)$ ismeretlen együttható függvények. Tekintettel a V_n (3.1) alakú előállítására, a (3.2) alakú közelítést a

$$\begin{aligned}(D_t u_n^*, \varphi_j)(t)+[u_n^*, \varphi_j]_L(t) &= (f, \varphi_j)\\ (u_n^*(\cdot, 0), \varphi_j) &= (u_0, \varphi_j)\end{aligned} \qquad j = 1, \dots n \tag{3.3}$$

feladat megoldásával kaphatjuk meg. Ez a (3.2) előállítást figyelembevéve, a keresett u_n^* függvény $\alpha_i(t)$ együtthatóira az

$$MD_t\alpha(t)+Q\alpha(t) = F(t) \quad 0 < t \leqq T, \tag{3.4}$$

$$M\alpha(0) = \alpha_0,$$

lineáris, közönséges, elsőrendű differenciálegyenletrendszer *Cauchy-feladatának* megoldását jelenti, ahol

$$\begin{gathered}\alpha = [\alpha_1(t), \dots, \alpha_n(t)]^T;\ F(t) = [F_1(t), \dots, F_n(t)]^T;\ F_j(t) = (f, \varphi_j),\\ \alpha_0 = [\alpha_{01}, \dots, \alpha_{0n}]^T;\ \alpha_{0j} = (u_0, \varphi_j);\ M = [m_{ij}]_{i,j=1}^n;\ m_{ij} = (\varphi_i, \varphi_j),\\ D_t\alpha = [D_t(\alpha_1(t), \dots, D_t\alpha_n(t)]^T;\ Q = [q_{ij}]_{i,j=1}^n;\ q_{ij} = [\varphi_i, \varphi_j]_L.\end{gathered} \tag{3.5}$$

A következő lépésben a (3.4) *Cauchy-feladat* megoldását tekintsük! Ehhez a $[0, T]$ intervallumon (az egyszerűség kedvéért) egy k lépésközű, egyenletes rácshálót definiálunk:

$$\{t_j = jk;\ k = T/K;\ K\text{: poz. egész};\ j = 0, 1, \dots K\}.$$

Jelölje α^j az $\alpha(t)$ vektor t_j-rácspontbeli közelítését. Ekkor a (3.4) feladat közelítő megoldására több differencia-séma is alkalmazható. Közülük a leggyakoribbak a következők:

a) *Explicit séma:*

$$M\frac{\alpha^{j+1}-\alpha^j}{k}+Q(t_j)\alpha^j = F(t_j) \quad j = 0, 1, \dots K \tag{3.6}$$

$$M\alpha^0 = \alpha_0,$$

b) *Euler-séma:*

$$M\frac{\alpha^{j+1}-\alpha^j}{k}+Q(t_{j+1})\alpha^{j+1}=F(t_j)\quad j=0,1,\ldots K \tag{3.7}$$

$$M\alpha^0=\alpha_0,$$

c) *Crank—Nicolson-séma:*

$$M\frac{\alpha^{j+1}-\alpha^j}{k}+Q\left(\frac{t_j+t_{j+1}}{2}\right)\frac{\alpha^{j+1}+\alpha^j}{2}=F\left(\frac{t_j+t_{j+1}}{2}\right)\quad j=0,1\ldots K \tag{3.8}$$

$$M\alpha^0=\alpha_0.$$

3.1. Megjegyzés. A (3.6) séma „explicit"-sége nem azt jelenti, hogy az egyes időrétegeken nem szükséges lineáris egyenletrendszert megoldanunk, hanem azt, hogy az időrétegenként megoldandó lineáris algebrai egyenletrendszer együttható-mátrixa Q-tól függetlenül ugyanaz az M mátrix, ami — a (3.5) jelölést figyelembe véve — állandó együtthatós, szimmetrikus, szigorúan pozitív definit. Így a módszer realizálása során elegendő csak az ilyen típusú lineáris egyenletrendszert megoldani, ami például az $M=S\cdot S^T$ *Cholesky-felbontással* (vagy valamilyen iterációs eljárással) viszonylag könnyen realizálható.

3.2. Megjegyzés. A (3.7) és (3.8) sémák implicit sémák, mégpedig olyan értelemben, hogy L operátortól függően időlépésenként egymástól különböző együttható mátrixszal rendelkező lineáris algebrai egyenletrendszereket kell megoldanunk.

3.3. Megjegyzés. A sémák nemcsak realizálásban, hanem stabilitásban és pontosságban is eltérnek egymástól. (Számunkra ezek a lényegesek.) A továbbiakban kimondjuk, hogy (3.6) és (3.7) séma elsőrendben, (3.8) séma másodrendben pontos, valamint a (3.6) séma k-ra nézve feltételesen stabil. Ezeket (és a többi, gyakorlatban is alkalmazható sémákat) a 4. szakaszban részletesen ismertetjük.

3.4. Megjegyzés. Mivel bázisfüggvények — a szakasz elején leírtaknak megfelelően — a p-ed fokú spline-tér bázisfüggvényei, ezért megadható az $e_n(x,t)=u_n^*(x,t)-u^*(x,t)$ függvényre hibabecslés. Ha $(\varphi_i(x))$ a lineáris spline-tér bázisfüggvényei, akkor megmutatható [19], hogy érvényesek a

$$\max_{0\leq t\leq T}\|e_n(t)\|+\left(\int_0^T\|e_n\|_L^2\,dt\right)^{1/2}=O(h), \tag{3.9}$$

$$\left(\int_0^T\|D_t e_n\|^2\,dt\right)^{1/2}+\max_{0\leq t\leq T}\|e_n\|_L=O(h), \tag{3.10}$$

$$\max_{0\leq t\leq T}\|e_n\|=O(h^{3/2}), \tag{3.11}$$

$$\left(\int_0^T\|e_n\|^2\,dt\right)^{1/2}=O(h^2) \tag{3.12}$$

becslések. Ezeket a becsléseket a következő szakaszokban tetszőleges p-ed fokú közelítésre és általános diszkretizációs módszerekre kiterjesztjük.

3.5. Megjegyzés. A (3.4) feladatnak mindig létezik megoldása. Ugyanis, mivel M és Q szigorúan pozitív definit mátrixok, ezért M^{-1} mindig létezik, s így (3.4) felírható a vele ekvivalens

$$D_t\alpha(t)+M^{-1}Q\alpha(t) = M^{-1}F(t) \quad 0 < t \leqq T, \tag{3.4a}$$

$$\alpha(0) = M^{-1}\alpha_0$$

alakban. A közönséges differenciálegyenletek elméletének megfelelően (3.4a) feladatnak mindig létezik egyetlen megoldása, amely felírható

$$\alpha(t) = \exp\{M^{-1}Qt\}M^{-1}\alpha_0 + \int_0^t \exp\{-M^{-1}Q(t-s)\}M^{-1}F(s)\,ds$$

alakban. Így u_n^* létezik.

Az előző megjegyzésekkel egyben megválaszoltuk a második szakasz végén feltett kérdéseket néhány speciális közelítésre.

Tekintsük tehát a (2.3)—(2.5) feladatot! Az előző fejezetben megmutattuk, hogy a feladat (2.3. definíció szerinti) általánosított megoldása a (2.15) feladat megoldásával nyerhető. Ha a feladat p, q, f, és u_0 függvényei olyanok, hogy a megoldás idő szerinti deriváltjára vonatkozó simasági feltétel nem teljesül, akkor a feladat általánosított megoldását 2.2. definíció értelemben meghatározhatjuk, azaz ebben az esetben a lényegesen munkaigényesebb (2.11) feladat megoldása szükséges.

Tekintsük tehát (2.15)-öt! Mivel (2.10) alapján feladatunkra H_L $H^1(0, \pi)$ sűrű altere, ezért tetszőleges, az $x=0$ helyen nulla értékű splinek tere megfelel a közelítésnek. (Ha H_L magasabb rendű simaságot igényel, akkor nem minden spline-tér felel meg, csak a megfelelő simasági rendűek.) Megválasztva bázisfüggvényeket (néhány konkrét alakot [8]-ban felsoroltunk) a (2.10)-ben definiált H_L-beli skaláris szorzattal közvetlenül meghatározhatók (3.5) alapján az M, Q, α_0 és $F(t)$ értékei. Ezután (3.6), (3.7), (3.8) séma valamelyikével megoldjuk a (3.4) alakú differenciálegyenlet-rendszerünket. A (2.3)—(2.5) feladat u^* megoldásának közelítését az (x, t_j) pontban úgy határozhatjuk meg (t_j : a $[0, T]$-n definiált rácsháló csomópontja, $x_i \in (0, \pi)$ tetszőleges pont), hogy a (3.6), (3.7), (3.8) sémák egyikével meghatározzuk α^j vektort és annak komponenseivel képezzük $u_n^*(x, t_j)$ függvényt (3.2) alapján, majd behelyettesítjük az $x=x_i$ értéket.

Ha a (2.3) egyenletben a $p=q=1$ esetet tekintjük, akkor lineáris spline-közelítés esetén az M és Q mátrixok alakját [8]-ban megadtuk. Magasabb fokú közelítésre a megfelelő mátrixok alakjait [9]-ben adtuk meg.

4. Lineáris, időtől független elliptikus részű, parabolikus típusú parciális differenciálegyenletek diszkretizációja

A továbbiakban az olyan lineáris parabolikus típusú feladatok diszkretizációját vizsgáljuk, ahol az elliptikus rész (a (2.1) feladat L operátora) nem függ az időtől. A feladat diszkretizációját valamely, T_n operátorral szimbolizált projekciós módszerrel hajtjuk végre. (Az előzőekben leírt *Galjorkin-módszer* is ide tartozik, de ezen kívül

még számos módszert reprezentálhat, például a legkisebb négyzetek módszerét, a kollokációs eljárásokat és további speciális módszereket). Tekintsük a következő feladatot:

$$D_t u = -Lu \quad (x, t)\in\Omega\times(0, T], \tag{4.1}$$

$$u(x, t) = 0 \quad (x, t)\in\Gamma\times(0, T], \tag{4.2}$$

$$u(x, 0) = u_0(x) \quad x\in\Omega \tag{4.3}$$

ahol

$$Lu = -\sum_{i,j=1}^{N} D_{x_i}(a_{ij}(x)D_{x_j})u + a_0(x)u \quad x\in\Omega\subset\mathbf{R}^N \tag{4.4}$$

a_{ij}, a_0 sima függvények; az $A(x)=[a_{ij}]_{i,j=1}^{N}$ $N\times N$ mátrix szimmetrikus, szigorúan pozitív definit Ω-n; $a_0(x)\geqq 0$ Ω-n adott, megfelelően sima függvény. Mint ismeretes [10], a (4.1)—(4.3) feladat megoldása előállítható

$$u(x, t) = \sum_{j=1}^{\infty} \beta_j \exp\{-\lambda_j t\}\varphi_j(x) \equiv \exp(-tL)u_0 \tag{4.5}$$

alakban, ahol $\beta_j=(u_0, \varphi_j)$; (λ_j, φ_j) $(j=1, 2, \ldots)$ az $Lw=\lambda w$ sajátérték-feladat teljes sajátérték- és sajátvektor rendszere.

Jelölje $\tilde{T}$ az

$$\begin{aligned} Lw &= f \quad x\in\Omega, \\ w &= 0 \quad x\in\Gamma \end{aligned} \tag{4.6}$$

elliptikus peremérték feladat megoldási operát, azaz

$$\tilde{T}f = w. \tag{4.7}$$

Ekkor a (4.1)—(4.3) feladat felírható

$$\begin{aligned} D_t\tilde{T}u + u &= 0, \\ u(0) &= u_0 \end{aligned} \tag{4.8}$$

alakban. A (4.8) feladat megoldására definiáljuk az (S_h) ($h=$ térbeli diszkretizáció pozitív, kis értékű paramétere) $H^0(\Omega)$-beli véges dimenziós altérsorozatot és adjunk meg egy (T_n) $(T_n: H^0(\Omega)\to S_h)\tilde{T}$ operátort approximáló operátor-sorozatot. Ezután (4.8) helyett a következő szemidiszkrét feladatok sorozatát oldjuk meg: keressük azon $u_n^*(x, t)\in S_h(t\geqq 0)$ elemet, amelyre

$$\begin{aligned} D_t T_n u_n^*(x, t) + u_n^*(x, t) &= 0, \\ u_n^*(x, t) &= u_{0,h}\in V_n \end{aligned} \tag{4.9}$$

ahol $u_{0,h}$ az u_0 függvény megfelelő V_n-beli approximációja.

4.1. Megjegyzés. Mivel a (2.16) általános alakban leírt feladatokat konkrét esetekben mindig valamelyik spline-térben oldjuk meg, ezért tértünk át a V_n tér S_h-val történő jelölésére.

Ahhoz, hogy a (4.9) feladat megoldásainak sorozata valóban a (4.8) feladat megoldásának közelítő sorozata legyen, a (T_n) operátorsorozatra bizonyos feltételeket kell

tennünk. (Többek között pontosítani kell az előzőekben használt „(T_n) approximálja $\tilde{T}$-t" kijelentést is.)

F_1 feltétel; Tegyük fel, hogy

a) T_n operátor $H^0(\Omega)$-ban önadjungált, pozitív definit és S_h-n szigorúan pozitív definit;

b) létezik olyan $r \geqq 2$ egész szám és C pozitív állandó, hogy

$$\sup_{\substack{w \in \mathring{H}^q(\Omega) \\ w \neq 0}} \frac{\|(T_n - \tilde{T})w\|}{\|w\|_q} \leqq C \cdot h^{q+2} \quad 0 \leqq q \leqq r-2 \tag{4.10}$$

ahol

$$\mathring{H}^s(\Omega) = \{w \in H^s(\Omega);\ L^j w|_\Gamma = 0;\ j \leqq [3/2]\}.$$

T_n definiálása a térbeli diszkretizációs módszer megadását jelenti. Egyik ilyen lehetséges módszer az előző szakaszokban tárgyalt Galjorkin-módszer. Írjuk le ezt a módszert jelen apparátusunkkal!

Legyen $S_h \subset \mathring{H}^1(\Omega)$ az $(r-1)$-ed fokú spline-tér (S_h elemei Γ-n nullák.) Ekkor a spline-függvények approximációs tulajdonságának következtében minden $w \in \mathring{H}^1(\Omega)$ függvényre érvényes az

$$\inf_{v \in S_h} \{\|w - v\| + h\|w - v\|_1\} \leqq C \cdot h^s \|w\|_s \quad 1 \leqq s \leqq r \tag{4.11}$$

ún. „inverz approximációs feltétel." Jelölje

$$\Phi(\varphi, \psi) = \int_\Omega \Big(\sum_{i,j=1}^{N} a_{ji} D_{x_j} \varphi \cdot D_{x_i} \psi + a_0 \varphi \psi \Big)\, dx \tag{4.12}$$

bilineáris funkcionált. (Vegyünk észre, hogy Φ funkcionál a (4.4) módon definiált L operátor esetén megegyezik a H_L tér skaláris szorzatával.) Legyen $u_n^* \in S_n$ olyan elem, amelyre

$$\Phi(u_n^*, v) = (f, v) \quad \forall v \in S_h. \tag{4.13}$$

Így (4.13) alapján az $f \in H^0(\Omega)$ elemhez hozzárendeljük (egyértelműen) az u_n^* elemet. Jelölje ezt a hozzárendelést a T_n operátor, azaz

$$T_n f = u_n^*.$$

Mutassuk meg, hogy T_n kielégíti az F_1 feltételt!

a) (4.13) alapján

$$\Phi(T_n f, v) = (f, v) \tag{4.13a}$$

és így $(T_n f, g) = (g, T_n f) = \Phi(T_n f, T_n g) = (f, T_n g)$; $\forall f, g \in H^0(\Omega)$ azaz T_n szimmetrikus és (L operátor együttható-függvényeinek tulajdonságai és (4.12) következtében) pozitív definit $H^0(\Omega)$-n. Mivel ha $f_h \in S_h$ és $T_n f_h = 0$, akkor $f_h = 0$, így T_n S_h-n szigorúan pozitív definit.

b) Legyen $u^* = \tilde{T} f$; $u_n^* = T_n f$ és ekkor a véges elemes közelítés elliptikus típusú feladatokra már kimutatott approximációs tulajdonsága következtében a (4.10) feltétel érvényes [8].

Mindezek alapján a (4.9) szemidiszkrét feladat felírható

$$(D_t u_n^*, v)+\Phi(u_n^*, v) = 0 \quad \forall v\in S_h;\ t>0 \tag{4.14}$$

alakban.

4.2. Megjegyzés. A (4.14) feladat természetesen megegyezik a (2.16) feladat felírásával.

4.3. Megjegyzés. Nem feltétlen szükséges az $S_h=V_n$ altérsorozatot a splineterek közül választani. Ugyanakkor, ha bármilyen más véges dimenziós altérsorozatot választunk, akkor annak a (4.11) „inverz approximációs feltételt" ki kell elégítenie. A módszer alkalmazásának alapvető kérdése, hogy milyen hibabecslés adható meg az $e_n(x, t)$ függvényre. A T_n operátorral jellemzett térbeli diszkretizáció utáni hibafüggvény Ω-beli normáját — az egyszerűbb jelölés érdekében — jelöljük $e_h(t)$-vel. A hibabecslés minőségét nagymértékben befolyásolja az $u_{0,n}$ elem u_0-tól való eltérésének mértéke. Mivel $u_{0,n}$ az u_0 adott elem S_h-beli approximációja, a továbbiakban feltesszük, hogy rögzített s esetén $(0\leqq s\leqq r)$ érvényes az

$$\|u_{0,h}-u_0\| \leqq C\cdot h^s \|u_0\|_s \tag{4.15}$$

becslés. Az ilyen megválasztás mindig lehetséges, mivel ha $P_0: H^0(\Omega)\to S_h$ képező ortogonális projektor és $u_{0,h}=P_0u_0$, akkor (4.15) becslés érvényes.

Ezután térjünk át a T_n térbeli diszkretizáció hibájának becslésére.

4.1 ÁLLÍTÁS ([3]). Tegyük fel, hogy T_n operátor kielégíti az F_1 feltételt és $u_{0,h}$ megválasztásánál a (4.15) feltétel teljesül. Ekkor létezik olyan C pozitív állandó, hogy

$$\|e_h(t)\| \leqq Ch^s\|u_0\|_s \quad 0\leqq t\leqq T, \quad 0\leqq s\leqq r. \tag{4.16}$$

4.2 ÁLLÍTÁS ([3]). Tegyük fel, hogy a 4.1 állítás feltételei teljesülnek. Ekkor

$$\|D_t^j e_h(t)\| \leqq C\cdot h^r t^{r/2-j}\|u_0\|. \tag{4.17}$$

Ezek az állítások $H^0(\Omega)$ normában adnak becslést a hibafüggvényre és annak idő szerinti deriváltjára. Ugyanakkor megadható becslés maximum-normában is. Jelölje

$$|w| = \|w\|_{C(\Omega)} = \sup_{x\in\Omega}|w|$$

$$|w|_s = \|w\|_{C^{(s)}(\Omega)} = \sup_{x\in\Omega}|D^k w| \quad 0<k\leqq s$$

és a T_n operátor-sorozatra definiáljunk egy feltételrendszert:

F_2 feltétel: tegyük fel, hogy a T_n sorozathoz létezik olyan C pozitív állandó és $\gamma(h)$ függvény, hogy megfelelően kicsiny h esetén

$$|T_n w| \leqq C|\tilde{T}w|_1; \quad \|T_n w\| \leqq C\|\tilde{T}w\|_1; \quad |(T_n-\tilde{T})w| \leqq \gamma(h)|\tilde{T}w|_r.$$

4.3 ÁLLÍTÁS. Tegyük fel, hogy a T_n operátor-sorozatra F_1 és F_2 feltételek, az $u_{0,n}$ megválasztására pedig a (4.15) érvényesek. Ekkor minden $t_0>0$ esetén létezik olyan C pozitív állandó, hogy

$$|e_h(t)| \leqq C(\gamma(h)+h^r)\|u_0\| \quad t_0\leqq t\leqq T. \tag{4.18}$$

4.1. Megjegyzés. Az $L=-\Delta$ és $r>2$ esetre NIETCHE [13] megvizsgálta az F_2 feltétel teljesülését. Megmutatta, hogy ha T_n az Ω tartomány egyenletes háromszögekre bontásával nyert, a *Galjorkin-módszert* a spline-függvényekre alkalmazó térbeli diszkretizáció (véges elemek módszere), akkor a $\gamma(h)=C\cdot h^r$ függvénnyel az F_2 feltétel teljesül, azaz (4.18) becslés jobb oldalán a $C\cdot h^r\|u_0\|$ felső korlát szerepel.

4.2. Megjegyzés. Ha (4.2) peremfeltétel helyett a második *(Neumann)* peremérték-problémát tekintjük, akkor V_n-re teljesülnie kell a (4.11)-nek megfelelő $\inf_{v\in V_n}\{\|w-v\|+h\|w-v\|_1\}\leqq C\cdot h^s\|w\|_s$ $1\leqq s\leqq r$ inverz approximációs feltételnek minden $w\in H^1(\Omega)$ elemre. Ez a $V_n=S_h$ (azaz spline-teres altér) megválasztásánál teljesül. Továbbá F_1 és F_2 feltételek teljesülése is szükséges. Ha T_n a (4.13a) által definiált *Galjorkin-módszer* és $L=-\Delta+I$ (I: az egység operátor), akkor $N=2$ esetre SCOTT [14] megmutatta, hogy nem lényeges korlátozások mellett az F_1 és F_2 feltételek a

$$\gamma(h)=\begin{cases} c\cdot h^2\log(h^{-1}) & r=2\\ c\cdot h^r & r>2\end{cases}$$

függvénnyel teljesülnek.

Térjünk át a T_n operátorral a térbeli változókban diszkretizált (4.9) feladat megoldására. A T_n operátor-sorozatra tett feltevéseink alapján nyilvánvalóan létezik T_n operátornak inverze. Jelölje L_n és ekkor (4.9) felírható

$$\begin{aligned} D_t u_n^*+L_n u_n^*&=0 \quad 0<t\leqq T\\ u_n^*(0)&=P_0u_0\in S_n\end{aligned} \tag{4.19}$$

alakban. Ekkor a (4.5) kifejezéssel megegyező módon a (4.19) megoldása is felírható

$$u_n^*(x,t)=\sum_{j=1}^{\infty}\beta_j\exp\{-t\Lambda_j\}\Phi_j\equiv\exp(-tL_n)P_0u_0 \tag{4.20}$$

alakban, ahol $\beta_j=(u_0,\Phi_j)$; (Λ_j,Φ_j) az L_n operátor teljes sajátérték- és sajátvektor rendszere. Jelölje U^m a (4.1)—(4.3) feladat megoldásának közelítését a $t=mk$ időrétegen ($k>0$ az idő szerinti diszkretizációs lépésköz). Az időrétegeken való közelítések meghatározására egylépéses iterációs eljárást alkalmazunk, amely

$$U^{m+1}=r(kL_n)U^m \quad m=0,1,\ldots;\quad mk\leqq T \tag{4.21}$$

alakú, ahol $r(\tau)$ függvény az $e^{-\tau}$ kifejezés bizonyos rendű approximációja. Tegyük fel, hogy

$$r(\tau)=e^{-\tau}+o(\tau^{\nu+1}) \quad (\tau\to 0) \tag{4.22}$$

(azaz $r(\tau)$ ν-ed rendű approximáció), továbbá, hogy az L_n sajátértékei nem gyökei $r(\tau)$-nek, azaz U^{m+1} egyértelműen meghatározható. Látható, hogy a (4.21) séma megadásához az $r(kL_n)$-t szükséges definiálnunk. Mivel a térbeli diszkretizációt T_n operátorral valósítjuk meg, ezért az L_n helyett a T_n ismert, és mivel ezek egymás inverzei, ezért a (4.21) séma megadásához a (4.22) approximációs tulajdonsággal rendelkező $r(k/\mu)$ függvényt szükséges definiálnunk. Legyen

$$r(k|\mu)=z_0\prod_j(\mu-\beta_j)/\prod_j(\mu-\gamma_j) \tag{4.23}$$

alakú, ahol z_0, β_j, γ_j ismeretlen együtthatók. Ekkor a (4.21) séma a

$$\prod_j (T_n-\gamma_j I)U^{m+1} = z_0 \prod_j (T_n-\beta_j I)U^m \tag{4.24}$$

alakban írható fel. Ez azt jelenti, hogy a (4.19) térben diszkretizált szemidiszkrét feladat időben való diszkretizációja során időrétegenként adott $F_h \in S_h$ mellett

$$(zT_n+\beta)w = (\gamma T_n+\delta I)F_h \tag{4.25}$$

típusú feladatokat kell megoldanunk, ahol γ, β, z, δ adott paraméterek. Ha T_n a *Galjorkin-módszert* reprezentálja, azaz a diszkretizációt (4.13a) írja le, akkor a (4.25) séma alakja a következő: Írjuk fel (4.25)-öt a vele ekvivalens általánosított alakban:

$$z(T_n w, \chi)+\beta\Phi(w, \chi) = \gamma\Phi(T_n F_h, \chi)+\delta\Phi(F_h, \chi) \quad \forall \chi \in S_h$$

és ekkor (4.13a)-t alkalmazva mindkét oldal első tagjára:

$$z(w, \chi)+\beta\Phi(w, \chi) = \gamma(F_h, \chi)+\delta\Phi(F_h, \chi) \quad \forall \chi \in S_h, \tag{4.26}$$

amely a (4.23) approximáció együtthatóinak megválasztásával időlépésenként közvetlenül definiálja a diszkretizációt.

Vegyük észre, hogy az $u_n^*(x, t)$ (4.20) előállításához hasonlóan U^m is megadható:

$$U^m = (r(kL_n))^m P_0 u_0 = \sum_{j=1}^{\infty} (r(k\Lambda_j))^m \beta_j \Phi_j \tag{4.27}$$

alakban. Ennek alapján közvetlenül belátható, hogy a (4.21) séma akkor stabil $H^0(\Omega)$-ban, ha a

$$\max_j |r(k\Lambda_j)| \leqq 1 \tag{4.28}$$

feltétel teljesül, ekkor ugyanis a *Parseval-egyenlőtlenséget* alkalmazva:

$$\|U^m\| \leqq \|P_0 u_0\| \leqq \|u_0\|;$$

ami a stabilitást jelenti. Ezért a továbbiakban olyan sémákat tekintünk, amelyekre a (4.28) tulajdonság érvényes. A könnyebb áttekinthetőség kedvéért soroljuk osztályokba a (4.22) és a (4.28)-t kielégítő sémák közül néhányat.

Megnevezés	Tulajdonság
I.	$r(\tau)<1 \quad 0<\tau<z_1 \quad (z_1>0)$
Ia.	$r(\tau)$ I. típusú és $k\Lambda_{max} \leqq z_2 \quad (0<z_2<z_1)$
II.	$\|r(\tau)\|<1 \quad (\tau>0)$
IIa.	$r(\tau)$ II. típusú és $k\Lambda_{max} \leqq z_3 \quad (0<z_3<\infty)$
III.	$\|r(\tau)\|<1 \quad (\tau>0)$; $\lim r(\tau)=0 \quad (\tau\to\infty)$

(Vegyük észre, hogy az I., II., III. egyre erősödő feltételrendszert jelentenek.) Mielőtt az ilyen típusú sémák konkrét megadásával foglalkoznánk, fogalmazzuk meg a rájuk vonatkozó hibabecsléseket [2].

4.4 ÁLLÍTÁS. Legyenek sémáink Ia vagy IIa típusúak. Ekkor $0<t=mk<T$ esetén létezik olyan C pozitív állandó, hogy

$$\|U^m - u_n^*(\cdot, mk)\| \leqq C \cdot k^\nu t^{-\nu} \|u_0\|. \tag{4.29}$$

KÖVETKEZMÉNY. Mivel 4.2. állításban (4.17) $j=0$ esetén $\|u_n^*(\cdot, mk) - u(\cdot, mk)\| \leqq C \cdot h^r t^{-r/2} \|u_0\|$, így a globális hibára $H^0(\Omega)$-ban a következő becslés érvényes:

$$\|U^m - u(\cdot, mk)\| \leqq C\{h^r t^{-r/2} + k^\nu t^{-\nu}\} \|u_0\| \quad 0 < t \leqq T. \tag{4.30}$$

Megjegyezzük, hogy a (4.30) becsléshez elegendő, ha $u_0 \in H^0(\Omega)$. Ha az u_0 simaságára vonatkozóan erősebb feltevést teszünk, akkor olyan becslés nyerhető, amely nem tartalmazza az idő szerinti változót.

4.5 ÁLLÍTÁS. Legyenek a sémáink Ia vagy II típusúak és legyen $u_0 \in H^{2\nu}(\Omega)$. Ekkor $0<t=mk\leqq T$ esetén létezik olyan C pozitív állandó, hogy

$$\|U^m - u^*(\cdot, mk)\| \leqq C\{h^r \|u_0\|_r + k^\nu \|u_0\|_{2\nu}\}. \tag{4.31}$$

A (4.30) becslés hibája, hogy a $t=0$ közelében elromlik. Ezért célszerű olyan (4.30) típusú becslést adni, amely $t=0$-ig bezárólag egyenletesen jó. Figyelembevéve a 4.1 állítás (4.16) becslését érvényes a következő:

4.6 ÁLLÍTÁS. Legyenek a sémáink Ia vagy II típusúak és $u_0 \in H^s(\Omega)$ $(0 \leqq s \leqq r)$. Ekkor $0 \leqq t = mk \leqq T$ esetén

$$\|U^m - u^*(\cdot, mk)\| \leqq C\{h^{\min(r,s)} + k^{\min(\nu, s/2)}\} \|u_0\|_s. \tag{4.32}$$

KÖVETKEZMÉNY. Ha a kezdeti állapotot leíró u_0 függvény maximálisan sima (azaz $u_0 \in H^r(\Omega)$) akkor

$$\|U^m - u^*(\cdot, mk)\| \leqq C\{h^r + k^{\min(\nu, r/2)}\} \|u_0\|_r \quad 0 \leqq mk \leqq T. \tag{4.33}$$

(Megjegyezzük, hogy 4.6 állítás lényegében 4.5 állítás kiterjesztése kevésbé sima kezdeti függvényekre és a teljes időintervallumra.)

A következő állítás a maximum-normában mond ki hibabecslést.

4.7 ÁLLÍTÁS. Tegyük fel, hogy $\{T_n\}$ az F_1 és F_2 feltételeket kielégítő diszkretizációs eljárás és legyen (4.21) Ia, IIa, vagy III. típusú séma. Ekkor $0<t_0\leqq mk<T$ esetén létezik olyan C pozitív állandó, hogy

$$|U^m - u^*(\cdot, mk)| \leqq C\{\gamma(h) + h^r + k^\nu\} \|u_0\|_r. \tag{4.34}$$

KÖVETKEZMÉNY. A *Galjorkin-típusú* (4.13a) sémákra a (4.34) érvényes. (Megjegyezzük, hogy a $\gamma(h)$ függvényeket néhány speciális operátor esetén 4.3 állítást követően megadtuk.)

A továbbiakban térjünk át a különböző típusú sémák konkrét definiálására. Ehhez a (4.22) és (4.28) tulajdonságokkal rendelkező $r(\tau)$ függvény megadása szükséges.

Először az $e^{-\tau}$ függvény *Padé-típusú approximációját* tekintsük. Mint ismeretes, az általános *Padé-típusú approximáció*

$$r_{p,q}(\tau) = \frac{n_{p,q}(\tau)}{d_{p,q}(\tau)} \tag{4.35}$$

alakú, ahol

$$n_{p,q}(\tau) = \sum_{j=0}^{q} \frac{(p+q-1)!q!}{(p+q)!j!(p-j)!}(-1)^j\tau^j, \tag{4.36}$$

$$d_{p,q}(\tau) = \sum_{j=0}^{p} \frac{(p+q-1)!p!}{(p+q)!j!(p-j)!}\tau^j \tag{4.37}$$

és ekkkor

$$r_{p,q}(\tau) = e^{-\tau}+o(\tau^{p+q+1}) \quad (\tau \to 0 \quad \text{esetén}). \tag{4.38}$$

Megmutatható, hogy a (4.35) által definiált (4.21) séma $v=p+q$-ad rendben pontos és

$p < q$ esetén I típusú,
$p = q$ esetén II típusú,
$p = q+1, q+2$ esetén III típusú

diszkretizációs séma.

A p és q értékeinek megadásával tetszőleges pontosságú sémák megadhatók. Tekintsünk néhány egyszerűbb, a gyakorlati alkalmazásban is sűrűn előforduló esetet és adjuk meg a (4.13a) *Galjorkin-típusú approximáció* esetén a sémákat!

a) $$r_{0,1} = 1-\tau.$$

Ez I típusú és $v=1$ rendben pontos séma: (4.21) alapján

$$U^{m+1} = U^m - kL_nU^m. \tag{4.39}$$

A számítási algoritmus a következő:

$$r_{0,1}(k/\mu) = 1-k/\mu = \frac{\mu-k}{\mu}$$

és (4.24) alapján: $T_nU^{m+1}=(T_n-kI)U^m=T_nU^m-kU^m$.
Ezt általános alakban felírva:

$$\Phi(T_nU^{m+1}, \chi) = \Phi(T_nU^m, \chi)-k\Phi(U^m, \chi) \quad \chi\in S_h$$

és (4.13a)-t alkalmazva

$$(U^{m+1}, \chi) = (U^m, \chi)-k\Phi(U^m, \chi) \quad \chi\in S_h. \tag{4.40}$$

(Vegyük észre, hogy (4.40) megegyezik (3.6) sémával.) Vizsgáljuk meg a (4.28) teljesülési feltételét, azaz, hogy (4.40) milyen típusú séma. A (4.13a)-ból közvetlenül következik a

$$(L_nu_n^*, \chi) = \Phi(u_n^*, \chi) \quad \chi\in S_h \tag{4.13b}$$

azonosság. Tegyük fel, hogy Φ bilineáris funkcionál korlátos $H^1(\Omega)$-ban és tetszőleges $\chi\in S_h$ elemre:

$$\|\chi\|_1 \leqq C\cdot h^{-1}\|\chi\| \tag{4.41}$$

ahol C χ-tól független pozitív állandó. (Eseteinkben ezek a feltételek teljesülnek: Φ a (4.12) módon történő megválasztásával korlátos H^1-ban valamint az S_h splinetéres definiálásával a spline-függvények tulajdonsága következtében (4.41) érvényes).

Ekkor (4.13b) és a feltételek következtében:

$$(L_n\chi, \chi) = \Phi(\chi, \chi) \leqq C_1\|\chi\|_1^2 \leqq C_2 h^{-2}\|\chi\|^2 \tag{4.42}$$

így L_n maximális sajátértékére: $\Lambda_{\max} \leqq C_2 h^{-2}$ azaz (4.40) Ia típusú séma a

$$kh^2 \leqq z_2 C_2^{-1} \quad (z_2 < 1) \tag{4.43}$$

feltétellel. Így (4.40) sémával (vagy a (3.6) sémával) a lépésközökre vonatkozó (4.43) korlátozó feltétellel lehet csak számolni. (Ez megegyezik a [9]-beli explicit differenciasémák stabilitásánál említett korlátozó feltétellel.) A (4.43) feltételt kielégítő lépésközök megválasztása esetén a közelítés pontossága $H^0(\Omega)$-normában $O(h^r+k)$ nagyságrendű.

b)
$$r_{1,0}(\tau) = \frac{1}{1+\tau}.$$

Ez III. típusú, $\nu=1$-rendben pontos séma. Mivel

$$r_{1,0}(k/\mu) = \frac{\mu}{k+\mu}$$

ezért az

$$(U^{m+1}, \chi)+k\Phi(U^{m+1}, \chi) = (U^m, \chi) \quad \chi\in S_h \tag{4.44}$$

implicit sémát generálja. (Ez megegyezik a (3.7) sémával. Ezen sémával, (4.40) sémától eltérően, a k és h lépésközre tett korlátozó feltételek nélkül számolhatunk és a közelítés pontossága $H^0(\Omega)$-normában $O(h^r+k)$.

c)
$$r_{1,1}(\tau) = \frac{1-\tau/2}{1+\tau/2}.$$

Ez II. típusú séma és $\nu=2$ rendben pontos. Mivel $r_{1,1}(k/\mu)=\dfrac{\mu-k/2}{\mu+k/2}$, ezért $r_{1,1}$ az

$$(U^{m+1}, \chi)+\frac{k}{2}\Phi(U^{m+1}, \chi) = (U^m, \chi)-\frac{k}{2}\Phi(U^m, \chi) \quad \chi\in S_h \tag{4.45}$$

sémát generálja. (Ez megegyezik a (3.8) sémával.) A (4.45) séma szintén korlátozó feltételek nélkül stabil és $O(h^r+k^2)$ nagyságrendben pontos a $H^0(\Omega)$-normában.

d) *Magasabb rendű Padé-típusú approximációk*

Gyakorlati jelentőséggel bírnak még azok a magasabb rendűek, ahol a (4.35) típusú approximáció nevezője másodfokú, azaz a következő:

$$r_{2,0}(\tau) = \frac{1}{1+\tau+\tau^2/2}; \quad r_{2,1}(\tau) = \frac{1-\tau/3}{1+2/3\tau+\tau^2/6}; \quad r_{2,2}(\tau) = \frac{1-\tau/2+\tau^2/12}{1+\tau/2+\tau^2/12}. \tag{4.46}$$

Ezek alkalmazása ugyanakkor eltér az előzőektől, mivel a nevezők gyökei komplex értékűek. Közvetlenül belátható, hogy a $j=0, 1, 2$ esetekre

$$r_{2,j}(k/\mu) = 1-k\,\mathrm{Re}\left(\frac{\gamma_j}{\mu-\beta_j k}\right) \tag{4.47}$$

ahol

$$\beta_j = \frac{1}{j+1}\left(1+i\frac{\sqrt{j+1}}{j+1}\right) \quad (i \equiv \sqrt{-1}); \quad \gamma_j = \left(1-i\frac{j}{2}\sqrt{j+1}\right). \tag{4.48}$$

Bevezetve az

$$A_{n,j} = T_n + k\beta_j I \tag{4.49}$$

operátort a diszkretizációs sémának

$$U^{m+1} = U^m - k\,\mathrm{Re}\,(\gamma_j A_{n,j}^{-1} U^m) \tag{4.50}$$

alakú, ahol $A_{n,j}^{-1}$ létezik, mivel $\mathrm{Re}\,\beta_j > 0$. Így a (4.50) sémával való számoláshoz a $W = k\gamma_j A_{n,j}^{-1} U^m$ típusú elem megkeresése szükséges, ami a

$$(W, \chi) + k\beta_j \Phi(W, \chi) = k\gamma_j \Phi(U^m, \chi) \tag{4.51}$$

feladat (egy komplex értékű, elliptikus típusú feladat) megoldását jelenti. Megjegyezzük, hogy (4.51) lényegében egy komplex együtthatós, lineáris algebrai egyenletrendszer, de a rendszer megoldása valós értékű.

A *Padé-típusú approximáción* kívül célszerű az $\{L_n^p\}$ p-ed rendű $(p > -1)$ *Laguerre-polinomokkal* való közelítés. Ezek alkalmazásával ugyanis az eddigiektől eltérő, ugyanakkor gyakorlatban jól alkalmazható (4.21) diszkretizáliós sémákat nyerhetünk.

Mint ismeretes, a *Laguerre-polinomokra* igazak a következők:

$$L_n^p(x) = (n+p)! \sum_{j=0}^{n} (-1)^j \frac{x^j}{j!\,(n-j)!\,(j+p)!}, \tag{4.52}$$

$$x^p e^{-x} \tag{4.53}$$

adott súlyfüggvénnyel ortogonálisak,

$$(1-t)^{-1-p} e^{-xt/1-t} = \sum_{n=0}^{\infty} L_n^p(x) t^n; \quad |t| < 1, \quad x > 0, \tag{4.54}$$

$$nL_n^p(x) = (-x+2n+p-1)L_{n-1}^p(x) - (n+p-1)L_{n-2}^p(x), \tag{4.55}$$

$$D_x L_n^p(x) = x^{-1}\{nL_n^p(x) - (n+p)L_{n-1}^p(x)\}; \quad n \geqq 1. \tag{4.56}$$

Legyen $\tau = xt/1-t$ és ekkor (4.54) $p=1$ esetén:

$$e^{-\tau} = (1-t)^2 \sum_{n=0}^{\infty} L_n^1(x) t^n. \tag{4.57}$$

Ezután (4.57)-re (4.55) összefüggést alkalmazva:

$$e^{-\tau} = 1 - \sum_{n=0}^{\infty} \frac{xL_n^1(x)}{n+1}\left(\frac{\tau}{x+\tau}\right)^{n+1}, \tag{4.58}$$

ami azt jelenti, hogy tetszőleges $b = 1/x$ pozitív számra

$$e^{-\tau} = 1 - \sum_{n=0}^{\infty} P_n(b)\left(\frac{\tau}{1+b\tau}\right)^{n+1} \quad (\tau > 0) \tag{4.59}$$

ahol

$$P_n(b) = (n+1)^{-1} b^n L_n^1(b^{-1}) \tag{4.60}$$

n-ed fokú polinom.

Mivel L_n^1 zérushelyei pozitívak és egymástól különbözőek, ezért (4.60) alapján ez P_n-re is igaz. Jelölje $b_n = \beta_n^{-1}$ a P_n polinom legnagyobb zérushelyét. (Azaz β_n az L_n^1 legkisebb zérushelye.) Legyen $\nu \geqq 2$ és

$$r_\nu(\tau) = 1 - \sum_{j=0}^{\nu-2} P_j(b_{\nu-1}) \left(\frac{\tau}{1+b_{\nu-1}\tau} \right)^{j+1} \tag{4.61}$$

racionális törtfüggvény, amelyre (4.59) és $B_{\nu-1}$ megválasztása következtében:

$$r_\nu(\tau) = e^{-\tau} + O(\tau^{\nu+1}). \tag{4.62}$$

Nyilvánvaló, hogy $r_\nu(0)=1$ és r_ν monoton csökkenő. Egyszerű eszközökkel belátható, hogy az $r_\nu(\tau) > -1$ tetszőleges $\tau > 0$ érték esetén és így a (4.61) kifejezéssel definiált $r_\nu(\tau)$ függvény egy II. típusú diszkretizációs sémát határoz meg. Tekintsük az r_ν által generált számítási algoritmust! Mivel

$$r_\nu(k/\mu) = 1 - \sum_{j=0}^{\nu-2} k^{j+1} P_j(b_{\nu-1}) (\mu + k b_{\nu-1})^{-(j+1)}$$

így (4.49) jelölést alkalmazva

$$U^{m+1} = U^m - \sum_{j=0}^{\nu-2} k^{j+1} P_j(b_{\nu-1}) A_{n,\nu-1}^{-(j+1)} U^m. \tag{4.63}$$

Ha T_n a *Galjorkin-módszer* operátora (azaz (4.13a) által definiált operátor), akkor a (4.63) gyakorlati realizálása a következő. Jelölje:

$$U_0^m = U^m; \quad U_{j+1}^m = k A_{n,\nu-1}^{-1} U_j^m \tag{4.64}$$

és ekkor

$$U^{m+1} = U^m - \sum_{j=0}^{\nu-1} P_j(b_{\nu-1}) U_{j+1}^m \tag{4.65}$$

azaz U^{m+1}-et U^m-ből (4.64) típusú $(\nu-1)$ darab lineáris rendszer megoldásával nyerhetjük. (Megjegyezzük, hogy ezen rendszerek mátrixai ugyanazok!)

Mivel (4.64) a (4.49) jelölést figyelembevéve felírható

$$(T_n + k b_{\nu-1} I) U_{j+1}^m = k U_j^m \tag{4.64a}$$

alakban, ezért a *Galjorkin-módszer* esetén a (4.64a) a

$$(U_{j+1}^m, \chi) + k b_{\nu-1} \Phi(U_{j+1}^m, \chi) = k \Phi(U_j^m, \chi) \quad \chi \in S_h \tag{4.66}$$

feladatok megoldását jelenti. Foglaljuk össze a számítási algoritmusunkat!

A módszer pontosságát meghatározó ν-t megválasztva meghatározzuk $b_{\nu-1}$ és $P_j(b_{\nu-1})$ $(j=0, \ldots \nu-2)$ értékeket. Ezután a $j=0, 1 \ldots\ldots \nu-2$ értékekre megoldjuk a (4.66) feladatokat. (Ez S_h bázisfüggvényeinek ismeretében $\nu-1$ számú lineáris algebrai egyenletrendszer megoldását jelenti. Az egyes rendszerek méretei $\dim S_h$). Ezt követően az előzőekben meghatározott $U_1^m, U_2^m, \ldots U_{\nu-1}^m$ értékekkel (4.65) formulával meghatározhatjuk U^{m+1} értékét.

Tekintsünk néhány speciális esetet!

e) $\nu=2$ megválasztás esetén

$$r_2(\tau) = 1-\tau(1+\tau/2)^{-1} = (1-\tau/2)(1+\tau/2)^{-1}$$

azaz a *Padé-típusú approximációnál* tárgyalt *Crank—Nicolson sémát* nyerjük (c) eset.)

f) $\nu=3$ megválasztás esetén: $r_3(\tau)=1-\tau\cdot(1+b_2\tau)^{-1}-(\sqrt{3}/6)\,\tau(1+b_2\tau)^{-2}$

$$b_2 = \frac{1}{2}\left(1+\frac{\sqrt{3}}{3}\right); \quad P_0 = 1; \quad P_1(b_2) = 0{,}288\,675.$$

(Ez az ún. *Calahan-séma*)

g) $\nu=4$ megválasztás esetén

$$b_3 = 1{,}068\,579; \quad P_0 = 1; \quad P_1(b_3) = 0{,}568\,579; \quad P_2(b_3) = 0{,}239\,948$$

($\nu=10$-ig bezárólag [2]-ben megtalálhatók a megfelelő értékek.)

Ha maximum-normában szeretnénk hibabecslést kapni és ugyanakkor a diszkretizációs lépésekre nem akarunk megkötéseket tenni, akkor III. típusú sémát szükséges definiálnunk. Ez a *Laguerre-típusú polinomokkal* is lehetséges. Az előzőekhez hasonlóan, de most $p=0$-val számolva, $b>0$ esetén

$$e^{-\tau} = (1+b\tau)^{-1}\sum_{j=0}^{\infty} Q_j(b)\left(\frac{\tau}{1+b\tau}\right)^j; \quad \tau \geqq 0 \tag{4.67}$$

ahol

$$Q_j(b) = b^j L_j^0(b^{-1}).$$

Jelölje most $\nu \geqq 1$ esetén

$$r_\nu(\tau) = \frac{1}{1+b_\nu\tau}\sum_{j=0}^{\nu-1} Q_j(b_\nu)\left(\frac{\tau}{1+b\tau}\right)^j; \quad \tau \geqq 0 \tag{4.68}$$

ahol b_ν a Q_ν polinom legnagyobb zérushelye.

Ekkor

$$r_\nu(\tau) = e^{-\tau}+o(\tau^{\nu+1}) \quad (\tau \geqq 0)$$

és

$$\lim_{\tau\to\infty} r_\nu(\tau) = 0.$$

Belátható, hogy $\tau>0$ esetén $0\leqq r(\tau)\leqq 1$, azaz (4.68) megválasztással a (4.21) séma III. típusú. Számítási algoritmusa a következő:

$$r_\nu(k/\mu) = \frac{\mu}{\mu+b_\nu k}\sum_{j=0}^{\nu-1} k^j Q_j(b_\nu)\left(\frac{1}{\mu+kb_\nu}\right)^j. \tag{4.69}$$

Jelölje

$$U_0^m = A_{n,\nu}^{-1}T_n U_n^m, \tag{4.70}$$

$$U_{j+1}^m = kA_{n,\nu}^{-1}U_j^m \quad j = 0, 1 \ldots \nu-2.$$

Ekkor (4.69) alapján

$$U^{m+1} = \sum_{j=0}^{\nu-1} Q_j(b_\nu)U_j^m. \tag{4.71}$$

Ha T_h a (4.13a) módon definiált *Galjorkin-módszer*, akkor (4.70) megoldása az

$$(4.72)\qquad \begin{aligned}(U_0^m, \chi)+kb_v\Phi(U_0^m, \chi) &= (U^m, \chi) \quad \chi\in S_h,\\ (U_{j+1}^m, \chi)+kb_v\Phi(U_{j+1}^m, \chi) &= k\Phi(U_j^m, \chi), \quad j=0, 1 \dots v-2\end{aligned}$$

$v-1$ számú lineáris feladat megoldását jelenti. Foglaljuk össze a számítási algoritmusunkat!

A módszer pontosságát meghatározó v megválasztása után meghatározzuk b_v és $Q_j(b_v)$ $(j=0, \dots v-1)$ értékeket. Ezután (4.72) alapján $j=0, \dots v-2$ értékekre megoldjuk a (4.72) lineáris feladatokat. (Ez S_h bázisfüggvényeinek ismeretében $v-1$ számú lineáris algebrai egyenletrendszer megoldását jelenti.) Végezetül (4.71) alapján meghatározzuk U^{m+1} értékét.

Tekintsünk néhány speciális esetet!

h) $v=2$ megválasztás esetén

$$b_2 = 1{,}707\,106; \quad Q_0 = 1; \quad Q_1(b_2) = 0{,}707\,106.$$

i) $v=3$ esetén: $b_3=2{,}405149$; $Q_0=1$; $Q_1(b_3)=1{,}405149$; $Q_2(b_3)=1{,}474\,445$ ($v=10$-ig bezárólag [2]-ben megtalálhatók a megfelelő értékek.)

Eddig a (4.19) térbeli változóban diszkretizált (lényegében lineáris, elsőrendű közönséges differenciálegyenletrendszert jelentő) feladat (4.21) egylépéses sémával történő megoldását tárgyaltuk. Ugyanakkor a rendszert megoldhatjuk többlépéses módszerekkel is, amelyek a gyakorlatban szintén jelentősek. Röviden foglaljuk össze a lineáris többlépéses módszereket (l. t. m.)

Legyenek:

$$(4.73)\qquad \varrho(\xi) = \sum_{j=0}^{v} \varrho_j \xi^j \quad (\varrho_v > 0),$$

$$\sigma(\xi) = \sum_{j=0}^{v} \sigma_j \xi^j$$

v-ed fokú polinomok és alkalmazzuk a (ϱ, σ) l. t. m.-t az

$$(4.74)\qquad \begin{aligned} D_t y(t) &= -\lambda y(t)\\ y(0) &= 1\end{aligned}$$

$(y(t)\colon \mathbf{R}^1 \to \mathbf{R}^1$ megfelelően sima, egyváltozós függvény) tesztfeladatra.

Jelölje y^m az $y(mk)$ $(k>0)$ közelítését, amit az

$$(4.75)\qquad \sum_{j=0}^{v} \varrho_j y^{m+j} = -\tau\lambda \sum_{j=0}^{v} \sigma_j y^{m+j}$$

egyenletből határozunk meg. (Az $y^0, y^1, \dots, y^{v-1}$ értékeket ismertnek tételezzük fel.)

Figyelembevéve $u_h^*(x, t)$ (4.20) alakú előállítását, a l. t. m-től is megköveteljük az A_0 stabilitást, azaz, hogy $y^m \to 0 (m\to\infty)$ minden pozitív λ esetén. (Az egylépéses módszereknél ezt a (4.28) feltétel biztosította). Mivel (4.75) átírható

$$(4.76)\qquad \sum_{j=0}^{v} (\varrho_j + \tau\sigma_j)\, y^{m+j} = 0$$

alakra és ennek a karakterisztikus egyenlete

$$\sum_{j=0}^{\nu} \gamma_j \xi^j = 0; \quad \gamma_j = \varrho_j + \tau\sigma_j$$

alakú, ezért a l. t. m. akkor A_0-stabil, ha a

$$P(\xi) = \varrho(\xi) + \tau\sigma(\xi) \tag{4.77}$$

polinom $\xi_j(\tau)$ gyökeire érvényes a

$$|\xi_j(\tau)| < 1 \quad \forall \tau > 0 \tag{4.78}$$

egyenlőtlenség. Így valamely (ϱ, σ) l. t. m. stabilitásának feltétele a (4.78) egyenlőtlenség.

Alkalmazzuk a (ϱ, σ) l. t. m.-t a (4.19) feladat megoldására. Legyen T_h a (4.13a) alakú *Galjorkin-módszer*. Ekkor a

$$(\sum_{j=0}^{\nu} \varrho_j U^{m+j}, \chi) + k\Phi(\sum_{j=0}^{\nu} \sigma_j U^{m+j}, \chi) = 0 \quad \chi \in S_h \tag{4.79}$$

alakú feladatot szükséges megoldanunk, feltéve, hogy U^0-t a (4.19) kezdeti feltételből, $u^1, \ldots u^{\nu-1}$-t pedig az előzőekben ismertetett valamely egylépéses módszerrel már meghatároztuk. Ekkor érvényes a következő.

4.8 ÁLLÍTÁS ([16]). Legyen a $(\varrho, \sigma)q$-ad rendű, A_0-stabil módszer olyan, hogy a σ polinom abszolút értékben egy értékű gyökei egyszeresek. Ekkor, ha a feladat $u^*(x, t)$ megoldása megfelelően sima, azaz, ha rögzített $t \in [0, T]$ esetén $u^* \in H^S(\Omega)$, akkor

$$\sup_{\nu \leq m \leq \infty} \|u^*(\cdot, mk) - U^m\| \leq C\Big[\sum_{j=0}^{\nu-1} \|u^*(\cdot, jk) - U^j\| + (h^r + k^q) \log \tau^{-1} \|u_0\|_s\Big]. \tag{4.80}$$

(Vegyük észre, hogy a l. t. m. pontosságát nagymértékben befolyásolja az U^j $(j=0, \ldots \nu-1)$ kezdeti közelítések megválasztása: ha q-ad rendű l. t. m.-t választunk, akkor a kezdeti értékeket is az ilyen, vagy az eggyel alacsonyabb rendben pontos egylépéses módszerrel célszerű meghatároznunk.)

A szakasz befejezéseként tegyünk néhány megjegyzést!

4.2. Megjegyzés. Ha véges dimenziós altér bázisfüggvényeit ismerjük, akkor (4.79) feladat a

$$\sum_{j=0}^{\nu} (\varrho_j M + \sigma_j kQ)\alpha^{m+j} = 0 \quad m = 0, 1, \ldots \tag{4.81}$$

feladat megoldását jelenti, ahol M, Q és α megegyezik az előző szakasz (3.5) jelöléseivel $[w, v]_L = \Phi(w, v)$ kiegészítéssel. Ez azt jelenti, hogy időlépésenként egy

$$B = \sum_{j=0}^{\nu} (\varrho_j M + \sigma_j kQ) \tag{4.82}$$

azonos alakú mátrixszal rendelkező lineáris algebrai egyenletrendszer megoldása szükséges. Ez a B mátrix pozitív definit, ritka és sávstruktúrájú, valamint kondício-

náltsági száma nem nő túlságosan gyorsan, véges elemes altér esetén [8]

$$\text{cond}\,(B) = O(kh^{-2}).$$

Így (4.81) megoldása numerikus szempontból aránylag egyszerű.

4.3. Megjegyzés. Eddigi állításainkat homogén jobb oldalú problémákra fogalmaztuk meg. Ha a

$$D_t u + Lu = F(x, t) \tag{4.83}$$

inhomogén jobb oldalú feladatot vizsgáljuk akkor, (4.79)

$$\left(\sum_{j=0}^{\nu} \varrho_j U^{m+j}, \chi\right) + k\Phi\left(\sum_{j=0}^{\nu} \sigma_j U^{m+j}, \chi\right) = k\left(\sum_{j=0}^{\nu} \sigma_j F^{m+j}(\cdot), \chi\right) \quad \chi \in S_h \tag{4.84}$$

alakú, ahol $F^m(x) = F(x, mk)$. Ekkor a (4.80) hibabecslés véges időintervallumra érvényes marad. (Végtelen időintervallumban $F_t(x, t)$ növekedése tett megszorító feltételek szükségesek hasonló becsléshez [16]. Ha a $\nu = 1$ esetet vizsgáljuk és figyelembevesszük, hogy az egylépéses módszereket véges időintervallumon vizsgáltuk, akkor a fentiekből következik, hogy az ott kimondott állítások a sémák (4.84) szerinti korrigálásával inhomogén egyenletekre is érvényesek.

4.4. Megjegyzés. Az eredeti (2.1), (2.2) probléma megoldására érvényes a következő becslés [18]

$$\|u^*(\cdot, t)\| \leqq e^{-\lambda_1 t}\|u_0\|; \quad t > 0 \tag{4.85}$$

ahol λ_1 az L operátor legkisebb sajátértéke. Ha a (ϱ, σ) l. t. m.-re kiegészítőleg feltesszük, hogy a $\varrho(\xi)$ polinom abszolút értékben egy értékű gyökei egyszeresek és a $\sigma(\xi)$ polinom gyökei abszolút értékben egynél kisebbek, akkor a (4.79) vagy a (4.84) sémával nyert (U^m) közelítések sorozata megőrzi ezt az „exponenciális lecsengést":

$$\|U^m\| \leqq C \exp\{-\tau_0 mk\} \max_{0 \leqq j \leqq \nu - 1} \|U^j\|; \quad \tau_0 > 0; \quad m \geqq \nu. \tag{4.86}$$

(Az ilyen tulajdonságú sémákat L_0-stabilnak nevezik, [17].)

4.5. Megjegyzés. Az $u^*(x, t)$ megoldás 4.8. állításban szereplő $u^*(x, t) \in H^S(\Omega)$ ($t > 0$ esetén) simasági feltétele teljesül, ha az u_0 kezdeti állapotot leíró függvényre $u_0 \in H^S(\Omega)$ és

$$u_0|_\Gamma = Lu_0|_\Gamma = \ldots = L^{[(s-1)/2]} u_0|_\Gamma = 0.$$

4.6. Megjegyzés. A (4.79) típusú l. t.m.-ek realizálása leegyszerűsödik, ha

$$\sigma_0 = \sigma_1 = \ldots = \sigma_{\nu-1} = 0.$$

Ekkor a sémánk:

$$\left(\sum_{j=0}^{\nu} \varrho_j U^{m+j}, \chi\right) + k\Phi(\sigma_\nu U^{m+\nu}, \chi) = 0 \quad \chi \in S_h, \tag{4.87}$$

$$(U^0, U^1, \ldots, U^{\nu-1} \text{ adottak})$$

alakú, és így a konkrét számítást jelentős (4.81) egyenlet is leegyszerűsödik:

$$\sum_{j=0}^{\nu} \varrho_j M\alpha^{n+j} + \sigma_\nu k Q \alpha^{n+\nu} = 0; \quad n = 0, 1... \tag{4.88}$$

$$\alpha^0, \alpha^1, \ldots \alpha^{\nu-1} \text{ adottak.}$$

5. Lineáris, időtől függő elliptikus részű parabolikus típusú parciális differenciálegyenletek diszkretizációja

Ebben a szakaszban a (4.1)—(4.3) feladat egy általánosabb kitűzésének diszkretizációját vizsgáljuk: feltesszük, hogy az egyenletben szereplő L elliptikus operátor időfüggő. Így tekintsük a

$$D_t u = -L(t)u \quad (x, t) \in \Omega \times (0, T], \tag{5.1}$$

$$u(x, t) = 0 \quad (x, t) \in \Gamma \times [0, T], \tag{5.2}$$

$$u(x, 0) = u_0(x) \quad x \in \Omega, \tag{5.3}$$

$$L(t)u = -\sum_{i,j=1}^{N} D_{x_i}(a_{ij}(x, t) D_{x_j} u) + a_0(x, t)u \tag{5.4}$$

feladatot. A továbbiakban néhány helyen (5.2) első peremfeltétel helyett az

$$\sum_{i,j=1}^{N} \tilde{n} a_{ij}(x, t) D_{x_j} u = 0 \quad (x, t) \in \Gamma \times [0, T] \tag{5.2a}$$

második peremfeltételt vizsgáljuk.

Tegyük fel, hogy a_0, a_{ij} $\bar{\Omega} \times [0, T]$-n elegendően sima függvények, az $A(x, t) = [a_{ij}]$ mátrix egyenletesen szigorúan pozitív definit, U_0 adott függvény. Ekkor $L(t)$ minden rögzített $0 < t < T$ esetén szigorúan pozitív definit, elliptikus típusú operátor $H^0(\Omega)$-n és dom $L = \{H^2(\Omega) \cap \mathring{H}^1(\bar{\Omega})$ térbeli az (5.1), (5.2), (5.3) feladatra, és a $H^2(\Omega) \cap \{w \in H^2(\bar{\Omega}); \sum_{i,j=1}^{N} \tilde{n} a_{ij} D_{x_j} w = 0, (x, t) \in \Gamma \times (0, T]$ (5.1), (5.2a), (5.3) feladatra.$\}$

Jelölje

$$\Phi(t)(\varphi, \psi) = \int_\Omega \Big(\sum_{i,j=1}^{N} a_{ij} D_{x_i} \varphi \cdot D_{x_j} \psi + a_0 \varphi \psi \Big) dx. \tag{5.5}$$

Ekkor az együttható függvényekre tett feltevéseink következtében $\Phi(t)$ bilineáris alak erősen koercitív $\mathring{H}^1(\Omega) \times \mathring{H}^1(\Omega)$-n az (5.1)—(5.2)—(5.3) feladat esetén és $H^1(\Omega) \times H^1(\Omega)$-n az (5.1)—(5.2a)—(5.3) feladat esetén. Így mindkét feladatosztálynak tetszőleges $u_0 \in H^0(\Omega)$ függvény esetén létezik általánosított megoldása [10].

Megtartva a 4. szakasz jelöléseit (csak kiemelve azok t-től való függését) jelölje $\tilde{T}(t)$: $H^0(\Omega) \to$ dom L az $L(t)$ operátor „megoldási operátorát", azaz

$$L(t)[\tilde{T}(t) f] = f \quad \forall f \in H^0(\Omega). \tag{5.6}$$

Jelölje továbbra is $\{S_h\} \subset H^0(\Omega)$ a véges dimenziós altérsorozatot, $\{T_n(t)\}$ $(0 \leqq t \leqq T)$

pedig olyan operátorsorozatot, hogy rögzített $0 \leqq t \leqq T$ esetén $T_n(t): H^0(\Omega) \to S_h$ képező operátor. A $\{T_n(t)\}$ operátorsorozatot úgy választjuk meg, hogy az S_h-n approximálja $\tilde{T}$ operátort, valamint az $L(t)$ operátor „jó tulajdonságai" mintegy átöröklődjenek. Ezt úgy biztosítjuk, hogy az $\{T_n(t)\}$ operátorsorozatnak ki kell elégítenie az alábbi feltételrendszert:

F_3 feltétel: legyen $\{T_n(t)\}$ olyan operátorsorozat, hogy tetszőleges rögzített $0 \leqq t \leqq T$ esetén

a) $T_n(t)$ operátor $H_0(\Omega)$-ban önadjungált, szigorúan pozitív definit.
Így létezik S_h-n inverz operátora és jelölje ezt $L_n(t)$, azaz

$$L_n(t)\varphi = T_n^{-1}(t)\varphi \quad \forall \varphi \in S_h.$$

b) Tetszőleges $f \in H^0(\Omega)$ esetén létezik olyan $r \geqq 2$ egész szám és $C(j)$ pozitív állandó hogy

$$\|(\tilde{T}^j(t) - T_n^j(t))f\| \leqq C(j)h^{l+2}\|f\|_l; \quad j \geqq 0; \quad 0 \leqq l \leqq r-2 \tag{5.7}$$

ahol

$$T_n^j(t) = (D_t)^j T_n(t)$$

c) S_h-n létezik olyan $\| \cdot \|_{S_h}$ norma, hogy

$$\|\varphi\|^2 \leqq C\|\varphi\|_{S_h}^2 \leqq C(L_n(t)\varphi, \varphi), \tag{5.8}$$

$$|(L_n^j(t)\varphi_1, \varphi_2)| \leqq C(j)\|\varphi_1\|_{S_h}\|\varphi_2\|_{S_h} \quad 0 \leqq t \leqq T; \quad \varphi_1, \varphi_2 \in S_h \tag{5.9}$$

ahol

$$L_n^j(t) = (D_t)^j L_n(t).$$

Megjegyezzük, hogy a (4.13a) típusú *Galjorkin-approximáció* operátorsorozata — mint az a korábbi eredményeinkből várható — kielégíti az F_3 feltételt [4]. Jelölje $P_l^h(t) = T_n(t)L(t)$: dom $L \to S_h$ az ún. „elliptikus projekciós operátort", $P_0: H^0(\Omega) \to S_h$ az ortogonális projektort. Ekkor tetszőleges $W \in H^{l+2}(\Omega) \cap$ dom L esetén

$$\|W - P_0 W\| \leqq \|W - P_l^h W\| \leqq Ch^{l+2}\|W\|_{l+2} \tag{5.10}$$

ugyanis

$$\|W - P_l^h W\| = \|\tilde{T}LW - T_n LW\| \leqq Ch^{l+2}\|LW\|_l$$

és mivel $L: H^{l+2}(\Omega) \to H^l(\Omega)$ korlátos operátor, ezért

$$\|LW\|_l \leqq \tilde{C}\|W\|_{l+2}, \tag{5.11}$$

amiből (5.10) közvetlenül adódik.

Tegyük fel, hogy az (5.1)—(5.4) feladat kezdeti feltételét leíró $u_0(x)$ függvény $H^r(\Omega)$-beli. Ekkor a feladat $u^*(x, t)$ megoldása rögzített t esetén $H^r(\Omega)$-beli és maga a feladat is korrekt kitűzésű [10], [18], azaz létezik olyan C pozitív állandó, hogy

$$\|u^*(\cdot, t)\|_r \leqq C\|u_0\|_r. \tag{5.12}$$

Jelölje (mint az előző szakaszban)

$$u_n^*(x, t) = P_l^h u^*(x, t).$$

Ekkor

$$\|u^*(\cdot, t)-u_n^*(\cdot, t)\| = \|\tilde{T}Lu^* - T_nLu^*\| \leqq Ch^{l+2}\|Lu^*\|_l \leqq \tilde{C}h^{l+2}\|u^*\|_{l+2}$$

(az utóbbi becslés (5.11) következtében). Így $l=r-2$ esetén, (5.12) figyelembevételével:

$$\|u^*(\cdot, t)-u_n^*(\cdot, t)\| \leqq Ch^r\|u_0\|_r \quad 0 \leqq t \leqq T. \tag{5.13}$$

(Vegyük észre, hogy (5.13) lényegében megegyezik az előző szakasz 4.1 állításával.)

Az u_n^* térbeli változókban diszkretizált közelítés deriváltjaira is nyerhető becslés.

5.1 ÁLLÍTÁS ([4]). Tegyük fel, hogy $\{T_n(t)\}$ kielégíti az F_3 feltételt. Legyen $u_n^*(x, t)$ rögzített $0 \leqq t \leqq T$ esetén $H^{l+2}(\Omega) \cap \text{dom } L$-beli elem az (5.1)—(5.4) feladat megoldásának elliptikus projekciója S_h-ban. Ekkor létezik olyan $C(j)$ pozitív állandó, hogy

$$\|D_t^j u^*(\cdot, t)-D_t^j u_n^*(\cdot, t)\| \leqq C(j)h^{l+2}\|u_0\|_{l+2+2j}; \quad 0 \leqq t \leqq T; \quad j \geqq 0. \tag{5.14}$$

(Vegyük észre, hogy (5.14) és 4.2. állítás becslésének nagyságrendjét biztosítja.)

A térbeli diszkretizációs operátorra esetenként a következő feltételt kötjük ki.

F_4 feltétel. Legyenek $\{T_n(t)\}$, $\{L_n(t)\}$ olyanok, hogy $j \geqq 0$ egész számhoz létezik olyan pozitív $C(j)$ állandó, hogy

$$\begin{aligned} \|L_n^j(t)T_n(s)f\| &\leqq C(j)\|f\| \quad \forall f \in H^0(\Omega); \quad 0 \leqq s \leqq T, \\ \|T_n(s)L_n^j f_h\| &\leqq C(j)\|f_h\| \quad \forall f_h \in S_h; \quad 0 \leqq s \leqq T. \end{aligned} \tag{5.15}$$

Megmutatható, hogy a (4.13a) típusú *Galjorkin-eljárás* operátora kielégíti az F_4 feltételt [4], [12].

A térbeli változók szerinti approximáció tehát azon $\{u_n^*\}_{0 \leqq t \leqq T} \subset S_n$ sorozat meghatározásából áll, amelyre

$$\begin{aligned} D_t u_n^* + L_n(t)u_n^* &= 0 \quad 0 < t \leqq T \\ u_n^*(0) &= u_{h,0} \end{aligned} \tag{5.16}$$

ahol $u_{h,0} \in S_h$ az u_0 függvény egy megfelelően jó S_h-beli közelítése. Általában

$$u_{h,0} = P_0 u_0 \tag{5.17}$$

azaz $u_{h,0}$ az u_0 S_h-beli ortogonális vetülete. Az $\{u_n^*\}$ approximációs sorozat $u^*(x, t)$ pontos megoldástól való eltérésére az 5.1. állítás becslése érvényes.

Térjünk át az (5.16) séma időváltozó szerinti diszkretizációjára. Ezt — az előzőeknek megfelelően — az $e^{-\tau}$ racionális approximációjával definiáljuk. A továbbiakban legyen

$$r(\tau) = P(\tau)(Q(\tau))^{-1}; \quad P(\tau) = \sum_{l=0}^{v} p_l \tau^l; \quad Q(\tau) = \sum_{l=0}^{v} q_l \tau^l$$

az $e^{-\tau}$ racionális approximációja, ahol P és Q relatív prím polinomok; továbbá kielégítik az alábbi feltételeket:

F_5 feltétel.

a) $Q(\tau)>0$ $\tau>0$ esetén és $Q(0)=1$.

b) $-1+\delta\leqq P(\tau)(Q(\tau))^{-1}<1$ valamely $\delta>0$ és minden τ esetén.

c) $|r(\tau)-e^{-\tau}|=O(\tau^{\nu+1})$; $\nu\geqq 1$ $(\tau\to 0)$

Az általánosság megszorítása nélkül legyen $p_0=\alpha_0=1$. Vegyük észre, hogy a 4. szakaszban definiált sémák nagy része eleget tesz F_5 feltételnek. Például

— (4.44) *implicit séma* $(\delta=1;\ \nu=1)$,

— (4.46) *Padé-típusú approximációk* $(\delta>0;\ \nu=2, 3, 4)$,

— *Calahan-séma* $(\delta>0;\ \nu=3)$ stb.

Megjegyezzük, hogy a (4.40) explicit sémára és a (4.45) *Crank—Nicolson sémára* F_5 feltétel $\nu=1, 2$; $\delta=0$-val érvényes, így a (b) feltétel nem teljesül, azaz a továbbiakban ezeket a sémákat nem tárgyaljuk.

Jelölje $k\,(0<k<1)$ továbbra is az idő szerinti diszkretizációs lépést. Először az $u_n^*(x, k)$ (azaz a térbeli változókban diszkretizált megoldás első időrétegen való) közelítését határozzuk meg. Ehhez a *kétpontos általánosított Taylor-formula* kiterjesztését alkalmazzuk. [12]

5.1 LEMMA. Legyen $g(t)$ megfelelően sima függvény a $[0, T]$ zárt intervallumon; P és Q F_5 feltételt kielégítő ν-edrendű polinomok. Ekkor tetszőleges $t\in[0, T]$ esetén:

$$(Q(-tD_t)g)(t)=(P(-tD_t)g)(0)+\int_0^t K(t, s)D_t^{\nu+1}g(s)\,ds, \tag{5.18}$$

ahol D_t a $[0, T]$-n értelmezett differenciáloperátor és

$$K(t, s)=\sum_{j=0}^{\nu}\frac{q_j}{(\nu-j)!}(-t)^j(t-s)^{\nu-j}. \tag{5.19}$$

Vegyük észre, hogy (5.16) alapján

$$\begin{aligned} D_t u_n^* &= -L_n(t)u_n^*, \\ D_t^2 u_n^* &= \left(L_n^2(t)-D_t L_n(t)\right)u_n^*. \end{aligned} \tag{5.20}$$

Az (5.18) összefüggést a $g(t)=u_n^*(x, s)(0\leqq s\leqq k)$ függvényre alkalmazva, az (5.20) figyelembevételével a következőt nyerjük:

$$\begin{aligned} &\left\{I+q_1kL_n(k)+q_2k^2\left(L_n^2(k)-D_tL_n(k)\right)\right\}u_n^*(x, k)= \\ &=\left\{I+p_1kL_n(0)+p_2k^2\left(L_n^2(0)-D_tL_n(0)\right)\right\}u_n^*(x, 0)+O(k^{\nu+1}D_t^{\nu+1}u_n^*). \end{aligned} \tag{5.21}$$

Feltéve, hogy (5.21) bal oldalán szereplő operátor invertálható, az $u_n^*(x, k)$ közelítését (5.21) alapján a következő módon határozhatjuk meg:

$$u_n^*(x, k)\sim\{I+q_1kL_n+q_2k^2(L_n^2-D_tL_n)\}^{-1}(k)\{I+p_1kL_n+p_2k^2(L_n^2-D_tL_n)\}(0)u_{h,0}. \tag{5.22}$$

Az (5.22) egy olyan egylépéses iterációs eljárást definiál, amelynek segítségével az előző időréteg eredményéből $u_n^*(x, mk)$ tetszőleges időrétegen való közelítését meg tudjuk határozni. Tehát az (5.22) algoritmus teljes definiálásához P, Q, $u_{h,0}$ és $\{D_t^j L_n(t)\}_{j=0,1}$ ismerete szükséges.

Az egyszerűbb jelölés céljából vezettük be a következő jelöléseket:

$$t_m = mk \ \{m = 0, 1 \dots \bar{M}\}; \quad D_t^j L_n(t_m) = L_m^{(j)}; \quad D_t^j T_n(t_m) = T_m^{(j)} \quad (j \geqq 0),$$

(5.23)

$$P_m = P(kL_m); \quad Q_m = Q(kL_m),$$

$$\tilde{P}_m = P_m - p_2 k^2 L_m^{(1)}; \quad \tilde{Q}_m = Q_m - q_2 k^2 L_m^{(1)}.$$

Ekkor, ha $u_n^*(x, k)$ közelítését V^m-vel jelöljük, akkor (5.22) alapján

$$\tilde{Q}_1 V^1 = \tilde{P}_0 u_{h,0} \tag{5.24}$$

és általában felírható.

$$\tilde{Q}_{m+1} V^{m+1} = \tilde{P}_m V^m \tag{5.25}$$

alakban.

Mint az (5.24)—(5.25) képletekből is látható, az algoritmus működésének alapfeltétele a $\tilde{Q}_m$; $(m=1, 2, \dots \bar{M})$ $H(\Omega) \to S_h$ operátorok invertálhatósága. A következő állítás ezt biztosítja.

5.2 ÁLLÍTÁS ([4]). Megfelelően kicsiny k időszerinti diszkretizációs lépésköz esetén léteznek olyan C_1, C_2 pozitív állandók, hogy

$$C_1(Q_m\varphi, \varphi) \leqq (\tilde{Q}_m\varphi, \varphi) \leqq C_2(Q_m\varphi, \varphi) \quad \forall \varphi \in S_h. \tag{5.26}$$

KÖVETKEZMÉNY. Miután Q_m szigorúan pozitív definit operátorok, így (5.26) összefüggés alapján $\tilde{Q}_m$ operátorok invertálhatók $m=1, 2, \dots, \bar{M}$ esetén.

Térjünk át az (5.24), (5.25) (most már bizonyítottan működő) algoritmus eredményeinek hibabecslésére, pontosabban az algoritmus által szolgáltatott közelítésnek és az (5.1)—(5.4) feladat pontos megoldásának a t_m időrétegen való elliptikus projekciójának eltérének becslésére. (Vegyük észre, hogy mindkettő (tehát V^m és $u_n^*(x, t_m)$) is S_h-beli.) Ha $V^m - u_n^*(x, t_m)$-re becslést adunk, akkor az 5.1 állítás segítségével globális hibabecslés is adható. Ez utóbbira ad választ a következő

5.3 ÁLLÍTÁS ([4]). Tegyük fel, hogy $u_0 \in H^\mu(\Omega)$; $\mu = \max\{2(v+1), r+2\}$ olyan, hogy az (5.1)—(5.4) feladat megoldására fennállnak az

$$\|u^*(\cdot, t)\|_{r+2} \leqq C\|u_0\|_{r+2},$$

$$\|D_t^{v+1} u^*(\cdot, t)\| \leqq C\|u_0\|_{2(v+1)} \tag{5.27}$$

összefüggések (C = pozitív állandó); valamint, ha $Q(x)$ másodfokú akkor $\{T_n(t)\}_{0 \leqq t \leqq T}$-re F_4 feltétel teljesül. Ekkor megfelelően kicsiny k esetén

$$\left\|Q_m^{1/2}(V^m - u^*(\cdot, t_m))\right\| \leqq C(h^r + k^v)\|u_0\|_\mu + C\|Q_0^{1/2}(V^0 - P_I u_0)\| \quad m = 1, 2 \dots \bar{M}. \tag{5.28}$$

Az (5.28) becslés (ami a $H^0(\Omega)$-beli becslésnél erősebb becslést jelent) minőségét tehát u_0 simasága és a $V^0 = u_{h,0}$ megválasztása határozza meg. A módszert jellemzően ez utóbbi a lényeges (mivel u_0 a' pirori adott). Célunk V^0 megválasztására olyan eljárást adni, amely $Q_0^{1/2}$ normában nem rontja el az (5.28) becslést, azaz $V^0 - P_I u_0$ elem $Q_0^{1/2}$ normában $O(h^r)$ nagyságrendű. (Megjegyezzük, hogy az optimális $V^0 = P_I u_0$ megválasztás csak elméleti lehetőség: P_I a T_n operátor ismeretét igényli, míg a számítási algoritmusunkban erre nincs szükség.)

5.4 ÁLLÍTÁS ([4]). Ha $u_0 \in H^{r+2}(\Omega) \cap \operatorname{dom} L$ és $L(0)u_0 \in \operatorname{dom} L$, akkor a

$$\begin{aligned} \tilde{Q}_0 V^{0,1} &= P(u_0 + q_1 kL(0)u_0 + q_2 k^2 (L^2(0) - D_t L^1(0))u_0), \\ Q_0 V^{0,2} &= P(u_0 + q_1 kL(0)u_0 + q_2 k^2 L^2(0)u_0) \equiv P(Q(kL(0))u_0) \end{aligned} \tag{5.29}$$

módon meghatározott $V^{0,1}$, $V^{0,2}$ elemekre érvényes a

$$\|Q_0^{1/2}(V^{0,j} - P_I u_0)\| \leqq Ch^r \|u_0\|_{r+2} \quad j = 1, 2 \tag{5.30}$$

becslés.

Mint látható, az (5.29)—(5.24)—(5.25) számítási eljárás realizálása időrétegenként egy lineáris algebrai egyenletrendszer megoldását igényli. Ezen rendszerek mátrixai lépésről lépésre változnak. Így a módszer meglehetősen munkaigényes. Célszerűnek látszik egy olyan iterációs eljárást definiálni, amely ezeket a lineáris egyenletrendszereket csak közelítőleg oldja meg, de a számítási munka csökken és a rendszerek közelítő megoldásából eredő hiba nem haladja meg a módszer (5.28) eredeti hibáját. Mindezek alapján (5.25) megoldásához valamilyen prekondicinált iterációs módszert (PIM) alkalmazunk, amelynek lényege a következő.

Tekintsük $H^{(N)}$ véges dimenziós *Hilbert-térben* adott $y \in H^{(N)}$ esetén az

$$Gx = y \tag{5.31}$$

egyenletet, ahol $G: H^{(N)} \to H^{(N)}$ pozitív definit, önadjungált operátor. Legyen $G_0: H^{(N)} \to H^{(N)}$ pozitív definit, önadjungált és könnyen invertálható operátor. A PIM lényege, hogy adott x^0 (az (5.31) egyenlet megoldásának egy közelítése) esetén generál egy olyan $\{x^{(i)}\}$ $(i \geqq 1)$ sorozatot, amely kielégíti a következő feltételeket:

F_6 feltétel:

a) $x^{(i+1)}$ meghatározásához $\{x^{(j)}\}_{j=0,i}$ ismerete, ezek G operátorra történő alkalmazása, valamint egy G_0 operátort tartalmazó lineáris feladat megoldása szükséges;
b) $x^{(i)} \to x$ egy olyan $\gamma(\xi)$ hányadosú geometriai sorozat gyorsaságával, amelyre

$$0 \leqq \gamma(\xi) < 1 \quad (\xi \in (0, 1]); \quad \gamma(1) = 1$$

c) az i-ik közelítésre érvényes az

$$\|G_0^{1/2}(x - x^{(i)})\|_{H^{(N)}} \leqq C\gamma^i \left(\frac{\lambda_0}{\lambda_1}\right) \|G_0(x - x^{(0)})\|_{H^{(N)}} \tag{5.32}$$

becslés, ahol λ_0, λ_1 a G operátor G_0 operátorra vonatkozó spektrumhatárai energetikai normában, azaz

$$\lambda_0 (G_0 z, z)_{H^{(N)}} \leqq (Gz, z)_{H^{(N)}} \leqq \lambda_1 (G_0 z, z)_{H^{(N)}} \quad \forall z \in H^{(N)}. \tag{5.33}$$

Ismeretesek az F_6 feltételt kielégítő iterációs eljárások, például a

$$G_0 x^{(i+1)} = \mu y + (G_0 - \mu G) x^{(i)} \tag{5.34}$$

egylépéses iteráció, amely a $\mu = 2(\lambda_0 + \lambda_1)^{-1}$ alakú megválasztás esetén a $\gamma(\xi) = (1 - \xi)(1 + \xi)^{-1}$ alakú hányados függvénnyel kielégíti az F_6 feltételt. [21]

Hasonlóan jó a konjugált gradiens módszer a

$$\gamma(\xi) = (1-\sqrt{\xi})(1+\sqrt{\xi})^{-1}$$

függvénnyel [5]. Ez utóbbi módszer lényeges előnye, hogy a gyakorlati alkalmazása során nincs szükség a λ_0, λ_1-et meghatározó (5.33) spektrum-becslésre.

A továbbiakban a PIM-t az (5.25) feladat megoldásához kívánjuk alkalmazni. Alapvető probléma a G_0 prekondicionált operátor megválasztása. Ezért térjünk át ezen kérdésre. Mint az (5.23)-ból is látszik, a megoldandó (5.25) lineáris egyenletrendszer operátorának alakja $Q(x)$ polinom megválasztásától függ. Így természetes módon G_0 prekondicionált operátort is $Q(x)$ alakjából határozzuk meg. Vizsgáljuk meg $Q(x)$ másodfokú polinom alakjáig bezárólag a különböző lehetséges eseteket.

a) $Q(x)=1+q_1x$ elsőfokú polinom. Az

$$G_{0,1} = I+q_1kL_0 \tag{5.35}$$

prekondicionált operátor megválasztása célszerűnek látszik, mivel $G_{0,1}$ struktúrája lényegében megegyezik L_0 operátoréval és ez utóbbi lényegében egy, a [8]-ban már részleteiben tárgyalt elliptikus típusú feladat megoldását jelenti.

b) $Q(x)=(1+\lambda x)^2$ teljes négyzet (pl. a *Calahan-módszer*). Ebben az esetben (5.35)-höz hasonlóan az

$$G_{0,2} = (I+k\lambda L_0)^2 \tag{5.36}$$

prekondicionált operátor megválasztása a célszerű, mivel ennek invertálása két, az a) lineáris esetben már említett elliptikus feladat egymás utáni alkalmazását jelenti.

c) $Q(x)$ nem teljes négyzet (pl. a (4.46) típusú *Padé-approximációk*). Ebben az esetben $Q(x)$ két elsőfokú, általában komplex együtthatós tényezők szorzatára bontható és az így nyert $G_{0,3}$ operátort alkalmazzuk prekondicionált operátorként. Megjegyezzük, hogy ebben az esetben is (a b) esethez hasonlóan) két, elliptikus típusú feladat egymás utáni megoldása szükséges, de a számítások elvégzéséhez a komplex aritmetika szükséges.

Megmutatható [4], hogy a $G_0 \equiv G_{0,j}$ ($j=1, 2, 3$) megválasztású PIM-ekre érvényes az (5.33) becslés, azaz léteznek olyan C_1, C_2 pozitív állandók, hogy

(5.37)

$$C_1(G_{0,j}\varphi, \varphi) \leqq (\tilde{Q}_m\varphi, \varphi) \leqq C_2(G_{0,j}\varphi, \varphi) \quad \forall \varphi \in S_h; \quad m = 1, 2, \ldots \overline{M}, \quad j = 1, 2, 3$$

(ahol $\overline{M}$ olyan pozitív állandó, hogy $\overline{M}k \leqq T$; $(\overline{M}+1)k > T$.)

Jelölje az (5.25) feladat PIM-es megoldásnál a $t_m = mk$-ik időrétegen az iterációk számát δ_m; a δ_m és $(\delta_m - 1)$-ik iterált eredmények eltérését β_m.

(Nyilvánvaló, hogy δ_m a β_m-től függ.) Ekkor, ha az elméletnek megfelelően az F_6 feltételt kielégítő PIM-t választunk az (5.37)-t kielégítő prekondicionált operátorral, akkor egy előre rögzített (β_m) hibasorozathoz létezik olyan (δ_m) lépésszámsorozat, hogy az m-ik időrétegen (azaz a $t_m = mk$ időpontban) $(0 \leqq m \leqq \overline{M})\delta_m$ számú iteráció elvégzése után nyert $U^{(\delta_m)} \subset S_h$ közelítésre:

$$\begin{aligned} &\|U^{(\delta_m)} - U^{(\delta_m-1)}\| \leqq \beta_m \quad 0 \leqq m \leqq \overline{M}, \\ &\|G_0^{1/2}(V^m - U^{(\delta_m)})\| \leqq \beta_m \|G_0^{1/2}(V^0 - U_m^{(0)})\| \end{aligned} \tag{5.38}$$

ahol $U_m^{(0)}$ a PIM alkalmazásához szükséges első közelítés, G_0 a $Q(x)$ polinom megválasztásától függően a $G_{0,j}$ $(j=1, 2, 3)$ prekondicionált operátorok egyike, V^m az (5.25) feladat pontos megoldása. (Megjegyezzük, hogy az egy időrétegre számított átlagos iterációs lépésszámra $\left(\frac{1}{\bar{M}} \sum_{m=1}^{\bar{M}} \delta_m\right)$ is megadható felső becslés [1].

Mint az a PIM leírásából és hibabecsléseinek kifejezéseiből is látható, a módszer realizálásának és minőségének alapvető kérdése az iterációs módszer időrétegenkénti kezdeti közelítéseinek megválasztása. Erre vonatkozóan definiáljunk egy megválasztási stratégiát, ahol $Z_{m+1}^{(\nu)}$-vel jelöljük a ν-ed rendű diszkretizáció (5.25) egyenletére alkalmazott PIM kezdeti közelítését az $(m+1)$-ik időrétegen.

$$
\begin{aligned}
&Z_{m+1}^{(1)} = U^m \quad 0 \leqq m \leqq \bar{M},\\
&Z_{m+1}^{(2)} = 2U^m - U^{m-1} \quad 1 \leqq m \leqq \bar{M},\\
&Z_{m+1}^{(3)} = 3U^m - 3U^{m-1} + U^{m-2} \quad 2 \leqq m \leqq \bar{M},\\
&Z_{m+1}^{(4)} = 4U^m - 6U^{m-1} + 4U^{m-2} - U^{m-3} \quad 3 \leqq m \leqq \bar{M}.
\end{aligned} \tag{5.39}
$$

(A gyakorlatban ennyi elegendő, hiszen negyedrendűnél magasabb pontosságú sémával nem szokásos számolni. Ha azonban mégis szükséges, akkor (5.39) felépítéséből (a binominális együtthatókkal való összefüggéséből) könnyen megadhatók egyéb közelítések.)

Mindezeket összefoglalva legyen PIMG a következő egylépéses, prekondicionált iterációs módszer.

a) A térbeli diszkretizáció után válasszuk meg az időbeli diszkretizációs sémát definiáló $r(\tau)$ F_5 feltételt kielégítő racionális polinomot. Ezzel P, Q polinomok és ν (a diszkretizációs pontossága) ismertekké válnak.
b) $Q(x)$ megválasztásának függvényében definiáljuk G_0 prekondicionált operátort, mint a $G_{0,j}$ $(j=1, 2, 3)$ egyikét.
c) Megválasztunk egy PIM-t.
d) Időrétegenként felépítjük az (5.25) lineáris egyenletrendszert.
e) Meghatározzuk U^0 értékét az (5.29)-ből kiszámított $V^{0,1}$ vagy $V^{0,2}$ egyikeként.
f) Az $m+1=1, 2, \ldots \nu-1$ értékekre a c)-ben a megválasztott egylépéses PIM-rel, $Z_{m+1}^{(\nu)}=U^m$ kezdeti közelítéssel; $\beta_{m+1}=k^\nu$ hibakorláttal meghatározzuk U^{m+1} értékeit.
g) Az $m+1=\nu, \nu+1, \ldots \bar{M}$ értékekre a c)-ben megválasztott egylépéses PIM-rel, (5.39) szerinti kezdeti közelítéssel $\beta_{m+1}=k$ hibakorláttal meghatározzuk U^{m+1} értékeit.

Az íly módon meghatározott $U^m (0 \leqq m \leqq \bar{M})$ megoldásnak az (5.1)—(5.4) feladat $u^*(x, t)$ megoldásának $t_m=mk$ helyen felvett értékétől való eltérésére ad becslést a következő állítás [4], amely egyben azt is bizonyítja, hogy a PIMG algoritmus alkalmazásával az eredmény pontosságára vonatkozó hibabecslés nagyságrendje nem változik.

5.5 ÁLLÍTÁS. Legyen $u_0 \in H^\mu(\Omega)$, $\mu=\max\{r+2, 2(\nu+1)\}$, k megfelelően kicsiny; valamint tegyük fel, hogy F_4 feltétel teljesül, ha $Q(x)$ másodfokú. Ekkor, ha U^m a PIMG algoritmus alkalmazásával nyert közelítés a $t_m=mk$ időrétegen, akkor létezik olyan C pozitív állandó, hogy

$$
\|Q_m^{1/2}(u^*(\cdot, mk) - U^m\| \leqq C(h^r + k^\nu)\|u_0\|_\mu \quad 0 \leqq m \leqq \bar{M}. \tag{5.40}
$$

5.1. Megjegyzés. Ebben a szakaszban terjedelmi okokból nem térünk ki részletesen a módszer stabilitásával összefüggő kérdésekre. Szükségesnek tartjuk azonban megjegyezni, hogy a 4. szakaszban targyalt eredmények nem vihetők át automatikusan az időfüggő elliptikus részű problémákra. Például, az időtől független elliptikus operátorú parabolikus feladatokra az *Euler-módszer* feltétel nélkül stabil, míg az időfüggő elliptikus operátorú parabolikus feladatok esetén feltételesen stabillá válhatnak, ha az elliptikus operátorának fő része az idő szerinti deriválásnál nem tűnik el ([11]).

A cikkben alkalmazott főbb jelölések jegyzéke

$\mathbf{R}$, $\mathbf{R}^1$	a valós számok halmaza
$\mathbf{R}^N$	az N dimenziós tér
Ω	$\mathbf{R}^N$-beli korlátos és zárt tartomány
Γ	az Ω tartomány elegendően sima pereme
$\bar{\Omega}$	az Ω tartomány lezárása
$[0, T]$	$\mathbf{R}$-beli korlátos halmaz, a t időváltozó változási tartománya
$H(\Omega)$	az Ω-n értelmezett függvények (absztrakt) *Hilbert-tere*
$H(\Omega\times(0, T))$	az $\Omega\times[0, T]$-n értelmezett függvények (absztrakt) *Hilbert-tere*
$((\,,\,))$	a $H(\Omega\times(0, T))$-n értelmezett skaláris szorzat
$H^S(\Omega)$	az $(u, v)_S=\sum_{\lvert\alpha\rvert\leqq s}\int_{\Omega}(D^\alpha u)(D^\alpha v)\,d\Omega$ skaláris szorzatú *Hilbert-tér (Szoboljev-tér)*
$\mathring{H}^S(\Omega)$	$H^S(\Omega)$ olyan altere, amelynek elemeire $L^j W\lvert_\Gamma=0$ $(j\leqq[s/2])$
$H^0(\Omega)$	az Ω-n négyzetesen integrálható függvények tere
$(\,,\,)$	a $H^0(\Omega)$-beli skaláris szorzat
$\|\ \|$	a $(\,,\,)$ skaláris szorzat által indukált norma
H_L	az L szigorúan pozitív definit operátor energetikai tere
$[\,,\,]_L$	a H_L-beli skaláris szorzat
$H^0(\Omega\times(0, T))$	az $\bar{\Omega}\times[0, T]$-n értelmezett, négyzetesen integrálható függvények tere
$H^1(\Omega\times(0, T))$	az $\bar{\Omega}\times[0, T]$-n értelmezett függvények *Szoboljev-tere*
V	$H(\Omega\times(0, T))$-beli sűrű altér
H_E	a (2.6) által definiált, $H(\Omega\times(0, T))$-ben sűrű altér
$H_{L\times T}$	$H^0(\Omega\times(0, T))$ olyan altere, amelynek elemei rögzített $x\in\Omega$ esetén $H^0(0, T)$; rögzített $t\in(0, T]$ esetén pedig H_L-beliek
$\tilde{H}_{L\times T}$	$H_{L\times T}$ olyan altere, amely elemeinek t szerinti parciális deriváltjai $H^0(\Omega\times(0, T))$-beliek
$\tilde{V}_0$	$H^1(\Omega\times(0, T))$ olyan altere, amely elemei a $t=T$ pontban nullák és tetszőleges, rögzített $t\in[0, T]$ pontban kielégítik a dom L fő peremfeltételeket
$\tilde{V}^*$	$\tilde{V}_0$ olyan altere, amely elemeinek t szerinti parciális deriváltjai $H^0(\Omega\times(0, T))$-beliek valamint az elemei előállíthatók $\alpha(t)\cdot\varphi(x)$ alakban
V_n	H_L-beli véges dimenziós altérsorozat
$E_n(t)$	$\tilde{H}_{L\times T}$-beli tetszőleges rögzített $t\in[0, T]$ esetén véges dimenziós altérsorozat

S_h	$H^0(\Omega)$-beli véges dimenziós altérsorozat
$(\varphi_i(x))$	egy rögzített spline-tér bázisfüggvényrendszere
$\mathscr{L}[\varphi_1 \ldots \varphi_n]$	a $\varphi_1(x) \ldots \varphi_n(x)$ elemek állandó együtthatós lineáris burka
$\mathscr{L}(t)[\varphi_1 \ldots \varphi_n]$	a $\varphi_1(x), \ldots \varphi_n(x)$ elemek t-től függő együttható függvényekkel képzett lineáris burka
$\Phi(u, v)$	a (4.12)-ben definiált bilineáris funkcionál
$\Phi(t)(u, v)$	az (5.5)-ben definiált, időtől függő bilineáris funkcionál
L	$H(\Omega) \to H(\Omega)$ képező, csak x-től függő lineáris, szigorúan pozitív definit operátor
dom L	az L operátor értelmezési tartománya
$L(t)$	tetszőleges rögzített $t \in [0, T]$ esetén L operátor tulajdonságával rendelkezik
$\tilde{T}, \tilde{T}(t)$	a megfelelő elliptikus feladat megoldási operátora (az L operátor inverze)
$T_n, T_n(t)$	a feladat térbeli diszkretizációját jelentő operátor (valamely diszkretizációs módszert reprezentálja)
$L_n, L_n(t)$	a $T_n, T_n(t)$ operátorok inverzei
$P_I^h(t)$	a $P_I^h(t) \equiv T_n(t) L_n(t)$ módon definiált, dom $L \to S_n$ képező projekciós operátor
P_0	a $H^0(\Omega) \to S_n$ képező ortogonális projektor
M, Q	a (3.5) módon definiált szimmetrikus, szigorúan pozitív definit mátrixok
k	a $[0, T]$ intervallum diszkretizációjának lépésköze
$\bar{M}$	olyan pozitív egész szám, amelyre $\bar{M}k \leqq T$; $(\bar{M}+1)k > T$
h	az Ω tartomány felbontását jellemző kis paraméter
$\vec{n}$	a Γ perem külső normálisa
Re z	a z komplex szám valós része
$u^*(x, t)$	a megoldandó feladat általánosított megoldása
$u_n^*(x, t)$	az u^* megoldás n-ik közelítése a térbeli diszkretizáció során
$e_n(x, t)$	a hibafüggvény; az u^* és az u_n^* eltérése
$e_h(t)$	az e_n függvény jelölése valamely $H(\Omega)$ normában
$\|W\|$	a $W(x, t)$ függvény térbeli változójára értelmezett maximum norma
$\|w\|_S$	$\sup_{\substack{x \in \Omega \\ 0 < k \leqq S}} \|D^k w\|$ módon definiált maximum-norma
V^m	az $u_n^*(x, t)$ közelítése a $t = m \cdot k$ helyen

IRODALOM

[1] AXELSSON, O., "On Preconditioning an Convergence Acceleration in Sparse Matrix Problems", CERN (European Organization for Nuclear Research), Geneva, 1974.

[2] BAKER, G. A., BRAMBLE, J. H. and THOMEE, V., "Single Step Galerkin Approximations for Parabolic Problems", *Math. Comp.* **31** (1977).

[3] BRAMBLE, J. H., SCHATZ, A. H., THOMEE, V. and WAHLBIN, L. B., "Some Convergence Estimates for semidiscrete Galerkin Type Approximations for Parabolic Equations", *SIAM J. Num. Anal.* **14** (1977).

[4] BRAMBLE, J. H. and SAMMON, P. H., "Efficient Higher Order Single Step Methods for Parabolic Problems", *Math. Comp.* **35** (1980).

[5] DANIEL, J. W., "The Conjugate Gradient Method for Linear and Nonlinear Operator Equations", *SIAM J. Num. Anal.* **4** (1967).

[6] FARAGÓ, I., „Parciális differenciálegyenletek megoldása véges elemek módszerével" Intézeti Tájékoztató 3, MüM SZÁMTI, 1979.
[7] FARAGÓ, I., „A transzport-jelenségek matematikai modellezése és numerikus vizsgálatának kérdései", Intézeti Tájékoztató 4, MüM SZÁMTI, 1980.
[8] FARAGÓ, I., „Véges elemek módszere elliptikus típusú feladatok megoldására" *Alk. Mat. Lapok.* **8** 1982.
[9] FARAGÓ, I. és GÁSPÁR, Cs., „Parciális differenciálegyenletek megoldásának numerikus módszerei hidrodinamikai alkalmazásokkal", BME, MTI, 1983.
[10] FRIEDMANN, A., *Partial Differential Equations* (Krieger, Huntington, New York, 1976).
[11] GEKELER, E., "A-priori Error Estimates of Galerkin Backward Differentiation Methods in Time-Inhomogeneous Parabolic Problems ", *Numer. Math.* **30** (1978).
[12] NASSIF, N. and DESCLOUX, J., "Stability study for time-dependent linear parabolic eauations" *Tipics in Numerical Analysis III.* (J. Miller Ed.) Academic Press, New York, 1977.
[13] NITSCHE, J. A., "L_{00}-convergence for Finite Element Approximation" 2. Conf. on Finite Elements, Rennes, 1975.
[14] SCOTT, R., "Optimal L_{00}-estimates for the Finite Method on Irregular Meshes." *SIAM J. Numer. Anal.* **15** (1978).
[15] STRANG, G. and FIX, G., *An Analysis of the Finite Element Method* (Prentice Hall Inc. Englewood Cliffs, N. Yersey, 1973).
[16] ZLAMAL, M., "Finite Element Multistep Discretizations of Parabolic Boundary Value Problems", *Math. Comp.* **29** (1975).
[17] ZLATEV, Z. and THOMSEN, P. G., "Application of Backward Differentiation Methods to the Finite Element Solution of Time-Dependent Problems", *Int. J. for Num. Meth. in. Eng.* **14** (1979).
[18] Ладыженская, О. А., Уральцева, Н. Н., *Линейные и нелинейные уравнения параболического типа* (Москва, Наука, 1967).
[19] Марчук, Г. И., Агошков, В. И., *Введение в проекционно-сеточные методы* (Москва, Наука, 1981).
[20] Михлин, С. Г., *Линейные уравнения в частных производных* (Москва, Высшая школа, 1977).
[21] Самарский, А. А., *Теория разностных схем* (Москва, Наука, 1982).

(Beérkezett: 1984. június 13.)

FARAGÓ ISTVÁN
GÖDÖLLŐI AGRÁRTUDOMÁNYI EGYETEM MATEMATIKAI INTÉZET
2103 GÖDÖLLŐ, PÁTER K. U. 1—3.

FINITE ELEMENT METHOD FOR SOLVING LINEAR PARABOLIC PROBLEMS

I. FARAGÓ

The paper presents the application of the finite element method for solving partial differential equations of parabolic type. We introduce the term "generalized solution" for such problems and by means of the semidiscrete *Galerkin type approximation* we give the numerical methods for their solution. After the treatment of the simplest problems there follows the discusion of the most general cases. We derive some error-estimates and define the algorithmic descriptions, which are very important for the computer implementation.

Alkalmazott Matematikai Lapok **11** (1985) 157–165

NÉGYZETES MÁTRIX LU FAKTORIZÁCIÓJÁNAK MÓDOSÍTÁSA DIÁDDAL VÁLTOZTATÁS ESETÉN

BARTALOS ISTVÁN

Szeged

A cikkben megmutatjuk, hogy bizonyos feltételek teljesülése esetén egy diáddal módosított négyzetes mátrix $\tilde{\mathbf{L}}\tilde{\mathbf{U}}$ faktorizációját legfeljebb $2n^2+O(n)$ szorzás és osztás árán megkaphatjuk, ha az eredeti mátrix $\mathbf{LU}$ faktorizációját ismerjük. Speciális esetként foglalkozunk a szimmetrikus diáddal módosított szimmetrikus mátrixok $\mathbf{LDL}^T$ faktorizációjának meghatározásával és megmutatjuk, hogy a módosított $\tilde{\mathbf{L}}\tilde{\mathbf{D}}\tilde{\mathbf{L}}^T$ faktorizáció legfeljebb $n^2+O(n)$ szorzás és osztás árán nyerhető.

0. Bevezetés: az ismert algoritmusok áttekintése

Gyakorlati problémák megoldásánál előfordul, hogy ismerjük egy $\mathbf{A}$ $n\times n$-es mátrix $\mathbf{LU}$ trianguláris faktorizációját és meg szeretnénk határozni az $\mathbf{A}$-tól kissé eltérő $\tilde{\mathbf{A}}$ négyzetes mátrix ugyanilyen típusú faktorizációját. Ez a cikk azzal az esettel foglalkozik, amikor $\tilde{\mathbf{A}}$ az $\mathbf{A}$-ból egyetlen diád hozzáadásával keletkezik, azaz $\tilde{\mathbf{A}}=\mathbf{A}+\mathbf{fg}^T$ alakú. Tárgyalja ezenkívül a szimmetrikus mátrixok $\mathbf{LDL}^T$ faktorizációjának módosítását is.

A szimmetrikus esettel foglalkozó cikkek közül [1] tárgyalja azokat a módszereket, amelyek közös vonása, hogy az $\mathbf{Lp}=\mathbf{f}$ lineáris egyenletrendszer megoldásaként előálló $\mathbf{p}$ vektor segítségével az $\tilde{\mathbf{A}}=\mathbf{A}+\alpha\mathbf{ff}^T=\mathbf{LDL}^T+\alpha\mathbf{ff}^T=\mathbf{L}(\mathbf{D}+\alpha\mathbf{pp}^T)\mathbf{L}^T$ átalakítást és a $\mathbf{D}+\alpha\mathbf{pp}^T=\bar{\mathbf{L}}\bar{\mathbf{D}}\bar{\mathbf{L}}^T$ faktorizációt hajtják végre. A Cl módszer származtatásakor a szerzők rekurzív úton megmutatják, hogy $\bar{\mathbf{L}}$ speciális szerkezetű, majd ezt a tényt felhasználva $n^2+O(n)$ szorzás műveletigényű rekurzív algoritmust definiálnak. A C2 módszer $\mathbf{D}+\alpha\mathbf{pp}^T$ faktorizációjának meghatározásához a $\mathbf{D}^{1/2}\mathbf{v}=\mathbf{p}$ lineáris egyenletrendszer megoldásaként adódó $\mathbf{v}$ vektort használja fel. A $\mathbf{D}+\alpha\mathbf{pp}^T=$ $=\mathbf{D}^{1/2}(\mathbf{I}+\alpha\mathbf{vv}^T)\mathbf{D}^{1/2}$ átalakítás után kihasználja azt a tényt, hogy $\mathbf{I}+\alpha\mathbf{vv}^T=(\mathbf{I}+\sigma\mathbf{vv}^T)\times$ $\times(\mathbf{I}+\sigma\mathbf{vv}^T)$ alakba írható és ezután $\mathbf{H}_i$ *Householder-mátrixokkal* való szorzásokkal állítja elő az $\hat{\mathbf{L}}=(\mathbf{I}+\sigma\mathbf{vv}^T)\mathbf{H}_1\mathbf{H}_2\ldots\mathbf{H}_{n-1}$ alsó trianguláris mátrixot. Az így konstruált algoritmus műveletigénye $3n^2/2+O(n)$ szorzás és $n+1$ négyzetgyökvonás. A C3, C4 és C5 módszerek *Givens-mátrixokkal* való szorzások útján triangularizálnak. Műveletigényük rendre: $5n^2/2+O(n)$ szorzás és $n+1$ négyzetgyökvonás, $9n^2/2+O(n)$ szorzás és $2n-1$ négyzetgyökvonás, illetve $2n^2+O(n)$ szorzás és $2n-1$ négyzetgyökvonás.

A [2]-ben tárgyalt módszerek közös vonása, hogy az $\mathbf{A}=\mathbf{LDL}^T$ faktorizációt n diád összegeként az $\mathbf{A}=\sum_{i=1}^{n} d_i\mathbf{l}_i\mathbf{l}_i^T$ alakban tekintik, ahol $\mathbf{l}_i$ az $\mathbf{L}$ i-edik oszlopát, d_i pedig a $\mathbf{D}$ i-edik főátlós-beli elemét jelöli. Ekkor $\tilde{\mathbf{A}}=\mathbf{A}+\alpha\mathbf{ff}^T=\sum_{i=1}^{n} d_i\mathbf{l}_i\mathbf{l}_i^T+$

$+\alpha\mathbf{f}\mathbf{f}^T$ alakú. Az első lépésben a $d_1\mathbf{l}_1\mathbf{l}_1^T+\alpha\mathbf{f}\mathbf{f}^T=\lambda\mathbf{x}\mathbf{x}^T+\mu\mathbf{y}\mathbf{y}^T$ átalakítást végzik el, ahol $x_1=1$ és $y_1=0$. Ezáltal az $\tilde{\mathbf{A}}=\mathbf{A}+\alpha\mathbf{f}\mathbf{f}^T=\tilde{\mathbf{L}}\tilde{\mathbf{D}}\tilde{\mathbf{L}}^T$ faktorizációból megkapják $\tilde{\mathbf{L}}$ első oszlopát és $\tilde{\mathbf{D}}$ első fődiagonálisbeli elemét. Így az eredeti feladatot eggyel kisebb méretűre vezetik vissza. Ezen módszerek műveletigénye is legalább $n^2+O(n)$ szorzás.

A [3]-ban tárgyalt módszernél az $\tilde{\mathbf{A}}=\mathbf{A}+\alpha\mathbf{f}\mathbf{f}^T$ kifejezést először skálázással az α előjelétől függően $\tilde{\mathbf{A}}=\mathbf{A}+\mathbf{v}\mathbf{v}^T$ vagy $\tilde{\mathbf{A}}=\mathbf{A}-\mathbf{v}\mathbf{v}^T$ alakban írják fel, majd az $\mathbf{L}\mathbf{p}=\mathbf{v}$ lineáris egyenletrendszer $\mathbf{v}$ megoldásvektora segítségével az $\tilde{\mathbf{A}}=\mathbf{L}(\mathbf{D}\pm\mathbf{p}\mathbf{p}^T)\mathbf{L}^T$ átalakítást végzik el. Bebizonyítják, hogy a $\mathbf{D}+\mathbf{p}\mathbf{p}^T=\bar{\mathbf{L}}\bar{\mathbf{D}}\bar{\mathbf{L}}^T$ és a $\mathbf{D}-\mathbf{p}\mathbf{p}^T=\bar{\mathbf{L}}\bar{\mathbf{D}}\bar{\mathbf{L}}^T$ faktorizációk speciális szerkezetűek és az [1]-ben tárgyalt C1 módszerhez hasonló úton konstruálnak legalább $n^2+O(n)$ szorzás műveletigényű módszereket.

[4]-ben BENETT az általános esettel foglalkozik: az $\mathbf{A}=\mathbf{L}\mathbf{D}\mathbf{U}$ faktorizáció ismeretében az $\tilde{\mathbf{A}}=\mathbf{A}+\mathbf{X}\mathbf{C}\mathbf{Y}^T=\tilde{\mathbf{L}}\tilde{\mathbf{D}}\tilde{\mathbf{U}}$ faktorizáció meghatározására ad módszert. Itt $\mathbf{X}$ és $\mathbf{Y}$ $n\times m$-es, $\mathbf{C}$ pedig $m\times m$-es mátrix. A módszer alapja az, hogy $\mathbf{L}$-et az $\mathbf{L}=\mathbf{L}_1\mathbf{L}_2\ldots\mathbf{L}_{n-1}$ és $\mathbf{U}$-t az $\mathbf{U}=\mathbf{U}_{n-1}\ldots\mathbf{U}_2\mathbf{U}_1$ alakú szorzatokra bontja, ahol $\mathbf{L}_i$ az egységmátrix i-edik oszlopának $\mathbf{L}$ i-edik oszlopával, $\mathbf{U}_i$ az egységmátrix i-edik sorának $\mathbf{U}$ i-edik sorával való helyettesítésével keletkezik. Az $\tilde{\mathbf{L}}_1$, $\tilde{\mathbf{d}}_1$ és $\tilde{\mathbf{U}}_1$ meghatározása után az $\tilde{\mathbf{L}}_1^{-1}\tilde{\mathbf{A}}\tilde{\mathbf{U}}_1^{-1}$ előállításával a feladatot eggyel kisebb méretűre vezeti vissza. Műveletigénye $m\ll n$ esetén $2mn^2+O(n)$ szorzás.

[5] összefoglalja a mátrixok faktorizációinak módosítására szolgáló módszereket, de műveletigény tekintetében nem szolgáltat jobb eljárást, mint az eddig felsoroltak.

Ezen cikk kiindulásául [6] szolgál. BERSENEV ebben a cikkében felhasználja azt a tényt, hogy ha $\mathbf{p}=[p_1, p_2, \ldots, p_n]^T$ tetszőleges n-elemű valós vektor és $\mathbf{I}$-vel jelöljük az $n\times n$-es egységmátrixot, akkor az

$$\mathbf{I}+\mathbf{p}\mathbf{p}^T=\mathbf{L}_0\mathbf{D}_0\mathbf{L}_0^T \tag{1.1}$$

összefüggés mindig fennáll, ahol

$$\alpha_0=1$$

$$\alpha_i=1+\sum_{j=1}^{i}p_j^2 \quad (i=1, 2, \ldots, n)$$

$$\mathbf{L}_0=\begin{bmatrix} \alpha_1 & & & 0 \\ p_2p_1 & \alpha_2 & & \\ \vdots & \vdots & \ddots & \\ p_np_1 & p_np_2 & \ldots & \alpha_n \end{bmatrix}$$

$$\mathbf{D}_0=\begin{bmatrix} \dfrac{1}{\alpha_0\alpha_1} & & & 0 \\ & \dfrac{1}{\alpha_1\alpha_2} & & \\ & & \ddots & \\ 0 & & & \dfrac{1}{\alpha_{n-1}\alpha_n} \end{bmatrix}.$$

Ennek segítségével megmutatta, hogy ha valamely $n\times n$-es valós elemű pozitív definit $\mathbf{A}$ mátrixnak ismeretes az $\mathbf{LL}^T$ faktorizációja — ahol $\mathbf{L}$ alsó trianguláris —, akkor bármely n-elemű valós $\mathbf{f}$ vektor esetén az $\tilde{\mathbf{A}}=\mathbf{A}+\mathbf{ff}^T=\tilde{\mathbf{L}}\tilde{\mathbf{L}}^T$ faktorizációja mindig megkapható úgy, hogy megoldjuk az $\mathbf{Lp}=\mathbf{f}$ lineáris egyenletrendszert és alkalmazzuk az (1.1) faktorizációt:

$$\tilde{\mathbf{A}} = \mathbf{A}+\mathbf{ff}^T = \mathbf{L}(\mathbf{I}+\mathbf{pp}^T)\mathbf{L}^T = \mathbf{LL}_0\mathbf{D}_0\mathbf{L}_0^T\mathbf{L}^T = (\mathbf{LL}_0\mathbf{D}_0^{1/2})(\mathbf{LL}_0\mathbf{D}_0^{1/2})^T. \tag{1.2}$$

BERSENEV az $\tilde{\mathbf{L}}=\mathbf{LL}_0\mathbf{D}_0^{1/2}$ kiszámításának összes műveletigényeként $O(n^2)$-et ad meg. A cikk elemzésekor kiderült, hogy a pontosabb műveletigény: n négyzetgyökvonás és $2n^2+O(n)$ szorzás és osztás.

1. Új algoritmusok

Ebben a cikkben először az (1.1) faktorizációt terjesztjük ki nemszimmetrikus esetre. Ennek felhasználásával megmutatjuk, hogy ha ismerjük egy $\mathbf{A}$ négyzetes mátrix $\mathbf{LU}$ faktorizációját — ahol $\mathbf{L}$ olyan alsó trianguláris mátrix, melynek fődiagonálisában csupa 1 áll és $\mathbf{U}$ felső trianguláris mátrix —, akkor adott $\mathbf{f}$ és $\mathbf{g}$ vektorok esetén bizonyos feltételek teljesülésekor az $\tilde{\mathbf{A}}=\mathbf{A}+\mathbf{fg}^T=\tilde{\mathbf{L}}\tilde{\mathbf{U}}$ faktorizációja megkapható legfeljebb $2n^2+O(n)$ szorzás és osztás árán.

Az (1.1) faktorizáció kiterjesztése nemszimmetrikus esetre módot adott arra is, hogy ha ismert az $\mathbf{A}$ szimmetrikus négyzetes mátrix $\mathbf{LDL}^T$ *Cholesky-féle faktorizációja,* akkor az $\tilde{\mathbf{A}}=\mathbf{A}+\alpha\mathbf{ff}^T$-nak az $\tilde{\mathbf{L}}\tilde{\mathbf{D}}\tilde{\mathbf{L}}^T$ *Cholesky-féle faktorizációja* legfeljebb $n^2+O(n)$ szorzás és osztás árán nyerhető legyen. Ez az algoritmus az [1]-ben C1 néven hivatkozott, eddig ismert egyik legjobb műveletigényű módszerrel ekvivalens.

A 2.2 és 3.2 tételek a diáddal módosított mátrixok faktorizációjának egzisztenciájára és unicitására vonatkozóan egyszerű, külön számolást nem igénylő kritériumokat adnak. A 2.2 tétel a nemszimmetrikus, míg a 3.2 tétel a szimmetrikus mátrixok esetét tárgyalja.

2. Nemszimmetrikus eset

Ebben a részben egy diáddal módosított nemszimmetrikus négyzetes mátrix $\mathbf{LU}$ faktorizációjának meghatározására adunk algoritmust. Ennek elméleti alapját a következő tétel szolgáltatja:

2.1. TÉTEL. Ha $\mathbf{p}=[p_1, p_2, \ldots, p_n]^T$ és $\mathbf{q}=[q_1, q_2, \ldots, q_n]^T$ olyan vektorok, amelyekre az

$$\alpha_i = 1+\sum_{j=1}^{i} p_j q_j \quad (j = 1, 2, \ldots, n) \tag{2.1}$$

számok egyike sem nulla, akkor az

$$\mathbf{I}+\mathbf{pq}^T = \mathbf{L}_0\mathbf{D}_0\mathbf{U}_0 \tag{2.2}$$

felbontás mindig létezik, ahol

$$\alpha_0 = 1$$

$$\mathbf{D}_0 = \operatorname{diag}\left[\frac{1}{\alpha_0\alpha_1}, \frac{1}{\alpha_1\alpha_2}, \ldots, \frac{1}{\alpha_{n-1}\alpha_n}\right]$$

$$\mathbf{L}_0 = \begin{bmatrix} \alpha_1 & & & & \\ p_2q_1 & \alpha_2 & & & \\ p_3q_1 & p_3q_2 & \alpha_3 & & 0 \\ \vdots & \vdots & \vdots & \ddots & \\ p_{n-1}q_1 & p_{n-1}q_2 & p_{n-1}q_3 & \cdots & \alpha_{n-1} \\ p_nq_1 & p_nq_2 & p_nq_3 & \cdots & p_nq_{n-1}\ \alpha_n \end{bmatrix}$$

$$\mathbf{U}_0 = \begin{bmatrix} \alpha_1 & p_1q_2 & p_1q_3 & \cdots & p_1q_{n-1} & p_1q_n \\ & \alpha_2 & p_2q_3 & \cdots & p_2q_{n-1} & p_2q_n \\ & & \alpha_3 & \cdots & p_3q_{n-1} & p_3q_n \\ & & & \ddots & \vdots & \vdots \\ & 0 & & & \alpha_{n-1} & p_{n-1}q_n \\ & & & & & \alpha_n \end{bmatrix}.$$

Bizonyítás. Legyen

$$\mathbf{S} = \mathbf{L}_0\mathbf{D}_0\mathbf{U}_0.$$

Elemeit jelöljük s_{ij}-vel $(1 \leqq i, j \leqq n)$

a) $i<j$ eset

$$s_{ij} = \frac{p_iq_1p_1q_j}{\alpha_1\alpha_0} + \frac{p_iq_2p_2q_j}{\alpha_2\alpha_1} + \frac{p_iq_3p_3q_j}{\alpha_3\alpha_2} + \ldots + \frac{p_iq_{i-1}p_{i-1}q_j}{\alpha_{i-1}\alpha_{i-2}} + \frac{p_iq_j}{\alpha_{i-1}} =$$

$$= p_iq_j\left(\frac{\alpha_1-\alpha_0}{\alpha_1\alpha_0} + \frac{\alpha_2-\alpha_1}{\alpha_2\alpha_1} + \frac{\alpha_3-\alpha_2}{\alpha_3\alpha_2} + \ldots + \frac{\alpha_{i-1}-\alpha_{i-2}}{\alpha_{i-1}\alpha_{i-2}} + \frac{1}{\alpha_{i-1}}\right) = p_iq_j.$$

b) $i=j$ eset

$$s_{ii} = \frac{p_iq_1p_1q_i}{\alpha_1\alpha_0} + \frac{p_iq_2p_2q_i}{\alpha_2\alpha_1} + \frac{p_iq_3p_3q_i}{\alpha_3\alpha_2} + \ldots + \frac{p_iq_{i-1}p_{i-1}q_i}{\alpha_{i-1}\alpha_{i-2}} + \frac{\alpha_i}{\alpha_{i-1}} =$$

$$= p_iq_i\left(\frac{\alpha_1-\alpha_0}{\alpha_1\alpha_0} + \frac{\alpha_2-\alpha_1}{\alpha_2\alpha_1} + \frac{\alpha_3-\alpha_2}{\alpha_3\alpha_2} + \ldots + \frac{\alpha_{i-1}-\alpha_{i-2}}{\alpha_{i-1}\alpha_{i-2}}\right) + \frac{\alpha_i}{\alpha_{i-1}} =$$

$$= p_iq_i\left(1 - \frac{1}{\alpha_{i-1}}\right) + 1 + \frac{p_iq_i}{\alpha_{i-1}} = 1 + p_iq_i.$$

c) $i>j$ eset

$$s_{ij} = \frac{p_iq_1p_1q_j}{\alpha_1\alpha_0} + \frac{p_iq_2p_2q_j}{\alpha_2\alpha_1} + \frac{p_iq_3p_3q_j}{\alpha_3\alpha_2} + \ldots + \frac{p_iq_{j-1}p_{j-1}q_j}{\alpha_{j-1}\alpha_{j-2}} + \frac{p_iq_j}{\alpha_{j-1}} =$$

$$= p_iq_j\left(\frac{\alpha_1-\alpha_0}{\alpha_1\alpha_0} + \frac{\alpha_2-\alpha_1}{\alpha_2\alpha_1} + \frac{\alpha_3-\alpha_2}{\alpha_3\alpha_2} + \ldots + \frac{\alpha_{j-1}-\alpha_{j-2}}{\alpha_{j-1}\alpha_{j-2}} + \frac{1}{\alpha_{j-1}}\right) = p_iq_j.$$

A 2.1 tétel alapján azonnal adódik az

$$\mathbf{A}+\mathbf{pq}^T=\mathbf{L}_1\mathbf{U}_1 \tag{2.3}$$

faktorizáció (feltéve, hogy $\alpha_i \neq 0$ $(i=1, 2, \ldots, n)$), ahol

$$\mathbf{L}_1=\begin{bmatrix} \frac{p_2q_1}{\alpha_1}+1 & & & \\ \frac{p_3q_1}{\alpha_1} & \frac{p_3q_2}{\alpha_2}+1 & & \\ \vdots & \vdots & \ddots & \\ \frac{p_{n-1}q_1}{\alpha_1} & \frac{p_{n-1}q_2}{\alpha_2} & \frac{p_{n-1}q_3}{\alpha_3} & \cdots & 1 \\ \frac{p_nq_1}{\alpha_1} & \frac{p_nq_2}{\alpha_2} & \frac{p_nq_3}{\alpha_3} & \cdots & \frac{p_nq_{n-1}}{\alpha_{n-1}}+1 \end{bmatrix}$$

és

$$\mathbf{U}_1=\begin{bmatrix} \alpha_1+p_1q_1 & p_1q_2 & p_1q_3 & \cdots & p_1q_{n-1} & p_1q_n \\ p_2q_1 & \alpha_2+p_2q_2 & p_2q_3 & \cdots & p_2q_{n-1} & p_2q_n \\ p_3q_1 & p_3q_2 & \alpha_3+p_3q_3 & \cdots & p_3q_{n-1} & p_3q_n \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ p_{n-1}q_1 & p_{n-1}q_2 & p_{n-1}q_3 & \cdots & \alpha_{n-1}+p_{n-1}q_{n-1} & p_{n-1}q_n \\ p_nq_1 & p_nq_2 & p_nq_3 & \cdots & p_nq_{n-1} & \alpha_n+p_nq_n \end{bmatrix}.$$

Ezek után tekintsük az

$$\mathbf{A}=\mathbf{A}+\mathbf{fg}^T=\tilde{\mathbf{L}}\tilde{\mathbf{U}} \tag{2.4}$$

faktorizáció meghatározását abban az esetben, amikor az $\mathbf{A}=\mathbf{LU}$ faktorizáció már ismert. Itt $\mathbf{L}$ és $\tilde{\mathbf{L}}$ olyan alsó trianguláris mátrixok, melyek fődiagonálisában csupa egyesek állnak, $\mathbf{U}$ és $\tilde{\mathbf{U}}$ felső trianguláris mátrixok, $\mathbf{f}=[f_1, f_2, \ldots, f_n]^T$ és $\mathbf{g}=[g_1, g_2, \ldots, g_n]^T$ pedig n-elemű vektorok.

Jelöljük $\mathbf{l}_i$-vel az $\mathbf{L}$, u_i-vel az $\mathbf{U}^T$ i-edik oszlopvektorát és oldjuk meg az

$$\mathbf{Lp}=\mathbf{f} \quad \text{és} \quad \mathbf{U}^T\mathbf{q}=\mathbf{g} \tag{2.5}$$

lineáris egyenletrendszereket visszahelyettesítéssel, ahol $\mathbf{p}=[p_1, p_2, \ldots, p_n]^T$ és $\mathbf{q}=[q_1, q_2, \ldots, q_n]^T$ jelöli a megoldásvektorokat. $\mathbf{p}$ és $\mathbf{q}$ elemeinek meghatározása folyamán egyúttal a

$$\mathbf{v}^{(i)}=\mathbf{f}-p_1\mathbf{l}_1-p_2\mathbf{l}_2-\ldots-p_i\mathbf{l}_i=p_{i+1}\mathbf{l}_{i+1}+p_{i+2}\mathbf{l}_{i+2}+\ldots+p_n\mathbf{l}_n \quad (i=0, 1, \ldots, n-1) \tag{2.6}$$

és a

$$\mathbf{w}^{(i)} = \mathbf{g} - q_1\mathbf{u}_1 - q_2\mathbf{u}_2 - \ldots - q_i\mathbf{u}_i = q_{i+1}\mathbf{u}_{i+1} + q_{i+2}\mathbf{u}_{i+2} + \ldots + q_n\mathbf{u}_n \quad (i = 0, 1, \ldots, n-1) \tag{2.7}$$

vektorokat is meghatároztuk.

Ezután a 2.1 tétel és a (2.3) faktorizáció alapján nyerjük, hogy

$$\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{f}\mathbf{g}^T = \mathbf{L}(\mathbf{I} + \mathbf{p}\mathbf{q}^T)\mathbf{U} = \mathbf{L}\mathbf{L}_1\mathbf{U}_1\mathbf{U} \tag{2.8}$$

amiből

$$\tilde{\mathbf{L}} = \mathbf{L}\mathbf{L}_1 \quad \text{és} \quad \tilde{\mathbf{U}} = \mathbf{U}_1\mathbf{U} \tag{2.9}$$

adódik.

Tekintsük ezután az

$$\mathbf{L}_1 = \bar{\mathbf{L}}_1\bar{\mathbf{D}}_1 + \mathbf{I} \tag{2.10}$$

előállítást, ahol

$$\bar{\mathbf{L}}_1 = \begin{bmatrix} 0 & & & & \\ p_2 & 0 & & & \\ p_3 & p_3 & 0 & & 0 \\ \vdots & \vdots & \vdots & \ddots & \\ p_{n-1} & p_{n-1} & p_{n-1} & \cdots & 0 \\ p_n & p_n & p_n & \cdots & p_n \quad 0 \end{bmatrix}$$

és

$$\bar{\mathbf{D}}_1 = \operatorname{diag}\left[\frac{q_1}{\alpha_1}, \frac{q_2}{\alpha_2}, \frac{q_3}{\alpha_3}, \ldots, \frac{q_{n-1}}{\alpha_{n-1}}, 0\right]$$

valamint az

$$\mathbf{U}_1 = \bar{\mathbf{D}}_2\bar{\mathbf{U}}_2 + \mathbf{I} \tag{2.11}$$

előállítást, ahol

$$\bar{\mathbf{D}}_2 = \operatorname{diag}\left[\frac{p_1}{\alpha_0}, \frac{p_2}{\alpha_1}, \frac{p_3}{\alpha_2}, \ldots, \frac{p_{n-1}}{\alpha_{n-2}}, \frac{p_n}{\alpha_{n-1}}\right]$$

$$\bar{\mathbf{U}}_2 = \begin{bmatrix} q_1 & q_2 & q_3 \cdots & q_{n-1} & q_n \\ & q_2 & q_3 \cdots & q_{n-1} & q_n \\ & & q_3 \cdots & q_{n-1} & q_n \\ & 0 & \ddots & \vdots & \vdots \\ & & & q_{n-1} & q_n \\ & & & & q_n \end{bmatrix}$$

és $\mathbf{I}$ jelöli az egységmátrixot.

(2.9)-ből a (2.10) és (2.11) felhasználásával nyerjük, hogy

$$\tilde{\mathbf{L}} = \mathbf{L} + \mathbf{L}\bar{\mathbf{L}}_1\bar{\mathbf{D}}_1 \tag{2.12}$$

és

$$\tilde{\mathbf{U}} = \mathbf{U} + \bar{\mathbf{D}}_2\bar{\mathbf{U}}_2\mathbf{U} \tag{2.13}$$

(2.6) és (2.7) alapján adódik, hogy

$$\mathbf{L}\bar{\mathbf{L}}_1 = [\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \ldots, \mathbf{v}_{n-1}, \mathbf{0}] \quad \text{és} \quad (\bar{\mathbf{U}}_2\mathbf{U})^T = [\mathbf{w}_0, \mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{n-2}, \mathbf{w}_{n-1}],$$

tehát a (2.5) egyenletrendszerek megoldása során egyúttal az $\mathbf{L}\bar{\mathbf{L}}_1$ és az $\bar{\mathbf{U}}_2\mathbf{U}$ mátrixokat is meghatároztuk.

Az $\mathbf{Lp}=\mathbf{f}$ lineáris egyenletrendszer megoldásához (és egyúttal a $\mathbf{v}^{(i)}$-k meghatározásához) $n^2/2-n/2$ szorzás, az $\mathbf{U}^T\mathbf{q}=\mathbf{g}$ lineáris egyenletrendszer megoldásához (és egyúttal a $\mathbf{w}^{(i)}$-k meghatározásához) $n^2/2-n/2$ szorzás és n osztás szükséges. Az α_i-k, $\bar{\mathbf{D}}_1$ és $\bar{\mathbf{D}}_2$ meghatározásának együttes műveletigénye n szorzás és $2(n-1)$ osztás. (2.10)-ben a $\bar{\mathbf{D}}_1$ diagonális mátrixszal szorzáshoz $n^2/2-n/2$, a (2.11)-ben $\bar{\mathbf{D}}_2$-sal szorzáshoz $n^2/2+n/2$ szorzás szükséges. Így a következő tételt nyertük:

2.2. TÉTEL. Ha ismert az $\mathbf{A}$ $n\times n$-es nemszinguláris mátrix $\mathbf{LU}$ faktorizációja, akkor tetszőleges olyan $\mathbf{f}$ és $\mathbf{g}$ n-elemű vektorok esetén, melyekből a (2.5) összefüggések által nyert $\mathbf{p}$ és $\mathbf{q}$ n-elemű vektorokra a (2.1) szerint képzett α_i-k egyike sem nulla, létezik az

$$\tilde{\mathbf{A}} = \mathbf{A}+\mathbf{fg}^T = \tilde{\mathbf{L}}\tilde{\mathbf{U}}$$

faktorizáció és megkapható legfeljebb $2n^2$ szorzás és $3n$ osztás árán.

3. Szimmetrikus eset

Ebben a részben egy szimmetrikus diáddal módosított szimmetrikus négyzetes mátrix $\mathbf{LDL}^T$ faktorizációjának meghatározására adunk algoritmust. Ennek elméleti alapját a következő lemma szolgáltatja:

3.1. LEMMA. Ha $\mathbf{D}=\operatorname{diag}[d_1, d_2, \ldots, d_n]$ olyan nemszinguláris mátrix, $\mathbf{p}=[p_1, p_2, \ldots, p_n]^T$ olyan vektor és α olyan skalár, hogy az

$$\alpha_i = 1+\alpha\sum_{j=1}^{i}\frac{p_j^2}{d_j} \quad (i = 1, 2, \ldots, n) \tag{3.1}$$

számok egyike sem nulla, akkor a

$$\mathbf{D}+\alpha\mathbf{pp}^T = \mathbf{L}_2\mathbf{D}_2\mathbf{L}_2^T \tag{3.2}$$

faktorizáció mindig létezik, ahol

$$\alpha_0 = 1$$

$$\mathbf{D}_2 = \operatorname{diag}\left[d_1\frac{\alpha_1}{\alpha_0}, d_2\frac{\alpha_2}{\alpha_1}, \ldots, d_n\frac{\alpha_n}{\alpha_{n-1}}\right]$$

$$\mathbf{L}_2 = \begin{bmatrix} 1 & & & \\ \frac{\alpha p_2 p_1}{d_1\alpha_1} & 1 & & 0 \\ \vdots & \vdots & \ddots & \\ \frac{\alpha p_n p_1}{d_1\alpha_1} & \frac{\alpha p_n p_2}{d_2\alpha_2} & \cdots & 1 \end{bmatrix}.$$

Bizonyítás. $\mathbf{D}+\alpha\mathbf{pp}^T$-ból $\mathbf{D}$-t kiemelve, majd a 2.1 tételt alkalmazva a lemma állítása igazolható.

Foglalkozzunk ezután az $\tilde{\mathbf{A}}=\mathbf{A}+\alpha\mathbf{ff}^T=\tilde{\mathbf{L}}\tilde{\mathbf{D}}\tilde{\mathbf{L}}^T$ *Cholesky-féle felbontásának* meghatározásával abban az esetben, amikor az $\mathbf{A}$ szimmetrikus és nemszinguláris mátrixnak ismert az $\mathbf{A}=\mathbf{LDL}^T$ *Cholesky-féle felbontása.* Itt $\mathbf{D}$ és $\tilde{\mathbf{D}}$ diagonális mátrixok, $\mathbf{L}$ és $\tilde{\mathbf{L}}$ pedig olyan alsó trianguláris mátrixok, melyek fődiagonálisában csupa egyesek állnak.

Jelöljük $\mathbf{l}_i$-vel az $\mathbf{L}$ i-edik oszlopvektorát és oldjuk meg az

$$\mathbf{Lp}=\mathbf{f} \tag{3.3}$$

lineáris egyenletrendszert visszahelyettesítéssel, ahol $\mathbf{p}=[p_1, p_2, \ldots, p_n]^T$ jelöli a megoldásvektort. $\mathbf{p}$ elemeinek meghatározása folyamán egyúttal a

$$\mathbf{v}^{(i)}=\mathbf{f}-p_1\mathbf{l}_1-p_2\mathbf{l}_2-\ldots-p_i\mathbf{l}_i=p_{i+1}\mathbf{l}_{i+1}+p_{i+2}\mathbf{l}_{i+2}+\ldots+p_n\mathbf{l}_n \tag{3.4}$$

$(i=0, 1, \ldots, n-1)$

vektorokat is meghatároztuk.

Ezután a 3.1 lemma segítségével a következőt nyerjük

$$\tilde{\mathbf{A}}=\mathbf{A}+\alpha\mathbf{ff}^T=\mathbf{L}(\mathbf{D}+\alpha\mathbf{pp}^T)\mathbf{L}^T=\mathbf{LL}_2\mathbf{D}_2\mathbf{L}_2^T\mathbf{L}^T \tag{3.5}$$

amiből

$$\tilde{\mathbf{L}}=\mathbf{LL}_2 \quad \text{és} \quad \tilde{\mathbf{D}}=\mathbf{D}_2 \tag{3.6}$$

adódik.

Tekintsük ezután az

$$\mathbf{L}_2+\mathbf{L}_2\mathbf{D}_2=\mathbf{I}+\alpha\mathbf{pp}^T \tag{3.7}$$

előállítást, ahol

$$\mathbf{L}_2=\begin{bmatrix} 0 & 0 & 0 & \ldots & 0 \\ p_2 & 0 & 0 & \ldots & 0 \\ p_3 & p_3 & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ p_{n-1} & p_{n-1} & p_{n-1} & \ldots & 0 \\ p_n & p_n & p_n & \ldots & p_n \end{bmatrix}$$

és

$$\mathbf{D}_2=\operatorname{diag}\left[\frac{\alpha p_1}{d_1\alpha_1}, \frac{\alpha p_2}{d_2\alpha_2}, \frac{\alpha p_3}{d_3\alpha_3}, \ldots, \frac{\alpha p_{n-1}}{d_{n-1}\alpha_{n-1}}, 0\right].$$

Így nyertük, hogy

$$\tilde{\mathbf{L}}=\mathbf{L}\mathbf{L}_2\mathbf{D}_2+\mathbf{L}. \tag{3.8}$$

Mivel (3.4) alapján

$$\mathbf{L}\mathbf{L}_2=[\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \mathbf{v}^{(3)}, \ldots, \mathbf{v}^{(n-1)}, \mathbf{0}]$$

adódik, ezért a (3.3) lineáris egyenletrendszer megoldása során egyúttal az $\mathbf{LL}_2$ szorzatot is meghatároztuk.

Az $\mathbf{Lp}=\mathbf{f}$ lineáris egyenletrendszer megoldásához (és egyúttal a $\mathbf{v}^{(i)}$-k meghatározásához) $n^2/2-n/2$ szorzás, az α_i-k meghatározásához $2n$ szorzás és n osztás szükséges. $\tilde{\mathbf{D}}=\mathbf{D}_2$ és $\overline{\mathbf{D}}_2$ meghatározásához további n szorzás és $2(n-1)$ osztás szükséges. (3.8)-ban a $\overline{\mathbf{D}}_2$-sal szorzás műveletigénye $n^2/2-n/2$ szorzás. Így a következő tételt nyertük:

3.2. Tétel. Ha ismert az $\mathbf{A}$ $n\times n$-es szimmetrikus és nemszinguláris mátrix $\mathbf{LDL}^T$ *Cholesky-féle faktorizációja,* akkor tetszőleges olyan α skalár és tetszőleges olyan $\mathbf{f}$ vektor esetén, melyből a (3.3) által nyert $\mathbf{p}$ vektorra a (3.1) szerint képzett α_i-k egyike sem nulla, létezik az

$$\tilde{\mathbf{A}} = \mathbf{A}+\alpha\mathbf{f}\mathbf{f}^T = \tilde{\mathbf{L}}\tilde{\mathbf{D}}\tilde{\mathbf{L}}^T.$$

Cholesky-féle faktorizációja és megkapható legfeljebb n^2+2n szorzás és $3n$ osztás árán.

Végezetül köszönetemet fejezem ki Dr. Móricz Ferenc egyetemi tanárnak, aki volt szíves a kiindulásul szolgáló cikkre figyelmemet felhívni és tanácsaival segítségemre lenni.

IRODALOM

[1] Gill, P. E., Golub, G. H., Murray, W., Saunders, M. A., "Methods for Modifying Matrix Factorizations", *MATH. COMP.* **28** (1974) 505—535.
[2] Fletcher, R., Powell, M. J. D., "On the Modification of LDL^T Factorizations", *MATH. COMP.* **28** (1974) 1067—1087.
[3] Gill, P. E., Murray, W., Saunders, M. A., "Methods for Computing and Modifying the LDV Factors of a Matrix", *MATH. COMP.* **29** (1975) 1051—1077.
[4] Benett, J. M., "Triangular Factors of Modified Matrices", *NUMER. MATH.* **7** (1965) 217—221.
[5] Gill, P. E., Murray, W., "Modification of Matrix Factorizations after a Rank-one Change", in: *The State of the Art in Numerical Analysis* Ed. *D. Jacobs* (Academic Press, London—New York—San Francisco, 1977) 55—83.
[6] Берсенёв, С. М., «О пересчете факторизации Холецкого», *ж. вычисл. Мат. и Мат. Физ.* **19** (1979) 1318—1319.
[7] Móricz, F., *Numerikus Analízis* II (Tankönyvkiadó, Budapest, 1975).
[8] Stoer, J., Bulirsch, R., *Introduction to Numerical Analysis* (Springer Verlag, New York—Heidelberg—Berlin, 1980).

(Beérkezett: 1984. július 5.)

(Átdolgozva beérkezett: 1984. november 1.)

BARTALOS ISTVÁN
JATE BOLYAI INTÉZET
6720 SZEGED, ARADI VÉRTANÚK TERE 1.

MODIFICATION OF THE LU FACTORIZATION OF SQUARE MATRICES AFTER CHANGING WITH A DIAD

I. Bartalos

It is shown that under certain circumstances the *LU* factorization of a square matrix after changing with a diad can be obtained in at most $2n^2+O(n)$ multiplications and divisions, provided the *LU* factorization of the original matrix is given. As a special case, the modification of the LDL^T factorization of symmetric matrices after changing with a symmetric diad is also dealt with. It is shown that the modified LDL^T factorization can be obtained in at most $n^2+O(n)$ multiplications and divisions.

Alkalmazott Matematikai Lapok **11** (1985) 167—170

EGY CSILLAPÍTOTT REZGŐMOZGÁS NEM-ATTRAKTÍV EGYENSÚLYI HELYZETTEL

KARSAI JÁNOS

Szeged

A dolgozatban olyan $\ddot{x}+a(t)\dot{x}+x=0$ alakú egyenletet konstruálunk, melyben $\int_0^\infty a\,dt=\infty$, de létezik az egyenletnek 0-hoz nem tartó megoldása $(t\to\infty)$.

Tekintsük az

$$\ddot{x}+a(t)\dot{x}+x=0 \tag{1}$$

differenciálegyenletet, melyben $a(t)$ nem-negatív, folytonos a $[0, \infty)$ intervallumon. Ez a csillapított rezgőmozgás differenciálegyenlete; ilyen alakú egyenlet írja le például a matematikai inga kis rezgéseit surlódó közegben.

Ismert, hogy ha az „összsúrlódás" véges, vagyis $\int_0^\infty a\,dt<\infty$, akkor az egyenlet minden megoldása oszcillál $[0, \infty)$-en, és az amplitúdó pozitív konstans felett marad [1]. Ha $\int_0^\infty a\,dt=\infty$, akkor mát létezik olyan megoldás, mely 0-hoz tart (oszcillálva vagy monoton módon) ha $t\to\infty$ [4].

Felmerül a következő kérdés: ha $\int_0^\infty a\,dt=\infty$, akkor az egyenlet minden megoldása 0-hoz tart-e, ha $t\to\infty$? A tapasztalat sugallja, hogy ha az $a(t)$ súrlódási együttható elég gyorsan akármilyen nagyra is növekedhet, akkor a súrlódás a rugalmassági erő ellenére egyenletesen távol tarthatja a mozgó pontot egyensúlyi helyzetétől [1, 5, 9]. Ezt az esetet az

$$\ddot{x}+(t^2+t+2/t)\dot{x}+x=0$$

egyenlet illusztrálja, melynek az $x(t)=1+1/t$ függvény megoldása. Módosítsuk hát az előző kérdést úgy, hogy feltételezzük $a(t)$ korlátosságát [1, 9]. A nem-oszcilláló megoldások (ha léteznek) ekkor mind az egyensúlyi helyzethez tartanak ha $t\to\infty$ [5, 6]. De a kérdésre adott válasz most is tagadó. Ugyanis, ha az $\int_0^\infty a\,dt$ függvény „nem megfelelően" növekszik, akkor létezhetnek olyan oszcilláló megoldások, melyek amplitúdója nem tart 0-hoz ha $t\to\infty$. Ez utóbbi állítást igazolandó, KERTÉSZ VIKTOR konstruált példát e folyóiratban megjelent cikkében [7].

Dolgozatunk tárgya egy a [7]-belinél egyszerűbb példa megszerkesztése. Megadunk olyan $a: [0, \infty) \to [0, 1]$ folytonos függvényt, melyre $\int_0^\infty a\,dt = \infty$, és az (1) egyenletnek mutatunk egy $x(t)$ oszcilláló megoldását, mely nem konvergál 0-hoz, ha $t \to \infty$.

Ha a monoton növekvő, végtelenbe divergáló $\{t_n\}$ sorozat adott, akkor legyen $a_n: [0, \infty) \to [0, 1]$ olyan folytonos függvény, amelyre

$$a_n(t) = \begin{cases} 0 & \text{ha} \quad t < t_n \quad \text{vagy} \quad t > t_n + 1/(n+1), \\ 1 & \text{ha} \quad t_n + 1/3(n+1) < t < t_n + 2/3(n+1). \end{cases}$$

Legyen továbbá $a(t) := \sum_{n=1}^{\infty} a_n(t)$. A $\{t_n\}$ sorozatot az alábbiak szerint határozzuk meg.

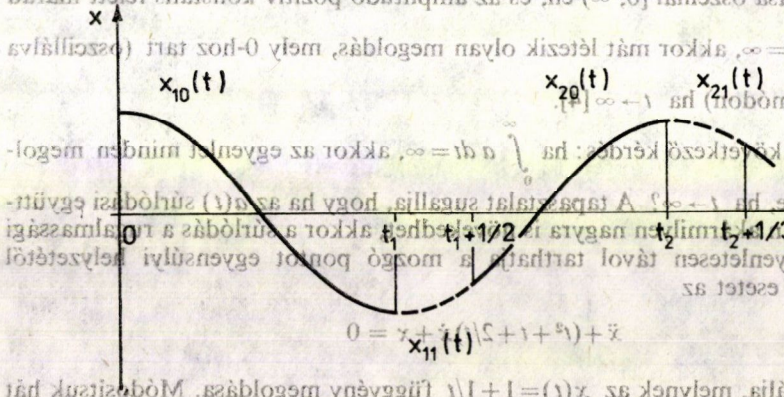
Tekintsük a harmonikus rezgőmozgás

$$\ddot{x} + x = 0 \tag{2}$$

mozgásegyenletének $x(0)=1$, $\dot{x}(0)=0$ kezdeti feltételekhez tartozó $x_{10}(t)$ megoldását a $[0, t_1]$ intervallumon, ahol t_1 az $x_{10}(t)$ O utáni első szélsőértékhelye. A $[t_1, t_1+1/2]$ intervallumon folytassuk $x_{10}(t)$-t differenciálható módon az

$$\ddot{x} + a_1(t)\dot{x} + x = 0$$

egyenlet $x_{11}(t)$ megoldásával, majd $[t_1+1/2, t_2]$-n újra a (2) egyenlet $x_{20}(t)$ megoldásával, ahol t_2 az $x_{20}(t)$ $t_1+1/2$ utáni első szélsőértékhelye (1. ábra).



1. ábra

Ha már t_n-et meghatároztuk, akkor a (2) egyenlet $x_{n0}(t)$, $[t_{n-1}+1/n, t_n]$-en már definiált megoldását folytassuk $[t_n, t_n+1/(n+1)]$-re az

$$\ddot{x} + a_n(t)\dot{x} + x = 0$$

egyenlet $x_{n1}(t)$ megoldásával, majd a $[t_n+1/(n+1), t_{n+1}]$ intervallumon a (2) egyen-

let $x_n(t)$ megoldásával differenciálható módon, ahol t_{n+1} az utóbbi folytatás első $t_n+1/(n+1)$-nél nagyobb szélsőértékhelye.

Ha az eljárást vég nélkül folytatjuk, akkor az $x_{10}(t)$, $x_{11}(t)$, ..., $x_{n0}(t)$, $x_{n1}(t)$,... darabokból az (1) egyenlet $[0, \infty)$-en értelmezett $x(t)$ megoldását nyerjük. Nyilvánvaló, hogy

$$\int_0^\infty a\,dt > \frac{3}{4}\sum_{n=1}^{\infty}\frac{1}{(n+1)^3} = \infty.$$

Elegendő bebizonyítanunk, hogy az $x(t)$ megoldásra $\lim_{t\to\infty} r^2(t) \neq 0$, ahol $r(t) := (x^2(t)+\dot{x}^2(t))^{1/2}$.

Az egyenleten elvégezve az

$$x(t) = r(t)\cos\varphi(t), \quad \dot{x}(t) = -r(t)\sin\varphi(t)$$

transzformációt, az

$$\dot{r} = -\frac{1}{2}a(t)r(t)\sin 2\varphi \qquad \dot{\varphi} = 1 - \frac{1}{2}a(t)\sin^2\varphi \tag{1'}$$

rendszert kapjuk. Mivel $\dot{r} \leqq 0$, $\lim_{t\to\infty} r(t) \geqq 0$ létezik és

$$r(\infty) = \exp\left\{-\frac{1}{2}\sum_{n=1}^{\infty}\int_{t_n}^{t_n+1/(n+1)} a(t)\sin^2\varphi(t)\,dt\right\}.$$

A $[t_n; t_n+1/(n+1)]$ intervallumon igaz az alábbi becslés:

$$\sin^2\varphi(t) < (\varphi(t)-\varphi(t_n))^2 = \left[\int_{t_n}^{t}\left(1-\frac{1}{2}a(\tau)\sin 2\varphi(\tau)\,d\tau\right)\right]^2 < \frac{1}{4}(t-t_n)^2.$$

Innen

$$r(\infty) > \exp\left\{-\frac{3}{4}\sum_{n=1}^{\infty}\frac{1}{(n+1)^3}\right\} > 0,$$

amivel a bizonyítás kész.

Megjegyezzük, hogy ha az $a(t)$ folytonosságának követelményétől eltekintünk, az eljárás sokkal egyszerűbb. Ugyanis ha

$$a_n(t) := \begin{cases} a(t), & \text{ha } t_n \leqq t < t_n+1/(n+1), \\ 0, & \text{egyébként,} \end{cases}$$

akkor csupán az

$$\ddot{y}+\dot{y}+y = 0$$

egyenletek megfelelő megoldásait kell felváltva egymáshoz illeszteni.

Megjegyezzük még, hogy egy ilyen egyenlet létezése közvetett úton is belátható. Ugyanis a

$$x'' + q(\tau)x = 0 \quad (x' = dx/d\tau) \tag{3}$$

egyenletben $(q\in C^1[0,\infty),\ q(\tau)>0,\ q'(\tau)>0)$ a τ független változó helyett az új

$$t(\tau):=\int_0^\tau q^{1/2}$$

független változót bevezetve (1) alakú egyenletet kapunk, és az $\int_0^\infty a(t)dt=\infty$ feltételnek a $\lim_{\tau\to\infty} q(\tau)=\infty$ feltétel felel meg. GALBRAITH, MC SHANE és PARRISH [3] — cáfolandó LEIGHTON [8] téves állítását — konstruáltak olyan (3) alakú egyenletet és annak $x(\tau)$ megoldását, melyben $\lim_{\tau\to\infty} q(\tau)=\infty$, de $\limsup_{\tau\to\infty} |x(\tau)|>0$.

IRODALOM

[1] BALLIEU, R. J. and PEIFFER, K., "Attractivity of the origin for the equation $\ddot{x}+f(t,x,\dot{x})\times\times|\dot{x}|^\alpha\dot{x}+g(x)=0$", *J. Math. Anal. Appl.* **65** (1978) 321—333.

[2] DE KLEINE, H. A., "A counterexample to a conjecture in second-order linear equations", *Michigan Math. J.* **17** (1970) 29—32.

[3] GALBRAITH, A. S., MC SHANE, E. J. and PARRISH, G. B., "On the solutions of linear second-order differential equations", *Proc. Nat. Acad. Sci. U.S.A.* **53** (1965) 247—249.

[4] HARTMAN, P., "On the theorem of Milloux", *Amer. J. Math.* **70** (1948) 395—399.

[5] HATVANI, L., "On the stability of the zero solution of certain secon-order nonlinear differential equations", *Acta Sci. Math. 32* (1971) 1—9.

[6] KARSAI, J., "On the asymptotic stability of the zero solution of the equation $\ddot{x}+g(t,x,\dot{x})\dot{x}++f(x)=0$", *Studia Sci. Math.* (megjelenés alatt).

[7] KERTÉSZ, V., „A csillapított rezgőmozgás differenciálegyenletének stabilitási vizsgálata", *Alk. Mat. Lapok 8* (1982) 323—339.

[8] LEIGHTON, W., "Behavior of solutions of a linear differential equation of second order", *Proc. Nat. Acad. Sci. U.S.A.* **52** (1964) 830—832.

[9] SMITH, R. A., "Asymptotic stability of $x''+a(t)x'+x=0$", *Quart. J. Math. Oxford 2* **12** (1961) 123—126.

[10] WILLETT, D., "On an example in second order linear ordinary differential equations", *Proc. Amer. Math. Soc.* **17** (1966) 1263—1266.

(Beérkezett: 1984. július 9.)

KARSAI JÁNOS
SZEGEDI ORVOSTUDOMÁNYI EGYETEM SZÁMÍTÓKÖZPONTJA
6720 SZEGED, PÉCSI U. 4/A.

A DAMPED OSCILLATION WITH NONATTRACTIVE EQUILIBRIUM POSITION

J. KARSAI

In the paper by a simple procedure an equation of form $\ddot{x}+a(t)\dot{x}+x=0$ is given, in which $\int_0^\infty a\,dt=\infty$ but it has an oscillatory solution not tending to zero as $t\to\infty$.

Alkalmazott Matematikai Lapok **11** (1985) 171—181

A CHEMOTON MATEMATIKAI MODELLJÉRŐL

CSENDES TIBOR

Szeged

A dolgozatban egy, a bevezetésben röviden ismertetett biológiai rendszer, a chemoton egy lehetséges matematikai modelljét állítjuk fel. A 2. és 3. fejezetben igazoljuk, hogy az ennek alapjául szolgáló differenciálegyenlet-rendszer megoldásaira érvényes az egzisztencia, unicitás, folytathatóság, a kezdeti értékektől való folytonos függés és a pozitivitás tulajdonsága. A 4. fejezet a modell egyensúlyi helyzeteit adja meg, és igazolja ezek egy részének instabilis voltát.

1. Bevezetés

A chemoton olyan, kémiai rendszerek sztöchiometriai kapcsolása révén előálló kémiai szuperrendszer, amely kielégíti az életre jellemző kritériumokat. Absztrakt modelljét GÁNTI állította fel [4, 5, 6]. Reakciókinetikai leírását BÉKÉS és munkatársai dolgozták ki [1, 2].

A chemoton három funkcionálisan összefüggő, önreprodukáló alrendszerből áll: egy autokatalitikus kémiai körfolyamatból, egy templát alrendszerből és az egészet magábafoglaló, a chemotont környezetétől elhatároló membrán-alrendszerből. Működése a belső anyagai között végbemenő kémiai és fizikai reakciókban, a belső anyagok mennyiségének változásaként valósul meg.

A chemoton a környezetből felvett magas energia-tartalmú tápanyagot (jelölése X) fogyasztva működik, miközben alacsony energiatartalmú salakanyagot (Y) ad le.

Az autokatalitikus körfolyamat a chemoton szempontjából külső anyagok kémiai átalakítása révén a belső anyagok szabályozott termelését valósítja meg. A ciklus termeli a templát- és a membrán-alrendszer nyersanyagát (V', illetve T'). Jelöljük az autokatalitikus kémiai körfolyamat belső anyagait A_1, A_2, A_3, A_4, A_5-tel. Egy A_i molekulának a körfolyamatban való egyszeri végighaladása során egy X fogy el, és egy-egy Y, V', T' és A_1 molekula keletkezik. A kémiai reakciók jelölése:

$$\begin{aligned} A_1+X &\rightleftharpoons A_2 \\ A_2 &\rightleftharpoons A_3+Y \\ A_3 &\rightleftharpoons A_4+V' \\ A_4 &\rightleftharpoons A_5+T' \\ A_5 &\rightleftharpoons A_1+A_1. \end{aligned}$$

Feltesszük ezekről a reakciókról, hogy reverzibilisek, azaz mindkét irányban lejátszódhatnak. A reakciósebességi állandók legyenek a megfelelő sorrendben k_1, k_1', ...

..., k_5, k_6 (a felső vessző a visszafelé irányuló reakciókhoz tartozó állandókat jelzi). Legyen $k_i \geqq k_i'$ ($i = 1, 2, \ldots, 5$). Tehát az X fogyasztását eredményező reakciók sebességi állandói legyenek nagyobbak, mint az X termelését eredményezőkéi. Feltesszük, hogy a chemoton környezete a tápanyagra nézve végtelen forrás, a salakanyagra nézve végtelen nyelő, ezek koncentrációja (mol/térfogat) állandónak tekinthető.

A nukleinsavak analógiájaként a chemotonban polimerizálódott szálak: n darab V monomerből álló, pV_n-nel jelölt homopolimerek szerepelnek. Ezek lemásolódása rendszerünkben polikondenzációs folyamattal megy végbe, amely során a V' molekulákból az R kondenzációs termék hasad le, és a monomerek polimer szállá kapcsolódnak össze a mintaként jelenlevő pV_n templátmolekulákon. A folyamat csak akkor indul be, ha a V' koncentrációja nagyobb egy meghatározott, $[V']^*$-gal jelölt küszöb-koncentrációnál. Feltételezzük, hogy a lemásolódott szálak azonnal szétválnak.

A folyamat jelölése

$$pV_nV_i + V' \rightarrow pV_nV_{i+1} + R \quad i = 1, 2, \ldots, n-2$$

$$pV_nV_{n-1} + V' \rightarrow 2pV_n + R.$$

Az első reakció reverzibilis, sebességi állandói k_6 és k_6', a többi reakció irreverzibilis, a sebességi állandó ezekre k_8.

A membrán-alrendszer a chemotont a környezettől elhatároló membránból és a membránt alkotó T molekulát létrehozó reakcióútból áll. Ezen alrendszer számára a másik kettő termeli a T' és R molekulákat. A T' molekula R-rel reagálva T-t ad, ami spontán módon beépül a membránba. A megfelelő reakciósebességi állandók k_7 és k_7'. A T-vel jelölt membránképző monomerek beépülési sebessége arányos a membrán felületének nagyságával és T koncentrációjával. Legyen az átányossági tényező k_9. A membrán a tárgyalt anyagok közül csak az X, Y és az oldószer számára átjárható, tehát az összes belső anyag számára átjárhatatlan. A membrán alaphelyzetben gömb alakú, de az ozmotikus nyomásviszonyok megváltozása esetén belapulhat. A chemoton felszínén (S) a membrán felszínét, térfogatán (Q) a membrán által közrezárt térfogatot értjük. A koncentrációkat erre a térfogatra vonatkozóan számítjuk.

A fenti három alrendszerből álló szupetrendszer a chemoton. Működését így képzelhetjük el: az autokatalitikus ciklus növeli V' koncentrációját, amíg ez a $[V']^*$-ot meghaladja. Ekkor a templát molekulákra leülnek az első monomerek. A pV_n-ek lemásolása közben R keletkezik, amivel reagálva a T' molekula T-t képez. Ez beépül a membránba, növelve annak felszínét. A chemoton térfogata mindig a membrán felszínének és az ozmotikus nyomásviszonyoknak megfelelő. Kezdetben a membrán annyi T molekulából áll, amennyi R keletkezik a templátok lemásolása során. Tehát a pV_n-ek replikációja után a membrán felülete is megduplázódik. A kétszeres felülethez gömbalak esetén $2^{3/2}Q^0$ térfogat tartozik. Mivel a belső anyagok mennyisége közel megkétszereződik ezalatt, az ozmotikus nyomás a kezdeti alá süllyed. Feltéve, hogy a kezdeti nyomás egyensúlyi, a membrán betüremkedik, befűződik, és a chemoton kettéosztódik. Feltételezzük, hogy a két utód egyenlő arányban osztozik a belső anyagokon. Ezután az egyik viselkedését követjük tovább.

2. A matematikai modell felállítása, megoldásának tulajdonságai

A bevezetésben tárgyalt rendszer matematikai modelljének felállításához a koncentrációkra vonatkozó *Guldberg—Waage differenciálegyenlet-rendszerből* indulhatunk ki:

$$
\begin{aligned}
d[A_1]/dt &= 2(k_5[A_5][A_1]-k_5'[A_1][A_1])-k_1[A_1][X]+k_1'[A_2] \\
d[A_2]/dt &= k_1[A_1][X]-k_1'[A_2]-k_2[A_2]+k_2'[A_3][Y] \\
d[A_3]/dt &= k_2[A_2]-k_2'[A_3][Y]+k_3[A_3]-k_3'[A_4][V] \\
d[A_4]/dt &= k_3'[A_3][V]-k_3[A_4]+k_4'[A_5][T] \\
d[A_5]/dt &= k_4[A_4]+k_4'[A_5][T]-k_5[A_5]+k_5'[A_1][A_1] \\
d[V]/dt &= k_3[A_3]-k_3'[A_4][V]+k_6[pV_n]-k_6'[V]\sum_{i=1}^{n-1}[pV_nV_i]
\end{aligned}
\tag{2.1}
$$

$$
\begin{aligned}
d[T]/dt &= k_4[A_4]-k_4'[A_5][T]+k_7[T'][R]-k_7'[T] \\
d[T']/dt &= k_7[T'][R]-k_7'[T]-k_9[T]S/Q \\
d[R]/dt &= k_7'[T]-k_7[T'][R]+k_8[V]\sum_{i=1}^{n-1}[pV_nV_i]
\end{aligned}
\tag{2.2}
$$

$$
\begin{aligned}
d[pV_n]/dt &= k_6'[pV_nV_1]-k_6[V][pV_n]+2k_8[V][pV_nV_{n-1}]-k_8[V][pV_n] \\
d[pV_nV_1]/dt &= k_6[V][pV_n]-k_6'[pV_nV_1]-k_8[V][pV_nV_1]+k_8[V][pV_nV_2] \\
d[pV_nV_i]/dt &= k_8[V][pV_nV_{i-1}]-k_8[V][pV_nV_i]+k_8[V][pV_nV_{i+1}], \quad i=2,3,\ldots,n-1 \\
dS/dt &= k_9[T]S\,S^0/(n-1).
\end{aligned}
\tag{2.3}
$$

Itt a szögletes zárójel az illető molekula koncentrációját jelöli: $[Z]=Z/Q$, ahol Q a rendszer térfogata. Mivel általános esetben a térfogat nem állandó, hanem az $\mathbf{u}$ belső anyagok és az S felszín adott $Q=Q(\mathbf{u}, S)$ függvénye, ezért (2.1) közvetlenül nem alkalmas rendszerünk leírására. A (2.1) differenciálegyenlet-rendszert a függelékben megadott új jelölésekkel a következő tömörebb alakban írhatjuk:

$$
\frac{d[\mathbf{u}]}{dt}=\mathbf{f}([\mathbf{u}], Q, S), \tag{2.4}
$$

ahol $\mathbf{u}=(u_1, u_2, \ldots, u_{n+9})^T$ oszlopvektor, és $(\ldots)^T$ a transzponálás jele. Megállapíthatjuk, hogy (2.2) $\mathbf{u}$ és S deriváltjára nézve nem megoldott differenciálegyenlet-rendszer. Alakítsuk át a

$$
\frac{d[\mathbf{u}]}{dt}=\mathbf{f}([\mathbf{u}], Q, S)
$$

differenciálegyenlet-rendszert

$$\frac{d[\mathbf{u}]}{dt} = \frac{d(\mathbf{u}/Q)}{dt} = \frac{1}{Q}\frac{d\mathbf{u}}{dt} - \frac{\mathbf{u}}{Q^2}\frac{dQ}{dt}$$

felhasználásával:

$$\frac{d\mathbf{u}}{dt} = \mathbf{f}\left(\frac{\mathbf{u}}{Q}, Q, S\right)Q + \frac{\mathbf{u}}{Q}\frac{dQ}{dt}.$$

Itt a differenciálegyenlet-rendszer jobb oldalán levő első tag a belső anyagok között lezajló kémiai reakciókból származó, a második tag pedig a kizárólag a térfogat növekedéséből származó anyagmennyiség-változásokat adja meg. Föltételeztük, hogy a membrán átjárhatatlan a belső anyagok számára, ezért a második taggal reprezentált anyagmennyiség-változás nem érvényesül rendszerünkben; ezt a tagot törölni kell egyenletünkből. A térfogat számítására a $Q(\mathbf{u}, S) = S^{3/2}/6\sqrt{\pi}$ függvényt használjuk, amely a gömbalaknak megfelelő térfogatot adja meg.

A Q-hoz tartozó, a (2.2)-nek megfelelő explicit differenciálegyenlet-rendszer:

$$\begin{aligned} \frac{d\mathbf{u}}{dt} &= \mathbf{f}\left(\frac{\mathbf{u}}{Q}\right)Q, \\ \frac{dS}{dt} &= k_9\frac{u_8 S S^0}{Q(n-1)}, \end{aligned} \tag{2.3}$$

ahol $Q = S^{3/2}/6\sqrt{\pi}$.

Ezek után megadhatjuk a matematikai modellt. Az állapothatározók u_i $(i=1, 2, \ldots, n+9)$, és S; időbeli változásukat a (2.3) differenciálegyenlet-rendszer írja le. Ennek paraméterei pozitív konstansok, illetve $n \in (2, 3, \ldots)$. A differenciálegyenlet-rendszer nem-folytonos jobboldalú lesz, mert k_6 és k_6' értéke az alábbiak szerint változik:

$$k_6 = \begin{cases} 0 & \text{ha } [V'] < [V']^* \text{ (azaz, ha } [u_6] < [V']^*), \\ k_6^* & \text{különben,} \end{cases}$$

$$k_6' = \begin{cases} 0 & \text{ha } [V'] < [V']^*, \\ k_6'^* & \text{különben,} \end{cases}$$

itt k_6^* és $k_6'^*$ pozitív konstansok. A fenti modell viselkedését a

$$\mathbf{u}(t_0) = \mathbf{u}^0, \quad S(t_0) = S^0 \tag{2.4}$$

kezdeti értékekből indulva vizsgáljuk, ahol S^0 a rendszerre jellemző pozitív állandó, u_i^0 $(i=1, 2, \ldots, n+9)$ nem-negatív konstans. Differenciálegyenlet-rendszerünk megoldásának olyan $[t_0, t_1)$-en értelmezett abszolút folytonos $(\mathbf{u}(t), S(t))$ függvényt nevezünk, amely kielégíti (2.4)-et, $[t_0, t_1)$-en majdnem mindenütt érvényes (2.3), és t_1 az első olyan t_0-nál nagyobb szám, hogy $S(t_1) = 2S^0$, illetve $t_1 = +\infty$, ha $S(t) < 2S^0$ minden $t > t_0$-ra. Tehát a megoldásokat az

$$U = \{(t, \mathbf{u}, S): t \geqq 0, \ u_i \geqq 0 \ \ (i = 1, 2, \ldots, n+9), \ \ 0 < S^0 \leqq S \leqq 2S^0)\}$$

halmazon vizsgáljuk.

A bevezetésben említettek szerint nem tekinthetünk a chemoton modelljének olyan rendszert, amelyben a három alrendszer valamelyike teljesen hiányzik, vagyis olyanokat, amelyekben $u_i=0, i=1, 2, \ldots, 5$ vagy $u_{10}=0$, $u_{10+i}=0$, $i=1, 2, \ldots, n-1$.

2.1. TÉTEL. A (2.3) differenciálegyenlet-rendszer megoldásaira minden U-ban levő korlátos G tartományon érvényes az egzisztencia, unicitás, a folytathatóság és a kezdeti értékektől való folytonos függés.

Bizonyítás. Az általunk vizsgált nem-folytonos jobboldalú differenciálegyenlet-rendszerekre a szokásos *(Peano, Picard—Lindelöf, Carathéodory) egzisztencia-*, illetve *unicitástételek* nem alkalmazhatók; így a tétel bizonyításához FILIPPOV [9] tételeit használhatjuk. Ezekből tételünk állítása akkor következik, ha a

$$\frac{d\mathbf{x}}{dt} = \mathbf{F}(t, \mathbf{x})$$

alakban írt differenciálegyenlet-rendszerre a G korlátos tartományon teljesülnek az alábbi feltételek:

E1 $\mathbf{F}$ mérhető és majdnem mindenütt értelmezve van,

E2 létezik olyan lokálisan integrálható $B(t)$ valós függvény, hogy $|\mathbf{F}(t, \mathbf{x})| \leqq B(t)$ majdnem mindenütt,

E3 léteznek olyan ε, K pozitív állandók, hogy majdnem minden $(t, \mathbf{x})$, $(t, \mathbf{z})$-re $|\mathbf{x}-\mathbf{z}|<\varepsilon$ esetén $(\mathbf{x}-\mathbf{z})(\mathbf{F}(t, \mathbf{x})-\mathbf{F}(t, \mathbf{z})) \leqq K|\mathbf{x}-\mathbf{z}|^2$.

A (2.3) differenciálegyenlet-rendszer jobb oldala U-n mindenütt értelmezve van, és egy nullamértékű halmazt kivéve akárhányszor differenciálható. Így az minden U-ban levő korlátos G tartományon mérhető. Itt $\mathbf{F}$ és elsőrendű parciális differenciálhányadosai is korlátosak, tehát E2 és E3 is teljesül. Ezzel a tételt igazoltuk.

A 2.1. tétel alapján az U tartományból indított megoldások dinamikus viselkedésének vizsgálatára alkalmas eszköz lehet a numerikus integrálás.

Megjegyzés. A tárgyalt matematikai modellek nem értelmezik az 1. fejezetben leírt ozmotikus térfogatváltozást és osztódást. Az utóbbi a számítógépes szimuláció során könnyen megvalósítható [3].

A jelen dolgozatban vizsgált modell eltér a korábbiaktól [3, 4, 6]; a módosítások az előzőeket is kidolgozó GÁNTI TIBOR javaslatára történtek.

3. A megoldások pozitivitásáról

Ebben a fejezetben azt igazoljuk, hogy a nem-negatív kezdeti értékű megoldások végig nem-negatívok lesznek értelmezési tartományukon. Ilyen kérdésekkel VOLPERT [8] foglalkozott, de tételei folytonos és az állapothatározók szerint mindenütt differenciálható jobboldalú differenciálegyenlet-rendszerekre vonatkoztak. Ezért ezeket közvetlenül nem használhatjuk rendszerünkre, általánosításukra van szükség.

Tekintsük a következő alakú differenciálegyenlet-rendszert:

$$\frac{dx_k}{dt} = \sum_{i=1}^{m} \gamma_{ik} F_i(t, \mathbf{x}) + G_k(t) \quad k = 1, 2, \ldots, n, \tag{3.1}$$

és az

$$x_k(t_0) = x_k^0 \quad k = 1, 2, \ldots, n \tag{3.2}$$

kezdeti értékeket. Teljesüljenek ezekre az alábbi feltételek:

F1 $F_i(t, \mathbf{x})$ $(i = 1, 2, \ldots, m)$ értelmezve van és mérhető a $t \geqq t_0$ féltér valamely P tartományán, valamint létezik olyan lokálisan integrálható $B(t)$ valós függvény, hogy $|\mathbf{F}(t, \mathbf{x})| \leqq B(t)$ majdnem mindenütt;

F2 $F_i(t, \mathbf{x}) \geqq 0$ ha $(t, \mathbf{x}) \in P$ és $x_k \geqq 0$ $(k = 1, 2, \ldots, n)$, $(i = 1, 2, \ldots, m)$;

F3 $G_k(t)$ integrálható minden véges intervallumon;

F4 a (3.1) differenciálegyenlet-rendszernek minden olyan (3.2) kezdeti értékkel, ahol $(t_0, \mathbf{x}^0) \in P$, pontosan egy megoldása létezik;

F5 ha $\gamma_{ik} < 0$, akkor minden $\mathbf{x}(t)$ megoldásra $F_i(t, \mathbf{x}(t)) = \varphi_i(t) x_k(t)$, ahol $\varphi_i(t)$ integrálható minden véges $[t_0, t_1]$ intervallumon.

Vannak olyan speciális differenciálegyenlet-rendszerek — és ilyenek az általunk vizsgáltak is —, amelyek esetén az F5 feltétel ellenőrzéséhez nem szükséges a megoldások ismerete. Ugyanis ha $F_i(t, \mathbf{x}) = H_i(t) x_1^{p_1^i} \ldots x_n^{p_n^i}$, ahol $H_i(t)$ lokálisan integrálható valós függvény, és p_j valós szám $(j = 1, 2, \ldots, n)$, akkor F5 teljesüléséhez elegendő, hogy ha $\gamma_{ik} < 0$, akkor $p_k^i \geqq 1$.

3.1. TÉTEL. Ha a (3.1) differenciálegyenlet-rendszerre a (3.2) kezdeti értékekkel érvényesek az F1—F5 feltételek, továbbá $x_k^0 > 0$ és $G_k(t) \geqq 0$ $(k = 1, 2, \ldots, n)$, akkor a (3.1)—(3.2) kezdetiérték-probléma tovább nem folytatható megoldása létezési intervallumának minden pontjában pozitív.

Bizonyítás. Tegyük fel, hogy a megoldás a $[t_0, t_1]$ intervallumon létezik, de ott nem pozitív mindenütt. Ekkor létezik olyan $t' \in [t_0, t_1]$, amelyre igaz, hogy

$$x_j(t') = 0 \tag{3.3}$$

valamely j-re, $1 \leqq j \leqq n$. Legyen

$$\varphi(t) = \sum_i \gamma_{ij} \varphi_i(t)$$

és

$$\psi(t) = \sum_i \gamma_{ij} F_i(t, \mathbf{x}(t)) + G_j(t).$$

Így $\psi(t) \geqq 0$ a $[t_0, t']$ zárt intervallumon. A (3.1) differenciálegyenlet-rendszerből kapjuk, hogy

$$\frac{dx_j(t)}{dt} \geqq \varphi(t) x_j(t) + \psi(t),$$

ahonnan

$$x_j(t) \geqq x_j(t_0) \exp\left\{\int_{t_0}^{t} \varphi(S)\, ds\right\} + \int_{t_0}^{t} \exp\left\{\int_{r}^{t} \varphi(s)\, ds\right\} \psi(r)\, dr.$$

Következésképpen $x_j(t') > 0$, ami ellentmond (3.3)-nak. Ezzel állításunkat beláttuk.

3.2. TÉTEL. Ha $x_k^0 \geqq 0$, $G_k(t) \geqq 0$ $(k=1, 2, \ldots, n)$, az F1—F5 feltételek teljesülnek, valamint a (3.1), (3.2) kezdetiérték-probléma P-n tovább nem folytatható megoldása folytonosan függ a kezdeti értékektől, akkor az létezési intervallumán nem-negatív.

Bizonyítás. A tétel állítása közvetlenül következik a 3.1. tételből és a megoldásoknak a kezdeti értékektől való folytonos függéséből.

A (2.3) differenciálegyenlet-rendszer esetén a 3.1. és a 3.2. tételben szereplő P legyen valamely, az U halmazban levő tartomány. Legyenek az $F_i(t, \mathbf{x})$ függvények a jobb oldalon levő tagok abszolút értékben, γ_{ik} a k-adik egyenletben szereplő $F_i(t, \mathbf{x})$ előjelének megfelelően $+1$ vagy -1, és $G_k(t)=0$ minden k-ra. Ekkor a fejezet elején levő F1—F3 feltételek teljesülnek. Az F4 feltétel a 2. fejezet eredményei szerint fennáll. Az F5 feltétel teljesülése pedig a (2.3), illetve a (2.1) egyenletekből olvasható le. Mivel a megoldásoknak a kezdeti értékektől való folytonos függését a 2. fejezetben igazoltuk, a 3.1. és a 3.2. tétel érvényes rendszerünkre. Tehát a pozitív kezdeti értékű megoldások pozitívak, a nem-negatív kezdeti értékűek pedig nem-negatívak maradnak azon az intervallumon, ahol az U halmazon értelmezve vannak. Ilyen értelemben tehát matematikai modellünk megfelelően írja le a kémiai modellben értelmezett történéseket, az egyes anyagfajták koncentrációja legalább nulla értéket vesz fel.

4. Egyensúlyi helyzetek

Vizsgáljuk meg, hogy a (2.3) differenciálegyenlet-rendszer egyensúlyi helyzetei hol vannak az U halmazon. Mivel ebben a fejezetben a leggyakrabban a (2.1) differenciálegyenlet-rendszert kell vizsgálni, ezért itt az állapothatározóknak ismét a hagyományos jelölését használjuk.

Az U halmazon S pontosan akkor állandó, ha T azonosan nulla. Ha az előzőek teljesülnek, akkor $\dot{T}=0$-ból $T'R=0$ következik, és hasonlóan a következőkre is

$$\text{ha még } \dot{R} = 0, \text{ akkor } V' \sum_{i=1}^{n-1} pV_n V_i = 0,$$

$$\dot{T}' = 0\text{-ból} \quad k_4 \frac{A_4}{Q^0} - k_4' \frac{A_5}{Q^0} \frac{T'}{Q^0} = 0,$$

$$\dot{A}_5 = 0\text{-ból} \quad k_5 \frac{A_5}{Q^0} - k_5' \frac{A_1}{Q^0} \frac{A_1}{Q^0} = 0,$$

$$\dot{A}_1 = 0\text{-ból} \quad k_1 \frac{A_1}{Q^0} [X] - k_1' \frac{A_2}{Q^0} = 0,$$

$$\dot{A}_2 = 0\text{-ból} \quad k_2 \frac{A_2}{Q^0} - k_2' \frac{A_3}{Q^0} [Y] = 0,$$

$$\dot{A}_3 = 0\text{-ból} \quad k_3 \frac{A_3}{Q^0} - k_3' \frac{A_4}{Q^0} \frac{V'}{Q^0} = 0,$$

és

$\dot{A}_1 = 0$ ezután már következik.

$\dot{V}' = 0$-ból $\quad k_6 \frac{pV_n}{Q^0} \frac{V'}{Q^0} - k_6' \frac{pV_n V_1}{Q^0} = 0,$

$p\dot{V}_n = 0$-ból $\quad V' pV_n V_{n-1} = 0,$

$p\dot{V}_n V_i = 0$-ból $\quad V' pV_n V_{i-1} = 0 \quad i = n-1,\ n-2,\ \ldots,\ 2$

adódik. Az utolsó két sorból következik, hogy

$$V' \sum_{i=1}^{n-1} pV_n V_i = 0.$$

Tehát a (2.3) differenciálegyenlet-rendszer egyensúlyi helyzetei az U halmazon ott vannak, ahol az alábbi egyenlőségek minden $t > t_0$-ra teljesülnek:

$$k_1 A_1[X] = k_1' A_2,$$

$$k_2 A_2 = k_2' A_3[Y],$$

$$k_3 A_3 = \frac{k_3'}{Q^0} A_4 V',$$

$$k_4 A_4 = \frac{k_4'}{Q^0} A_5 T',$$

$$k_5 A_5 = \frac{k_5'}{Q^0} A_1^2,$$

$$\frac{k_6}{Q^0} V' pV_n = k_6' pV_n V_1,$$

$$T' R = 0,$$

$$T = 0,$$

$$V' pV_n V_i = 0 \quad i = 1, 2, \ldots, n-1.$$

A lehetséges egyensúlyi helyzeteket az alábbiak szerint csoportosíthatjuk.
a) Ha $T' > 0$ és $k_6, k_6' > 0$ (azaz $V' \cong [V']^* Q^0 > 0$), akkor

$$R,\ T,\ pV_n,\ pV_n V_i = 0, \quad i = 1, 2, \ldots, n-1,$$

$$V' \cong [V']^* Q^0,$$

$$T' > 0$$

és $A_i=0$, $i=1, 2, \ldots, 5$, vagy

$$A_1 = \frac{k_1 k_2 k_3 k_4 k_5 [X] Q^{03}}{k_1' k_2' k_3' k_4' k_5' [Y] V' T'}, \tag{4.1}$$

$$A_2 = \frac{k_1^2 k_2 k_3 k_4 k_5 [X]^2 Q^{03}}{k_1'^2 k_2' k_3' k_4' k_5' [Y] V' T'}, \tag{4.2}$$

$$A_3 = \frac{k_1^2 k_2^2 k_3 k_4 k_5 [X]^2 Q^{03}}{k_1'^2 k_2'^2 k_3' k_4' k_5' [Y]^2 V' T'}, \tag{4.3}$$

$$A_4 = \frac{k_1^2 k_2^2 k_3^2 k_4 k_5 [X]^2 Q^{04}}{k_1'^2 k_2'^2 k_3'^2 k_4' k_5' [Y]^2 V'^2 T'}, \tag{4.4}$$

$$A_5 = \frac{k_1^2 k_2^2 k_3^2 k_4^2 k_5 [X]^2 Q^{05}}{k_1'^2 k_2'^2 k_3'^2 k_4'^2 k_5' [Y]^2 V'^2 T'^2}. \tag{4.5}$$

b) Ha $V' \geqq [V']^* Q^0$ és $T'=0$, akkor

$$A_i, T', T, pV_n, pV_n V_j = 0 \quad i = 1, 2, \ldots, 5; \quad j = 1, 2, \ldots, n-1,$$

$$V' \geqq [V']^* Q^0,$$

$$R \geqq 0.$$

c) Ha $[V']^* Q^0 > V' > 0$ és $T' > 0$, akkor $A_i = 0$, $(i = 1, 2, \ldots, 5)$ vagy a (4.1)—(4.5) egyenletek szerint,

$$T, R, pV_n V_i = 0 \quad i = 1, 2, \ldots, n-1,$$

$$[V']^* Q^0 > V' > 0,$$

$$pV_n \geqq 0,$$

$$T' > 0.$$

d) Ha $[V']^* Q^0 > V' > 0$ és $T'=0$, akkor

$$A_i, T', T, pV_n V_j = 0 \quad i = 1, 2, \ldots, 5; \quad j = 1, 2, \ldots, n-1,$$

$$R, pV_n \geqq 0,$$

$$[V']^* Q^0 > V' > 0.$$

e) Ha $V'=0$ és $T' > 0$, akkor

$$A_i, V', T, R = 0 \quad i = 1, 2, \ldots, 5,$$

$$pV_n, pV_n V_i \geqq 0 \quad i = 1, 2, \ldots, n-1,$$

$$T' > 0.$$

f) Ha $V'=0$ és $T'=0$, akkor

$$A_i, V', T', T = 0 \quad i = 1, 2, \ldots, 5,$$

$$R, pV_n, pV_nV_i \geqq 0 \quad i = 1, 2, \ldots, n-1.$$

4.1. TÉTEL. A (2.3) differenciálegyenlet-rendszer $\mathbf{u}=\mathbf{0}$, $S^0 \leqq S^1 \leqq 2S^0$ egyensúlyi helyzetei instabilisak.

Bizonyítás. Mint ismeretes [7], LJAPUNOVnak az első közelítés alapján való stabilitásvizsgálatról szóló tétele szerint: ha az A mátrixnak van pozitív valós részű sajátértéke, és $\mathbf{R}(\mathbf{x})=o(|\mathbf{x}|)$ $(|\mathbf{x}|\to 0)$, akkor a

$$\frac{d\mathbf{x}}{dt} = \mathbf{A}\mathbf{x}+\mathbf{R}(\mathbf{x})$$

differenciálegyenlet-rendszer $\mathbf{x}=\mathbf{0}$ megoldása instabilis. Hajtsuk végre az $S'=S-S^1$ transzformációt, ekkor a (2.3) differenciálegyenlet-rendszer $\mathbf{u}=\mathbf{0}$, $S=S^1$ megoldásának a transzformált rendszer $\mathbf{u}=\mathbf{0}$, $S'=0$ triviális megoldása felel meg. Esetünkben $\mathbf{A}\mathbf{x}$-be a (2.3) jobb oldalán levő, az állapothatározókban lineáris tagok tartoznak, $R_i(\mathbf{x})$-ben pedig $(i=1, 2, \ldots, n+10)$ a $C\dfrac{u_ju_k}{(S'+S^1)^{3/2}}$, illetve $C\dfrac{u_jS'}{(S'+S^1)^{3/2}}$ alakú tagok összege szerepel, ahol C konstans. Nyilvánvaló, hogy ilyen $\mathbf{R}(\mathbf{x})$ teljesíti a *Ljapunov-tétel* feltételét. Az A mátrix karakterisztikus polinomja:

$$-\lambda^7\left(-k_7'-k_9\frac{S^1}{Q}-\lambda\right)\det \mathbf{D},$$

ahol

$$\mathbf{D} = \begin{vmatrix} -k_1X-\lambda & k_1' & 0 & 0 & 2k_5 \\ k_1X & -k_1'-k_2-\lambda & k_2'Y & 0 & 0 \\ 0 & k_2 & -k_2'Y-k_3-\lambda & 0 & 0 \\ 0 & 0 & k_3 & -k_4-\lambda & 0 \\ 0 & 0 & 0 & k_4 & -k_5-\lambda \end{vmatrix}$$

és $Q=\dfrac{(S'+S^1)^{3/2}}{6\sqrt{\pi}}$.

A determináns meghatározásával az alábbi polinomot kapjuk:

$$-\lambda^7\left(-k_7'-k_9\frac{S^1}{Q}-\lambda\right)\{2k_1[X]k_2k_3k_4k_5+k_1[X]k_1'(k_2'[Y]+k_3+\lambda)(k_4+\lambda)(k_5+\lambda)+$$

$$+(k_1[X]+\lambda)k_2k_2'[Y](k_4+\lambda)(k_5+\lambda)-$$

$$-(k_1[X]+\lambda)(k_1'+k_2+\lambda)(k_2'[Y]+k_3+\lambda)(k_4+\lambda)(k_5+\lambda)\}.$$

Leolvasható, hogy a kapcsos zárójelben levő polinomban a konstans tag pozitív, a λ^5 együtthatója pedig -1 lesz. Mivel ez valós együtthatós polinom, ezért ebből az következik, hogy van legalább egy pozitív valós gyöke. Ezzel a tétel állítását igazoltuk.

A fejezet eredményei alapján állíthatjuk, hogy a vizsgált modellnek minden egyensúlyi helyzete olyan, hogy a három alrendszer valamelyike teljesen hiányzik. Tehát a chemotonnak nincs egyensúlyi helyzete, vagyis szerkezete kizárja a rendszer stagnálását. A 4.1. tétel szerint az itt tárgyal fiktív egyensúlyi helyzetek közeléből indított megoldások általában nem konvergálnak ezekhez az egyensúlyi helyzetekhez.

Köszönetnyilvánítás. Köszönetemet fejezem ki HATVANI LÁSZLÓnak értékes tanácsaiért és a kézirattal kapcsolatos kritikai észrevételeiért.

5. Függelék

A chemoton-elméletben szokásos jelölések helyett az egyszerűbb írásmód kedvéért újakat vezettem be a 2. fejezetben. Az alábbi táblázat tartalmazza az egyes állapothatározók kétféle jelölését:

$$\begin{array}{cccccccccccc} A_1 & A_2 & A_3 & A_4 & A_5 & V' & T' & T & R & pV_n & pV_nV_i \\ u_1 & u_2 & u_3 & u_4 & u_5 & u_6 & u_7 & u_8 & u_9 & u_{10} & u_{10+i} \end{array}$$

$$i = 1, 2, \ldots, n-1.$$

IRODALOM

[1] BÉKÉS, F., "Simulation of kinetics of proliferating chemical systems", *Biosystems* **7** (1975) 189—195.
[2] BÉKÉS, F., HIDVÉGI, M. and KORPÁDI, M., "Computer simulation of the time-dependent behaviour of chemotons, chemical super-systems having the criteria of life" in: *Proc. of the Symp. on Simulation of Systems in Biology and Medicine* Ed. *M. Kotva,* Vol. **2** (1979) 93—101.
[3] CSENDES, T., "A simulation study on the chemoton", *Kybernetes* **13** (1984) 79—85.
[4] GÁNTI, T., *Az élet principiuma* (Gondolat Könyvkiadó, Budapest, 1971, 1978).
[5] GÁNTI, T., "Organization of chemical reactions into dividing and metabolizing units: the chemotons", *Biosystems* **7** (1975) 15—21.
[6] GÁNTI, T., *A Theory of Biochemical Supersystems and its Application to the Artificial Biogenesis* (University Park Press, Baltimore, 1979).
[7] SZTYEPANOV, V. V., *A differenciálegyenletek tankönyve* (Tankönyvkiadó, Budapest, 1952).
[8] Вольперт, А. И., «Дифференциальные уравнения на графах», *Мат. сборник* **88** (1972) 578—588.
[9] Филиппов, А. Ф., «Дифференциальные уравнения с разрывной правой частью», *Мат. сборник* **51** (1960) 99—128.

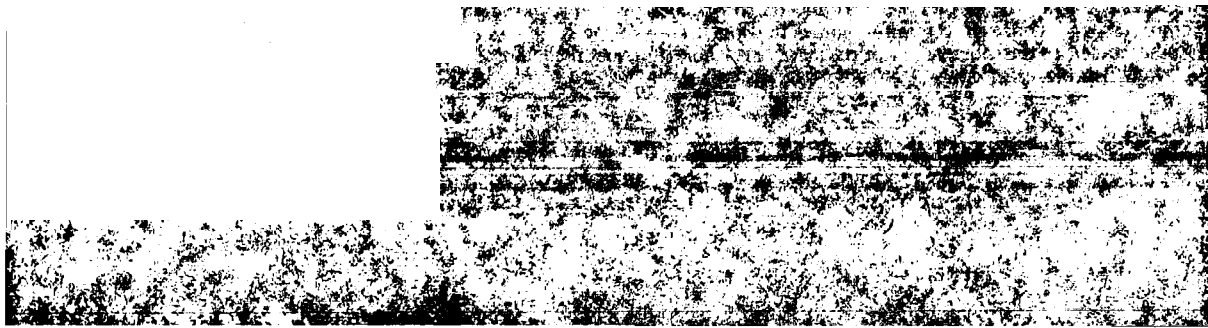
(Beérkezett: 1984. május 20.)

CSENDES TIBOR
JATE KALMÁR LÁSZLÓ KIBERNETIKAI LABORATÓRIUM
6720 SZEGED, ÁRPÁD TÉR 2.

ON THE MATHEMATICAL MODEL OF THE CHEMOTON

T. CSENDES

In this paper a mathematical model of a biological system, the chemoton is established. In the second and third chapter we examine the properties (existence, uniqueness, continuability, continuous dependence on the initial values and positivity) of the solutions of the differential systems in the model. In the fourth chapter the states of equilibrium are given and it is proved that some of them are unstable.



Alkalmazott Matematikai Lapok **11** (1985) 183—189

LINEÁRIS REGRESSZIÓ EGYÜTTHATÓINAK MAXIMUM LIKELIHOOD BECSLÉSE

HUHN EDIT

Szeged

A dolgozatban a szokásos lineáris modell ismeretlen paramétereinek maximum likelihood becslését vizsgáljuk abban az esetben, amikor a hibatagok *Gauss—ARMA sorozatot* alkotnak.

1. Bevezetés

Tekintsük az

$$\zeta(t) = \sum_{k=1}^{n} \alpha_k(t)\vartheta_k + \xi(t), \quad t = 0, \pm 1, \ldots \tag{1.1}$$

folyamatot, ahol $\vartheta = (\vartheta_1, \ldots, \vartheta_n)^T$ az ismeretlen paraméter $(-\infty < \vartheta_k < \infty, k = 1, \ldots, n)$ az $\alpha_1(t), \ldots, \alpha_n(t)$ ismert függvények. Legyen $\zeta_N = (\zeta(0), \ldots, \zeta(N-1))^T$, $\xi_N = (\xi(0), \ldots, \xi(N-1))^T$. Tegyük fel, hogy $E\{\xi_N\} = 0$ és $\operatorname{cov}(\xi_N \xi_N) = \operatorname{cov}(\xi_N, \xi_N) = \Sigma_N$. Amennyiben a Σ_N^{-1} mátrix nem ismert, akkor ϑ legkisebb négyzetes becslése a függő változó $\zeta(0), \ldots, \zeta(N-1)$ megfigyelt értékei alapján

$$\vartheta_N^* = (\alpha^T \alpha)^{-1} \alpha^T \zeta_N, \tag{1.2}$$

ahol $\alpha = \{\alpha_{ik}\}_{N \times n}$, $\alpha_{ik} = \alpha_k(i)$. Ez a ϑ_N^* becslés nyilván torzítatlan és kovariancia mátrixa

$$E\{(\vartheta_N^* - \vartheta)(\vartheta_N^* - \vartheta)^T\} = (\alpha^T \alpha)^{-1} \alpha^T \Sigma_N \alpha (\alpha^T \alpha)^{-1}. \tag{1.3}$$

Ha a Σ_N^{-1} mátrix ismert a ϑ *Gauss—Markov becslése* (lásd [1])

$$\hat{\vartheta}_N = (\alpha^T \Sigma_N^{-1} \alpha)^{-1} \alpha^T \Sigma_N^{-1} \zeta_N, \tag{1.4}$$

ami szintén torzítatlan és kovariancia mátrixa

$$E\{(\hat{\vartheta}_N - \vartheta)(\hat{\vartheta}_N - \vartheta)^T\} = (\alpha^T \Sigma_N^{-1} \alpha)^{-1}. \tag{1.5}$$

Ismert, hogy a $\hat{\vartheta}_N$ *Gauss—Markov becslés* optimális a ϑ összes lineáris torzítatlan becslései között abban az értelemben, hogy ha $\tilde{\vartheta}_N = C\zeta_N$ valamely torzítatlan becslése ϑ-nak, akkor a

$$\operatorname{cov}(\tilde{\vartheta}_N, \tilde{\vartheta}_N) - \operatorname{cov}(\hat{\vartheta}_N, \hat{\vartheta}_N)$$

ámtrix pozitív szemidefinit (lásd [1]).

Látható, hogy ha a $\xi(t)$ változók azonos eloszlásúak és függetlenek, akkor az (1.2) legkisebb négyzetes becslés és az (1.4) *Gauss—Markov becslés* megegyeznek. Ha a $\xi(t)$-k normális eloszlásúak, akkor az (1.4) alatti $\hat{\vartheta}_N$ egyúttal maximum likelihood becslése ϑ-nak, ugyanis ekkor a likelihood függvény

$$L(\zeta_N) = \frac{d\mu_\zeta^\vartheta}{d\mu_\zeta^0}(\zeta_N) = \exp\left\{\vartheta^T\alpha^T\Sigma_N^{-1}\zeta_N - \frac{1}{2}\vartheta^T\alpha^T\Sigma_N^{-1}\alpha\vartheta\right\}, \tag{1.6}$$

ahol $d\mu_\zeta^\vartheta/d\mu_\zeta^0$ a ζ_N^ϑ-nak megfelelő μ_ζ^ϑ mértéknek a μ_ζ^0 mérték szerinti *Radon—Nikodym deriváltja*, ahol a μ_ζ^0 a ζ_N^0 $(\vartheta=0)$ változók által generált mérték. Könnyen látható, hogy (1.4) maximalizálja (1.6)-ot. Ebben az esetben $\hat{\vartheta}_N$ nyilván normális eloszlású mivel ζ_N is az.

Gyakori az a feltevés, hogy $\xi(t)$ normális eloszlású *ARMA* (p, q) folyamat. LIPCER és SIRJAJEV [6] azt az esetet vizsgálták, amikor $\xi(t)$ *Gauss—ARMA* $(q-1, q)$ folyamat. Megmutatták, hogy a $\hat{\vartheta}_N$ torzítatlan és effektív becslése ϑ-nak, de a Σ_N^{-1} mátrixot explicite nem adták meg. GRENANDER és SZEGŐ [3] valamint GRENANDER és ROSENBLATT [2] $n=1$, $\alpha_1(t)=\vartheta e^{it\lambda_0}$, λ_0 ismert valós konstans esetén határozták meg a minimális szórású torzítatlan becslést.

Ismert, hogy az (1.3), illetve az (1.5) kovariancia mátrixok aszimptotikusan ekvivalensek (lásd [1]). Kis elemszám esetén a kisebb szórás miatt előnyösebb az (1.4) *Gauss—Markov becslés* alkalmazása. Ha $\xi(t)$ *Gauss—ARMA* $(p, 1)$, $MA(1)$, vagy $AR(p)$ folyamat, akkor Σ_N^{-1} elemeit meg tudjuk határozni (lásd [4]).

Például ha $\xi(t)$ $MA(1)$ folyamat, azaz

$$\xi(t) = \varepsilon(t) + b_1\varepsilon(t-1) \quad (t = 0, \pm 1, \ldots),$$

ahol $\varepsilon(t)$ normális eloszlású, $E\{\varepsilon(t)\}=0$, $E\{\varepsilon(t)\varepsilon(s)\}=\delta_{ts}\sigma_\varepsilon^2$, akkor $\hat{\varepsilon}(t)=$ $=E\{\varepsilon(t)|\xi(0), \ldots, \xi(t)\}$ és $\gamma(t)=E\{|\varepsilon(t)-\hat{\varepsilon}(t)|^2\}$ a *Kálmán szűrés* módszerével meghatározható. A szűrőegyenletek ebben az esetben:

$$\hat{\varepsilon}(t) = \frac{1}{\sigma_\varepsilon}[1+b_1^2\gamma(t-1)]^{-1}[\xi(t)-b_1\hat{\varepsilon}(t-1)], \quad \hat{\varepsilon}(0) = \frac{1}{\sigma_\varepsilon}\frac{1}{1+b_1^2}\xi(0),$$

$$\gamma(t) = 1-[1+b_1^2\gamma(t-1)]^{-1}, \quad \gamma(0) = \frac{b_1^2}{1+b_1^2}.$$

A fenti egyenleteket megoldva adódik, hogy $|b_1|\neq 1$ esetén

$$\gamma(t) = \frac{b_1^{2(t+1)}(b_1^2-1)}{b_1^{2(t+2)}-1},$$

$$\hat{\varepsilon}(t) = (-b_1)^t\frac{1-b_1^2}{1-b_1^{2(t+2)}}\xi(0) + \frac{1}{1-b_1^{2(t+2)}}\sum_{s=1}^{t}(-b_1)^{t-s}(1-b_1^{2(s+1)})\xi(s).$$

Tudjuk továbbá, hogy

$$\tilde{\varepsilon}(t) = \xi(t)-b_1\hat{\varepsilon}(t-1), \quad t = 1, 2, \ldots$$

normális eloszlású fehér zaj és $E\{\tilde{\varepsilon}(t)\}=0$, $E\{|\tilde{\varepsilon}(t)|^2\}=\sigma_\varepsilon^2+b_1^2\gamma(t-1)$ (lásd [6]).

A fentieket felhasználva a $\xi(0), \ldots, \xi(N-1)$ változók együttes sűrűségfüggvénye:

$$f(x_0, \ldots, x_{N-1}) = \left\{(2\pi)^N \sigma_\varepsilon^{2N} \frac{1-b_1^{2(N+1)}}{1-b_1^2}\right\}^{-1/2} \times$$

$$\times \exp\left\{-\frac{1}{2\sigma_\varepsilon^2}\left[\sum_{j=1}^{N-1} \frac{b_1^{2(j+1)}-1}{b_1^{2(j+2)}-1}\left(x_j - b_1 \sum_{s=0}^{j-1} M_s^{(j-1)} x_s\right)^2 + \frac{x_0^2}{1+b_1^2}\right]\right\},$$

$$M_s^{(j-1)} = (-b_1)^{j-1-s} \frac{1-b_1^{2(s+1)}}{1-b_1^{2(j+1)}}.$$

2. A maximum likelihood becslés tulajdonságai

Ha $\tilde{\vartheta}_N$ tetszőleges torzítatlan becslése ϑ-nak, akkor a *Cramer—Rao egyenlőtlenség* szerint

$$E\{(\tilde{\vartheta}_N - \vartheta)(\tilde{\vartheta}_N - \vartheta)^T\} \geqq I_N(\vartheta)^{-1}, \tag{2.1}$$

ahol

$$I_N(\vartheta) = E\left\{\left(\frac{\partial}{\partial \vartheta} \ln L(\zeta_N)\right)\left(\frac{\partial}{\partial \vartheta} \ln L(\zeta_N)\right)^T\right\} =$$

$$= E\{\alpha^T \Sigma_N^{-1}(\zeta_N - \alpha\vartheta)(\zeta_N - \alpha\vartheta)^T \Sigma_N^{-1}\alpha\} = \alpha^T \Sigma_N^{-1}\alpha.$$

Ha valamely $\tilde{\vartheta}_N$ torzítatlan becslés kovariancia mátrixa megegyezik a (2.1)-ben szereplő $I_N(\vartheta)^{-1}$-gyel, akkor az illető becslést effektívnek nevezzük (lásd [1]). (1.4) alapján látható, hogy $\hat{\vartheta}_N$ effektív becslése ϑ-nak.

2.1. TÉTEL. Tegyük fel, hogy az (1.1)-ben szereplő $\alpha_k(t)$ függvények a

$$\alpha_k(t) = \int_{-\pi}^{\pi} e^{it\lambda} d\mu_k(\lambda) \quad (k = 1, \ldots, n) \tag{2.2}$$

Lebesgue—Stieltjes integrállal állíthatók elő, ahol a $\mu_k(\lambda)$-k a $(-\pi, \pi)$-n korlátos, nem csökkenő, tiszta ugró függvények (ha komplex értékűek, akkor a valós és képzetes részeikről tesszük fel ugyanezt), és a $\mu_i(\lambda)$-nak, illetve a $\mu_k(\lambda)$-nak megfelelő mértékek szingulárisak, ha $i \neq k$. Ekkor az (1.4) alatti $\hat{\vartheta}_N$ konzisztens becslése ϑ-nak.

Bizonyítás. A fenti állítás bizonyításához be kell látni, hogy

$$\lim_{N\to\infty} E\{(\hat{\vartheta}_N - \vartheta)(\hat{\vartheta}_N - \vartheta)^T\} = \lim_{N\to\infty} I_N(\vartheta)^{-1} = 0.$$

Először megmutatjuk, hogy ϑ-nak létezik lineáris, aszimptotikusan torzítatlan, konzisztens becslése.

Tegyük fel, hogy ϑ-t a $\zeta(t)$ folyamat $t = -N, \ldots, 0, 1, \ldots, N$ időpontokban megfigyelt $\zeta = (\zeta(-N), \ldots, \zeta(0), \ldots \zeta(N))^T$ értékei alapján akarjuk becsülni. Tekintsük a

$$\tilde{\vartheta}_N = C_N \zeta, \quad C_N = \{c_{ik}^{(N)}\}_{n \times (2N+1)}$$

lineáris becslést. Ahhoz, hogy $\tilde{\vartheta}_N$ aszimptotikusan torzítatlan becslés legyen a

$$\lim_{N\to\infty} \mathbf{C}_N\boldsymbol{\alpha} = \mathbf{I}_{n\times n},$$

azaz a

$$\lim_{N\to\infty} \sum_{t=-N}^{N} c_{jt}^{(N)}\alpha_k(t) = \lim_{N\to\infty} \int_{-\pi}^{\pi} P_N^{(j)}(e^{i\lambda})\, d\mu_k(\lambda) = \delta_{jk}, \quad j, k = 1, \ldots, n$$

egyenlőségnek kell teljesülnie, ahol $P_N^{(j)}(z) = \sum_{t=-N}^{N} c_{jt}^{(N)} z^t$. Legyen a $\tilde{\vartheta}_N$ becslés kovariancia mátrixa

$$\boldsymbol{\Gamma}_N = \{\gamma_{ik}^{(N)}\} = E\{(\tilde{\vartheta}_N - \mathbf{C}_N\boldsymbol{\alpha}\vartheta)(\tilde{\vartheta}_N - \mathbf{C}_N\boldsymbol{\alpha}\vartheta)^T\}.$$

A kérdés az, hogy $\lim_{N\to\infty} \boldsymbol{\Gamma}_N = 0$ — vagy ami ezzel ekvivalens: $\lim_{N\to\infty} \gamma_{kk}^{(N)} = 0$, $k = 1, \ldots, n$ — milyen feltételek mellett teljesül.

$$\boldsymbol{\Gamma}_N = \mathbf{C}_N E\{\boldsymbol{\xi}\boldsymbol{\xi}^T\}\mathbf{C}_N^T$$

és

$$\gamma_{kk}^{(N)} = \sum_{j=-N}^{N} \sum_{l=-N}^{N} c_{kj}^{(N)} \overline{c_{kl}^{(N)}} \int_{-\pi}^{\pi} e^{i(j-l)\lambda} dF(\lambda) = \int_{-\pi}^{\pi} |P_N^{(k)}(e^{i\lambda})|^2 dF(\lambda),$$

ahol $F(\lambda)$ a $\xi(t)$ folyamat spektrális eloszlásfüggvénye.

Vegyük a $(-\pi, \pi)$ intervallumnak egy B_m: $-\pi < \lambda_1^{(m)} < \ldots < \lambda_m^{(m)} < \pi$ beosztását. Ekkor a *Cauchy—Schwartz egyenlőtlenség* alkalmazásával kapjuk, hogy tetszőleges $\delta > 0$ és $\varepsilon > 0$ esetén, ha N elég nagy és a B_m beosztás elég sűrű,

$$1 - \delta \leqq \left[\int_{-\pi}^{\pi} |P_N^{(k)}(e^{i\lambda})|\, |d\mu_k(\lambda)|\right]^2 \leqq \left[\sum_{\nu=1}^{m-1} |P_N^{(k)}(e^{i\lambda_\nu^{(m)}})|\, |\Delta\mu_k(\lambda_\nu^{(m)})|\right]^2 + \varepsilon \leqq$$

$$\leqq \sum_{\nu=1}^{m-1} |P_N^{(k)}(e^{i\lambda_\nu^{(m)}})|^2 \Delta F(\lambda_\nu^{(m)}) \sum_{j=1}^{m-1} \frac{|\Delta\mu_k(\lambda_j^{(m)})|^2}{\Delta F(\lambda_j^{(m)})} + \varepsilon,$$

ahol $\Delta\mu_k(\lambda_\nu^{(m)})$, $\Delta F(\lambda_\nu^{(m)})$ a $\mu_k(\lambda)$, illetve az $F(\lambda)$ változása a ν-edik részintervallumon.

$$\lim_{m\to\infty} \sum_{j=1}^{m-1} |P_N^{(k)}(e^{i\lambda_j^{(m)}})|^2 \Delta F(\lambda_j^{(m)}) = \int_{-\pi}^{\pi} |P_N^{(k)}(e^{i\lambda})|^2 dF(\lambda) < \infty,$$

és a

$$H_m^{(k)} = \sum_{j=1}^{m-1} \frac{|\Delta\mu_k(\lambda_j^{(m)})|^2}{\Delta F(\lambda_j^{(m)})}$$

sorozat nem csökkenő. Ez utóbbi sorozat véges vagy végtelen határértéke legyen

$$H^{(k)} = \int_{-\pi}^{\pi} \frac{|d\mu_k(\lambda)|^2}{dF(\lambda)}, \tag{2.3}$$

ami az ún. *Hellinger integrál*. A fentiekből látható, hogy

$$\liminf_{N\to\infty} \gamma_{kk}^{(N)} \geqq \frac{1}{H^{(k)}}, \quad k = 1, \ldots, n, \tag{2.4}$$

tehát csak akkor várható, hogy konzisztens becslés létezik, ha $H^{(k)} = \infty$ teljesül.

A $\mu_k(\lambda)$ függvényekre tett feltevéseink mellett a (2.3) integrálok divergálnak, mivel esetünkben az $F(\lambda)$ abszolút folytonos. Ezek után megmutatjuk, hogy (2.4)-ben az egyenlőség teljesül. Tekintsük a *következő* lépcsős függvényeket:

$$\varphi_m^{(k)}(\lambda) = \frac{1}{H_m^{(k)}} \frac{\overline{\Delta\mu_k(\lambda_\nu^{(m)})}}{\Delta F(\lambda_\nu^{(m)})}, \quad \text{ha} \quad \lambda_\nu^{(m)} < \lambda \leqq \lambda_{\nu+1}^{(m)}.$$

Ezeket a függvényeket a $\lambda_\nu^{(m)}$ osztópontok tetszőlegesen kicsiny környezeteit kivéve egyenletesen közelíthetjük trigonometrikus polinomokkal. Válasszuk a B_m beosztást úgy, hogy a $\lambda_\nu^{(m)}$ osztópontok ne essenek egybe a $\mu_k(\lambda)$-k szakadási helyeivel. Ekkor

$$\int_{-\pi}^{\pi} \varphi_m^{(k)}(\lambda) d\mu_j(\lambda) = \delta_{jk} \tag{2.5}$$

és

$$\int_{-\pi}^{\pi} |\varphi_m^{(k)}(\lambda)|^2 dF(\lambda) = \frac{1}{H_m^{(k)}}. \tag{2.6}$$

Legyen $P_{N_m}^{(k)}(z) = \sum_{j=-N_m}^{N_m} c_{kj}^{(N_m)} z^j$ olyan, hogy

$$|\varphi_m^{(k)}(\lambda) - P_{N_m}^{(k)}(e^{i\lambda})| < \varepsilon_m \tag{2.7}$$

és itt $m \to \infty$, $\varepsilon_m \to 0$ esetén $N_m \to \infty$.

Tekintsük a

$$\tilde{\vartheta}_k = \sum_{j=-N_m}^{N_m} c_{kj}^{(N_m)} \zeta(j), \quad \tilde{\vartheta}_{N_m} = C_{N_m}\zeta \tag{2.8}$$

becslést. Ekkor

$$E\{\tilde{\vartheta}_{N_m}\} = \mathbf{C}_{N_m}\boldsymbol{\alpha}\vartheta \quad \text{és} \quad \mathbf{C}_{N_m}\boldsymbol{\alpha} \to \mathbf{I}_{n\times n}, \tag{2.9}$$

ami (2.5) és (2.7) alapján könnyen látható. A $\boldsymbol{\Gamma}_{N_m}$ kovariancia mátrix egy főátlóbeli eleme

$$\gamma_{kk}^{(N_m)} = \sum_{j=-N_m}^{N_m} \sum_{l=-N_m}^{N_m} c_{kj}^{(N_m)} \overline{c_{kl}^{(N_m)}} \int_{-\pi}^{\pi} e^{i(j-l)\lambda} dF(\lambda) = \int_{-\pi}^{\pi} |P_{N_m}^{(k)}(e^{i\lambda})|^2 dF(\lambda)$$

alakú és (2.6), (2.7) szerint, valamint figyelembe véve, hogy $\lim_{m\to\infty} H_m^{(k)} = \infty$, adódik, hogy

$$\gamma_{kk}^{(N_m)} \to 0, \quad \text{azaz} \quad \lim_{N_m\to\infty} \boldsymbol{\Gamma}_{N_m} = \mathbf{0}.$$

Az előzőekből látható, hogy a $\mathbf{C}_{N_m}$ mátrix elemei ϑ-tól függetlenek. Így (2.9) alapján a $\tilde{\vartheta}_{N_m}$ becslés torzítása $\boldsymbol{\Delta}_{N_m}\vartheta$ alakú, ahol $\boldsymbol{\Delta}_{N_m} = \{\Delta_{ij}^{(N_m)}\}$ ϑ-tól független és $\Delta_{ij}^{(N_m)} \to 0$,

$N_m \to \infty$. Tehát $\tilde{\vartheta}_{N_m}$ torzítatlan becslése $\vartheta + \Delta_{N_m}\vartheta$-nak és így a *Cramer—Rao egyenlőtlenség* szerint

$$\Gamma_{N_m} \geqq \mathbf{D}(\vartheta)\mathbf{I}_{N_m}(\vartheta)^{-1}\mathbf{D}(\vartheta)^T,$$

ahol

$$\mathbf{D}(\vartheta) = \{D_{ij}(\vartheta)\}_{n\times n},$$

$$D_{ij}(\vartheta) = \frac{\partial}{\partial \vartheta_j}\left[\vartheta_i + \sum_{k=1}^{n} \Delta_{ik}^{(N_m)}\vartheta_k\right] = \delta_{ij} + \Delta_{ij}^{(N_m)},$$

azaz

$$\Gamma_{N_m} \geqq (\mathbf{I} + \Delta_{N_m})I_{N_m}\vartheta)^{-1}(\mathbf{I} + \Delta_{N_m})^T \geqq 0,$$

amiből $\Gamma_{N_m} \to 0$ és $\Delta_{N_m} \to 0$ miatt következik, hogy

$$\mathbf{I}_{N_m}(\vartheta)^{-1} \to 0, \quad N_m \to \infty.$$

Ez pedig éppen a maximum likelihood becslés konzisztenciáját jelenti a vizsgált esetben. Példák:

a) Legyen

$$\zeta(t) = \vartheta e^{it\lambda_0} + \xi(t),$$

azaz

$$\alpha(t) = e^{it\lambda_0} = \int_{-\pi}^{\pi} e^{it\lambda}\, d\mu(\lambda)$$

és

$$\mu(\lambda) = \begin{cases} 0, & -\pi \leqq \lambda \leqq \alpha_0 \\ 1, & \lambda_0 < \lambda \leqq \pi, \end{cases}$$

és λ_0 adott valós konstans (lásd [2]).

b) Másik, gyakorlati szempontból fontos eset az ún. trigonometrikus regresszió, amikor $\zeta(t)$ (1.1) alakú és

$$\alpha_1(t) = 1,$$

$$\alpha_k(t) = \cos \alpha_k t, \quad k = 2, \ldots, n,$$

és a λ_k-k különböző ismert pozitív állandók. Az $\alpha_k(t)$ függvények (2.2) alakba írhatók a következő $\mu_k(\lambda)$-ákkal:

$$\mu_1(\lambda) = \begin{cases} 0, & -\pi \leqq \lambda \leqq 0, \\ 1, & 0 < \lambda \leqq \pi, \end{cases}$$

$$\mu_k(\lambda) = \begin{cases} 0, & -\pi \leqq \lambda \leqq -\lambda_k, \\ 1/2, & -\lambda_k < \lambda \leqq \lambda_k, \\ 1, & \lambda_k < \lambda \leqq \pi. \end{cases}$$

Könnyen látható, hogy a fenti esetekben teljesülnek a $\hat{\vartheta}_N$ konzisztenciáját biztosító feltételek.

Végül köszönetet mondok ARATÓ MÁTYÁSnak a dolgozat elkészítése során nyújtott segítségéért.

IRODALOM

[1] ANDERSON, T. W., *The Statistical Analysis of Time Series* (John Wiley & Sons, New York—London—Sidney—Toronto, 1971).
[2] GRENANDER, U. and ROSENBLATT, M., *Statistical Analysis of Stationary Time Series* (Almqvist & Wiksell, Stockholm, 1956).
[3] GRENANDER, U. and SZEGŐ, P., *Toeplitz Forms and Their Applications* (University of California Press, Berkeley and Los Angeles, 1958).
[4] HUHN, E., „ARMA folyamatok egzakt sűrűségfüggvénye", *Alk. Mat. Lapok* **10** (1983) megjelenés alatt.
[5] ZACKS, S., *The Theory of Statistical Inference* (John Wiley, New York, 1971).
[6] Липцер, Р. Ш., Ширяев, А. Н., *Статистика случайных процессов* (Наука, 1974).

(Beérkezett: 1984. június 15.)

(Átdolgozva beérkezett: 1984. szeptember 15.)

HUHN EDIT
SZEGEDI ORVOSTUDOMÁNYI EGYETEM
SZÁMÍTÁSTECHNIKAI KÖZPONT
6720 SZEGED, PÉCSI U. 4/A.

MAXIMUM LIKELIHOOD ESTIMATION OF LINEAR REGRESSION

E. HUHN

The usual linear regression model is considered. The disturbances are assumed to be generated by a *Gaussian autoregressive-moving average process* and the properties of the maximum likelihood estimator are studied.

Alkalmazott Matematikai Lapok **11** (1985) 191—200

BINÁRIS SOROK PÁRHUZAMOS KISZOLGÁLÁSA

IVÁNYI ANTAL ÉS PERGEL JÓZSEF

Budapest

Két típusú igényt tartalmazó prioritásos sorok kiszolgálását vizsgáljuk. Meghatározzuk a kiszolgálási állapotokat jellemző *homogén Markov-lánc* ergodikus eloszlását abban az esetben, amikor a kiszolgálási igények azonos eloszlású, független valószínűségi változók. A sorok számának és az igénytípusok előfordulási valószínűségének függvényében megadjuk az időegységenként kiszolgált igények számának várható értékét. Arra a meglepő eredményre jutunk, hogy ha az igények előfordulási valószínűségei különbözőek, akkor a sorok számának növelésekor az időegységenként kiszolgált igények száma nem kettőhöz, hanem annál kisebb értékhez tart.

1. Bevezetés

Nagyteljesítményű számítógépek működésének modellezése során merült fel a következő feladat [1, 2], melyet a könnyebb érthetőség érdekében a következőképpen fogalmazunk meg.

Tegyük fel, hogy egy benzinkútnál normál és szuper benzin kapható. A gépkocsik a fontosságuknak megfelelő sorban várnak a tankolásra — például az első sorban a mentőautók, a másodikban a diplomáciai, a harmadikban pedig a magán gépkocsik.

A gépkocsik tankolása a diszkrét 0, 1, 2, ... időpontokban (például 6^{00}, 6^{05}, 6^{10}-kor stb.) kezdődik és egy időegységet (például öt percet) vesz igénybe.

Mivel a kútnál egy-egy csap van a normál és a szuper benzin részére, ezért időegységenként legfeljebb 1-1 gépkocsi kaphat normál, ill. szuper benzint.

A gépkocsik fontosságát figyelembe véve az első sor elején álló gépkocsit minden időegységben kiszolgálják. Lehetőleg vele együtt kiszolgálják a mögötte álló gépkocsit is. Ha ez nem lehetséges (mert ugyanolyan típusú benzinre vár, mint az előtte álló), akkor a második, ... sor elején álló gépkocsit szolgálják ki. Ily módon általában két, esetenként azonban (ha az első két mentőautó, az első diplomáciai és magángépkocsi azonos típusú benzinre vár) csak egy gépkocsit szolgálnak ki egy időegység alatt.

A gépkocsisor állapotát minden időegységben a kiszolgálási vektor segítségével jellemezzük, amely a sorok elején álló gépkocsi által igényelt benzintípust adja meg.

Dolgozatunkban meghatározzuk a kiszolgálási állapotok határeloszlását tetszőleges számú gépkocsisor esetén — feltételezve, hogy a gépkocsik benzinigénye azonos eloszlású, független valószínűségi változókkal adható meg.

Ennek az eloszlásnak a segítségével azután meghatározzuk a tankolási sebességet, amit lényegében az időegységenként várhatóan kiszolgált gépkocsik számával definiálunk.

Végül arra a némiképp meglepő eredményre jutunk, hogy ha a gépkocsik különböző valószínűséggel igényelnek normál, ill. szuper benzint, akkor a kiszolgálási sebesség értéke a sorok számának növelésekor nem a természetesnek látszó határértékhez, azaz kettőhöz tart, hanem annál kisebb értékhez.

2. A feladat megfogalmazása

Legyen $r \geqq 1$ és legyenek

$$\begin{matrix} g_{11}, & g_{12}, & \cdots \\ \vdots & & \\ g_{r1}, & g_{r2}, & \cdots \end{matrix}$$

kiszolgálandó igényekből álló végtelen bináris sorozatok, azaz $g_{ij} \in \{0, 1\}$ $(i = 1, \ldots, r; j = 1, 2, \ldots)$.

A sorozatok elemeit a következő algoritmus szerint szolgáljuk ki [3, 4].

A kiszolgálási algoritmus definíciója.

1. lépés. Legyen $t = 1$.

2. lépés. Legyen $\mathbf{K}_t = (g_{11}, \ldots, g_{r1})$.

3. lépés. A t időpontban három eset lehetséges:

a) ha $g_{12} \neq g_{11}$, akkor a g_{11} és g_{12} igényeket szolgáljuk ki;

b) ha $g_{12} = g_{11}$ és van olyan k index, amelyre $1 \leqq k < r$, $g_{12} = g_{11} = g_{21} = \ldots = g_{k1}$, $g_{k+1,1} \neq g_{11}$, akkor a g_{11} és $g_{k+1,1}$ igényeket szolgáljuk ki;

c) ha $g_{12} = g_{11} = g_{21} = \ldots = g_{r1}$, akkor a g_{11} igényt szolgáljuk ki.

4. lépés. Elhagyjuk a kiszolgált igényeket, a megmaradó igények második indexét az i-edik sorban $(i = 1, \ldots, r)$ annyival csökkentjük, ahány igényt a t-edik időegységben az i-edik sorban kiszolgáltunk.

5. lépés. Hozzáadunk t-hez egyet, és a 2. lépéstől folytatjuk a kiszolgálást.

Legyenek most

$$\begin{matrix} \xi_{11}, & \xi_{12}, & \cdots \\ \vdots & & \\ \xi_{r1}, & \xi_{r2}, & \cdots \end{matrix} \tag{2.1}$$

egymástól független valószínűségi változók, melyek közös eloszlása

$$P(\xi_{ij} = 0) = p \quad \text{és} \quad P(\xi_{ij} = 1) = q, \tag{2.2}$$

ahol $i = 1, \ldots, r$; $j = 1, 2, \ldots$; $0 < p < 1$; $q = 1 - p$.

Tegyük fel, hogy a (2.1) valószínűségi változók realizációit a fenti algoritmussal szolgáljuk ki.

Legyen η_t $(t = 1, 2, \ldots)$ az a valószínűségi változó, amelyet a K_t kiszolgálási állapot lehetséges értékeinek eloszlása határoz meg.

Nem nehéz megmutatni, hogy ebben az esetben az $\eta_1, \eta_2, \ldots$ sorozat *homogén ergodikus Markov-lánc*. Ugyanis, ha $\eta_t = (j_1, \ldots, j_r)$, akkor

$$P(g_{12} = g_{13} = \ldots = g_{1,r+2} = g_{22} = g_{32} = \ldots = g_{r2} = 0) = p^{2^r},$$

és ekkor $\eta_{t+r}=(0, 0, \ldots, 0)$, azaz a csupa nullából álló kiszolgálási állapot r lépés alatt bármely állapotból legalább $p^{2^r}>0$ valószínűséggel elérhető, és ez a *Markov-tétel* [5] szerint elegendő $\eta_1, \eta_2, \ldots$ ergodicitásához.

Egy korábbi dolgozatunkban [4] a $p=q=0,5$ esetben meghatároztuk a kiszolgálási állapotok ergodikus eloszlását. Most ugyanezt tetszőleges $0<p<1$ értékre előállítjuk.

3. Az ergodikus valószínűségek néhány tulajdonsága

Ha r sorozatot dolgozunk fel, akkor 2^r különböző kiszolgálási állapot lehetséges.

Jelöljük A_i-vel $(i=1, \ldots, r)$ azt az i hosszúságú bináris sorozatot, melynek első $i-1$ eleme 0, az i-edik eleme pedig egyes; legyen B_i az az i hosszúságú sorozat, amelyben $i-1$ egyest egy nulla követ. Legyen A_{jk} $(1 \leqq j<k \leqq r)$ olyan k hosszúságú bináris sorozat, amelyben a j-edik és k-adik helyeken egyes áll, a többi helyen nulla, és B_{jk} olyan k hosszúságú sorozat, amelyben a j-edik és k-adik helyen nulla, a többi helyen egyes áll.

Legyen Γ_k^r az $r-k$ hosszúságú bináris sorozatok halmaza, és G_k ezen halmaz egy eleme.

Használni fogjuk az 1^i és 0^i jelölést, ahol 1^i az i darab egyesből, a 0^i pedig az i darab nullából álló sorozatot jelenti.

Az A_kG_k és B_kG_k típusú jelölés a megfelelő bináris sorozatok konkatenációját jelöli. A G_k sorozatokat mindig k hosszúságú sorozatok r hosszúságú sorozattá való kiegészítésére fogjuk használni.

Az általánosság megszorítása nélkül feltehető, hogy $p \geqq q$. Vezessük be a $p^2/q^2=t$ és a $T_i=1+t+\ldots+t^i$ $(i=0, 1, \ldots, r)$ jelöléseket.

Az ergodikus valószínűségeket Q-val jelölve igaz a következő.

1. LEMMA. Minden $i=1, \ldots, k (1 \leqq k \leqq r)$ indexre fennáll, hogy minden $G_i \in \Gamma_i^r$ sorozatra

$$Q(A_iG_i) = r_iQ(1^iG_i), \tag{3.1}$$

$$Q(B_iG_i) = s_i(Q0^iG_i), \tag{3.2}$$

ahol

$$r_i = \frac{t^{i-1}}{T_{i-1}}, \tag{3.3}$$

$$s_i = \frac{1}{T_{i-1}}. \tag{3.4}$$

Bizonyítás. Ha $i=1$, akkor (3.1) és (3.2) alakja $Q(1G_i)=r_1Q(1G_i)$, ill. $Q(0G_i)=s_1Q(1G_i)$, azaz $r_1=s_1=1$, ahogyan az (3.3)-ból és (3.4)-ből is adódik.

Tegyük fel most, hogy $k \geqq 2$, továbbá (3.1), (3.2), (3.3) és (3.4) fennáll $i=1, 2, \ldots, (k-1)$-re. Megmutatjuk, hogy akkor $i=k$-ra is teljesülnek.

Először egy rekurzív képletet vezetünk le az r_i és s_i értékekre, majd feloldjuk a rekurziót.

Mivel az A_kG_k kiszolgálási állapothoz az A_kG_k és az $A_{jk}G_k$ $(j=1, \ldots, k-1)$ állapotokból juthatunk, mégpedig az előbbiből $2pq$ (az első sorozatban 1, majd nulla

következik, vagy az első sorozatban 0, a k-adik sorozatban 1 következik), az utóbbiakból p^2 (az első és a j-edik elem után is 0 következik) valószínűséggel, így szükségképpen

$$Q(A_kG_k) = 2pqQ(A_kG_k)+p^2\sum_{j=1}^{k-1} Q(A_{jk}G_k).$$

Mivel $1=p^2+2pq+q^2$, így innen

$$(p^2+q^2)Q(A_kG_k) = p^2\sum_{j=1}^{k-1} Q(A_{jk}G_k). \tag{3.5}$$

Hasonlóképpen kapható

$$(p^2+q^2)P(B_kG_k) = q^2\sum_{j=1}^{k-1} P(B_{jk}G_k). \tag{3.6}$$

Legyen most $k\geqq 3$, $1\leqq j\leqq k-2$, és adott A_{jk} és G_k esetén G_{jk} az az $r-j$ hosszúságú sorozat, amelyben $k-1-j$ darab nulla után egy egyes, majd G_k elemei következnek. Ekkor $Q(A_{jk}G_k)=Q(A_jG_{jk})$. Alkalmazva (3.1)-et, az indukciós hipotézis alapján $Q(A_jG_{jk})=r_jQ(B_{j+1}G_{j+1,k})$, ahol $G_{j+1,k}$-ban $k-2+j$ darab nulla után egy egyes, majd G_k elemei következnek. Most (3.2)-t felhasználva

$$Q(A_{jk}G_k) = r_js_{j+1}Q(A_kG_k) \quad (j = 1, \ldots, k-2). \tag{3.7}$$

Hasonló módon

$$Q(B_{jk}G_k) = s_jr_{j+1}Q(A_kG_k) \quad (j = 1, \ldots, k-2). \tag{3.8}$$

Ha $j=k-1$, akkor ugyanilyen módon kapható

$$Q(A_{k-1,k}G_k) = r_{k-1}Q(1^kG_k), \tag{3.9}$$

$$Q(B_{k-1,k}G_k) = s_{k-1}Q(0^kG_k). \tag{3.10}$$

(3.7)-et $j=1, \ldots, (k-2)$-re és (3.9)-et behelyettesítve (3.5)-be, és az egyenletet átrendezés után (3.1)-gyel összehasonlítva kapjuk, hogy

$$r_k = \frac{r_{k-1}p^2}{p^2+q^2-p^2\sum_{j=1}^{k-2} r_js_{j+1}} \quad (k = 2, \ldots, r). \tag{3.11}$$

(Itt és a következő képletben a szumma értéke, $k=2$-re nulla.)

Hasonlóképpen (3.8), (3.10) és (3.6) alapján

$$s_k = \frac{s_{k-1}q^2}{p^2+q^2-q^2\sum_{j=1}^{k-2} s_jr_{j+1}} \quad (k = 2, \ldots, r). \tag{3.12}$$

Felhasználva, hogy $r_1=s_1=1$, (3.11)-ből és (3.12)-ből azt kapjuk, hogy $r_2=\frac{p^2}{p^2+q^2}=\frac{t}{t+1}=$ és $s_2=\frac{q^2}{p^2+q^2}=\frac{1}{t+1}$.

Tehát (3.3) és (3.4) fennáll $i=1$-re és $i=2$-re. Megmutatjuk, hogy ha teljesülnek $i=1, \ldots, (k-1)$-re, akkor $i=k$-ra is. Ehhez felhasználjuk, hogy $d=0, 1, \ldots,$ ese-

tén fennáll a következő azonosság:

$$\frac{t^{d+1}}{T_d T_{d+1}} = \frac{1}{T_d} - \frac{1}{T_{d+1}}. \tag{3.13}$$

(3.12) számlálóját és nevezőjét osztva q^2-tel, az $s_j r_{j+1}$ szorzatokat és s_{k-1}-et az indukciós feltevés és (3.17) alapján helyettesítve

$$s_k = \frac{1/T_{k-2}}{t+1-\sum_{j=1}^{k-2} t^j/T_{j-1}+t^{j+1}/T_j}.$$

Elvégezve a nevezőben az összevonásokat $s_k = \frac{1/T_{k-1}}{1+t^{k-1}/T_{k-2}}$, ahonnan a törtet bővítve a kívánt $s_k = \frac{1}{(1+\ldots+t^{k-2})+t^{k-1}}$ formula adódik.

Hasonlóképpen vezethető le (3.11) és (3.12) alapján az r_k-ra vonatkozó $r_k = t^{k-1}/T_{k-1}$ összefüggés.

Most meghatározzuk a 0^k és 1^k alakú kiszolgálási vektorok ergodikus valószínűségeit.

2. LEMMA. Minden $i=1, 2, \ldots$ indexre igaz

$$Q(0^i) = t^i/T_i \quad \text{és} \quad Q(1^i) = 1/T_i. \tag{3.14}$$

Bizonyítás. Először vizsgáljuk meg egyetlen sorozat feldolgozását. Ekkor két lehetséges feldolgozási állapot van, amelyek egy nullát illetve egy egyest tartalmaznak. Az átmenetvalószínűségek mátrixa:

	0	1
0	p^2+2pq	q^2
1	p^2	q^2+2pq

1. táblázat. Átmenetvalószínűségek $r=1$ esetén

Az 1. táblázat alapján felírt

$$Q(0^1) = (p^2+2pq)Q(0^1)+p^2Q(1^1),$$
$$Q(0^1)+Q(1^1) = 1$$

egyenletrendszert megoldva $Q(0^1)=\frac{t}{t+1}$, $Q(1^1)=\frac{1}{t+1}$, azaz a (3.14) egyenlőtlenségek $i=1$-re teljesülnek.

Minden $j=1, 2, \ldots$ indexre igazak a

$$Q(0^j) = Q(0^{j+1})+Q(0^j 0^1)$$

és

$$Q(1^j) = Q(1^{j+1})+Q(0^j 1^1)$$

rekurzív formulák. Mivel $0^j 1^1 = A_{j+1}$, és $1^j 0^1 = B_{j+1}$, így az 1. lemmát alkalmazva

$i=j+1=r$ esetén

$$Q(0^j) = Q(0^{j+1})+r_{j+1}Q(1^{j+1}),$$
$$Q(1^j) = Q(1^{j+1})+s_{j+1}Q(0^{j+1}).$$

Tegyük fel, hogy a (3.14) egyenlőtlenségek teljesülnek $i=1, \ldots, j$-re, és $j<r$. Megmutatjuk, hogy ekkor $i=(j+1)$-re is fennállnak.

Helyettesítsük $Q(0^j)$ és $Q(1^j)$ értékét az indukciós feltevés alapján (3.14) szerint, r_{j+1}-et és s_{j+1}-et pedig az 1. lemma alapján (3.3) és (3.4) szerint. Ekkor

$$\frac{t^j}{T_j} = Q(0^{j+1})+\frac{t^j}{T_j}Q(1^{j+1}), \tag{3.15}$$

$$\frac{1}{T_j} = Q(1^{j+1})+\frac{1}{T_j}Q(0^{j+1}). \tag{3.16}$$

Megoldva a (3.15)—(3.16) egyenletrendszert azt kapjuk, hogy

$$Q(1^{j+1}) = \frac{T_j-t^j}{(T_j)^2-t^j} = \frac{T_{j-1}}{(T_j)^2-t^j}, \tag{3.17}$$

és

$$Q(0^{j+1}) = \frac{t^j(T_j-1)}{(T_j)^2-t^j} = \frac{t^{j+1}T_{j-1}}{(T_j)^2-t^j}. \tag{3.18}$$

Külön indukcióval belátható, hogy $(T_j)^2-t^j=T_{j-1}T_{j+1}$. Ezt felhasználva (3.17)-ből $Q(1^{j+1})=1/T_{j+1}$, és (3.18)-ból $Q(0^{j+1})=t^{1+j}/T_{j+1}$.

4. Az ergodikus eloszlás

Az eddig megismert összefüggések segítségével most már tetszőleges kiszolgálási vektor ergodikus valószínűségét ki tudjuk számítani.

1. TÉTEL. Legyen $(j_1, \ldots, j_r)$ tetszőleges r-dimenziós ($r\geqq 1$) kiszolgálási állapot. Ekkor

$$Q(j_1, \ldots, j_r) = \frac{(t^r)^{1-j_r}}{T_r} \prod_{\substack{2\leq u\leq r \\ j_{u-1}=0 \\ j_u=1}} \frac{t^u}{T_{u-1}} \prod_{\substack{2\leq v\leq r \\ j_{v-1}=1 \\ j_v=0}} \frac{1}{T_{v-1}}, \tag{4.1}$$

ahol $t=p^2/q^2$, $T_i=1+t+\ldots+t^i$ $(i=1, 2, \ldots, r)$.

Ha egy kiszolgálási állapotban a j_k $(k=2, \ldots, r)$ érték a j_{k-1} értéktől különbözik, akkor azt mondjuk, hogy az állapot k-adik helyén váltás van: ha $j_{k-1}=0$ és $j_k=1$, akkor $0-1$ típusú, ellenkező esetben $1-0$ típusú váltásról beszélünk. Azt is mondhatjuk, hogy a váltáshoz az első esetben r_k, a második esetben s_k tartozik.

A tételben szereplő tört $j_r=0$ esetén $t^r/T_r=Q(0^r)$, $j_r=1$ esetén pedig $1/T_r=Q(1^r)$, a produktumok tényezői pedig a $0-1$ váltásokhoz tartozó r_u-k illetve az $1-0$ váltásokhoz tartozó s_v-k. Tehát tetszőleges kiszolgálási állapot ergodikus

valószínűségét úgy kapjuk meg, hogy az utolsó komponens (j_r) által meghatározott alapértéket szorozzuk az állapot mindazon komponenseihez tartozó r_u ill. s_v értékekkel, amelyeket tőlük különböző komponens előz meg.

1. tétel bizonyítása. Ha $j_1=\ldots=j_r=0$ vagy $j_1=\ldots=j_r=1$, akkor a produktumok üresek, így értékük egy, és (4.1) az 1. lemma alapján igaz.

Ha a $j_1, \ldots, j_r$ elemek nem mind azonosak, tegyük fel, hogy az $a<b<\ldots<z$ helyeken levő elemek különböznek az őket megelőző elemektől.

Legyen például a $j_{a-1}=0$, $j_a=1$; akkor $(j_1, \ldots, j_r)=(A_aG_a)$, ahol $G_a=(j_{a+1}, \ldots, j_r)$. Ekkor az 1. lemma alapján $Q(A_aG_a)=r_cQ(1^aG_a)$. A következő váltási hely a b-edik hely, ezért $r_aQ(1^aG_a)=r_aQ(B_bG_b)$, ahol $G_b=(j_{b+1}, \ldots, j_r)$. Ekkor a 2. lemma alapján $Q(A_aG_a)=r_as_bQ(0^bG_b)$.

Hasonló módon járunk el minden olyan helyen, ahol az őt megelőzőtől különböző elem van, és így a tételben szereplő két produktumhoz jutunk.

Ha $j_r=0$, akkor az utolsó váltási hely $1-0$ típusú, így (3.2)-t alkalmaztuk utoljára, ezért a keresett Q-értéket $Q(0^r)=t^r/T_r$ segítségével fejeztük ki. Mivel ekkor $1-j_r=1$, így $Q(0^r)=(t^r)^{1-j_r}/T_r$ is fennáll.

Ha $j_r=1$, akkor az utolsó váltási hely $0-1$ típusú, így (3.1)-et alkalmaztuk utoljára. Mivel most $1-j_r=0$, így $Q(1^r)$ számlálójában az 1 így is írható: $(t^r)^{1-j_r}$.

Tehát (4.1) valóban mindkét esetben helyes.

A speciális $p=q=0{,}5$ esetben $t=1$, és ezért $r_i=s_i=1/i$ $(i=1, 2, \ldots)$ és $Q(0^r)=Q(1^r)=\dfrac{1}{r+1}$.

Ekkor például a (0010100) kiszolgálási állapotra $r_3s_5\dfrac{1}{r+1}=\dfrac{1}{3}\dfrac{1}{5}\dfrac{1}{8}$ adódik.

5. A kiszolgálási sebesség

A második részben definiált kiszolgálási algoritmus szerint időegységenként legalább egy, és legfeljebb két igényt szolgálunk ki: az a) és b) esetben kettőt, a c) esetben egyet. A kiszolgálási sebességet konkrét sorok esetén az időegységenként kiszolgált igények számának átlagával, véletlen sorok esetén pedig a megfelelő várható értékek átlagával jellemezzük.

Ha a (2.1)—(2.2) igénysorozatok realizációit a fenti algoritmussal szolgáljuk ki, jelöljük $v_t(r, p)$-vel $(t=1, 2, \ldots)$ a t-edik időegységben kiszolgált igények számát.

Az adott algoritmus esetén $1\leqq v_t(r, p)\leqq 2$ és így $1\leqq M[v_t(r, p)]\leqq 2$ $(t=1, 2, \ldots)$.

A kiszolgálási sebességet ebben az esetben — azaz r darab, a $0<p<1$ és $q=1-p$ valószínűségekkel jellemezhető sor kiszolgálására — az

$$S(r, p)=\lim_{t\to\infty}\frac{\sum_{i=1}^{t} M[v_i(r, p)]}{t} \tag{5.1}$$

összefüggéssel definiáljuk.

A sebességet a kiszolgálási vektorok és a v_t értékek határeloszlása közötti szoros kapcsolat alapján számítjuk ki.

Ha ugyanis a kiszolgálási vektorban van két különböző elem, akkor a megfelelő időegységben biztosan két igényt szolgálunk ki.

Csak akkor nem tud az algoritmus két igényt kiszolgálni, ha

a) a kiszolgálási vektor csupa nullából áll, és az első sorban a nulla után újabb nulla következik, vagy

b) a kiszolgálási vektor csupa egyesből áll, és az első sorban az egyes után újabb egyes következik.

Amint a kiszolgálási vektorok eloszlása tart a határeloszláshoz, úgy tart a fenti két esemény együttes valószínűsége a $pQ(0^r)+qQ(1^r)$ értékhez, azaz

$$\lim_{t\to\infty} P(v_t(r,p)=1)=p\frac{t^r}{T_r}+q\frac{1}{T_r},$$

és így

$$\lim_{t\to\infty} M[v_t(r,p)]=1\left(\frac{pt^r}{T_r}+\frac{q}{T_r}\right)+2\left(1-\frac{pt^r}{T_r}-\frac{q}{T_r}\right). \tag{5.2}$$

A sebességre nézve igaz a következő.

2. TÉTEL. A (2.1)—(2.2) igénysorozatok realizációinak a 2. részben leírt algoritmussal való kiszolgálása esetén az $S(r,p)$ kiszolgálási sebességre fennáll

$$S(r,p)=2-\frac{pt^r+q}{T_r}. \tag{5.3}$$

Bizonyítás. Felhasználjuk, hogy ha egy h_t $(t=1,2,\ldots)$ számsorozat határértéke H, akkor az $m_t=\sum_{i=1}^{t} h_t/t$ $(t=1,2,\ldots)$ számsorozat is konvergens, és H a határértéke.

Ezt az állítást alkalmazva az $a_t=M[v_t(r,p)]$ számsorozatra, m_t éppen a sebesség (5.1) definíciójában szereplő tört.

Mivel a határértéket (5.2) alapján ismerjük, és m_t határértéke ugyanaz, így az (5.2) kifejezés egyúttal az $S(r,p)$ sebességet is megadja. (5.2)-ből a zárójelek felbontása és összevonás után (5.3)-at kapjuk.

A tétel segítségével megvizsgálhatjuk, hogyan változik a kiszolgálási sebesség, ha a sorok számát növeljük. A H_p kiszolgálási határsebességet a következőképpen definiáljuk:

$$H_p=\lim_{r\to\infty} S(r,p).$$

1. KÖVETKEZMÉNY. A kiszolgálási határsebesség

$$H_p=\frac{1}{\max(p,q)}.$$

Bizonyítás. Ha $p=0{,}5$, akkor $q=0{,}5$, $t=\left(\frac{p}{q}\right)^2=1$, $T_r=r+1$. Ezeket az értékeket (5.3)-ba helyettesítve leolvasható, hogy $H_p=2$. Ha például $p>0{,}5$, akkor $t>1$, és $T_r=\frac{t^{r+1}-1}{t-1}$. Ezeket az értékeket (5.3)-ba helyettesítve

$$S(r,p)=2-\frac{pt^r(t-1)}{t^{r+1}-1}-\frac{q(t-1)}{t^{r+1}-1}.$$

Innen leolvasható, hogy r értékének növelésekor a $H_p = 2 - p(t-1)/t = 1/p$ határértéket kapjuk.

Ha $p < 1/2$, akkor szerepcserével $H_p = 1/q$ adódik.

A tétel alapján meghatározható, milyen sebességgel történik az egyes sorok kiszolgálása. Az i-edik sor kiszolgálási sebessége legyen

$$S^1(p) = S(1, p), \quad S^i(p) = S(i, p) - S(i-1, p) \quad (i = 2, 3, \ldots). \tag{5.4}$$

2. KÖVETKEZMÉNY. Az $S^i(p)$ $(i = 1, 2, \ldots)$ kiszolgálási sebességek

$$S^1(p) = 2 - (p^3 + q^3)/(p^2 + q^2), \quad S^j(p) = \frac{T_j(pt^{j-1} + q) - T_{j-1})pt^j + q)}{T_{j-1}T_j} \quad (j = 2, 3, \ldots), \tag{5.5}$$

ami a $p = 1/2$ esetben

$$S^1(0{,}5) = 1{,}5 \quad \text{és} \quad S^j(0{,}5) = \frac{1}{j(j+1)} \quad (j = 2, 3, \ldots). \tag{5.6}$$

Bizonyítás. Ha (5.3)-ba $r = 1$-et helyettesítünk, akkor a törtet bővítve q^2-tel megkapjuk $S^1(p)$ (5.5)-beli alakját.

Ha (5.4)-be behelyettesítjük $S(j, p)$-t (5.3) alapján, akkor közös nevezőre hozás és összevonás után megkapjuk $S^j(p)$ (5.5)-beli alakját.

Végül (5.5)-be $p = q = 0{,}5$-et helyettesítve (5.6) adódik.

6. Összefoglalás

Dolgozatunkban olyan igénysorok párhuzamos kiszolgálásával foglalkoztunk, amelyek egymástól különböző prioritással rendelkeztek, és két típusú igényt tartalmaztak.

Egy adott kiszolgálási algoritmus esetén meghatároztuk az egyes időpontokban a kiszolgálandó sorok állapotát jellemző, az egyes sorok elején levő igényekből álló kiszolgálási vektorok határeloszlását, és ennek segítségével az egyes sorokra külön, és a sorokra együtt vonatkozó kiszolgálási sebességet.

Lényegesen nehezebbnek látszik a probléma akkor, ha a benzinkútnál többfajta benzin kapható, azaz a kiszolgálandó sorok kettőnél több típusú igényt tartalmazhatnak.

Természetesen a cikkben vizsgálttól különböző kiszolgálási algoritmusok is alkalmazhatók.

Például számos területre alkalmazható az, a cikkben vizsgálthoz hasonlóan ugyancsak KÁTAI IMRÉTŐL származó feladat, amelyben a soroknak nincs prioritása, és a kiszolgálási algoritmus a sorokra vonatkozó teljes vagy részleges információ alapján a feldolgozási sebesség maximalizálására törekszik.

A szerzők köszönetüket fejezik ki KÁTAI IMRE akadémikusnak a feladat kitűzéséért és a konzultációért.

IRODALOM

[1] BURNETT, J. and COFFMAN, E. G., "A combinatorial problem related to interleaved memory systems", *Journal of ACM* **20** (1973) 39—45.
[2] IVÁNYI, A. és KÁTAI, I., „Átfedéses memóriájú számítógépek teljesitményéről", *Alkalmazott Matematikai Lapok* **3** (1977) 1—11.
[3] IVÁNYI, A. and KÁTAI I., "Processing of random sequences with priority", *Acta Cybernetica* **4** (1) (1978) 85—101.
[4] IVÁNYI, A. and PERGEL, J., "Parallel processing of 0—1 sequences", *Annales Univ. Sci. Budapest., Sectio Computatorica* **4** (1983) 85—95.
[5] PRÉKOPA, A., *Valószínűségelmélet* (Műszaki Könyvkiadó, Budapest, 1974).
[6] Ивани, А., «О минимизации непроизводительных затрат при параллельном выполнении программ», *Программирование* **10** (1) (1984) 63—68.

(Beérkezett: 1984. augusztus 1.)

IVÁNYI ANTAL ÉS PERGEL JÓZSEF
ELTE MATEMATIKAI INTÉZET
1088 BUDAPEST, MÚZEUM KRT. 6—8.

PARALLEL PROCESSING OF BINARY QUEUES

A. IVÁNYI AND J. PERGEL

IVÁNYI and KÁTAI described an algorithm processing of sequences of input signals. They determined the asymptotic speed of the algorithm for any finite set of possible input signals and for one or two sequences (priority classes), if the input signals are independent random variables with uniform distributfon over the set of possible values.

In an earlier paper we determined the asymptotic speed for any number of priority classes, if the set of possible input signals contains two elements having the probalility 0,5.

Here we consider the problem when the probability of the possible input signals is different from 0,5. Surprisingly we shall find that the limit of the asymptotic speed, when the number of the priority classes tends to infinity, is less than 2.

Alkalmazott Matematikai Lapok **11** (1985) 201—237

A KÜLFÖLDI SZAKIRODALOMBÓL

A FEIGENBAUM-UNIVERZALITÁS ÉS A TERMODINAMIKAI FORMALIZMUS[1]

J. B. VUL, JA. G. SZINAJ, K. M. HANYIN

TARTALOM

1. A perióduskettőződéses bifurkációsorozat

1.1. Ez a cikk a *Feigenbaum-univerzalitás* fogalmát ismerteti, amely a közelmúlt egyik jelentős felfedezése a dinamikai rendszerek elméletében. A felfedezés a matematikai és a fizika határára esik: a problémafelvetés lényegében matematikai, a megoldás módja pedig, amely a renormálási csoportok elméleti fizikából jól ismert módszerén alapul, a fizikából származik. E módszert elsősorban a statisztikus mechanikában és a kvantummezők elméletében alkalmazzák.

A probléma, amelyből FEIGENBAUM kiindult, egészen általános jellegű: paramétertől függő dinamikai rendszerekben hogyan játszódik le az átmenet stabilis típusú mozgásból (amelyet természetes módon laminárisnak nevezhetünk) instabilis típusú mozgásba (amelyet gyakran a turbulencia fogalmával társítunk). A felfedezés túllépi e kérdés kereteit, ugyanis teljesen új módszert nyújt a dinamikai rendszerek mikroszkopikus, azaz lokális viselkedésének kutatására.

A *Feigenbaum-univerzalitás* fogalma közvetlenül a perióduskettőződéses bifurkációsorozatokra vonatkozik. A bifurkációk tradícionális elméletében rendszerint egy dinamikai rendszercsalád lokális viselkedését vizsgálják a paraméter bifurkációs értékének környezetében. Esetünkben egészen új probléma merül fel: egy dinamikai rendszercsalád lokális viselkedése a paraméter egy olyan értékének környezetében, ahol végtelen sok bifurkációs paraméterérték torlódik. Úgy tűnik, az egyik első mun-

[1] A fordítás a szerzők és a kiadó hozzájárulásával készült az Е. Б. Вул, Я. Г. Синай и К. М. Ханин: «Универсальность Фейгенбаума и термодинамический формализм» Успехи Математических Наук 237 (1984) №. 3. 1—37. dolgozatról.

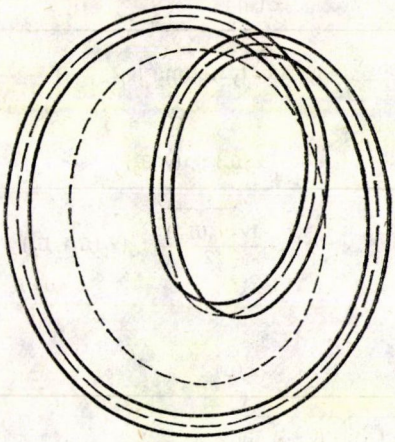
ka ezen a területen M. METROPOLIS, M. L. STEIN és P. R. STEIN [1] cikke. Ebben szerepel az az észrevétel, hogy egy szakasz önmagára való egyparaméteres leképezéscsaládjainak széles osztályánál a paraméter növekedésével a trajektóriák bonyolultabbakká válnak: a stabilis periodikus pálya instabilis lesz, emellett egy kétszer akkora periódusú stabilis periodikus pálya keletkezik, amely az instabilis ciklusok pontjainak kivételével minden pontra nézve attraktív tulajdonságú. Megemlítünk egy jóval korábbi munkát is [2]. M. FEIGENBAUM (zsebkalkulátorral dolgozva) észrevette, hogy a paraméter azon egymásutáni értékei, ahol a [0, 1] szakasz önmagára való $x \mapsto \mu x(1-x)$, $0 \leqq \mu \leqq 4$ leképezéseire ilyen bifurkációk lépnek fel, egy δ^{-1} hányadosú geometriai sorozat sebességével konvergálnak, ahol $\delta = 4{,}6692\ldots$, a nevezetes Feigenbaum-konstans. Ezután hasonló számítást végzett az $f(x; \mu) = \mu \sin(\pi x)$ leképezéscsaládra vonatkozóan, és ismét ugyanolyan hányadosú geometriai növekedést észlelt. Így merült fel az a hipotézis, hogy δ értéke független a konkrét leképezés családtól. Ugyancsak M. FEIGENBAUM állította fel δ univerzalitásának elméletét (l. [3]—[6]).

Hasznos lehet, ha kialakítunk egy szemléletes képet arról a jelenségről, amely a perióduskettőződések végtelen bifurkációsorozatánál játszódik le. Külön vázoljuk fel a folytonos, illetve a diszkrét idő esetét.

Tekintsünk egy sima (C^∞ osztályú) R^n-beli közönséges differenciálegyenlet rendszert, amely egy paramétertől, μ-től függ:

$$\frac{dx_i}{dt} = f_i(x_1, \ldots, x_n; \mu). \tag{1.1}$$

Tegyük fel, hogy $\mu = \mu_0$ esetén az (1.1) rendszernek létezik $x(t; \mu_0) = (x_1(t; \mu_0), \ldots, x_n(t; \mu_0))$ stabilis periodikus megoldása, amelynek periódusa $T = T(\mu_0)$, azaz $x(t+T; \mu_0) = x(t; \mu_0)$. Esetünkben a stabilitás azt jelenti, hogy a rendszer karakterisztikus számai a komplex egységkör belsejében vannak. Tegyük fel, hogy a μ paramétert valamely μ_1 értékig növelve ez a trajektória elveszti stabilitását, azaz a karakterisztikus számok egyike -1-gyel lesz egyenlő, ugyanakkor a többi karakterisztikus szám abszolút értéke 1-nél kisebb marad. Ekkor néhány, a magasabbrendű deriváltakra vonatkozó egyszerű egyenlőtlenség következményeképpen a μ_1-en áthaladva perióduskettőződéses bifurkáció jön létre, ami azt jelenti, hogy a korábban stabilis periodikus trajektória instabilissá válik, ezzel egyidőben egy másik stabilis periodikus trajektória keletkezik, amelynek periódusa kétszer akkora. A paraméter további növelésekor mindegyik periodikus trajektória kissé eltolódik a térben, de megőrzi stabilitási típusát. Tegyük fel, hogy a $\mu = \mu_2 > \mu_1$ értéknél játszódik le a következő perióduskettőződéses bifurkáció, amelynek eredményeképpen két instabilis és egy stabilis periodikus trajektória keletkezik; ez utóbbi periódusa megközelítőleg kétszer akkora, mint az előző stabilis periodikus trajektóriáé (l. az 1.1. ábrát), és körülbelül négyszer akkora, mint az első stabilis periodikus trajektória periódusa. Most már világos, hogy mi történik n hasonló bifurkáció után: létrejön egy maximális periódusú stabilis periodikus trajektória, amelyet n instabilis periodikus trajektória vesz körül; ezeknek periódusai körülbelül 1/2 kvóciensű geometriai sorozatot alkotnak. Leírhatjuk az $n \to \infty$ határátmenetkor keletkező objektumot is: a *Van Dantzig szolenoiddal* topológiailag azonos invariáns halmazhoz jutunk, amelynek tetszőleges környezetében találhatók instabilis periodikus trajektóriák. Minden ilyen

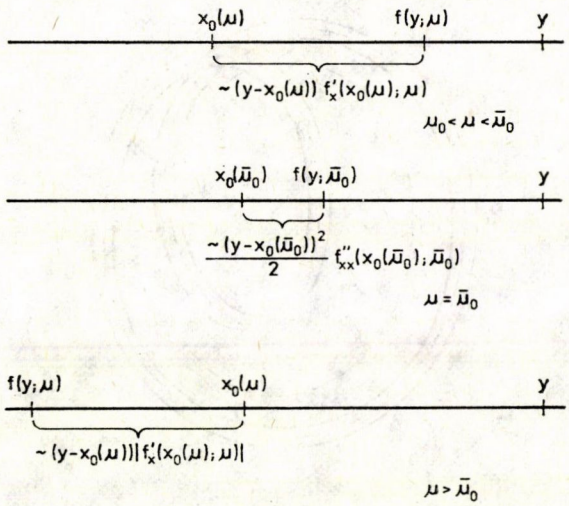


1.1. ábra

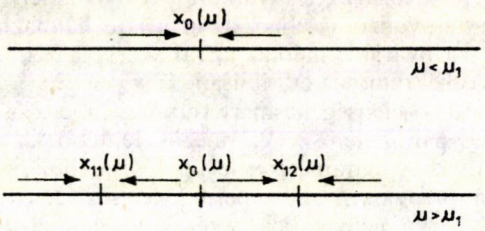
elég kicsiny környezetben az összes olyan trajektória, amely egy nyílt, mindenütt sűrű részhalmazból indul, vonzódik ehhez az invariáns halmazhoz. Ebben a tekintetben tehát az imént leírt invariáns halmaz attraktor. Nehezen képzelhető el, hogyan találhatunk hasonló objektumot analitikusan.

1.2. Most ugyanezt a jelenséget ismertetjük egydimenziós leképezésekre. Megjegyezzük, hogy az egyparaméteres vektormezők fentebb vizsgált esete a szokásos módon visszavezethető az egyparaméteres leképezéscsaládok esetére a Poincaré-leképezés segítségével. Látni fogjuk, hogy a periódusketttőződéses bifurkáció egydimenziós karakterű, így az egydimenziós leképezések elegendően teljes képet nyújtanak a jelenség sajátosságairól. Tekintsük a valós számegyenesnek önmagára való C^2-beli f leképezését. Eszerint az $f(x)$ függvényérték az a pont, ahová az x pont kerül az f transzformáció hatására. Az x_0 pontot az f leképezés stabilis fixpontjának nevezzük, ha létezik olyan U környezete, hogy minden $y \in U$ pontra $\lim_{n\to\infty} f^{(n)}(y) = x_0$. Itt és a továbbiakban $f^{(n)}$ az f leképezés n-szeres szuperpozícióját jelöli. Az x_0 fixpont stabilitására nézve $f(x_0) = x_0$ és $|f'(x_0)| < 1$ nyilvánvaló elégséges feltételt jelent. Tekintve, hogy a bennünket érdeklő kérdések lokális jellegűek, feltesszük, hogy az x_0 pont U vonzási tartománya az egész R^1 valós számegyenes. Tegyük fel továbbá, hogy az f leképezése függ egy valós μ paramétertől is, azaz egyparaméteres leképezéscsaláddal van dolgunk. Most az x_0 fixpont is függ μ-től: $x_0 = x_0(\mu)$. Legyen a $\mu = \mu_0$ kezdeti értéknél x_0 stabilis, továbbá legyen $0 < f'_x(x_0; \mu) < 1$. Kezdjük el növelni μ-t. Lehetséges, hogy ekkor $f'_x(x_0; \mu)$ csökken: $\mu = \bar{\mu}_0$-nál 0, majd negatív lesz, és valamely, $\mu = \mu_1$-nél felveszi a -1 értéket. Az 1.2. ábrán vázoltuk az $f(x; \mu)$ leképezés jellegét, feltéve, hogy $f''_{xx}(x_0; \bar{\mu}_0) \neq 0$.

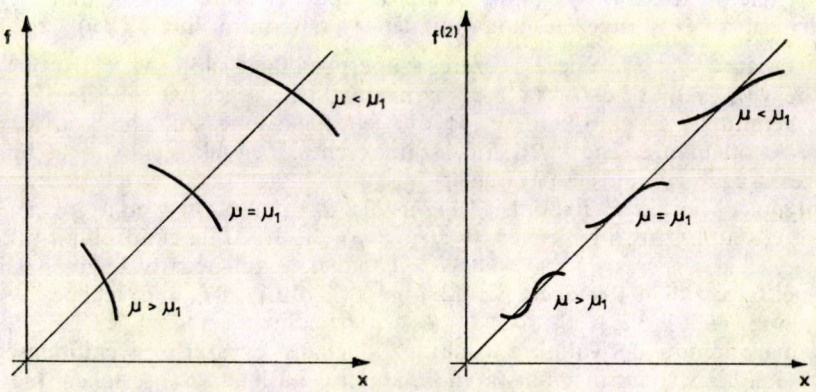
A másodrendű deriváltakra vonatkozó néhány egyszerű egyenlőtlenség teljesülése esetén az x_0 pont instabilissá válik, amikor μ áthalad a μ_1 értéken, megjelenik továbbá két pont: $x_{11}(\mu)$ és $x_{12}(\mu)$, amelyek egy 2 periódusú stabilis periodikus trajektóriát reprezentálnak. Ezt a jelenséget kétféleképpen ábrázoltuk (1.3. és 1.4. ábra).



1.2. ábra



1.3. ábra



1.4. ábra

Megjegyezzük, hogy $\mu-\mu_1\to 0$ esetén az $|x_{11}(\mu)-x_0(\mu_1)|$ és az $|x_{12}(\mu)-x_0(\mu_1)|$ különbségek nagyságrendje $\sqrt{\mu-\mu_1}$, ugyanakkor

$$|x_0(\mu)-x_0(\mu_1)| = O(\mu-\mu_1).$$

Tehát egy periódusketttőződéses bifurkáció bekövetkezésekor a korábban stabilis x_0 fixpont instabilissá válik, körülötte pedig létrejön egy 2 periódusú stabilis periodikus trajektória. A paraméter további növekedésekor ez az eredeti x_0 fixpont instabilis fixpont marad, minden más pont a 2 periódusú stabilis periodikus trajektóriához vonzódik. A paraméter valamely $\mu=\mu_2$ értékénél a 2 periódusú trajektória elveszti stabilitását, mivel

$$\left.\frac{\partial f^{(2)}(x;\ \mu_2)}{\partial x}\right|_{x=x_{11}} = \left.\frac{\partial f^{(2)}(x;\ \mu_2)}{\partial x}\right|_{x=x_{12}} = -1.$$

Az $f^{(2)}$-re vonatkozó hasonló meggondolások és ábrák alapján azt kapjuk, hogy a 2 periódusú trajektória instabilissá válik, környezetében pedig egy 4 periódusú periodikus trajektória keletkezik. Tehát a paraméterértékek egy végtelen $\{\mu_n\}$ sorozatához jutunk: $\mu=\mu_n$-nél a 2^{n-1} periódusú trajektória elveszti stabilitását és egy 2^n periódusú trajektória. Most már elképzelhető, hogy mi történik a $\mu=\mu_\infty=\lim_{n\to\infty}\mu_n$ értéknél: az $f(x;\mu_\infty)$ leképezés *Cantor-típusú F invariáns halmazát* végtelen sok instabilis periodikus trajektória veszi körül, ezek periódusa 2^n. Az $f(x;\mu_\infty)$ leképezés hatására az instabilis trajektóriák és ősképeik pontjai kivételével minden pont F-hez vonzódik. A *Feigenbaum-univerzalitás* fogalma első közelítésben a $\{\mu_n\}$ sorozat egyetemes viselkedésére utal: $\mu_\infty-\mu_n\sim C\delta^{-n}$, ahol a C konstans az f leképezéscsaládtól függ, δ pedig univerzális és a fent említett *Feigenbaum-konstanssal* azonos. Látni fogjuk, hogy a *Feigenbaum univerzalitás* ennél tágabb fogalom: az F attraktor szerkezete, egyebek között *Hausdorff-dimenziója,* az $f^{(n)}$ iteráltak viselkedése $\mu=\mu_\infty$ környezetében nem függ definit módon f-től.

Bemutatunk egy sor számítási eredményt, amely konkrét leképezésekre és vektormezőkre vonatkozik. E számítógépes vizsgálatok a periódusketttőződéses bifurkációsorozat és az univerzalitás jelenlétére utalnak.

Az 1.1. táblázat a $[-1, 1]$ intervallum önmagára való kvadratikus, $f(x;\mu)=1-\mu x^2$ leképezéscsaládjára vonatkozó numerikus eredményeket tartalmaz (1. [8]).

1.1. TÁBLÁZAT

n	μ_n	$\Delta_n=\mu_n-\mu_{n-1}$	$\delta_n=\frac{\mu_{n-2}-\mu_{n-1}}{\mu_{n-1}-\mu_n}$
1	0,75		
2	1,25	0,5	
3	1,3680989394	0,1180989394	4,233738275
4	1,3940461566	0,0259472172	4,551506949
5	1,3996312389	0,0055850823	4,645807493
6	1,4008287424	0,0011975035	4,663938185
7	1,4010852713	0,0002565289	4,668103672
8	1,401140214699	0,000054943399	4,668966942
9	1,401151982029	0,000011767330	4,669147462
10	1,401154502237	0,000002520208	4,669190003

1.2. TÁBLÁZAT

n	r_n	$\Delta_n = r_{n-1} - r_n$	$\delta_n = \frac{r_{n-2} - r_{n-1}}{r_{n-1} - r_n}$	n	r_n	$\Delta_n = r_{n-1} - r_n$	$\delta_n = \frac{r_{n-2} - r_{n-1}}{r_{n-1} - r_n}$
0	100,7952	—	—	3	99,54712	0,08139	4,319
1	99,9800	0,8152	—	4	99,52934	0,01778	4,578
2	99,62851	0,35149	2,319	5	99,525533	0,003807	4,670

Az 1.2 táblázatban a nevezetes *Lorenz-modell* bifurkációs paraméterértékeit adjuk meg. A modellt a következő differenciálegyenlet rendszer írja le:

$$\dot{x} = -\sigma x + \sigma y,$$

$$\dot{y} = -xz + rx - y,$$

$$\dot{z} = xy - bz.$$

Itt $\sigma = 10$, $b = 8/3$ és r a paraméter (l. [50]).

A cikk felépítése a következő. A 2. fejezetben bemutatunk néhány alapvető tényt az egydimenziós leképezések elméletéből. A 3. fejezetben ismertetjük a *Feigenbaum-univerzalitás* magyarázatát, lényegében a [3]—[8] munkák alapján. Ennek a fejezetnek számos eredménye új. Az 5. fejezetben *Feigenbaum-leképezések* kis sztochasztikus perturbációjával foglalkozunk. A 6. fejezetben P. COLLET, J.-P. ECKMANN és H. KOCH [9] eredményeit ismertetjük, amelyek a többdimenziós esettel és más általánosításokkal kapcsolatosak.

2. A folytonos egydimenziós leképezések általános elméletéről

2.1. Néhány évvel ezelőtt úgy tűnt, hogy a folytonos egydimenziós leképezések elmélete a dinamikai rendszerek általános elméletének olyan része, amely csekély érdeklődésre tarthat számot. Úgy gondolták, hogy az ilyen leképezések felépítése meglehetősen egyszerű, emellett a rájuk vonatkozó eredmények tipikusan egydimenziós jellegűek, nincs természetes általánosításuk több dimenzióra. Mindkét megállapítás tévesnek bizonyult: kiderült, hogy a folytonos egydimenziós leképezések szerkezete egyáltalán nem triviális, számos rendkívül szép és váratlan eredmény látott napvilágot, amelyek közül nem egy természetes módon vihető át a többdimenziós esetre. Meg kell azonban jegyeznünk, hogy A. N. SARKOVSZKIJ már 20 évvel ezelőtt rámutatott az elmélet mélységére, a leképezések szerkezetének bonyolultságára (l. [10], [11]). Ebben a paragrafusban ismertetjük a folytonos egydimenziós leképezések elméletének azokat az eredményeit, amelyek szükségesek a továbbiak megértéséhez. A témában számos publikáció jelent meg, így az alábbiak semmiképpen sem tekinthetők a kérdéskör kimerítő tárgyalásának; a hivatkozások listája sem a teljesség igényével készült.

D. SINGER [12] fontos felfedezése, hogy a *Schwarz-derivált* jól használható a folytonos egydimenziós leképezések elméletében. Az $f \in C^3$ függvény *Schwarz-deriváltját* a következőképpen értelmezzük:

$$Sf = \frac{f'''}{f'} - \frac{3}{2}\left(\frac{f''}{f'}\right)^2.$$

Lényeges tulajdonsága, hogy ha $Sf_1<0$ és $Sf_2<0$, akkor $S(f_1 \circ f_2)<0$. A negatív *Schwarz-deriválttal* rendelkező egydimenziós leképezések mély sajátosságaira sikerült fényt deríteni. Így például D. SINGER kimutatta (l. [12]), hogy egy szakasz önmagába való olyan f leképezésére, amelyre $Sf<0$, teljesül a következő: minden stabilis periodikus trajektória vonzza a szakasz egyik végpontjának trajektóriáját, vagy valamely kritikus x_c pont trajektóriáját (azaz olyan pontét, ahol $f'(x_c)=0$). Az egydimenziós leképezések elmélete rendszerint úgynevezett unimodális leképezésekkel foglalkozik.

2.1. Definíció. Egy szakasz önmagába való folytonos leképezését unimodálisnak nevezzük, ha a szakasz egy belső x_c pontjában a leképezésnek szélsőértéke van, továbbá x_c előtt, illetve után a leképezés szigorúan monoton.

SINGER fenti eredményéből következik, hogy negatív *Schwarz-deriválttal* rendelkező unimodális leképezéseknél legfeljebb egy stabilis periodikus trajektória létezik. Meg kell azonban jegyeznünk, hogy ez az állítás csak néhány kiegészítő feltétellel igaz, ugyanis az általános esetben létezhetnek további stabilis fixpontok. J. GUCKENHEIMER ([13]) bebizonyította, hogy ha egyetlen stabilis periodikus trajektória van, akkor a hozzá nem vonzódó pontok halmaza nullmértékű. Ismerünk olyan konkrét példákat, amelyek azt mutatják, hogy ez a halmaz lehet *Cantor-típusú*, amelyen a dinamika instabilis, tehát sztochasztikus.

A. N. SARKOVSZKIJ az egydimenziós leképezések nagyon lényeges tulajdonságát tárta fel. Tekintsük a természetes számok következő rendezését:

$$1 \prec 2 \prec 2^2 \ldots \prec 2^n \prec \ldots \prec \ldots \prec 2^m \cdot 7 \prec 2^m \cdot 5 \prec 2^m \cdot 3 \prec \ldots$$

$$\ldots \prec 2 \cdot 7 \prec 2 \cdot 5 \prec 2 \cdot 3 \prec \ldots \prec \ldots \prec 7 \prec 5 \prec 3.$$

Ezt a rendezést *Sarkovszkij-rendezésnek* nevezzük.

SARKOVSZKIJ TÉTELE (l. [10]). Ha az $[a, b]$ szakasz önmagába való unimodális leképezésének létezik k periódusú trajektóriája, akkor létezik periodikus trajektória minden olyan periódussal, amely k-t a *Sarkovszkij-rendezésben* megelőzi.

Ma már *Sarkovszkij tételének* egészen elemi bizonyításai is léteznek (l. [14]). Egy ilyen bizonyítás éppen megjelenés alatt áll a *„Kvant"* című folyóiratnál. E témához kapcsolódik T. LI és J. A. YORKE [15] munkája, amelyben a *Sarkovszkij-tétel* egyik speciális esetéről van szó: a szakasz önmagába való unimodális leképezésénél egy 3 periódusú trajektória létezése maga után vonja tetszőleges periódusú trajektória létezését. Ezt a jelenséget a szerzők a sztochasztikus viselkedéssel asszociálják.

A *Feigenbaum-elmélet* szempontjából elsősorban a *Sarkovszkij-rendezés* első szeletének megfelelő leképezések fontosak, amelyeknek minden $n \geqq 0$ esetén van 2^n periódusú trajektóriájuk. Erről a sokat kutatott területről mindenekelőtt JU. SZ. BARKOVSZKIJ és G. M. LEVIN [16], valamint M. MISIUREWICZ [17] munkáját kell kiemelnünk. Az utóbbi cikk az egydimenziós leképezések alábbi osztályával foglalkozik:

2.2. Definíció. A $[-1, 1]$ szakasz önmagába való f páros leképezése a G osztályba tartozik, ha:

1. $f \in C^1([-1, 1]), f' \in C^3((-1, 0) \cup (0, 1))$,
2. $f(-1)=-1, f'(-1)>1$,

3. $f'(x)\neq 0$, ha $x\neq 0$, $Sf(x)<0$, ha $x\neq 0$,
4. f-nek van 2^n periódusú trajektóriája minden $n\geqq 0$ esetén,
5. f-nek nincs más periodikus trajektóriája.

A G osztályba tartozó leképezésekhez minden $n\geqq 1$-re konstruálható szakaszoknak olyan $\Delta_k^{(n)}$, $0\leqq k<2^n$ rendszere, amely az alábbi tulajdonságokkal rendelkezik:

I. A $\Delta_n^{(k)}$ szakaszok páronként diszjunktak, $0\leqq k<2^n$.
II. $f(\Delta_n^{(k)})=\Delta_{k+1}^{(n)}$, $0\leqq k<2^n-1$, $f(\Delta_{2^n-1}^{(u)})\subset\Delta_0^{(n)}$.
III. Minden $\Delta_k^{(n-1)}$ szakasz pontosan két n-ed rangú szakaszt tartalmaz, ezek $\Delta_k^{(n)}$ és $\Delta_{k+2^{n-1}}^{(n)}$.

Tekintsük az $F=\bigcap_{n\geqq 1}\bigcup_{k=1}^{2^n-1}\Delta_k^{(n)}$ halmazt. Könnyen igazolható, hogy F invariáns az f leképezésre és homeomorf a *Cantor-halmazzal.* Bizonyítható továbbá, hogy majdnem minden $x\in[-1, 1]$-re $d(f^{(l)}(x), F)\to 0$ ha $l\to\infty$ (azok a kivételes pontok, amelyek az iteráció során instabilis periodikus trajektóriára kerülnek). Később látni fogjuk, hogy a periódusketttőződéses bifurkációval kapcsolatos leképezések lényegében a G osztályhoz tartoznak és attraktoruk szintén *Cantor-típusú.* Most annyit jegyzünk meg, hogy a $\Delta_k^{(n)}$ szakaszok rendszere lehetővé teszi az $f|_F$ dinamikai rendszer metrikus és spektrális tulajdonságainak leírását (l. [17]).

2.1. TÉTEL. 1. Az f leképezésnek egyetlen μ_0 invariáns mértéke van, amely F-re koncentrált. Erre $\mu_0(\Delta_k^{(n)})=2^{-n}$.

2. Az (F, μ_0) mértéktér f leképezése ergodikus, és diszkrét spektruma a következő számokból áll:

$$\exp\{2\pi i(2r+1)2^{-n}\},\quad n\geqq 1,\quad 0\leqq r\leqq 2^{n-1}-1.$$

Bizonyítás. 1. Mivel a $\Delta_k^{(n)}$ szakaszok páronként diszjunktak, így $\mu_0(\Delta_k^{(n)})$ független k-tól, továbbá $\bigcup_{k=0}^{2^n-1}\Delta_k^{(n)}\supset F$. Ebből következik a tétel első állítása.

2. Minden $n\geqq 1$-re megadjuk az $\exp\{2\pi i2^{-n}\}$ sajátértékhez tartozó $e^{(n)}(x)$ sajátfüggvényt. Legyen $e^{(n)}(x)=\exp\{2\pi ik2^{-n}\}$, ha $x\in\Delta_k^{(n)}$, $0\leqq k\leqq 2^n-1$. Minthogy $f^{(2n)}(\Delta_0^{(n)})\subset\Delta_0^{(n)}$, ez a meghatározás korrekt. Nem nehéz belátni, hogy az $e_r^{(n)}=(e^{(n)}(x))^{2r+1}$, $0\leqq r\leqq 2^{n-1}-1$ függvények az U_f operátor $\exp\{2\pi i(2r+1)2^{-n}\}$ sajátértékeihez tartozó sajátfüggvényei és bázist alkotnak $\mathcal{L}^2(F, \mu_0)$-ban. A tételt bebizonyítottuk.

Képezzük egy olyan $\varphi\in C^1([-1, 1])$ függvény *Fourier-együtthatóit,* amelyre $\int\varphi d\mu_0=0$:

$$c_n=(U_f^n\varphi, \varphi)_{\mu_0}=\int_0^1 e^{2\pi i\omega n}\,d\varrho(\omega).$$

Itt ϱ a φ függvény spektrálmértéke (l. [18]). A 2.1. tételből következik, hogy

$$\varrho=\sum_{n=1}^{\infty}\sum_{r=0}^{2^{n-1}-1}|\varrho_r^{(n)}|^2\,\delta\left(\omega-\frac{2r+1}{2^n}\right),$$

ahol a $\varrho_r^{(n)}$-ek állandók. Most becslést adunk ezek csökkenésére.

2.2. TÉTEL.

$$|\varrho_r^{(n)}| \leqq \frac{1}{2} \max_{x\in[-1,1]} |\varphi'(x)| \max_{0\leqq k\leqq 2^{n-1}-1} |\Delta_k^{(n-1)}|.$$

Bizonyítás. Írjuk fel explicit alakját:

$$\varrho_r^{(n)} = \int_F \varphi(x)\overline{e_r^{(n)}(x)}\,d\mu_0(x) = \sum_{k=0}^{2^{n-1}-1} \int_{\Delta_k^{(n-1)}} \varphi(x)\overline{e_r^{(n)}(x)}\,d\mu_0(x) =$$

$$= \sum_{k=0}^{2^{n-1}-1} \left[\int_{\Delta_k^{(n)}} \varphi(x)\overline{e_r^{(n)}(x)}\,d\mu_0(x) + \int_{\Delta_{k+2^{n-1}}^{(n)}} \varphi(x)\overline{e_r^{(n)}(x)}\,d\mu_0(x) \right] =$$

$$= \sum_{k=0}^{2^{n-1}-1} \{\varphi(x_k^{(n-1)})[\overline{e_r^{(n)}}|_{\Delta_k^{(n)}} + \overline{e_r^{(n)}}|_{\Delta_{k+2^{n-1}}^{(n)}}]2^{-n}\} +$$

$$+ \sum_{k=0}^{2^{n-1}-1} \left[\int_{\Delta_k^{(n)}} (\varphi(x)-\varphi(x_k^{(n-1)}))\overline{e_r^{(n)}(x)}\,d\mu_0 + \int_{\Delta_{k+2^{n-1}}^{(n)}} (\varphi(x)-\varphi(x_k^{(n-1)}))\overline{e_r^{(n)}(x)}\,d\mu_0 \right].$$

Itt $x_k^{(n-1)}$ a $\Delta_k^{(n-1)}$ szakasz felezőpontja, $e_r^{(n)}|_{\Delta_k^{(n)}}$ pedig az $e_r^{(n)}(x)$ függvény értéke a $\Delta_k^{(n)}$ szakaszon. Mivel

$$e_r^{(n)}|_{\Delta_k^{(n)}} = (e^{(n)})^{2r+1}|_{\Delta_k^{(n)}} = e^{2\pi i(2r+1)k2^{-n}},$$

$$e_r^{(n)}|_{\Delta_{k+2^{n-1}}^{(n)}} = (e^{(n)})^{2r+1}|_{\Delta_{k+2^{n-1}}^{(n)}} = -e^{2\pi i(2k+1)k2^{-n}},$$

ezért az első összeg 0-val egyenlő. A másodikra

$$|\varphi(x)-\varphi(x_k^{(n-1)})| \leqq \frac{1}{2} \max_{x\in[-1,1]} |\varphi'(x)| \max_{0\leqq k\leqq 2^{n-1}-1} |\Delta_k^{(n-1)}|.$$

Ebből közvetlenül adódik a tétel állítása.

A most bizonyított tétel sima függvények diszkrét spektrumát írja le: létezik egy δ-amplitúdó az 1/2 pontban, két δ-amplitúdó az $1/2^2$ és a $3/2^2$ pontban, kisebb amplitúdók az $1/2^3$, $3/2^3$, $5/2^3$, $7/2^3$ pontokban, és így tovább. Összegezett intenzitásuk a $(2r+1)2^{-n}$ pontokban kielégíti a következő becslést:

$$\sum_{r=1}^{2^{n-1}-1} |\varrho_r^{(n)}|^2 \leqq \text{const} \max_{0\leqq k\leqq 2^{n-1}-1} (|\Delta_k^{(n-1)}|^2)2^n.$$

2.2. Most megvizsgáljuk, hogy a μ_n bifurkációs paraméterértékek μ_∞ határértéke milyen értelemben tekinthető a sztochasztikus viselkedés kezdetének. Legyen f az $[a, b]$ szakasz önmagába való leképezése. A ν mértéket invariánsnak nevezzük, ha $\nu(C)=\nu(f^{-1}C)$ minden $C\subset[a, b]$ esetén.

2.3. Definíció. Az f leképezés sztochasztikus jellegű, ha létezik abszolút folytonos ν_0 invariáns mértéke, amelyre nézve a $\bigwedge_n f^{-n}(\mathcal{F})$ σ-algebra véges sok atomból áll. Itt $\mathcal{F}$ az $[a, b]$ szakasz *Borel-halmazainak*, $f^{-n}(\mathcal{F})$ pedig az $f^{-n}C$, $C\in\mathcal{F}$ alakú halmazok σ-algebráját jelöli.

Megvilágítjuk a fenti definíciót. Ha a $\bigwedge_n f^{-n}(\mathscr{F})$ σ-algebra atomjainak száma r, akkor léteznek olyan $C_1, C_2, \ldots, C_r$ részhalmazok, amelyekre $C_i \cap C_j = \emptyset$, ha $i \neq j$, $f(C_i) = C_{i+1}$, ha $i < r$ és $f(C_r) \subset C_1$. Az $f^{(r)}$ leképezés nem ergodikus, ergodikus komponensei a C_i, $1 \leqq i \leqq r$ halmazok. Az $f^{(r)}|_{C_i}$ leképezés azonban már keverő tulajdonságú. A leképezés pontos endomorfizmus (l. [18]), amiből egyebek között az következik, hogy spektruma végtelen rendű *Lebesgue-típusú*, minden rendben keverő és entrópiája pozitív. Több esetben sikerült a korrelációcsökkenésre becslést adni (l. például [19]—[21]). Megjegyezzük, hogy ha a sztochasztikus jellegű unimodális f leképezés *Schwarz-deriváltja* negatív, akkor nem lehet stabilis periodikus trajektóriája.

Az első példát olyan sztochasztikus jellegű leképezésre, amelynek van kritikus pontja, J. VON NEUMANN és S. M. ULAM [22] találta. Ez a [0, 1] szakasz önmagára való $f(x) = 4x(1-x)$ leképezése. A megfelelő ν_0 invariáns mérték sűrűségfüggvénye $p(x) = \text{const}\,(x(1-x))^{1/2}$. Látható, hogy ez a négyzetgyökhöz hasonlóan viselkedik a 0, 1 pontokban, amelyek az 1/2 kritikus pont trajektóriáját alkotják.

A következő lépést D. RUELLE munkája jelentette (l. [23]), aki megmutatta, hogy a [0, 1] szakasz önmagába való $f(x; \mu) = \mu x(1-x)$ leképezése a μ paraméter bizonyos μ_0 értékénél szintén sztochasztikus jelleget ölt. A μ_0 érték úgy keletkezik, hogy az 1/2 kritikus pont két lépés alatt egy 3 periódusú instabilis ponttá válik. Jóval előbb L. A. BUNYIMOVICS kapott analóg eredményt az $x \to \{\lambda \sin 2\pi x\}$ leképezésre vonatkozóan (l. [24]). A közelmúltban A. I. OGNYEV (l. [25]) és M. MISIUREWICZ (l. [26]) igazolta a következőt: tegyük fel, hogy az f leképezés kielégíti az $Sf < 0$ feltételt, továbbá a kritikus x_c pont véges sok iteráció után instabilis periodikus trajektóriára kerül. Ekkor f sztochasztikus jellegű. (Ténylegesen M. MISIUREWICZ valamivel általánosabb tételt bizonyított.)

Ebben a kérdéskörben a legnehezebb és legjelentősebb eredmény M. V. JAKOBSON nevéhez fűződik. Legyen $f(x; \mu) = \mu x(1-x)$.

JAKOBSON TÉTELE (l. [27]). A paraméterértékeknek az az $\mathscr{A}$ halmaza, ahol f sztochasztikus jellegű, pozitív *Lebesgue-mértékű*.

Valójában [27]-ben általánosabb alakú leképezéscsalád szerepel. *Jakobson tételében* az invariáns mérték sűrűségfüggvénye egy mindenütt sűrű halmazon gyökös jellegű, tehát rendkívül bonyolult.

Most már meg tudjuk magyarázni, hogy μ_∞ milyen értelemben tekinthető a sztochasztikus viselkedés kezdetének. A μ_∞ pont tetszőleges U jobb oldali környezetében vannak olyan μ paraméterértékek, amelyeknél teljesülnek az *Ognyev—Misiurewicz-tétel* feltételei, ezért $f(x; \mu)$ sztochasztikus jellegű. A *Jakobson-tétel* következménye továbbá, hogy $U \cap \mathscr{A}$ *Lebesgue-mértéke* pozitív, noha nem ismeretes, hogy μ_∞ sűrűsödési pontja-e az $\mathscr{A}$ halmaznak. Az elmondottakból következik, hogy $\mu > \mu_\infty$ esetén a sztochasztikus viselkedés mintegy villanásszerűen jelentkezik, miközben a strukturálisan stabilis rend az uralkodó; ennek főszereplői a majdnem minden pontot magukhoz vonzó stabilis periodikus trajektóriák. Ezzel kapcsolatban említjük meg a híres hipotézist, amely szerint a paraméterértékek stabilis rendnek megfelelő halmaza mindenütt sűrű a paramétertartományban. Numerikus vizsgálatok azt mutatják, hogy μ növekedésekor egyre gyakrabban lép fel a sztochasztikus viselkedés (l. [28]).

3. A renormálási transzformáció

Ebben a paragrafusban kifejtjük a *Feigenbaum-univerzalitás* fogalmának magyarázatát.

3.1. Tekintsük a $[-1, 1]$ szakasz önmagába való sima $f(x; \mu)$ egyparaméteres leképezéscsaládját. Kizárólag unimodális leképezésekkel foglalkozunk. Feltesszük továbbá, hogy az egyetlen kritikus pont $x_c(\mu)=0$ maximumhely minden μ-re. Ez nem jelenti az általánosság megszorítását, hiszen tetszőleges $f(x; \mu)$ unimodális leképezéscsalád a következő átalakítással ilyen alakra hozható: $\tilde{f}(.; \mu)=S_\mu^{-1}\circ f(.; \mu)\circ S_\mu$. Megjegyezzük, hogy az $\tilde{f}=S^{-1}\circ f\circ S$ transzformáció — ahol S a szakasz önmagára való diffeomorfizmusa — nem befolyásolja a periodikus trajektóriák létezését és stabilitási jellegét.

Tegyük fel, hogy μ növelésekor periódusketttőződéses bifurkációknak egy végtelen sorozata jön létre. A korábbiaknak megfelelően jelölje $\mu_1, \mu_2, \ldots$ a paraméter bifurkációs értékeit. Emlékeztetünk arra, hogy $\mu=\mu_n$-nél keletkezik egy periodikus trajektória, amelynek periódusa 2^n. Vizsgáljuk meg kissé részletesebben, hogy mi történik a $\mu_n<\mu<\mu_{n+1}$ intervallumban. A paraméter ilyen értékeire az $f^{(2^n)}(x; \mu)$ leképezésnek van stabilis $x(\mu)$ fixpontja. A μ_n-hez közeli μ-k esetén a $\frac{\partial}{\partial x} f^{(2^n)}(x; \mu)|_{x=x(\mu)}$ derivált értéke közel van $+1$-hez, és ha $\mu\uparrow\mu_{n+1}$, akkor

$$\frac{\partial}{\partial x} f^{(2^n)}(x; \mu)|_{x=x(\mu)} \to -1,$$

továbbá a $\mu=\mu_{n+1}$ pontban a periodikus trajektória elveszti stabilitását. Létezik egy olyan $\mu_n<\bar{\mu}_n<\mu_{n+1}$, ahol $\frac{\partial}{\partial x} f^{(2^n)}(x; \bar{\mu}_n)|_{x=x(\bar{\mu}_n)}=0$. ilyenkor azt mondjuk, hogy az $f^{(2^n)}(x; \bar{\mu}_n)$ leképezésnek szuperstabilis fixpontja van. A mondott feltételek értelmében $x(\bar{\mu}_n)=0$. Technikailag kissé kényelmesebb a $\bar{\mu}_n$ értékek sorozatát követni.

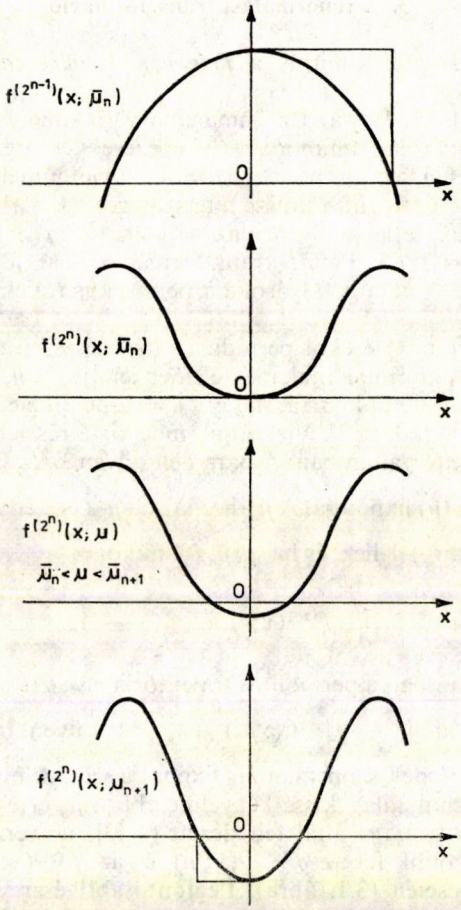
Térjünk rá a FEIGENBAUM által felfedezett (l. [3]) univerzalitás bizonyításának alapgondolatára. Rajzoljuk fel az $f^{(2^{n-1})}(x; \mu)$ és az $f^{(2^n)}(x; \mu)$ leképezések grafikonját $\bar{\mu}_n\leqq\mu\leqq\bar{\mu}_{n+1}$ esetén (3.1. ábra). Legfontosabb észrevételünk a következő: az $f^{(2^{n-1})}(x; \bar{\mu}_n)$ és az $f^{(2^n)}(x; \bar{\mu}_{n+1})$ leképezések nagy n-ekre a nullák valamely n-től függő környezetében aszimptotikusan megegyeznek skálatranszformáció és az x tengelyre való tükrözés erejéig. Ténylegesen a μ paraméter eltolása és normálása, az x tengelyre való tükrözés, valamint skálatranszformáció után az összes

$$f^{(2^{n-1})}(x; \mu), \quad \bar{\mu}_n \leqq \mu \leqq \bar{\mu}_{n+1} \quad \text{és} \quad f^{(2^n)}(x; \mu), \quad \bar{\mu}_{n+1} \leqq \mu \leqq \bar{\mu}_{n+2}$$

leképezéscsalád aszimptotikusan egybeesik.

Tehát sorozatosan megkettőzve a leképezéscsaládot és végrehajtva a paraméter normálását, valamint a skálatranszformációt határértékként olyan leképezéscsaládot kapunk, amely invariáns a bevezetett transzformációkra. Természetes gondolat, hogy a határértéket a kiindulási leképezéscsaládtól függetlenül teljes egészében meghatározzák a végrehajtott transzformációk. Többek között e meggondolásokból származik a *Feigenbaum-elmélet* pontos tartalma.

Ha feltevéseink ésszerűek, akkor a G-osztályú leképezések is 2^n iteráció, x megfelelő normálása és $n\to\infty$ esetén ugyanarra a leképezésre vezetnek, a kezdeti leké-



3.1. ábra

pezés megválasztásától függetlenül. Első lépésként meg kell találnunk ezt az univerzális leképezést. Most ezt végezzük el.

3.2. Definiáljuk a renormálási transzformációt, amely a $[-1, 1]$ szakaszt önmagába képező leképezések terében hat. Legyen $f(x)$ a $[-1, 1]$ szakasz önmagába való páros unimodális leképezése, amelynek $x=0$ maximumhelye. Legyen

$$\alpha = \alpha(f) = -\frac{f(0)}{f(f(0))}.$$

f a $[-\alpha^{-1}, \alpha^{-1}]$ szakaszt az $[f(\alpha^{-1}), f(0)]$ szakaszra képezi le, amelynek képe $[f(f(0)), f(f(\alpha^{-1}))]$.

Tegyük fel, hogy teljesülnek az alábbi feltételek:

$$\alpha > 0, \quad f(f(\alpha^{-1})) < \alpha^{-1}, \quad \alpha^{-1} < f(\alpha^{-}), \quad f(0) > 0.$$

Ekkor $[f(f(0)), f(f(\alpha^{-1}))] \subset [-\alpha^{-1}, \alpha^{-1}]$ és $[f(\alpha^{-1}), f(0)] \cap [-\alpha^{-1}, \alpha^{-1}] = \emptyset$. Emiatt a $h(x) = -\alpha f(f(\alpha^{-1}x))$ leképezés ismét a $[-1, 1]$ szakasz önmagába való unimodális leképezése, amelyre $h(0) = f(0)$. Most definiáljuk a T renormálási transzformációt:

$$(Tf)(x) = -\alpha f(f(\alpha^{-1}x)), \quad \alpha = -\frac{f(0)}{f(f(0))}.$$

Megjegyezzük, hogy ha T értelmes valamely f leképezésre, akkor értelmes az f-hez közeli leképezésekre is. Ebben a paragrafusban lényegében T és a hozzá hasonló transzformációk tulajdonságaival foglalkozunk. Most megvilágítjuk, hogy milyen módon kapcsolódik T a periódusketttőződéses bifurkációsorozat univerzalitásához.

Tekintsük a $[-1, 1]$ szakasz önmagába való (nem feltétlenül páros) $f(x)$ leképezéseinek terét, amelyekre $f(x) \in C^2([-1, 1])$, az $x = 0$ maximumhely és $f(0) = \text{const}$. Legyen például $\text{const} = 1$. Ez a tér invariáns T-re vonatkozóan. Kiderül, hogy T-nek ebben a térben van $g(x)$ fixpontja, ugyanis a linearizált transzformáció $DT(g)$ g-beli spektruma egy sajátérték kivételével az egységkör belsejébe esik. Ez az egyetlen 1-nél nagyobb sajátérték is a *Feigenbaum-konstans:* $\delta = 4{,}6692\ldots$ Ezért a g fixponton áthalad egy $\Gamma^{(u)}(g)$ egydimenziós instabilis szeparatrix; ez olyan leképezésekből áll, amelyek T hatására távolodnak g-től. Átmegy továbbá g-n egy $\Gamma^{(s)}(g)$ stabilis szeparatrix, amelynek kodimenziója 1; ez olyan leképezésekből áll, amelyek T hatására vonzódnak g-hez. Az instabilis $\Gamma^{(u)}(g)$ szeparatrix a δ sajátértéknek felel meg.

Már korábban megjegyeztük, hogy periódusketttőződéses bifurkáció akkor játszódik le, amikor a leképezés fixpontbeli deriváltja áthalad a -1-en. Legyen Σ_1 az olyan leképezések 1-kodimenziójú hiperfelülete a függvénytérben, amelyek deriváltja a fixpontban -1. Ha egy leképezéscsalád transzverzálisan metszi a Σ_1 felületet, akkor periódusketttőződéses bifurkáció jön létre. Legyen

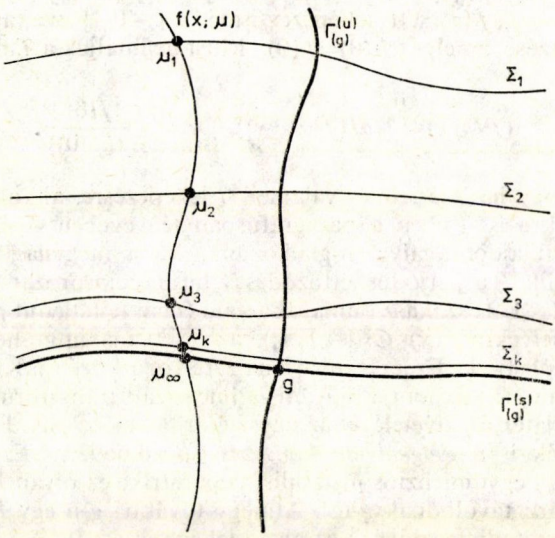
$$\Sigma_2 = T^{-1}\Sigma_1, \ldots, \Sigma_k = T^{-1}\Sigma_{k-1}.$$

Megjegyezzük, hogy ha az $f(x; \mu)$ leképezéscsalád a Σ_k felületet metszi, akkor a 2^{k-1} periódusú stabilis trajektóriából 2^k periódusú stabilis trajektória lesz, azaz a leképezéscsalád és Σ_k metszetének megfelelő paraméterértékek szintén bifurkációs értékek. Valóban, ha $f(x; \mu)$ metszi Σ_k-t, akkor $T^{k-1}f$ metszi Σ_1-et, eszerint periódusketttőződéses bifurkáció játszódik le és egy 2^k periódusú trajektória keletkezik.

A $\Gamma^{(u)}(g)$ instabilis szeparatrix transzverzálisan metszi Σ_1-et és így minden Σ_k-t is, $k \geq 2$-re (l. a 3.2. ábrát).

Látható, hogy a Σ_k felületek $\Gamma^{(s)}(g)$-hez konvergálnak, mivel nagy k esetén Σ_{k+1} és $\Gamma^{(s)}(g)$ távolsága δ-szor kisebb Σ_k és $\Gamma^{(s)}(g)$ távolságánál. Emiatt tetszőleges, a $\Gamma^{(u)}(g)$ valamely környezetében levő $f(x; \mu)$ leképezéscsalád bifurkációs értékei is kielégítik a $\mu_\infty - \mu_k \sim \text{const}\, \delta^{-k}$ összefüggést, ahol const a leképezéscsaládtól függ, értékét a $\frac{\partial}{\partial \mu} f(x; \mu)|_{\mu=\mu_\infty}$ határozza meg. A g leképezés univerzális abban az értelemben, hogy $\lim_{k\to\infty} T^k f(x; \mu_\infty) = g$. Természetesen a $\Gamma^{(u)}(g)$ instabilis szeparatrix is hasonló univerzális tulajdonságokkal rendelkezik (l. később).

A továbbiakban kissé formálisabb értelmezését adjuk meg a *Feigenbaum-univerzalitásnak*. Megfogalmazzuk a T transzformációnak azokat a tulajdonságait, amelyek az univerzalitáshoz vezetnek.



3.2. ábra

1. A T transzformációnak létezik g fixpontja.

2. A linearizált $DT(g)$ transzformációnak egyetlen olyan sajátértéke van, amelynek abszolút értéke 1-nél nagyobb; $\delta = 4{,}6692\ldots$

3. A δ-nak megfelelő instabilis szeparatrix metszi a Σ_1 felületet.

Ma már a T transzformáció 1, 2, 3 tulajdonságai bizonyítottnak tekinthetők.

A bizonyítás O. E. LANFORD [29], valamint M. CAMPANINO és H. EPSTEIN [30] munkájában található. Bizonyos értelemben ezek különböznek a szokásos matematikai bizonyításoktól. Mindkét munka számítógépes vizsgálatok eredményére támaszkodik, és ellenőrzésükre nincs más mód, mint a számítások megismétlése. Konkrétabban, a kutatási irány a következő. Legyen $\tilde{g}$ páros $2m$-ed fokú polinom. Ekkor $-\alpha\tilde{g}(\tilde{g}(\alpha^{-1}x))$ $4m^2$ fokú polinom. A feladat $-\alpha\tilde{g}(\tilde{g}(\alpha^{-1}x))$-hez közeli $\tilde{g}$-ot találni. Egy lehetőség a csonkított $\tilde{g}_1(x)$ polinom vizsgálata, azaz $-\alpha\tilde{g}(\tilde{g}(\alpha^{-1}x))$ magasabb fokú tagjainak elhanyagolása. Ezután olyan $\tilde{g}$ polinomot keresünk, amelyre $\tilde{g} = \tilde{g}_1$. Egy másik lehetőség, hogy előre kiválasztott pontokban egyenlővé tesszük $\tilde{g}(x)$ és $-\alpha\tilde{g}(\tilde{g}(\alpha^{-1}x))$ értékét. Ezen a módon dolgozott O. E. LANFORD számítógép segítségével. [29]-ben találhatók számításának eredményei, amelyek néhol elérik a 10^{-40} pontosságot. A T transzformáció valódi fixpontjának nagy pontosságú $\tilde{g}$ közelítése után elvégezhető az egzakt linearizálás $\tilde{g}$ környezetében és annak igazolása a Newton-módszerrel, hogy a renormálási transzformációnak az adott környezetben van fixpontja. [31]-ben a $g(x)$ komplex síkra való analitikus folytatásának tulajdonságait tanulmányozták. Felírjuk g kifejtésének néhány tagját:

$$g(x) \approx 1 - 1{,}52763x^2 + 0{,}104815x^4 - 0{,}0267057x^6 + \ldots$$

$$\alpha = \alpha(g) = 2{,}50290\ldots$$

α is a renormálási transzformációhoz kapcsolódó univerzális konstans, amely a skálázás mértékét jellemzi. A továbbiakban α-val mindig a 2,50290... számot jelöljük.

Általában a T transzformációnak sok más fixpontja is van. Tekintsük például az $f(|x|^\gamma)$, $\gamma > 1$ alakú leképezések terét, ahol f a $[0, 1]$ szakasz valamely környezetében analitikus függvény. Ez a tér invariáns a T-re vonatkozóan és található benne fixpont. Nyilvánvaló, hogy minden $\gamma > 1$ esetén létezik a fenti alakú fixpont és a térben a linearizált transzformációnak egyetlen 1-nél nagyobb sajátértéke van. Természetesen ez a $\delta(\gamma)$ érték jellemzi a perióduskettőződéses bifurkációsorozat univerzális tulajdonságait arra a leképezéscsaládra, amely a kritikus pontban $|x|^\gamma$ jellegű. Ha $\gamma = 1+\varepsilon$, ahol ε kicsi, alkalmazható a perturbációelmélet az 1—3 tulajdonságok szigorú bizonyítására (l. [32]). Ebben az esetben $\delta(1+\varepsilon) \to 2$, ha $\varepsilon \to 0$. Az alkalmazások szempontjából legfontosabb $\varepsilon = 1$, azaz $\gamma = 2$ eset tárgyalása nehezen vihető végig a perturbációelmélet alapján.

3.3. A páros leképezések tere invariáns T-re. A g fixpont és a δ sajátértékhez tartozó $g(\delta)$ sajátfüggvény szintén páros. Ebben a pontban a $DT(g)$ linearizált operátor tulajdonságaival foglalkozunk, megmutatjuk, hogy spektrumának a nem páros leképezésekhez tartozó része triviális abban az értelemben, hogy explicit módon megadható.

T-t az analitikus leképezések $M(\mathcal{D})$ *Banach-terén* értelmezett transzformációnak tekintjük, ahol $\mathcal{D}$ a $[-1, 1]$ szakasz valamely komplex környezete.

A T transzformációnak létezik a következő tulajdonságokat kielégítő fixpontja:

a) $g(x) = -\alpha g(g(\alpha^{-1}x))$, $\alpha = 2{,}50290...$,

b) $g(0) = 1$, $g'(x) > 0$, ha $x < 0$, $g'(x) < 0$ ha $x > 0$,

c) $g(x) = g(-x)$.

Jelölje $\tilde{T}$ azt a transzformációt, amely T-hez hasonló, de $\alpha(g) = 2{,}50290...$ rögzített:

$$(\tilde{T}f)(x) = -\alpha f(f(\alpha^{-1}x)), \quad \alpha = 2{,}50290...$$

Nyilvánvalóan $\tilde{g}(x)$ a $\tilde{T}$ transzformációnak is fixpontja. Most felírjuk a $D\tilde{T}(g)$, $DT(g)$ linearizált transzformációk explicit alakját:

$$(D\tilde{T}(g))h(x) = -\alpha g'(g(\alpha^{-1}x))h(-\alpha^{-1}x) - \alpha h(g(\alpha^{-1}x)),$$

$$(DT(g))h(x) = (D\tilde{T}(g))h(x) + c(h)e_1,$$

$$c(h) = (\alpha^2 - 1)h(0) - \alpha h(1), \quad e_1 = g'(x)x - g(x).$$

Látni fogjuk, hogy e_1 az 1 sajátértékhez tartozó sajátvektora mind a $D\tilde{T}(g)$, mind pedig a $DT(g)$ operátornak (nem nehéz igazolni, hogy $c(e_1) = 0$). A $D\tilde{T}(g) - DT(g)$ különbség az e_1 sajátirányra való projekció, ezért a két operátor spektruma egybeesik.

Vizsgáljuk meg $D\tilde{T}(g)$ spektrumát (l. [9]).

Montel tételéből és a $D\tilde{T}(g)$ explicit alakjából következik, hogy az operátor kompakt, tehát spektruma diszkrét.

Legyen $S_n(\varepsilon)$ a következő koordinátatranszformáció:

$$S_n(\varepsilon): \; x \mapsto x + \varepsilon x^n, \quad n \geqq 0.$$

Elegendően kis ε-ra a $[-1, 1]$ szakasz valamely környezetében létezik az $S_n^{-1}(\varepsilon)$ inverz.

Definiáljuk az e_k, $k \geqq 0$ vektorokat:

$$e_k = \frac{d}{d\varepsilon} S_k^{-1}(\varepsilon) \circ g \circ S_k(\varepsilon)|_{\varepsilon=0} = g'(x)x^k - (g(x))^k.$$

Megmutatjuk, hogy az e_k, $k \geqq 0$ vektorok a $D\tilde{T}(g)$ operátor $(-\alpha^{-1})^{k-1}$ sajátértékeihez tartozó sajátvektorai. Jelölje A az $Ax = -\alpha^{-1}x$ lineáris transzformációt. Ekkor

$$(D\tilde{T}(g))e_k = \frac{d}{d\varepsilon} \tilde{T}(S(\varepsilon) \circ g \circ S_k(\varepsilon))|_{\varepsilon=0},$$

$$\tilde{T}(S_k^{-1}(\varepsilon) \circ g \circ S_k(\varepsilon)) = A^{-1} \circ (S_k^{-1}(\varepsilon) \circ g \circ S_k(\varepsilon)) \circ (S_k^{-1}(\varepsilon) \circ g \circ S_k(\varepsilon)) \circ A =$$

$$= A^{-1} \circ S_k^{-1}(\varepsilon) \circ A \circ (A^{-1} \circ g \circ g \circ A) \circ A^{-1} \circ S_k(\varepsilon) \circ A =$$

$$= A^{-1} \circ S_k^{-1}(\varepsilon) \circ A \circ g \circ A^{-1} \circ S_k(\varepsilon) \circ A.$$

Felhasználtuk, hogy $A^{-1} \circ g \circ g \circ A = g$. Megjegyezzük, hogy $A^{-1} \circ S_k(\varepsilon) \circ A = S_k(\varepsilon(-\alpha^{-1})^{k-1})$. Emiatt

$$(D\tilde{T}(g))e_k = \frac{d}{d\varepsilon} S_k^{-1}(\varepsilon(-\alpha^{-1})^{k-1}) \circ g \circ S_k(\varepsilon(-\alpha^{-1})^{k-1})|_{\varepsilon=0} = (-\alpha^{-1})^{k-1} e_k.$$

Az e_k, $k \geqq 0$ sajátvektorok a koordinátatranszformációhoz kapcsolódnak, így nem lényegesek az univerzalitás sajátosságainak vizsgálatában. Könnyű bebizonyítani, hogy az e_k, $k \geqq 0$ vektorok által kifeszített altér kodimenziója végtelen (l. [8], [9]). Az $e_0 = g'(x) - 1$ vektornak megfelelő sajátérték $(-\alpha) = -2,5\ldots$ Ezt a vektort az $f'(0) = 0$ feltétel zárja ki. Az 1 sajátértékhez tartozó e_1 vektor pedig az $f(0) = \text{const}$ feltétel miatt zárható ki. Az e_k, $k \geqq 2$ vektoroknak megfelelő sajátértékek abszolút értéke pedig 1-nél kisebb.

Már említettük, hogy a páros függvények $M^{(e)}(\mathscr{D})$ tere invariáns a $D\tilde{T}(g)$ operátorra. Ebbe a térbe esik az operátor spektrumának lényeges része, speciálisan a δ sajátértékhez tartozó $g(\delta)$ sajátvektor. Megmutatjuk, hogy tetszőleges $f \in M(\mathscr{D})$ függvény egyértelműen fejthető az alábbi sorba:

$$f = \sum_{k=0}^{\infty} f_{2k} e_{2k} + f^{(e)},$$

ahol $f^{(e)}$ páros függvény. Legyen

$$f(x) = \sum_{k=0}^{\infty} c_k x^k.$$

Mivel

$$\sum_{k=0}^{\infty} f_{2k} e_{2k} = \left(\sum_{k=0}^{\infty} f_{2k} x^{2k}\right) g'(x) - \sum_{k=0}^{\infty} f_{2k} (g(x))^{2k},$$

$$\left(\sum_{k=0}^{\infty} f_{2k} x^{2k}\right) g'(x) = \sum_{k=0}^{\infty} c_{2k+1} x^{2k+1},$$

$$f^{(e)} = \sum_{k=0}^{\infty} c_{2k} x^{2k} + \sum_{k=0}^{\infty} f_{2k} (g(x))^{2k}.$$

Felhasználtuk, hogy $g'(x)$ páratlan függvény. Tehát az f_{2k}, $k=0, 1, \ldots$ együtthatókat és az $f^{(e)}$ függvényt egyértelműen meghatározzák az alábbi összefüggések:

$$\sum_{k=0}^{\infty} f_{2k} x^{2k} = \frac{x}{g'(x)} \sum_{k=0}^{\infty} c_{2k+1} x^{2k},$$

$$f^{(e)} = \sum_{k=0}^{\infty} c_{2k} x^{2k} + \frac{g(x)}{g'(g(x))} \sum_{k=0}^{\infty} c_{2k+1} (g(x))^{2k}.$$

3.4. Mint láttuk, a *Feigenbaum-univerzalitás* bizonyításához meg kell találnunk a T transzformáció fixpontját, meg kell konstruálnunk a fixpont instabilis szeparatrixát és meg kell győződnünk arról, hogy a szeparatrix metszi a periódusketttőződés felületét.

Az alábbiakban ismertetjük az instabilis szeparatrix numerikus meghatározásának módját (l. [33]). Tekintsük az egyparaméteres egydimenziós leképezéscsaládok terében értelmezett T^* transzformációt. Legyen $f(x; \mu)$ a szakasz önmagába való páros, μ-vel parametrizált leképezéscsaládja. Tegyük fel, hogy $\mu=0$-nál az $x=0$ kritikus pont képe 1, az 1-é pedig 0, azaz:

$$f(0; 0) = 1, \quad f(1; 0) = 0.$$

Legyen $f^{(2)}(x; \mu)=f(f(x; \mu); \mu)$ és legyen $\bar{\mu}(f)$ a paraméter olyan minimális pozitív értéke, amelyre az $x=0$ pont az $f^{(2)}(x; \mu)$ leképezésnek 2 periódusú pontja, azaz $f^{(2)}(f^{(2)}(0; \bar{\mu}(f)); \bar{\mu}(f))=0$. Most definiálhatjuk a T^* transzformációt:

$$(T^*f)(x; \mu) = -\alpha(f) f^{(2)}(\alpha^{-1}(f)x; \ \bar{\mu}(f)(1+\mu)),$$

ahol $\alpha(f)=-(f^{(2)}(0; \bar{\mu}(f)))^{-1}$ a $(T^*f)(0; 0)=1$ normálási feltételből adódik.

Nem nehéz bizonyítani, hogy a $\tilde{T}$ transzformáció g fixpontjának instabilis szeparatrixa skálázás erejéig megegyezik a T^* stabilis fixpontjával. Jelölje $g(x; \mu)$ a T^* transzformáció fixpontját, $\bar{\mu}=\bar{\mu}(g)$ és $\alpha=\alpha(g)$ pedig a megfelelő állandókat. Az instabilis szeparatrixon úgy választjuk a paraméterezést, hogy a $g(x)$ leképezés a $\mu=\bar{\mu}/(1-\bar{\mu})$ értéknek feleljen meg. Megjegyezzük, hogy α új értéke megegyezik a régivel, továbbá $\bar{\mu}=\delta^{-1}$.

A $g(x; \mu)$ fixpont meghatározható numerikusan. Ekkor $g(x; \mu)$-t x és μ polinomjaként kapjuk meg:

$$g(x; \mu) = \sum_{i=0}^{k} \sum_{j=0}^{m} g_{ij} x^{2i} \mu^j.$$

A T^* transzformációval együtt tekintsük a $T^*_{k,m}$ transzformációt, amely úgy keletkezik, hogy a kettőződés után elhanyagoljuk az $x^{2p}\mu^q$ tagokat, ha $p>k$ vagy $q>m$. $T^*_{k,m}$ segítségével iterálva valamely $f(x; \mu)=\sum_{i=0}^{k} \sum_{j=0}^{m} f_{ij} x^{2i} \mu^j$ kezdeti leképezéscsaládot, határértékként a $g^{(k,m)}(x; \mu)$ fixpont közelítését kapjuk. A $k=4$, $m=5$ és $k=5$, $m=6$ esetre vonatkozó $g_{ij}^{(k,m)}$ együtthatókat a 3.1. táblázatban tüntettük fel. Megjegyezzük, hogy $k=5$, $m=6$ esetén a $(\bar{\mu}; \alpha)$ állandó értéke a véges feladatban (0,214115; 2,50306), míg a pontos érték (0,214169; 2,50290). A $T^*_{k,m}$ transzformációk másodrendű deriváltakra vonatkozó becslést is magában foglaló részletesebb vizsgálata lehetővé teszi, hogy szigorúan bebizonyítsuk T^* fixpontjának létezését, to-

3.1. TÁBLÁZAT

i \ j	0	1	2	3	4	5	6
0	1,000−0	1,334−0	3,040−1	8,675−3	−3,309−3	−3,781−4	—
	1,000−0	1,282−0	2,811−1	6,771−3	−3,456−3	−5,274−4	−3,345−5
1	−1,032−0	−2,626−1	9,840−3	9,577−3	1,515−3	1,096−4	—
	−1,046−0	−2,569−1	9,685−3	8,878−3	1,371−3	1,061−4	2,526−6
2	4,441−2	−1,562−2	−7,950−3	−1,390−3	−1,205−4	−2,575−6	—
	4,577−2	−1,582−2	−7,684−3	1,209−3	−4,435−5	1,952−5	3,095−6
3	4,374−3	2,691−3	5,241−4	7,531−6	−1,721−5	−2,736−6	—
	4,676−3	2,777−3	4,969−4	−2,561−5	−3,567−5	−8,755−6	−8,744−7
4	−3,193−4	−8,448−5	2,037−5	1,655−5	4,072−6	4,072−7	—
	−3,515−4	−8,995−5	2,576−5	2,205−5	6,912−6	1,242−6	1,095−7
5	—	—	—	—	—	—	—
	4,539−6	−8,033−6	−6,907−6	−2,594−6	−6,235−7	−1,024−7	−8,975−9

(*Magyarázat:* A táblázat minden helyén az első szám a $k=4$, $m=5$, a második szám pedig a $k=5$, $m=6$ esetnek felel meg; a 2,561-5 jelölés jelentése: $2{,}561\times10^{-5}$.)

vábbá explicit kritériumok adódnak a perióduskettőződéses bifurkációsorozat megjelenésére egydimenziós leképezéscsaládok esetében.

A T^* transzformáció és a $g(x;\mu)$ fixpont felhasználásával egyszerűen magyarázható az univerzalitás jelensége. Legyen az $f(x;\mu)$ leképezéscsalád a $g(x;\mu)$ fixpont vonzási tartományában. Tekintsük a következő sorozatot: $f_0(x;\mu)=f(x;\mu)$, $f_1(x;\mu)=T^*f_0(x;\mu)$, $f_n(x;\mu)=T^*f_{n-1}(x;\mu)=-\alpha(n)f_{n-1}(f_{n-1}(\alpha(n)^{-1}x;\bar\mu(n)(1+\mu))$; $\bar\mu(n)(1+\mu))$. Ha $n\to\infty$, akkor $f^{(n)}(x;\mu)\to g(x;\mu)$, $\alpha(n)\to\alpha$, $\bar\mu(n)\to\bar\mu$. Legyen $\bar\mu_n$ a szuperstabilis 2 periódusú trajektóriáknak megfelelő paraméterértékek sorozata, $n=1,2,\ldots$ T^* definíciójából közvetlenül adódik, hogy

$$\bar\mu_{n+1}-\bar\mu_n=\bar\mu(1)\bar\mu(2)\ldots\bar\mu(n).$$

Tehát

$$\frac{\bar\mu_{n+1}-\bar\mu_n}{\bar\mu_n-\bar\mu_{n-1}}=\bar\mu(n)\to\bar\mu=\delta^{-1},\quad\text{ha}\quad n\to\infty.$$

Megemlítjük, hogy az instabilis szeparatrix egy másik konstrukciója található a [34], [35] munkákban.

4. A Feigenbaum-attraktor tulajdonságai

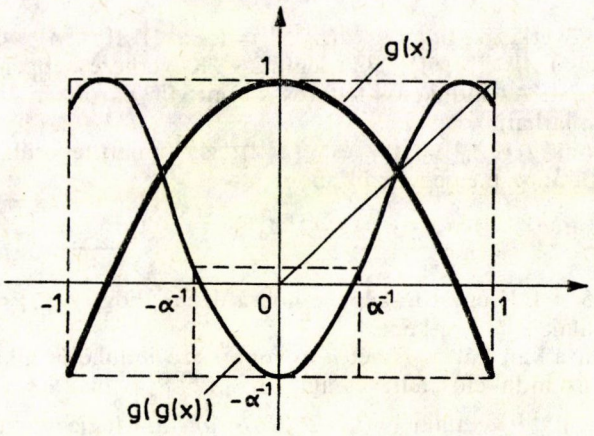
4.1. Az előző paragrafusban láttuk, hogy a *Feigenbaum-elméletben* alapvető szerepet játszik a T transzformáció fixpontja, amely kielégíti az alábbi egyenletet:

$$g(x)=-\alpha g(g(\alpha^{-1}x)),\quad \alpha=g(1). \tag{4.1}$$

Ezt az egyenletet a továbbiakban renormálási egyenletnek nevezzük. Elsőként P. CVITANOVIC és M. J. FEIGENBAUM jutott el hozzá ([36], [3]). Hasonló egyenlet szerepel [37]-ben is. A (4.1) egyenletnek megfelelően a $(-\alpha^{-1},\alpha^{-1})$ intervallumon vett

$g(g(x))$ függvény grafikonja a vízszintes tengelyre való tükrözés és α skálafaktorral történő nyújtás után keresztezi a $g(x)$ függvény grafikonját. A $g(x)$ függvény hozzávetőleges alakját vázoltuk a 4.1. ábrán.

g-vel fogjuk jelölni a $[-1, 1]$ szakasz önmagába való leképezését is, amelyet az adott $g(x)$ függvény határoz meg. Azonnal látni fogjuk, hogy a leképezés a 2. fejezetben bevezetett G osztályba tartozik és a megfelelő $\Delta_k^{(n)}$, $0 \leqq k < 2^n$ szakaszok, amelyek most explicite megadhatók, egy sor fontos tulajdonsággal rendelkeznek.



4.1. ábra

Világos, hogy g-nek van instabilis $x_0^{(0)}$ fixpontja. Jelölje $\Delta_0^{(1)}$ a $[-\alpha^{-1}, \alpha^{-1}]$ szakaszt. Ekkor a $g(\Delta_0^{(1)}) \equiv \Delta_1^{(1)} = [g(\alpha^{-1}), 1]$ szakasz nem metszi $\Delta_0^{(1)}$-et, és az instabilis $x_0^{(0)}$ pont $\Delta_0^{(1)}$ és $\Delta_1^{(1)}$ közé esik. Továbbá a renormálási egyenletből azonnal következik, hogy $g(\Delta_1^{(1)}) \subset \Delta_0^{(1)}$: ha $x \in [-1, 1]$, akkor $\alpha^{-1}x \in [-\alpha^{-1}, \alpha^{-1}]$ és $g(g(\alpha^{-1}x)) = -\alpha^{-1}g(x) \in [-\alpha^{-1}, \alpha^{-1}]$, mivel $|g(x)| < 1$. Legyen most $\Delta_0^{(n)} = [-\alpha^{-n}, \alpha^{-n}]$, $\Delta_k^{(n)} = g^{(k)}(\Delta_0^{(n)})$, $0 \leqq k < 2^n$.

4.1. TÉTEL. A $\Delta_k^{(n)}$, $0 \leqq k < 2^n$ szakaszok a következő tulajdonságokkal rendelkeznek:

1°. $g^{(2^n)}(\Delta_0^{(n)}) \subset \Delta_0^{(n)}$.

2°. Minden $\Delta_k^{(n-1)}$ szakasz csak a $\Delta_k^{(n)}$ és a $\Delta_{k+2^{n-1}}^{(n)}$ szakaszt tartalmazza, $0 \leqq k < 2^{n-1}$.

3°. $\Delta_k^{(n)}$ különböző k-kra diszjunkt, $0 \leqq k < 2^n$.

4°. $\Delta_{2k+1}^{(n)} \subset \Delta_0^{(1)}$, $\Delta_{2k}^{(n)} \subset \Delta_0^{(1)}$.

5°. Minden $\Delta_{2k}^{(n+1)}$ szakasz a következő két transzformáció kompozíciójával keletkezik $\Delta_k^{(n)}$-ből: α^{-1} együtthatóval történő zsugorítás és az $x \to -x$ tükrözés.

6°. Jelölje $|\Delta|$ a szakasz hosszát. Ekkor

$$|\Delta_0^{(n)}| = \max_{0 \leqq k < 2^n} |\Delta_k^{(n)}|, \quad |\Delta_1^{(n)}| = \min_{0 \leqq k < 2^n} |\Delta_k^{(n)}|.$$

Bizonyítás. Az 1° tulajdonság a renormálási egyenlet következménye, ugyanis $g^{(2^n)}(\alpha^{-n}x) = (-1)^n \alpha^{-n} g(x)$.

A 2° és 3° tulajdonságot n-re vonatkozó indukció segítségével egyszerre bizonyítjuk. Tegyük fel, hogy n-re teljesülnek és igazoljuk $n+1$-re. A renormálási egyenletből, felhasználva, hogy g páros:

$$\underbrace{g\circ\ldots\circ g}_{2^n}(\alpha^{-(n+1)}x) = \underbrace{g\circ\ldots\circ g}_{2^n-2}\big(-\alpha^{-1}g(\alpha^{-n}x)\big) =$$

$$= \underbrace{g\circ\ldots\circ g}_{2^n-2}\big(\alpha^{-1}g(\alpha^{-n}x)\big) = \ldots = -\alpha^{-1}\underbrace{g\circ\ldots\circ g}_{2^{n-1}}(\alpha^{-n}x) = \ldots = (-\alpha^{-1})^n g(\alpha^{-1}x).$$

Ebből azonnal következik, hogy $g^{(2^n)}(\Delta_0^{(n+1)}) = (-\alpha^{-1})^n\Delta_1^{(1)} \subset \Delta_0^{(n)}$ és $g^{(2^n)}(\Delta_0^{(n+1)}) \cap \cap \Delta_0^{(n+1)} = \emptyset$. Tehát $\Delta_0^{(n+1)}$ és $\Delta_{2^n}^{(n+1)}$ kielégíti 2°-t. Következésképpen $\Delta_k^{(n)}$ tartalmazza $\Delta_k^{(n+1)}$-t és $\Delta_{k+2^n}^{(n+1)}$-t. Az indukciós feltevés szerint $\Delta_k^{(n)}$ páronként diszjunkt, amiből 2° és 3° könnyen látható.

A 4° tulajdonság $g(\Delta_0^{(1)}) = \Delta_1^{(1)}$ és $g(\Delta_1^{(1)}) \subset \Delta_0^{(1)}$ miatt teljesül.

5° bizonyításához megjegyezzük, hogy

$$\Delta_{2k}^{(n+1)} = \underbrace{g\circ\ldots\circ g}_{2k}(\Delta_0^{(n+1)}) = \underbrace{g\circ\ldots\circ g}_{2k}(\alpha^{-1}\Delta_0^{(n)}) = -\alpha^{-1}\underbrace{g\circ\ldots\circ g}_{2k}(\Delta_0^{(n)}) = -\alpha^{-1}\Delta_k^{(n)},$$

ami ekvivalens 5°-tel. Kiegészítésképpen hozzáfűzzük, hogy $\Delta_{2k+1}^{(n+1)} = g(\Delta_{2k}^{(n+1)})$, azaz a fentebb leírt alak $\Delta_{2k}^{(n+1)}$ g-képe.

6° bizonyítása van hátra. Ismét n-re vonatkozó indukciót alkalmazunk. Az 5° tulajdonság és az indukciós feltevés alapján $|\Delta_0^{(n+1)}| = \max\limits_{0\leqq k\leqq 2^n} |\Delta_{2k}^{(n+1)}|$. Tegyük fel, hogy $|\Delta_{2k+1}^{(n+1)}| > |\Delta_0^{(n+1)}|$ valamely 0 és 2^n közötti k-ra. Megjegyezzük, hogy $\Delta_{2k+1}^{(n+1)} \subset \subset \Delta_1^{(1)}$ (4° tulajdonság), továbbá g alakjából közvetlenül következik, hogy $|g'(x)| > 1$, ha $x \in \Delta_1^{(1)}$. Ezért $|\Delta_{2k+2}^{(n+2)}| = |g(\Delta_{2k+1}^{(n+1)})| > |\Delta_{2k+1}^{(n+1)}| > |\Delta_0^{(n+1)}|$, ami ellentmond az imént bizonyítottaknak.

Most belátjuk, hogy $\min\limits_{0\leqq k\leqq 2^{n+1}} |\Delta_k^{(n+1)}| = |\Delta_1^{(n+1)}|$ (ismét n szerinti teljes indukcióval). 5° alapján

$$|\Delta_{2k}^{(n+1)}| = \alpha^{-1}|\Delta_k^{(n)}| \geqq \alpha^{-1}|\Delta_1^{(n)}| > |\Delta_1^{(n+1)}|.$$

Az utolsó egyenlőtlenség közvetlenül igazolható. Ebből és 5°-ből

$$\min_{0\leqq k<2^n} |\Delta_{2k}^{(n+1)}| = \alpha^{-1} \min_{0\leqq k<2^n} |\Delta_k^{(n)}| = \alpha^{-1}|\Delta_1^{(n)}| = |\Delta_2^{(n+1)}| > |\Delta_1^{(n+1)}|.$$

Tegyük fel, hogy $|\Delta_{2k+1}^{(n+1)}| < |\Delta_1^{(n+1)}|$ valamely 0 és 2^n közötti k-ra. Ekkor g konvexitása miatt

$$\min_{x\in\Delta_1^{(n+1)}} |g'(x)| > \max_{x\in\Delta_{2k+1}^{(n+1)}} |g'(x)|,$$

tehát $|\Delta_{2k+2}^{(n+1)}| = |g(\Delta_{2k+1}^{(n+1)})| < |g(\Delta_1^{(n+1)})| = |\Delta_2^{(n+1)}|$, ami ellentmond a korábban igazoltaknak. A tétel bizonyítását befejeztük.

4.1. Definíció. Az $F = \bigcap\limits_{n\geqq 1} \bigcup\limits_{k=0}^{2^n-1} \Delta_k^{(n)}$ zárt halmazt *Feigenbaum-attraktornak* nevezzük.

Látható, hogy a $\Delta_k^{(n)}$ szakaszok rendszere hasonló tulajdonságoknak tesz eleget, mint a G-osztályú leképezéseknél bevezetett szakaszrendszer. A 2.1. tétel alapján tehát:

1. A g leképezésnek egyetlen invariáns mértéke van, amelyre nézve g ergodikus, szigorúan ergodikus és diszkrét spektruma a következő számokból áll:

$$\exp\{2\pi i(2r+1)2^{-n}\}, \quad n=1,2,\ldots, \quad 0\leqq r\leqq 2^{n-1}-1.$$

2. Legyen $\varphi\in C^1([-1,1])$, $\int \varphi d\mu_0=0$. Írjuk fel a *Fourier-együtthatókat* és spektrálelőállításukat:

$$c_n=(U_g^n\varphi,\varphi)_{\mu_0}=\int_0^1 e^{2\pi i\omega n}d\varrho(\omega).$$

Ekkor

$$\varrho=\sum_{n=1}^{\infty}\sum_{r=0}^{2^{n-1}-1}|\varrho_r^{(n)}|^2\delta\left(\omega-\frac{2r+1}{2^n}\right).$$

A 6° tulajdonság és a 2.2. tétel alapján

$$|\varrho_r^{(n)}|\leqq\max_{x\in[-1,1]}|\varphi'(x)|\alpha^{(1-n)}.$$

Ténylegesen a 2.2. tétel bizonyításából erősebb becslés adódik:

$$|\varrho_r^{(n)}|\leqq\frac{1}{2}\max|\varphi'(x)|\Big(\sum_{0\leqq k<2^{n-1}}|\Delta_k^{(n-1)}|\Big)2^{-(n-1)}.$$

Bebizonyítható, hogy

$$\Big(\sum_{0\leqq k<2^n}|\Delta_k^{(n)}|\Big)2^{-n}=O(\sigma^n),$$

ahol $\sigma\approx 0{,}29$ (l. a 4.2. tételt). A δ-amplitúdók részletesebb vizsgálata megtalálható az [5], [38], [68] munkákban. Kimutatták, hogy kis r-ekre teljesül az $\ln|\varrho_r^{(n)}|\sim n\ln\bar{\sigma}$ aszimptotika, ahol $\bar{\sigma}=0{,}15\ldots$, $\ln\Big(\sum_{r=0}^{2^{n-1}-1}|\varrho_r^{(n)}|^2\Big)\sim n\ln\tilde{\sigma}$, $\tilde{\sigma}=0{,}095\ldots$

A továbbiakban szükségünk lesz a g leképezés instabilis periodikus pontjainak szerkezetére vonatkozó információra. Már láttuk, hogy létezik egyetlen, a $\Delta_0^{(1)}$ és $\Delta_1^{(1)}$ közötti $x_0^{(0)}$ instabilis fixpont. A renormálási egyenletből következik, hogy $g^{(2^n)}(\alpha^{-n}x)=(-\alpha^{-1})^n g(x)$, amiből látható, hogy minden $\Delta_0^{(n)}$ szakasz tartalmaz $\Delta_0^{(n+1)}$-t és $\Delta_{2^n}^{(n+1)}$-t elválasztó instabilis 2^n periódusú $x_0^{(n)}$ pontot. Ebből és a 2° tulajdonságból következik, hogy az $x_k^{(n)}=g^{(k)}(x_0^{(n)})$ pont elválasztja a $\Delta_k^{(n)}$ belsejében levő $\Delta_k^{(n+1)}$ és $\Delta_{k+2^n}^{(n+1)}$ szakaszokat. Következésképpen ha $\Delta_{k'}^{(n)}$ és $\Delta_{k''}^{(n)}$ szomszédos szakaszok, akkor közöttük található periodikus $x_k^{(m)}$ pont, $0\leqq m<n$, $0\leqq k<2^m$. Valóban, ha $\Delta_{k'}^{(n)}$ és $\Delta_{k''}^{(n)}$ ugyanabban a $\Delta_k^{(n-1)}$ szakaszban van, akkor az állítást bebizonyítottuk. Ellenkező esetben áttérünk a $\Delta_{k_1'}^{(n-1)}$ és $\Delta_{k_1''}^{(n-1)}$ szakaszokra, amelyek tartalmazzák $\Delta_{k'}^{(n)}$-t, illetve $\Delta_{k''}^{(n)}$-t. Ekkor ezek vagy egy $\Delta_{k_1}^{(n-2)}$ szakaszban vannak, amikor is a bizonyítás kész, vagy áttérünk a $\Delta_{k_2'}^{(n-2)}$ és $\Delta_{k_2''}^{(n-2)}$ szakaszokra, amelyek tartalmazzák $\Delta_{k_1'}^{(n-1)}$-t, illetve $\Delta_{k_1''}^{(n-1)}$-t, és folytatjuk az eljárást.

Most megmutatjuk, hogy a $[-1,1]$ szakaszon a g leképezésnek nincs más periodikus pontja. g alakjából következik, hogy a $[-1,-\alpha^{-1}]$ szakaszban nem lehet periodikus pont, hiszen $g([-1,-\alpha^{-1}])=[-\alpha^{-1},g(\alpha^{-1})]\subset[-\alpha^{-1},1]$ és $g([-\alpha^{-1},1])\subset[-\alpha^{-1},1]$. Továbbá az $[\alpha^{-1},g(\alpha^{-1})]$ szakaszon a $g'(x)$ derivált abszolút értéke nagyobb 1-nél és $g([\alpha^{-1},g(\alpha^{-1})])\subset[-\alpha^{-1},1]$. Ebben az esetben az $x_0^{(0)}$ kivételével az $[\alpha^{-1},g(\alpha^{-1})]$ szakasz minden pontja véges sok lépés után $\Delta_0^{(1)}$-ba

vagy $\Delta_1^{(1)}$-be kerül. Következésképpen az $[\alpha^{-1}, g(\alpha^{-1})]$ szakasz belsejében $x_0^{(0)}$-n kívül nincs periodikus pont. A $\Delta_0^{(n)}$ szakaszra és a $g^{(2^n)}$ leképezésre vonatkozó hasonló meggondolások alapján adódik a kívánt eredmény.

Ezek a megfontolások arra is utalnak, hogy milyen értelemben tekinthető az F halmaz attraktornak. Legyen $x\in[-1, 1]$ olyan pont, amelyre $g^{(m)}(x)\neq x_k^{(n)}$ semmilyen m, n és k esetén. Ekkor minden $n\geqq 0$-ra van olyan r, hogy $g^{(r)}(x)\subset\Delta_k^{(n)}$ valamely k-ra, $0\leqq k<2^n$, azaz $d(g^{(r)}(x), F)\to 0$, ha $r\to\infty$.

Megmutatjuk, hogy a $\Delta_k^{(n)}$ szakaszrendszer segítségével szimbolikus dinamika konstruálható a g leképezésre. (Természetesen ez végrehajtható minden G-osztályú leképezésre is.) Legyen $x\in F$ és $x\in\Delta_{k_n}^{(n)}$ $(n=1, 2, \ldots)$. Ekkor a 2° tulajdonságból következik, hogy $k_{n+1}=k_n$, vagy $k_{n+1}=k_n+2^n$. Írjuk fel a k_n kettes számrendszerbeli alakját: $k_n=\varepsilon_0+\varepsilon_1 2+\ldots+\varepsilon_{n-1}2^{n-1}$, ahol $\varepsilon_i=0$ vagy 1. Ekkor $k_{n+1}=k_n+\varepsilon_n 2^n$. Mivel $x=\bigcap_{n\geqq 1}\Delta_{k_n}^{(n)}$, így egy kölcsönösen egyértelmű φ leképezést kapunk: $x\leftrightarrow(\varepsilon_0, \varepsilon_1, \ldots, \varepsilon_n, \ldots)$, amely az F attraktort a diadikus sorozatok E terébe viszi. Ez a tér úgy tekinthető, mint a diadikus egészek *Abel-csoportja*. Minthogy $g(\Delta_k^{(n)})=\Delta_{k+1}^{(n)}$ ha $0\leqq k\leqq 2^n-1$, így φ a g leképezést az $\bar{\varepsilon}=(1, 0, 0, \ldots)$ elem hozzáadásának műveletébe transzformálja. Jelölje a μ_0 mérték transzformáltját μ_0^*. Könnyű megmutatni, hogy a μ_0^* mértékre vonatkozóan az $\varepsilon_0, \varepsilon_1, \varepsilon_2, \ldots$ koordinátasorozat független valószínűségi változók sorozata, amelyek a 0 és 1 értéket veszik fel 1/2 valószínűséggel. 4.2. Most definiálunk három olyan paramétert, amelyek az F attraktor metrikus tulajdonságait jellemzik.

4.2. TÉTEL. I. Létezik olyan $\gamma<0$ szám, hogy μ_0-majdnem minden $x\in F$-re

$$\lim_{n\to\infty}\frac{1}{n}\ln|\Delta_s^{(n)}(x)|=\gamma,$$

ahol $\Delta_s^{(n)}(x)$ x-et tartalmazó n-ed rangú szakasz.

II. Létezik olyan β_0 szám, hogy

$$\lim_{n\to\infty}\frac{1}{n}\ln\Big[\sum_{k=0}^{2^n-1}|\Delta_k^{(n)}|^{\beta_0}\Big]=0.$$

Ez a szám az F attraktor *Hausdorff-dimenziója*.

III. Léteznek olyan λ, $\sigma>0$ számok, hogy

$$\lim_{n\to\infty}\frac{1}{n}\ln\Big[1+\sum_{k=1}^{2^n-1}\prod_{i=k}^{2^n-1}(g'(x_i))^2\Big]=2(\ln\lambda-\ln\alpha),$$

$$x_i=g^{(i)}(x_0),\quad x_0\in\Delta_0^{(n)};$$

$$\lim_{n\to\infty}\frac{1}{n}\ln\left[\frac{1}{2^n}\Big(\sum_{k=1}^{2^n-1}|\Delta_k^{(n)}|\Big)\right]=\ln\sigma.$$

Az első konstans mutatja, hogy hogyan csökken a μ_0-tipikus x pontokat tartalmazó $\Delta_k^{(n)}$ szakaszok hossza. Ami II-t illeti, az állítható, hogy $s_0\geqq-\frac{\ln 2}{\gamma}$. Az utóbbi összefüggésben azonban nincs egyenlőség, és ez azt mutatja, hogy F a *Hausdorff-dimenzióját* a μ_0-ra vonatkozóan nem tipikus pontjai halmazának terhére éri el.

A λ állandó a g leképezés kis sztochasztikus perturbációinak analízisekor játszik fontos szerepet (1.5. fejezet). Az állandó spektrális becsléseknél használható (l. később).

Bizonyítás. 1. A renormálási egyenletből fontos aszimptotikus összefüggés következik, amelyet 1-univerzalitásnak fogunk nevezni. Az 1 ponttal határos $\Delta_1^{(n)}$ szakasz hosszának nagyságrendje α^{-2n}. Az 1 pont környezetében bevezetjük az $X^{(n)}$ renormált koordinátát: $x=1-\alpha^{-2n}X^{(n)}$. Ha x $\Delta_1^{(n)}$ határai között változik, akkor $X^{(n)}$ a $[0, d_n]$ intervallumba esik, ahol $d_n \to c$, ha $n\to\infty$, és c a következő összefüggésből származtatható: $g(x)=1-cx^2+O(x^4)$, ha $x\to 0$. Továbbá $g^{(2^n)}(\Delta_1^{(n)})\subset\Delta_1^{(n)}$ és célunk az $X^{(n)}$ renormált változó segítségével leírni ezt a leképezést. Adott $X^{(n)}$ érték esetén található olyan $X_0^{(n)}$, hogy $X_0^{(n)}\alpha^{-n}\in\Delta_0^{(n)}$ és $g(X_0^{(n)}\alpha^{-n})=1-\alpha^{-2n}X^{(n)}$. g aszimptotikus kifejezéséből következik, hogy $X_0^{(n)}\sim\sqrt{X^{(n)}c^{-1}}$, ha $n\to\infty$. Továbbá $g^{(2^n)}(1-\alpha^{-2n}X^{(n)}=g^{(2^n+1)}(\alpha^{-n}X_0^{(n)})=g\big(g^{(2^n)}(X_0^{(n)}\alpha^{-n})\big)$. A renormálási egyenletből következik, hogy $g^{(2^n)}(X_0^{(n)}\alpha^{-n})=(-\alpha^{-1})^n g(X_0^{(n)})$. Ezért $g\big(g^{(2^n)}(X_0^{(n)}\alpha^{-n})\big)=$ $=g\big((-\alpha^{-1})^n g(X_0^{(n)})\big)=g\big(\alpha^{-n}g(X_0^{(n)})\big)=1-c\alpha^{-2n}\big(g(X_0^{(n)})\big)^2+O(\alpha^{-4n})$. Így ha

$$g^{(2^n)}(1-\alpha^{-2n}X^{(n)}) = 1-\alpha^{-2n}Y^{(n)},$$

akkor

$$Y^{(n)} = c\big(g(X_0^{(n)})\big)^2+O(\alpha^{-2n}) = c\big(g\big(\sqrt{X^{(n)}c^{-1}}\big)\big)^2+O(\alpha^{-2n}).$$

2. Bebizonyítjuk a tétel első állítását. Rögzítsük k-t és tekintsük az 1 környezetében levő 2^k számú $\Delta_{p\cdot 2^{n-k}+1}^{(n)}$, $0\leqq p<2^k$ szakaszt. Az $X^{(n-k)}$ változókban e szakaszok végpontjai, következésképpen hosszuk is exponenciális sebességgel konvergálnak, ha $n-k\to\infty$. Legyen $\Gamma_r^{(n)}=\Delta_1^{(n-r-1)}\setminus\Delta_1^{(n-r)}$. Világos, hogy $\Gamma_r^{(n)}$ tartalmaz 2^r-számú $\Delta_i^{(n)}$ szakaszt.

Technikai szempontból kényelmesebb az F attraktor $\Delta_1^{(1)}$-be eső részét vizsgálni. Rögzítsük az $x\in F\cap\Delta_1^{(1)}$ pontot. Található olyan $\Delta_{p\cdot 2^{n-k}+1}^{(n)}\subset\Delta_1^{(n-k)}$ szakasz, amelyre $\Delta_s^{(n)}(x)=g^{(t)}(\Delta_{p\cdot 2^{n-k}+1}^{(n)})$, $0\leqq t<2^{n-k}$. Legyen $\Delta_{s'}^{(n-1)}(x)=g^{(t)}(\Delta_{p'\cdot 2^{n-k-1}}^{(n-1)})\supset\Delta_s^{(n)}(x)$, és válasszunk ki egy tetszőleges $\bar{x}\in\Delta_{2\cdot p^{n-k}+1}^{(n)}$ pontot. A $g^{(t)}$ leképezés homeomorfizmus $\Delta_{p\cdot 2^{n-k}+1}^{(n)}$ és $\Delta_s^{(n)}(x)$ között. Ezért

$$|\Delta_s^{(n)}(x)| = \int\limits_{\Delta_{p\cdot 2^{n-k}+1}^{(n)}} \prod_{i=0}^{t-1}|g'(x_i)|\,dx_0 =$$

$$= \int\limits_{\Delta_{p\cdot 2^{n-k}+1}^{(n)}} \Big(\prod_{r=k}^{n-2}\Big(\prod_{i:\,x_i\in g^{-1}(\Gamma_r^{(n)})}|g'(x_i)|\Big)\Big)\Big(\prod_{i:\,x_i\in\Delta_1^{(1)}}|g'(x_i)|\Big)\,dx_0 =$$

$$= \prod_{i=0}^{t-1}|g'(\bar{x}_i)| \int\limits_{\Delta_{p\cdot 2^{n-k}+1}^{(n)}} \Big(\prod_{r=k}^{n-2}\Big(\prod_{i:\,x_i\in g^{-1}(\Gamma_r^{(n)})}(|g'(x_i)|/|g'(\bar{x}_i)|)\Big)\Big)\times$$

$$\times \prod_{i:\,x_i\in\Delta_1^{(1)}}\big(|g'(x_i)|/|g'(\bar{x}_i)|\big)\,dx_0.$$

Megjegyezzük, hogy

$$\left|\frac{g'(x_i)}{g'(\bar{x}_i)}\right| = \left|1+\frac{g'(x_i)-g'(\bar{x}_i)}{g'(\bar{x}_i)}\right|$$

és $|g'(x_i)-g'(\bar{x}_i)|\leqq C_1|x_i-\bar{x}_i|\leqq 2C_1\alpha^{-n}$ (a 4.1. tétel 6° tulajdonság értelmében), ahol $C_1=\max_{-1\leqq x\leqq 1}|g''(x)|$. Ha $\bar{x}_i\in g^{-1}(\Gamma_r^{(n)})$, teljesül az $|g'(\bar{x}_i)|\geqq C_2\alpha^{-(n-r)}$ egyenlőtlenség. Emellett $|g'(\bar{x}_i)|\geqq C_3$ ha $\bar{x}_i\in\Delta_1^{(1)}$. Ezért

$$1-2C_1C_2^{-1}\alpha^{-r}\leqq\left|\frac{g'(x_i)}{g'(\bar{x}_i)}\right|\leqq 1+2C_1C_2^{-1}\alpha^{-r},$$

ha $g(x_i)\in\Gamma_r^{(n)}$, és

$$1-2C_1C_3^{-1}\alpha^{-n}\leqq\left|\frac{g'(x_i)}{g'(\bar{x}_i)}\right|\leqq 1+2C_1C_3^{-1}\alpha^{-n},$$

ha $x_i\in\Delta_1^{(1)}$. Felhasználva, hogy azon i-k száma, amelyekre $g(x_i)\in\Gamma_r^{(n)}$, legfeljebb 2^r

$$(1-2C_1C_3^{-1}\alpha^{-n})^{2^n}\prod_{r=k}^{n-2}(1-2C_1C_2^{-1}\alpha^{-r})^{2^r}\leqq\frac{\Delta_s^{(n)}(x)}{\left(\prod_{i=0}^{t-1}|g'(\bar{x}_i)|\right)|\Delta_{p\cdot 2^{n-k+1}}^{(n)}|}\leqq$$

$$\leqq(1+2C_1C_3^{-1}\alpha^{-n})^{2^n}\prod_{r=k}^{n-2}(1+2C_1C_2^{-1}\alpha^{-r})^{2^r}.$$

Tekintve, hogy $2\alpha^{-1}<1$, egyszerű átalakítás után kapjuk, hogy

$$\exp\{-\text{const}\,(2\alpha^{-1})^k\}\leqq\frac{\Delta_s^{(n)}(x)}{\left(\prod_{i=0}^{t-1}|g'(\bar{x}_i)|\right)|\Delta_{p\cdot 2^{n-k+1}}^{(n)}|}\leqq\exp\{\text{const}\,(2\alpha^{-1})^k\}.$$

Analóg egyenlőtlenségek vezethetők le $\Delta_{s'}^{(n-1)}(x)$re is, mivel kiválaszthatók ugyanezek az $\bar{x}_i$ pontok. Tehát

$$(4.2)\qquad \exp\{-\text{const}\,(2\alpha^{-1})^k\}\leqq\frac{|\Delta_s^{(n)}(x)|}{|\Delta_{s'}^{(n-1)}(x)|}:\frac{|\Delta_{p\cdot 2^{n-k+1}}^{(n)}|}{|\Delta_{p'\cdot 2^{n-k-1+1}}^{(n-1)}|}\leqq\exp\{\text{const}\,(2\alpha^{-1})^k\}.$$

A hátralevő rész meglehetősen egyszerű. Az $x\in F\cap\Delta_1^{(1)}$ pont és a $\Delta_{s_n}^{(n)}(x)$ szakaszsorozat birtokában (most kimutatjuk az s-nek n-től való explicit függését) található olyan $p_n(x)=p_n$ számsorozat, hogy

$$\Delta_{s_n}^{(n)}(x)=g^{(t_n)}(\Delta_{p_n\cdot 2^{n-k}+1}^{(n)}),$$

$0\leqq t<2^{n-k}$, $0\leqq p_n<2^k$. Megjegyezzük, hogy a $p_n(x)$ számok k-függő valószínűségi változósorozatot reprezentálnak, azaz p_{n_1} és p_{n_2} független, ha $|n_1-n_2|>k$. Valóban, legyen x szimbolikus előállítása $(1,\varepsilon_1,\ldots,\varepsilon_{n-k},\ldots,\varepsilon_{n-1},\ldots)$. Ekkor

$$s_n=1+\varepsilon_1 2+\ldots+\varepsilon_{n-i}2^{n-1},\quad p_n=\varepsilon_{n-k}+\varepsilon_{n-k+1}2+\ldots+\varepsilon_{n-1}2^{k-1}.$$

A k-függőségre vonatkozó állítás abból következik, hogy az ε_i változók függetlenek a μ_0^* invariáns mérték szerint.

Írjuk át $|\Delta_{s_n}^{(n)}(x)|$-et más alakba:

$$|\Delta_s^{(n)}(x)|=|\Delta_{s_k}^{(k)}(x)|\prod_{m=k+1}^{n}\left(\frac{|\Delta_{s_m}^{(m)}(x)|}{|\Delta_{s_{m-1}}^{(m-1)}(x)|}\right).$$

Jelölje

$$\frac{1}{n}\ln|\Delta_{s_n}^{(n)}(x)| \equiv \gamma^{(n)}(x)$$

és

$$\frac{1}{n}\ln\prod_{m=k+1}^{n}\left(\frac{|\Delta_{p_m\cdot 2^{m-k}+1}^{(m)}|}{|\Delta_{p_{m-1}\cdot 2^{m-k-1}+1}^{(m-1)}|}\right) \equiv \gamma_k^{(n)}(x).$$

Felhasználva (4.2)-t

(4.3)

$$|\gamma^{(n)}(x)-\gamma_k^{(n)}(x)| \leqq \frac{1}{n}\max_{0\leqq i<2^k}\left|\ln|\Delta_i^{(k)}|\right|+\frac{1}{n}\,\mathrm{const}\,(2\alpha^{-1})^k(n-k) \leqq \mathrm{const}\,(2\alpha^{-1})^k.$$

Átalakítva $\gamma_k^{(n)}(x)$-et

$$\gamma_k^{(n)}(x)=\sum_{i=0}^{2^k-1}\frac{1}{n}\sum_{m:\,p_m=i}\ln\left(\frac{|\Delta_{p_m\cdot 2^{m-k}+1}^{(m)}|}{|\Delta_{p_{m-1}\cdot 2^{m-k-1}+1}^{(m-1)}|}\right).$$

Az l-univerzalitásból következik, hogy rögzített k és $m\to\infty$ esetén

$$\ln\left(\frac{|\Delta_{p_m\cdot 2^{m-k}+1}^{(m)}|}{|\Delta_{p_{m-1}\cdot 2^{m-k-1}+1}^{(m-1)}|}\right)$$

exponenciális sebességgel konvergál, és a határérték csak a $p_m=i$ értéktől függ. Jelölje a határértéket $U_k(i)$. Továbbá a nagy számok erős törvénye alapján minden k-ra a $p_m=i$ értéknek megfelelő rész $(0\leqq i<2^k)$ 2^{-k}-hoz tart μ_0 -majdnem minden $x\in F\cap\Delta_1^{(1)}$ pontra. Így μ_0-majdnem minden $x\in F\cap\Delta_1^{(1)}$ pontra és minden k-ra

$$\gamma_k^{(n)}(x)\xrightarrow[n\to\infty]{}\frac{1}{2^k}\sum_{i=0}^{2^k-1}U_k(i)\equiv\gamma_k. \tag{4.4}$$

(4.3) és (4.4) miatt μ_0 -majdnem minden $x\in F\cap\Delta_1^{(1)}$ pontra léteznek a $\lim_{n\to\infty}\gamma^{(n)}(x)$, $\lim_{k\to\infty}\gamma_k$ határértékek és e két határérték megegyezik. Valóban,

$$|\gamma^{(n)}(x)-\gamma^{(m)}(x)| \leqq |\gamma^{(n)}(x)-\gamma_k^{(n)}(x)|+|\gamma_k^{(n)}(x)-\gamma_k^{(m)}(x)|+|\gamma_k^{(m)}(x)-\gamma^{(m)}(x)|.$$

Az első és a harmadik tagot k megfelelő választásával tetszőlegesen kicsivé tehetjük. Rögzített k-ra a második tag 0-hoz tart, ha $n,\ m\to\infty$.

Hasonlóan bizonyítható, hogy létezik a $\lim_{k\to\infty}\gamma_k$ határérték, és hogy a két határérték megegyezik. Megjegyezzük, hogy

$$|\gamma_k-\gamma_l| \leqq \mathrm{const}\,\left((2\alpha^{-1})^k+(2\alpha^{-1})^l\right),$$

így konstruktív eljárást kaptunk a γ állandó kiszámítására. A renormálási egyenletből triviálisan következik, hogy mivel a határérték létezik $F\cap\Delta_1^{(1)}$ μ_0-tipikus pontjaira, ezért létezik a teljes F attraktor μ_0-tipikus pontjaira is. Numerikus számítások alapján γ körülbelüli értéke $-1{,}34$.

3. Megfogalmazzuk az előző pontban kapott eredményeket a φ leképezés által adott szimbolikus előállításra vonatkozóan. Tekintsük a $\Delta_{p\cdot 2^{n-k}+1}^{(n)}\subset\Delta_{p'2^{n-k-1}+1}^{(n-1)}$,

$0 \leqq p < 2^k$ szakaszokat. Ezek φ-képe

$$(\underbrace{1, 0, \ldots, 0}_{n-k}, \varepsilon^{(k)}, \varepsilon^{(k-1)}, \ldots, \varepsilon^{(1)}) \quad \text{és} \quad (\underbrace{1, 0, \ldots, 0}_{n-k}, \varepsilon^{(k)}, \varepsilon^{(k-1)}, \ldots, \varepsilon^{(2)}).$$

Jelölje

$$\ln\left(\frac{|\Delta^{(n)}_{p\cdot 2^{n-k}+1}|}{|\Delta^{(n-1)}_{p'2^{n-k-1}+1}|}\right) \equiv U_k^{(n)}(\varepsilon^{(1)}, \varepsilon^{(2)}, \ldots, \varepsilon^{(k)}).$$

Rögzített $\varepsilon^{(1)}, \ldots, \varepsilon^{(k)}$-ra az l-univerzalitás értelmében az $U_k^{(n)}(\varepsilon^{(1)}, \varepsilon^{(2)}, \ldots, \varepsilon^{(k)})$ függvénynek van határértéke, ha $n \to \infty$. Jelölje $U_k(\varepsilon^{(1)}, \ldots, \varepsilon^{(k)})$ a határfüggvényt. Az l-univerzalitás bizonyításában szereplő meggondolások lehetővé teszik az alábbi becslés igazolását:

$$|U_k^{(n)}(\varepsilon^{(1)}, \ldots, \varepsilon^{(k)}) - U_k(\varepsilon^{(1)}, \ldots, \varepsilon^{(k)})| \leqq \text{const}\, \alpha^{-2(n-k)}\alpha^{2k}. \tag{4.5}$$

Az $U_k(\varepsilon^{(1)}, \ldots, \varepsilon^{(k)})$ függvények konvergálnak, ha $k \to \infty$, a határértéket jelölje $U(\varepsilon^{(1)}, \varepsilon^{(2)}, \ldots, \varepsilon^{(n)}, \ldots)$. Létezése könnyen következik az előző pont eredményeiből, ahol lényegében a következő becslést kaptuk:

$$|U(\varepsilon^{(1)}, \ldots, \varepsilon^{(k)}, \ldots) - U_k(\varepsilon^{(1)}, \ldots, \varepsilon^{(k)})| \leqq \text{const}\, (2\alpha^{-1})^k. \tag{4.6}$$

A 2. pontban végrehajtott alapvető számítás azt mutatja, hogy ha $\Delta_s^{(n)} \subset \Delta_{s'}^{(n-1)} \subset \Delta_1^{(1)}$ és $\varphi(\Delta_s^{(n)}) = (1, \varepsilon_2, \ldots, \varepsilon_n)$, $\varphi(\Delta_{s'}^{(n+1)}) = (1, \varepsilon_2, \ldots, \varepsilon_{n-1})$, akkor

$$\exp\{-\text{const}\,(2\alpha^{-1})\} \leqq \frac{|\Delta_s^{(n)}|}{|\Delta_{s'}^{(n-1)}|} \exp\{-U_k^{(n)}(\varepsilon_n, \ldots, \varepsilon_{n-k})\} \leqq \exp\{\text{const}\,(2\alpha^{-1})^k\}.$$

Figyelembe véve (4.5)-öt és (4.6)-ot

$$\exp\{-\text{const}\,((2\alpha^{-1})^k + \alpha^{4k-2n})\} \leqq \frac{|\Delta_s^{(n)}|}{|\Delta_{s'}^{(n-1)}|} \exp\{-U(\varepsilon_n, \ldots, \varepsilon_2, 1, 0, 0, \ldots)\} \leqq$$

$$\leqq \exp\{\text{const}\,((2\alpha^{-1})^k + \alpha^{4k-2n})\},$$

ahol $1 \leqq k \leqq n$. Legyen $k = [n/3]$. n-nel átszorozva

$$\text{const} \leqq \frac{|\Delta_s^{(n)}|}{\exp\left\{\sum_{s=1}^{n} U(\varepsilon_s, \varepsilon_{s-1}, \ldots, \varepsilon_2, 1, 0, \ldots, 0, \ldots)\right\}} \leqq \text{const}. \tag{4.7}$$

A (4.7) összefüggés alapján természetes az U függvényt — a statisztikus mechanika terminológiáját használva — a *Feigenbaum-attraktor potenciáljának* nevezni. Megjegyezzük, hogy az U potenciál szorosan kapcsolódik ahhoz a $\sigma(t)$ függvényhez, amelyet Feigenbaum tanulmányozott ([5], [6]). A továbbiakban megmutatjuk, hogyan határozható meg U segítségével a β_0, λ és σ konstans. Megjegyezzük, hogy a γ állandót nyilvánvalóan definiálja az alábbi formula:

$$\gamma = \int U(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n, \ldots)\, d\mu^*(\varepsilon).$$

4. Most bebizonyítjuk a 4.2. tétel második állítását. A bizonyítás a statisztikus mechanikából jól ismert állításokon alapul. Az $U(\varepsilon^{(1)}, \ldots, \varepsilon^{(k)}, \ldots)$ függvény $\varepsilon^{(1)}$ és a többi koordináta kölcsönhatási potenciáljának tekinthető, és így alkalmazhatók a

Gibbs-határeloszlások elméletének ismert tételei (l. pl. [39], [40]). Tetszőleges β-ra tekintsük a $\Delta_k^{(n)}$ szakaszokon az l_β mértéket, ahol $l_\beta(\Delta_k^{(n)})=|\Delta_k^{(n)}|^\beta$.

(4.7) alapján

$$\text{const} \leqq \frac{l_\beta(\Delta_k^{(n)})}{\exp\{\beta \sum_{s=1}^{n} U(\varepsilon_s, \ldots, \varepsilon_2, 1, 0, \ldots, 0, \ldots)\}} \leqq \text{const.}$$

Ez az összefüggés igazolja β jelölését. A statisztikus mechanika ismert tételei szerint (l. [41], [42]) levezethető, hogy létezik az

$$f(\beta) = \lim_{n\to\infty} \frac{1}{n} \ln\Big[\sum_{\varepsilon_2,\ldots\varepsilon_n} \exp\{\beta \sum_{s=1}^{n} U(\varepsilon_s, \ldots, \varepsilon_2, 1, 0, \ldots, 0, \ldots)\}\Big]$$

határérték, amelyet szabad energiának nevezünk. Mivel $U<0$, $f(\beta)$ sima, β-nak monoton fogyó függvénye. Emellett $f(0)=\ln 2$, $\lim_{\beta\to\infty} f(\beta)=-\infty$. Tehát egyértelműen létezik olyan $\beta_0>0$, amelyre $f(\beta_0)=0$. Megmutatható, hogy erre a β_0-ra

$$\sum_{\varepsilon_2,\ldots,\varepsilon_n} \exp\{\beta_0 \sum_{s=1}^{n} U(\varepsilon_s, \ldots, \varepsilon_2, 1, 0, \ldots, 0, \ldots)\}$$

n-ben egyenletesen két pozitív korlát között marad, ezért

$$\text{const} \leqq \sum_{k=0}^{2^{n-1}-1} l_{\beta_0}(\Delta_{2k+1}^{(n)}) \leqq \text{const.}$$

Az összes $\Delta_k^{(n)}$, $0\leqq k<2^n$ szakaszra vett összeg is korlátos, mivel

$$\text{const} \leqq \frac{\sum_{k=0}^{2^{n-1}-1} l_\beta(\Delta_{2k}^{(n)})}{\sum_{k=0}^{2^{n-1}-1} l_\beta(\Delta_{2k+1}^{(n)})} \leqq \text{const.}$$

Ebből azonnal következik, hogy az F attraktor *Hausdorff-dimenziója* legfeljebb β_0. Másik oldali becslést akkor kapunk, ha F-nek azt a részhalmazát tekintjük, amely a $\beta_0 U$ potenciálú *Gibbs-határeloszlás* szerinti tipikus pontokból áll. Nem nehéz megmutatni, hogy az attraktor βU potenciálú Gibbs-határeloszlás szerinti tipikus részhalmazának Hausdorff-dimenziója $\beta-f(\beta)/f'(\beta)$. Következésképpen a *Hausdorff-dimenzió* $\beta=\beta_0$-nál éri el maximumát és értéke β_0. Következésképpen a $\beta_0 U$ potenciálhoz tartozó *Gibbs-határeloszlás* jelöli ki az F attraktor „legmasszívabb" részét a *Hausdorff-dimenzió* alapján. Numerikus számítás a $\beta_0\approx 0{,}54$ értéket adja. Ezzel kapcsolatban megemlítjük a [43] munkát, amelyben a *Feigenbaum-attraktor Hausdorff-dimenziójának* jóval pontosabb numerikus becslése található.

5. Most bebizonyítjuk a tétel utolsó két állítását. Alakítsuk át a λ állandót definiáló összeget:

$$S = 1+\sum_{k=1}^{2^n-1}\prod_{i=k}^{2^n-1}(g'(x_i))^2 = \prod_{i=1}^{2^n-1}(g'(x_i))^2\Big(1+\sum_{k=1}^{2^n-1}\prod_{i=1}^{k}(g'(x_i))^{-2}\Big) =$$

$$= |\Delta_1^{(n)}|^2 \prod_{i=1}^{2^n-1}(g'(x_i))^2\Big[|\Delta_1^{(n)}|^{-2}+\sum_{k=1}^{2^n-1}|\Delta_k^{(n)}|^{-2}\prod_{i=1}^{k}(g'(x_i))^{-2}\Big].$$

A 2. pontban tulajdonképpen beláttuk, hogy

$$\frac{\text{const}}{|\Delta_{k+1}^{(n)}|^2} \leqq \frac{1}{|\Delta_1^{(n)}|^2} \prod_{i=1}^{k} (g'(x_i))^{-2} \leqq \frac{\text{const}}{|\Delta_{k+1}^{(n)}|^2}.$$

Ezért teljesül a const $\tilde{S}_n \leqq S_n \leqq$ Const $\tilde{S}_n$ becslés, ahol

$$\tilde{S}_n = |\Delta_{2^n}^{(n)}|^2 \sum_{k=1}^{2^n-1} |\Delta_k^{(n)}|^{-2}, \quad \Delta_{2^n}^{(n)} = g(\Delta_{2^n-1}^{(n)}).$$

Ugyanúgy, mint az előző pontban, azt kaptuk, hogy

$$\lim_{n\to\infty} \frac{1}{n} \ln \Big[\sum_{k=1}^{2^n-1} |\Delta_k^{(n)}|^{-2} \Big] = f(-2),$$

ahol $f(-2)$ az U potenciálnak megfelelő szabad energia $\beta = -2$-nél. Mivel $|\Delta_{2^n}^{(n)}| = O(\alpha^{-n})$,

$$\lim_{n\to\infty} \frac{1}{n} \ln S_n = \lim_{n\to\infty} \frac{1}{n} \ln \tilde{S}_n = f(-2) - 2 \ln \alpha, \quad \text{azaz} \quad \lambda = e^{1/2 f(-2)}.$$

Már megemlítettük, hogy a λ állandó, amelynek értéke körülbelül 6,6, a g leképezés kis sztochasztikus perturbációinak feladatához kapcsolódik. Ezzel a kérdéssel részletesebben a következő fejezetben foglalkozunk.

Vegyük szemügyre végül σ-t. A határérték létezése a (4.7) összefüggésből következik. Ugyanebből triviálisan adódik, hogy $\sigma = 1/2\, e^{f(1)}$. Numerikusan $\sigma \approx 0{,}29$.

A 4.2. tételt bebizonyítottuk.

Azokra a leképezésekre, amelyek a stabilis $\Gamma^{(s)}(g)$ szeparatrixon helyezkednek el (g a renormálási egyenletet kielégítő fixpont), szintén meghatározható a $\Delta_k^{(n)}$ szakaszok rendszere és az U potenciál. Nem nehéz megérteni, hogy ez a potenciál megegyezik az imént meghatározottal. Tehát mind az U potenciál, mind a segítségével kifejezett γ, β_0, λ, σ konstansok univerzálisak.

5. A g leképezés kis sztochasztikus perturbációi

5.1. A g leképezés kis sztochasztikus perturbációinak problémája nagy érdeklődésre tarthat számot. Ennek az az oka, hogy reális rendszerekben a periódusketttőződéses bifurkációsorozat megfigyelése elkerülhetetlenül együtt jár bizonyos mértékű zaj jelenlétével. Természetes tehát azt vizsgálni, hogyan befolyásolják ezek a zajok a g leképezés tulajdonságait. Az említett kérdéseket J. P. CRUTCHFIELD, M. NAUENBERG, J. RUDNICK [44], B. SHRAIMAN, C. E. WAYNE, P. C. MARTIN [45] és mások ([46], [47]) tanulmányozták. Az alábbiakban az ide tartozó eredmények kvalitatív leírását adjuk.

A g leképezéssel együtt tekintsünk most egy Markov-láncot is, amelyet a következő összefüggés definiál:

$$x_{n+1} = g(x_n) + \xi_{n+1},$$

ahol a ξ_n-ek független, azonos eloszlású valószínűségi változók. Feltesszük, hogy ξ_n eloszlása a $[-\varepsilon, \varepsilon]$ intervallumra koncentrálódik, ahol $\varepsilon > 0$ paraméter és az elosz-

lás a $p_\varepsilon(x)$ sűrűséggel adott, amelyre

$$\int_{-\varepsilon}^{\varepsilon} x p_\varepsilon(x)\,dx = 0, \qquad \int_{-\varepsilon}^{\varepsilon} x^2 p_\varepsilon(x)\,dx \sim C\varepsilon^2$$

ha $\varepsilon \to 0$ és C állandó. Például feltehető, hogy $p_\varepsilon(x) = 1/2\varepsilon$, ami a $[-\varepsilon, \varepsilon]$ szakaszon az egyenletes eloszlásnak felel meg. A bevezetett *Markov-láncnak* van $q(x;\varepsilon)$ sűrűségű stacionárius eloszlása, amely $\varepsilon \to 0$ esetén gyengén konvergál a μ_0 mértékhez (l. [48]). Feladatunk $q(x;\varepsilon)$ viselkedésének tanulmányozása kicsiny, de nem 0 ε-ok esetén.

Ha nem volna sztochasztikus járulék, akkor majdnem minden $x_0 \in [-1, 1]$ pont $x_n = g^{(n)}(x_0)$ képe konvergálna az F attraktorhoz, azaz $d(x_n, F) \to 0$ ha $n \to \infty$. A véletlen járulék elkeni a pontok képét a $\Delta_k^{(n)}$ szakaszokon. Most ezt a folyamatot vizsgáljuk meg részletesebben.

Legyen $x_0 \in \Delta_0^{(n)}$. Ekkor

$$x_{2^n} = \xi_{2^n} + g\big(\xi_{2^n-1} + g(\xi_{2^n-2} + \ldots + g(\xi_1 + g(x_0)\ldots)\big).$$

Mivel minden ξ_i kicsi, felírhatjuk az x_{2^n} pont ξ_i-k szerinti lineáris kifejtését, azaz

$$x_{2^n} = \bar{x}_{2^n} + \xi_{2^n} + \sum_{i=1}^{2^n-1} \xi_i \prod_{s=i}^{2^n-1} g'(\bar{x}_s) + Q_{2^n}(\xi;\, x_0), \tag{5.1}$$

ahol $\bar{x}_s = g^{(s)}(x_0)$, $Q_{2^n}(\xi, x_0)$ pedig a hibatag. A

$$\xi_{2^n} + \sum_{i=1}^{2^n-1} \xi_i \prod_{s=i}^{2^n-1} g'(\bar{x}_s) \equiv \zeta_{2^n}(x_0)$$

összeg valószínűségi változó, várható értéke nulla, szórása

$$\mathcal{D}\zeta_{2^n}(x_0) = \mathcal{D}\xi\Big(1 + \sum_{i=1}^{2^n-1} \prod_{s=i}^{2^n-1} (g'(\bar{x}_s))^2\Big). \tag{5.2}$$

Ezt az összeget vizsgáltuk a 4. paragrafusban, ahol megmutattuk, hogy

$$\text{const}\,\alpha^{-2n}\lambda^{2n} \leqq 1 + \sum_{i=1}^{2^n-1} \prod_{s=i}^{2^n-1} (g'(\bar{x}_s))^2 \leqq \text{const}\,\alpha^{-2n}\lambda^{2n}, \quad \lambda \approx 6{,}6.$$

Az (5.2) összeg egy más megközelítése található [44]-ben. Legyen $x_0 = (-\alpha^{-1})^n z_0$, ahol $z_0 \in [-1, 1]$. Jelölje

$$D_n(z_0) = 1 + \sum_{i=1}^{2^n-1} \prod_{s=i}^{2^n-1} (g'(\bar{x}_s))^2.$$

Ekkor a (4.1) renormálási egyenlet felhasználásával könnyen jutunk az alábbi rekurzív összefüggéshez:

$$D_{n+1}(z_0) = (g'(g(\alpha^{-1}z_0)))^2 D_n(\alpha^{-1}z_0) + D_n\big(g(\alpha^{-1}z_0)\big). \tag{5.3}$$

Jelölje $\mathcal{L}$ az (5.3)-nak megfelelő lineáris operátort, azaz

$$\mathcal{L}f(z) = (g'(g(\alpha^{-1}z_0)))^2 f(\alpha^{-1}z) + f\big(g(\alpha^{-1}z)\big).$$

Ekkor $D_n(z_0)=\mathscr{L}^n 1$, ahol 1 jelöli az azonosan 1 függvényt. Mivel $\mathscr{L}$ pozitív operátor, létezik nem elfajult, pozitív maximális $\lambda(\mathscr{L})$ sajátértéke és az ennek megfelelő pozitív $l(z)$ sajátfüggvény. Ezért $n\to\infty$ esetén

$$D_n(z_0) \sim \lambda(\mathscr{L})^n l(z_0),$$

ahol $\lambda(\mathscr{L})=\alpha^{-2}\lambda^2$. Természetesen mindkét eljárás ugyanarra az eredményre vezet, de az eredmény a termodinamikai formalizmus 4. fejezetben kifejtett eszközeinek segítségével közvetlenül adódik.

Legyen most $n_1=n_1(\varepsilon)$ olyan, amelyre $\varepsilon\lambda^{n_1(\varepsilon)}\sim\text{const}$ ha $\varepsilon\to 0$. Ennek az az értelme, hogy az $x_0\in\Delta_0^{(n_1)}$ kezdeti pontra $\bar{x}_{2^{n_1}}\in\Delta_0^{(n_1)}$, és a véletlen járulék lineáris része elkeni a pont képét egy olyan szakaszon, amelynek hossza $|\Delta_0^{(n_1)}|$ hosszával megegyező nagyságrendű. Más szavakkal: 2^{n_1} lépés alatt játszódik le az „emlékezet elvesztése" a $\Delta_0^{(n_1)}$ szakasz hosszával megegyező nagyságrendben. Megjegyezzük, hogy $n_1(\varepsilon)$ invariáns a kezdeti pont elhelyezkedésére nézve. Valóban, legyen $x_0\in\Delta_k^{(n_1)}$, $0\leq k<2^{n_1}$. Ekkor, mint korábban $x_{2^{n_1}}=\bar{x}_{2^{n_1}}+\zeta_{2^{n_1}}(x_0)+Q_{2^{n_1}}(\xi;x_0)$. A 4.2. tétel III. állításának bizonyításakor alkalmazott meggondolások alapján könnyű megmutatni, hogy

$$\mathscr{D}\left(\zeta_{2^{n_1}}(x_0)\right)=\sigma(|\Delta_k^{(n_1)}|^2\varepsilon^2\lambda^{2n_1})=\sigma(|\Delta_k^{(n_1)}|^2),$$

azaz a $\zeta_{2^{n_1}}(x_0)$ valószínűségi változó szórásának nagyságrendje a $\Delta_k^{(n_1)}$ szakasz hosszának négyzete. 5.2. Most megmutatjuk, hogy a *Markov-lánc staticonárius eloszlása* bizonyos értelemben $\bigcup_{k=0}^{2^{n_1}-1}\Delta_k^{(n_1)}$-ra koncentrált. Ennek érdekében először megbecsüljük a $Q_{2^n}(\xi;x_0)$ maradéktagot.

Legyen az egyszerűség kedvéért $x_0\in\Delta_0^{(n)}$, $x_i=\bar{x}_i+L_i+Q_i$, ahol $\bar{x}_i=g^{(i)}(x_0)$, L_i a véletlen perturbáció lineáris része és Q_i a hibatag. Feltesszük, hogy $|L_i|\leq\tau|\Delta_i^{(n)}|$ minden $0\leq i<2^n$-re. Megmutatjuk, hogy ha τ elég kicsi, akkor $|Q_i|\leq C_1\tau^2|\Delta_i^{(n)}|$, ahol C_1 a kezdeti ponttól és a ξ valószínűségi változó realizációjától független konstans. Írjuk fel az L_i és Q_i mennyiségekre vonatkozó rekurzív összefüggéseket:

$$\begin{aligned} L_{i+1} &= L_i g'(\bar{x}_i)+\xi_{i+1}, \\ Q_{i+1} &= L_i\left(g'(\tilde{x}_i)-g'(\bar{x}_i)\right)+Q_i g'(\tilde{x}_i), \end{aligned} \tag{5.4}$$

ahol $|\tilde{x}_i-\bar{x}_i|\leq|L_i+Q_i|$. (5.4)-ből következik, hogy

$$Q_j=\sum_{i=1}^{j-1}\left[L_i\frac{\left(g'(\bar{x}_i)-g'(\bar{x}_i)\right)}{g'(\tilde{x}_i)}\prod_{s=i}^{j-1}g'(\tilde{x}_s)\right].$$

Legyenek az $x_i'\in\Delta_i^{(n)}$, $0\leq i<2^n$ pontok olyanok, amelyekre $|g'(x_i')|\cdot|\Delta_i^{(n)}|=|\Delta_{i+1}^{(n)}|$, tegyük fel továbbá, hogy $|Q_i|\leq\chi|\Delta_i^{(n)}|$, $0\leq i<j$. Ekkor $|g'(\tilde{x}_i)-g'(\bar{x}_i)|\leq$ $\leq M(\tau+\chi)|\Delta_i^{(n)}|$, ahol $M=\max_{x\in[-1,1]}g''(x)$, tehát

$$|Q_j|\leq M\tau(\tau+x)\sum_{i=1}^{j-1}\left[\frac{|\Delta_i^{(n)}|^2}{|g'(\tilde{x}_i)|}\left(\prod_{s=i}^{j-1}g'(x_s')\right)\left(\prod_{s=i}^{j-1}\frac{|g'(\tilde{x}_s)|}{|g'(x_s')|}\right)\right]=$$

$$=M\tau(\tau+\chi)|\Delta_j^{(n)}|\sum_{i=1}^{j-1}\left[\frac{|\Delta_i^{(n)}|}{|g'(\tilde{x}_i)|}\prod_{s=i}^{j-1}\frac{|g'(\tilde{x}_s)|}{|g'(x_s')|}\right].$$

Most becslést adunk a $\prod_{s=i}^{j-1} \frac{|g'(\tilde{x}_s)|}{|g'(x'_s)|}$ szorzatra:

$$\prod_{s=1}^{j-1} \frac{|g'(\tilde{x}_s)|}{|g'(x'_s)|} = \prod_{s=i}^{j-1} \left|1+\frac{g'(\tilde{x}_s)-g'(x'_s)}{g'(x'_s)}\right| \leqq \prod_{s=i}^{j-1} \left|1+\frac{M|\Delta_s^{(n)}|(1+\tau+\chi)}{g'(x'_s)}\right| \leqq$$

$$\leqq \exp\left(\sum_{s=i}^{j-1} M(1+\tau+\chi)\frac{|\Delta_s^{(n)}|}{|g'(x'_s)|}\right) \leqq \exp\left(MR_1(1+\tau+\chi)\right),$$

ahol $R_1=\max_n \left(\sum_{s=0}^{2^n-1} \frac{|\Delta_s^{(n)}|}{|g'(x'_s)|}\right)$. Mivel $\sum_{i=1}^{j-1} \left(\frac{|\Delta_i^{(n)}|}{|g'(\tilde{x}_i)|}\right)$, $2 \leqq j \leqq 2^n$ nem nagyobb egy abszolút R_2 konstansnál, ezért

$$|Q_j| \leqq MR_2\tau(\tau+\chi)\exp\left(MR_1(1+\tau+\chi)\right)|\Delta_j^{(n)}|.$$

Legyen $\chi=C_1\tau^2$, ahol $C_1=2MR_2\exp(MR_1)$. Ekkor elegendően kis τ-ra

$$MR_2\tau(\tau+\chi)\exp\left(MR_1(1+\tau+\chi)\right) \leqq \chi$$

következésképpen

$$|Q_j| \leqq \chi|\Delta_j^{(n)}| = C_1\tau^2|\Delta_j^{(n)}|, \quad 0 \leqq j \leqq 2^n. \tag{5.5}$$

Tehát Q_j kis hibatagnak tekinthető mindaddig, amíg L_i, $1 \leqq i \leqq j$ kicsi $|\Delta_i^{(n)}|$-hez viszonyítva, $1 \leqq i \leqq j$. Megjegyezzük, hogy analóg állítás igaz abban az esetben, ha $x_0 \in \Delta_k^{(n)}$, $0 \leqq k < 2^n$. Legyen most $n=n_1-m$. Jelölje

$$\tilde{\Delta}_i^{(n_1-m)} = \{x; d(x, \Delta_i^{(n_1-m)}) \leqq 4\alpha^{-2m}|\Delta_i^{(n_1-m)}|\}, \quad 0 \leqq i \leqq 2^{n_1-m},$$

$$F(m) = \bigcup_{i=0}^{2^{n_1-m-1}} \tilde{\Delta}_i^{(n_1-m)}.$$

5.1. TÉTEL Legyen μ_ε az

$$x_{i+1} = g(x_i)+\xi_{i+1}$$

Markov-lánc stacionárius eloszlása és legyen $n_1=[-\ln\varepsilon/\ln\lambda]$. Ekkor elég nagy m-re

$$\mu_\varepsilon(F(m)) \geqq 1-\exp\left(-\text{const}\,(\lambda\alpha^{-2})^m\right).$$

Megjegyzés. A *Feigenbaum-attraktorra* $\lambda\alpha^{-2}>1$.

Az 5.1. tétel bizonyítása. Először becslést adunk annak valószínűségére, hogy $\max_{0 \leqq i \leqq 2^{n_1-m}} \frac{|L_i|}{|\Delta_i^{(n_1-m)}|} > \alpha^{-2m}$. Egy *Kolmogorov-típusú egyenlőtlenség* alapján

$$P\left(\max_{0 \leqq i \leqq 2^{n_1-m}} \frac{|L_i|}{|\Delta_i^{(n_1-m)}|} > \alpha^{-2m}\right) \leqq 2P\left(\frac{|L_{2^{n_1-m}}|}{|\Delta_{2^{(n_1-m)}}|} > \text{const}\,\alpha^{-2m}\right). \tag{5.6}$$

Mivel a $\mathcal{D}\left(\frac{L_{2^{n_1-m}}}{|\Delta_{2^{(n_1-m)}}^{(n_1-m)}|}\right)$ szórás nagyságrendje λ^{2m} és $\alpha^{-2}>\lambda^{-1}$, alkalmazhatjuk *Bern-*

stein nagy eltérésekre vonatkozó exponenciális becslését (l. [49]). Könnyen meggyőződhetünk arról, hogy elegendően nagy m-ekre

$$P\left(\frac{|L_{2^{n_1-m}}|}{|\Delta_{2^{(n_1-m)}}^{(n_1-m)}|} > \text{const}\,\alpha^{-2m}\right) < \exp\left(-\text{const}\,(\lambda\alpha^{-2})^m\right). \tag{5.7}$$

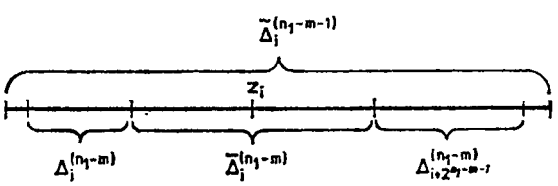
Tekintsük a *Markov-lánc* N-lépéses p_N átmenetvalószínűségeit $N=B^m 2^{n_1-m}$ esetén, itt B nagy állandó. Megmutatjuk, hogy elég nagy m-ekre

$$\min_{x_0\in F(m+1)} p_N(x_0, F(m)) \geqq 1-\exp\left(-\text{const}\,(\lambda\alpha^{-2})^m\right). \tag{5.8}$$

Legyen $x_0\in\tilde{\Delta}_i^{(n_1-m-1)}\subset F(m+1)$. Jelölje $y_s=x_{s\cdot 2^{n_1-m}}$, $0\leqq s\leqq B^m$. Az (5.5)–(5.7) becslésekből következik, hogy $\left(1-B^m\exp\left(-\text{const}\,(\lambda\alpha^{+2})^m\right)\right)$-nél nagyobb valószínűséggel teljesül az alábbi egyenlőtlenség:

$$|y_s-g^{(2^{n_1-m})}(y_{s-1})| \leqq \alpha^{-2m}\min\left(|\Delta_i^{(n_1-m)}|, |\Delta_{i+2^{n_1-m-1}}^{(n_1-m)}|\right), \quad 1\leqq s\leqq B. \tag{5.9}$$

Tegyük fel, hogy $y_{s'}\in\Delta_i^{(n_1-m)}(\Delta_{i+2^{n_1-m-1}}^{(n_1-m)})$ valamely s'-re. Ekkor az (5.9) egyenlőtlenség miatt minden $s'\leqq s\leqq B^m$-re $y_s\in\tilde{\Delta}_i^{(n_1-m)}(\tilde{\Delta}_{i+2^{n_1-m-1}}^{(n_1-m)})$. Azt kell még megmutatnunk, hogy $\left(1-\exp\left(-\text{const}\,(\lambda\alpha^{-2})^m\right)\right)$-nél nagyobb valószínűséggel van olyan $s'(y_0)\leqq B^m$, amelyre $y_{s'}\in\Delta_i^{(n_1-m)}\cup\Delta_{i+2^{n_1-m-1}}^{(n_1-m)}$. Elegendő azt az esetet vizsgálnunk, amikor $y_0\in\Delta_i^{(n_1-m-1)}-\Delta_i^{(n_1-m)}-\Delta_{i+2^{n_1-m-n}}^{(n_1-m)}\equiv\bar{\Delta}_i^{(n_1-m)}$ (l. az 5.1. ábrát). A g leképezés-



5.1. ábra

nek van a $\bar{\Delta}_i^{(n_1-m)}$ szakasz belsejében instabilis z_i fixpontja, így azok az y_0 pontok, amelyek z_i-nek egy m szerint exponenciálisan kis környezetén kívül esnek, $\text{const}\cdot m$ idő alatt kijutnak $\bar{\Delta}_i^{(n_1-m)}$-ből, azaz $y_{\text{const}\cdot m}\in\Delta_i^{(n_1-m)}\cup\Delta_{i+2^{n_1-m-1}}^{(n_1-m)}$. Másrészt y_s, $1\leqq s\leqq B^m$ a sztochasztikus perturbációk előjelének fluktuációja miatt kikerül a z_i pont tetszőleges, exponenciálisan kicsi környezetéből. Pontosabban elég nagy B esetén a z_i pont exponenciálisan kis környezetéből B^m idő alatt történő kijutás valószínűsége nagyobb, mint $\left(1-\exp\left(-\text{const}\,(\lambda\alpha^{-2})^m\right)\right)$, tehát az (5.8) becslést bebizonyítottuk. Ezt felhasználva

$$\mu_\varepsilon(F(m)) = \int p_N(x, F(m))\,d\mu_\varepsilon(x) \geqq \left(1-\exp\left(-\text{const}\,(\lambda\alpha^{-2})^m\right)\right)\mu_\varepsilon(F(m+1)).$$

Mivel $\mu_\varepsilon(F(n_1))=1$, végül azt kapjuk, hogy

$$\mu_\varepsilon(F(m)) \geqq \prod_{i=m}^{\infty}\left(1-\exp\left(-\text{const}\,(\lambda\alpha^{-2})^i\right)\right) \geqq 1-\exp\left(-\text{const}\,(\lambda\alpha^{-2})^m\right).$$

6. A Feigenbaum-univerzalitás több dimenzióban és néhány egyéb általánosítás

6.1. Eddig egydimenziós leképezéscsaládokkal foglalkoztunk. Numerikus vizsgálatok ugyanakkor arról tanúskodnak, hogy a $\delta=4{,}6692$ konstans olyan többdimenziós dinamikai rendszereknél is jellemzi a perióduskettőződéses bifurkációsorozat aszimptotikus viselkedését, amelyek nem kapcsolódnak az egydimenziós leképezésekhez. A $\mu_\infty-\mu_n\sim\text{const}\cdot\delta^{-n}$ univerzális törvénynek eleget tevő perióduskettőződéses bifurkációsorozatot találtak a *Lorenz-modell* [50] (1.1. fejezet), a *Navier—Stokes egyenletek* [51], az *Hénon-leképezés* [52] és más rendszerek [53] numerikus vizsgálata során. Ennek a ténynek a magyarázatát adta meg P. COLLET, J.-P. ECKMANN és H. KOCH [9]-ben. Elnagyoltan a következőről van szó: tekintsünk egy olyan többdimenziós leképezéscsaládot, amely egy irányban úgy hat, mint az általunk vizsgált egydimenziós leképezéscsaládok, a többi irányban pedig erősen összehúzó. Ekkor a paraméter változásával fellép a perióduskettőződéses bifurkációsorozat és ugyanannak az aszimptotikus törvénynek tesz eleget, mint az egydimenziós eset.

Rátérünk a pontosabb megfogalmazásra. Tekintsük a $\mathbf{C}^1\oplus\mathbf{C}^{n-1}$ direkt összegként előállított $\mathbf{C}^n$ teret. A $z\in\mathbf{C}^n$ vektorokat (z_0, z) alakba írjuk, ahol $z_0\in\mathbf{C}^1$ és $z\in\mathbf{C}^{n-1}$. Legyen D $\mathbf{C}^n$ nyilt részhalmaza. Jelölje $K(D)$ a $D\to\mathbf{C}^n$ korlátos, analitikus leképezések *Banach-terét*, a norma

$$\|h\| = \sup\{\|h(z)\|, \quad z\in D\}.$$

Jelölje $D(\Delta)$, $\Delta>0$ a következő halmazt:
$D(\Delta)=\{z\in\mathbf{C}^n: \|z-(y_0, 0)\|<\Delta$ valamely $y_0\in[-1, 1]$-re$\}$. Legyen a Feigenbaum-leképezés $g(x)=f(x^2)$ és legyen

$$\varphi(z) = (f(\zeta(z)), 0),$$

ahol $\zeta(z)=z_0^2-\gamma\cdot z$, γ pedig olyan $\mathbf{C}^{n-1}$-beli vektor, amelynek normája 2-nél kisebb. A $z\mapsto\varphi(z)$ leképezés analitikus a $D(\Delta)$ tartományon, ha Δ elég kicsi, azaz φ a $K_\Delta\equiv K(D(\Delta))$ térben van.

Jelölje A az

$$Az = (-\alpha^{-1}z_0, \alpha^{-2}z)$$

lineáris transzformációt, ahol $\alpha=2{,}50290\ldots$ és definiáljuk a renormálási transzformációt

$$T: G\mapsto A^{-1}\circ G\circ G\circ A,$$

ahol $G: \mathbf{C}^n\to\mathbf{C}^n$. Könnyű igazolni, hogy elég kis Δ-ra a T transzformáció a φ középpontú $b\cdot\Delta$ sugarú gömböt K_Δ belsejébe képezi le. Itt $b>0$ abszolút konstans. φ definíciójából azonnal következik, hogy T-nek fixpontja. Az univerzalitási elmélet alkalmazása érdekében a T transzformáció φ-beli differenciáljának spektrumát kell megvizsgálnunk. Nem nehéz belátni, hogy elég kis $\Delta>0$-ra $DT(\varphi)$ a K_Δ tér kompakt operátora. A [9] munka alapvető eredménye, hogy a $DT(\varphi)$ operátor spektrumának egységkörön kívül eső része explicite leírható. Az egydimenziós esethez hasonlóan létezik egy sajátirány, amely a $\delta=4{,}6692\ldots$ sajátértéknek felel meg, ettől eltekintve a spektrum egységkörön kívüli része lényegtelen, mivel a $\mathbf{C}^n$ tér változóhelyettesítésével kapcsolatos (l. 3. fejezet).

Tehát a többdimenziós probléma bizonyos értelemben egydimenziósra redukálódik.

6.2. Befejezésül ismertetünk néhány, a fenti kérdéskörhöz kapcsolódó általánosítást. Az [54], [55] munkákban területtartó, azaz *Hamilton-típusú dinamikai rendszereknek* megfelelő leképezéscsaládok perióduskettőződéses bifurkációit tanulmányozták. Ebben az esetben a bifurkáció úgy jelentkezik, hogy a paraméter változásakor elliptikus periodikus trajektóriák hiperbolikussá válnak, ugyanakkor megjelenik egy kétszer akkora periódusú elliptikus trajektória. A bifurkációs paraméterértékek most is eleget tesznek a $\mu_\infty - \mu_n \sim \text{const}\, \tilde{\delta}^{-n}$ univerzális aszimptotikának, ahol $\tilde{\delta} \approx 8{,}72$. Elvileg a szituáció hasonlít az általunk tárgyalt esethez. A tanulmányozandó objektumok: a megfelelő renormálási egyenlet, a fixpont és a $\tilde{\delta}$ állandót is meghatározó spektrum.

Ismét más, a feladat többdimenziós jellegét kihasználó általánosítás azoknak a jóval bonyolultabb bifurkációsorozatoknak vizsgálata, ahol a periódus háromszorozódik, négyszereződik, s.i.t. A többdimenziós esetben a kétparaméteres leképezéscsaládhoz tartozó ilyen bifurkációsorozatok topologikusan stabilisak ([56], [57], [58]). Tekintsük kissé részletesebben a periódust megháromszorozó bifurkációk esetét. Legyen $f(x; \mu)$ $\mathbf{C}^1 \to \mathbf{C}^1$ komplex paraméteres leképezéscsalád. Legyen U_0 azoknak a paraméterértékeknek a tartománya, amelyekre a leképezésnek létezik stabilis fixpontja. E tartomány határán azok a paraméterértékek vannak, amelyekre a fixpontbeli derivált az egységkörre esik. U_0-lal érintkezik két kisebb $U_1^{(1)}$ és $U_2^{(2)}$ tartomány, amelyekben levő paraméterértékekre létezik 3 periódusú stabilis trajektória. Az érintkezési pontokat az jellemzi, hogy ezekre a paraméterértékekre a fixpontbeli derivált értéke $-\frac{1}{2} \pm \frac{\sqrt{3}}{2} i$. Itt lép fel a bifurkáció és keletkezik a 3 periódusú stabilis trajektória. Az $U_1^{(1)}$ és $U_2^{(2)}$ tartomány csatlakozik két-két olyan tartományhoz, amelyek 9 periódusú stabilis trajektóriáknak felelnek meg, s.i.t. Tekintsük a bifurkációs paraméterértékek következő sorozatát: $\mu = \mu_n$-nél keletkezik egy 3^n periódusú stabilis trajektória, miközben az $f^{(3^{n-1})}(z; \mu)$ leképezés fixpontbeli deriváltja egy rögzített harmadik egységgyökkel egyenlő, például $-\frac{1}{2} + \frac{\sqrt{3}}{2} i$. Ekkor $\mu_n \to \mu_\infty$ és $\mu_\infty - \mu_n \sim \text{const} \cdot (\delta^{(1)}(3))^n$, ahol $\delta^{(1)}(3) \approx 4{,}600 + 8{,}981 i$ univerzális állandó, amely nem függ az $f(z; \mu)$ leképezéscsaládtól. A másik egységgyöknek megfelelő $\tilde{\mu}_n$ sorozathoz tartozó állandó $\delta^{(2)}(3) = \overline{\delta^{(1)}(3)}$.

Cikkünk nem öleli fel a *Feigenbaum-univerzalitás* teljes problematikáját. Összefoglalónk keretein kívül maradt az érdekes [59] munka, amely a *Sarkovszkij-rendezés* más szeleteinek univerzalitási típusaival foglalkozik. Kutatók sora szentelt figyelmet az ún. intermittencia univerzális tulajdonságainak (l. [60]—[63]) és a kör önmagába való sima leképezéseinek. Úgy tűnik, a legfrissebb probléma az invariáns tóruszok megsemmisülése a KAM-elméletben (l. [64]—[67]).

Köszönetet mondunk A. I. GOLBERGnak, M. MISIUREWICZnek, JA. B. PESZINnek, M. I. RABINOVICSnak, E. A. SZATAJEVnek, M. FEIGENBAUMnak, P. CVITANOVICnak, A. N. SARKOVSZKIJnak és M. V. JAKOBSONNAK cikkünk témájának megvitatásáért.

IRODALOM

[1] M. METROPOLIS, M. L. STEIN, P. R. STEIN, On finite limit sets for transformations of the unit interval. — J. Combinatorial Theory (A), 1973, **15**:1, p. 25—44.

[2] P. J. MYRBERG. Interation von quadratwurzeloperationen. — Ann. Acad. Sci. Fennicae, 1985, **259**, p. 1—16.

[3] M. J. FEIGENBAUM, Quantitative universality for a class of nonlinear transformations. — J. Stat. Phys., 1978, **19**:1, p. 25—52.

[4] M. J. FEIGENBAUM. The universal metric properties of nonlinear transformations. — J. Stat. Phys., 1979, **21**: 6, p. 669—706.

[5] M. J. FEIGENBAUM, The transition to aperiodic behavior in turbulent systems. — Comm. Math. Phys., 1980, **77**:1, p. 65—86.

[6] M. J. FEIGENBAUM, Universal behavior in nonlinear systems. — Los Alamos Science, 1980, **1**:1, p. 4—27. (Русск. пер.: М. Фейгенбаум. Универсальность в поведении нелинейных систем. — УФН, 1983, 141:2, с. 343—374).

[7] O. E. LANFORD III, Smooth transformations of intervals. — Seminaire Bourbaki 1980/1981, **563**, (Lect. Notes in Math. Berlin; Heidelberg; New York: Springer-Verlag, 1981, **901**, p. 36—54.)

[8] P. COLLET, J.-P. ECKMANN, Interated maps on the interval as dynamical systems. — Basel; Boston; Stuttgart: Birkhäuser, 1980.

[9] P. COLLET, J.-P. ECKMANN, H. KOCH, Period doubling bifurcations for families of maps on R^n. — J. Stat. Phys., 1980, **25**:1, p. 1—14.

[10] А. Н. Шарковский, Сосуществование циклов непрерывного отображения прямой в себя. — УМЖ, 1964, **16**:1, с. 61—71.

[11] А. Н. Шарковский, О циклах структуре непрерывного стображения. — УМЖ, 1965, 17:3, с. 104—111.

[12] D. SINGER, Stable orbits and bifurcations of maps of the interval. — SIAM Journ. on Appl. Math., 1978, **35**:2, p. 260—267.

[13] J. GUCKENHEIMER, Sensitive dependence to initial conditions for one-dimensional maps. — Comm. Math. Phys., 1979, **70**:2, p. 133—160.

[14] L. BLOCK, J. GUCKENHEIMER, M. MISIUREWICZ, L. S. YOUNG, Periodic points and topological entropy of one dimensional maps. — Lect. Notes in Math., **819**: Global theory of dynamical systems, Northwestern 1979. — Berlin; Heidelberg; New York: Springer-Verlag, 1980, p. 18—34.

[15] T. LI, J. A. YORKE, Period three implies chaos. — Amer. Math. Monthly, 1975, **82**:10, p. 985—992.

[16] Ю. С. Барковский, Г. М. Левин, О предельном канторовском множестве. — УМН, 1980, **35**:2, с. 101—202.

[17] M. MISIUREWICZ, Structure of mappings of an interval with zero entropy. — Publ. Math. I.H.E.S., 1981, **53**, p. 5—16.

[18] И. П. Корнфельд, Я. Г. Синай, С. В. Фомин, Эргодическая теория. — М.: Наука, 1980.

[19] Л. А. Бунимович, Я. Г. Синай, Скорость убывания корреляций в одномерных экологических моделях. — В кн.: Термодинамика и кинетика биологических процессов. — М.: Наука, 1980.

[20] E. HOFBAUER, G. KELLER, Ergodic properties of invariant measures for piecewice monotonic transformations. — Math. Zeitschrift, 1982, **180**:1, p. 119—140.

[21] М. Бланк, Оценка скорости убывания корреляций в одномерных динамических системах. — Функц. анализ, 1984, **18**:1, с. 61—62.

[22] S. M. ULAM, J. VON NEUMANN, On combinations of stochastic and deterministic processes. — Bull. Amer. Math. Soc., 1947, **53**:11, p. 1120.

[23] D. RUELLE, Applications conservant une measure absolument continue par rapport a dx sur [0, 1]. — Comm. Math. Phys., 1977, **55**:1, p. 47—51.

[24] Л. А. Бунимович, Об одном преобразовании окружности. — Матем. заметки, 1970, **8**:2, с. 205—206.

[25] А. И. Огнев, Метрические свойства некоторого класса отображений отрезка. — Матем. заметки , 1981, **30**:5, с. 723—736.

[26] M. MISIUREWICZ, Absolutely continuous measures for certain maps of an interval. — Publ. Math. I.H.E.S., 1981, **53**, p. 17—61.

[27] M. V. JAKOBSON, Absolutely continuous invariant measures for one-parameter families of one-dimensional maps. — Comm. Math. Phys., 1981, **81**:1, p. 39—88.
[28] R. SHAW, Strange attractors, chaotic behavior and information flow. — Zeitschrift für Naturforsch. A. 1981, **36a**:1, p. 80—112.
[29] O. E. LANFORD III., A computer assisted proof of the Feigenbaum conjectures. — Bull. Amer. Math. Soc., 1982, **6**:3, p. 427—434.
[30] M. CAMPANINO, H. EPSTEIN, On the existence of Feigenbaum fixed-point. — Comm. Math. Phys., 1981, **79**:2, p. 261—302.
[31] H. EPSTEIN, J. LASCOUX, Analyticity properties of the Feigenbaum function. — Preprint I.H.E.S. P/81/27, 1981.
[32] P. COLLET, J.-P. ECKMANN, O. E. LANFORD III, Universal properties of maps of an interval. — Comm. Math. Phys., 1980, **76**:3, p. 211—254.
[33] Е. Б. Вул, К. М. Ханин, О неустойчивой сепаратрисе неподвижной точки Фейгенбаума. — УМН, 1982, **37**:5, с. 173—174.
[34] H. DAIDO, Theory of the period-doubling phenomenon of one-dimensional mappings based on the parameter dependence. — Phys. Lett., 1981, **83A**:6, p. 246—250.
[35] H. DAIDO, Period-doubling bifurcations and associated universal properties including parameter dependence. — Progress of Theor. Ohys., 1982, **67**:6, p. 1698—1723.
[36] P. CVITANOVIC, Universality in chaos (or, Feigenbaum for cyclists). — Preprint NORDITA, 1983.
[37] P. COULLET, C. TRESSER, Itérations d'endomorphismes et groupe de renormalisation. — J. de Physique Colloque, 1978, **39**:C5, p. C5-25—C5-28; supplement au **39**:8.
[38] M. J. FEIGENBAUM, The onset spectrum of turbulence. — Phys. Lett., 1979, **74A**: 6, p. 375—378.
[39] Р. Л. Добрушин, Случайные гиббсовские поля для решетчатых систем с парным взаимодействием. — Функц. анализ, 1968, 2:4, с. 31—43.
[40] Я. Г. Синай, Теория фазовых переходов. Строгие результаты. — М.: Наука, 1980.
[41] Д. Рюэль. Статистическая механика. Строгие результаты. — М.: Мир, 1971.
[42] D. Ruelle. Thermodynamic formalism. — Addison — Wesley Publishing Company, 1978.
[43] P. GRASSBERGER, On the Hausdorff dimension of fractal attractors. — J. Stat. Phys., 1981, **26**:1, p. 173—179.
[44] J. P. CRUTCHFIELD, M. NAUENBERG, J. RUDNICK, Scaling for externa noise at the onset of chaos. — Phys. Rev. Lett., 1981, **46**:14, pp. 933—935.
[45] B. SHRAIMAN, C. E. WAYNE, P. C. MARTIN, Scaling theory for nois period doubling transitions to chaos. — Phys. Rev. Lett., 1981, **46**:14, p. 935—939.
[46] J. P. CRUTCHFIELD, J. D. FARMER, B. A. HUBERMAN, Fluctuations an simple chaotic dynamics. — Phys. Reports, 1982, **92**:2, p. 45—82.
[47] T. KAI, Lyapunov number for a noisy 2^n cycle. — J. Stat. Phys:, 1982, **29**:2, p. 329—343.
[48] Р. З. Хасьминский, Устойчивость систем дифференциальных уравнений при случайных возмущениях их параметров. — М.: Наука, 1969.
[49] В. В. Петров, Суммы независимых случайных величин. — М.: Наука, 1972.
[50] V. FRANCESCHINI, A., Feigenbaum sequence of bifurcations in the Lorenz model. J. Stat. Phys., 1980, **22**:3, p. 397—406.
[51] V. FRANCESCHINI, C. TEBALDI, Sequences of infinite bifurcations and turbulence in a five-mode truncation of the Navier—Stokes equations. — J. Stat. Phys., 1979, **21**: 6, p. 707—726.
[52] B. DERRIDA, A. GERVOIS, Y. POMEAU, Universal metric properties of bifurcations of endomorphisms. — J. Physics A, 1979, **12**:3, p. 269—296.
[53] F. T. ARECCHI, F. LISI, Hopping mechanism generating $1/f$ noise in nonlinear systems. — Phys. Rev. Lett., 1982, **49**:2, p. 94—98.
[54] P. COLLET, J.-P. ECKMANN, H. KOCH, On universality for area-preserving maps of the plane. — Physica D, 1981, **3D**:3, p. 457—467.
[55] J. M. GREENE, R. S. MACKAY, F. VIVALDI, M. J. FEIGENBAUM, Universal behaviour in families of area-preserving maps. — Physica D, 1981, **3D**:3, p. 468—486.
[56] А. И. Гольберг, Я. Г. Синай, К. М. Ханин, Универсальные свойства для последовательностей бифуркаций утроения периода. — УМН, 1983, 38:1, с. 159—160.
[57] P. CVITANOVIC, J. MYRHEIM, Universality for period n-tuplings in compex mappings. — Phys. Lett., 1983, **94A**:8, p. 329—333.
[58] B. MANDELBROT, On the quadratic mapping $z \to z^2 - \mu$ for compex μ and z: the fractal structure of its $\mathcal{M}$ set, and sclaing. — Physica D, 1983, **7D**: 1—3, p. 224—239.

[59] С. Коляда, А. Г. Сивак, Универсальные константы для однопараметрических семейств отображений: — В кн.: Осцилляция и устойчивость решений дифференциально-функциональных уравнений. — Киев: Институт математики АН УССР, 1982.
[60] Y. POMEAU, P. MANNEVILLE, Intermittent transition to turbulence in dissipative dynamical systems. — Comm. Math. Phys., 1980, **74**:2, p. 189—197.
[61] P. MANNEVILLE, Y. POMEAU, Different ways to turbulence in dissipative dynamical systems. — Physica D., 1980, **1D**:2, p. 219—226.
[62] J. E. HIRSCH, M. NAUENBERG, D. J. SCALAPINO, Intermittency in the presence of noise: a renormalization group formulation. — Phys. Lett., 1982, **87A**:8, p. 391—393.
[63] B. HU, J. RUDNICK, Exact solutions to the Feigenbaum renormalization-group equations for intermittency. — Phys. Rev. Lett., 1982, **48**:24, p. 1645—1648.
[64] M. J. FEIGENBAUM, L. P. KADANOFF, S. J. SHENKER, Quasi-periodisity in dissipative systems: a renormalization group analysis. — Physica D, 1982, **5D**:2—3, p. 370—386.
[65] S. J. SHENKER, Scaling behavior in a map of a circle onto itself: empirica/results. — Physica D, 1982, **5D**:2—3, p. 405—411.
[66] S. J. SHENKER, L. P. KADANOFF, Critical behavior of a KAM surface: I. Empirical results. — J. Stat. Phys., 1982, **27**:4, p. 631—656.
[67] S. OSTLUND, D. RAND, J. SETHNA, E. SIGGIA, Universal properties of the transition from quasi-periodicity to chaos in dissipative systems. — Physica D, 1983, 8**D**:3, p. 303—342.
[68] M. NAUENBERG, J. RUDNICK, Universality and the power spectrum at the onset of chaos. — Phys. Rev. B, 1981, **24**:1, p. 493—495.

Forditotta:
KAJTÁR LÁSZLÓ
ELTE TTK NUMERIKUS
ANALÍZIS TANSZÉK
1088 BUDAPEST
MÚZEUM KRT. 6—8.

A kiadásért felelős az Akadémiai Kiadó és Nyomda főigazgatója
Műszaki szerkesztő: Sándor István
A kézirat nyomdába érkezett: 1985. január 22. — Terjedelem: 21 (A/5 ív)
85-327 — Szegedi Nyomda — F. v.: Surányi Tibor 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

1985/3-4

A MAGYAR TUDOMÁNYOS AKADÉMIA MATEMATIKAI ÉS FIZIKAI TUDOMÁNYOK OSZTÁLYÁNAK KÖZLEMÉNYEI

11. KÖTET

AKADÉMIAI KIADÓ, BUDAPEST

ALKALMAZOTT MATEMATIKAI LAPOK

A MAGYAR TUDOMÁNYOS AKADÉMIA MATEMATIKAI ÉS FIZIKAI TUDOMÁNYOK OSZTÁLYÁNAK KÖZLEMÉNYEI

XI. 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 128 forint. Belföldi megrendelések az Akadémiai Kiadó, 1055 Budapest V., Alkotmány u. 21. címen (pénzforgalmi jelzőszám 215—11 488), *Stúdium* (1368 Budapest, Váci utca 22., Tel.: 185-881) és a *Magiszter* 1052 Budapest, Városház utca 1., Tel.: 382-440) Akadémiai Kiadó Könyvesboltjaiban, 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 11 (1985) 239—296

KLASZTERÁLÁS ALKALMAZÁSA HÁLÓS ADATBÁZIS LOGIKAI TERVEZÉSE SORÁN

MEZEY GYULA

Budapest

A cikk adatbázis szegmens- és area-tervezésével foglalkozik. A normalizált és funkcionálisan elemzett elvi adatmodell egyedtípusait rekordtípusokká, illetve részrekordokká (azaz szegmensekké) képezik le, majd később e rekord (részrekord) típusokat nagyobb egységekbe (areak) fogják össze.

A cikk olyan módszert ismertet, ahol mindkét lépés során klaszteranalízist használnak.

A szegmenstervezéshez agglomeratív hierarchikus klaszterálást és egy nem-metrikus távolságmértéket, az area-tervezéshez pedig táblázat-átrendező eljárást alkalmaznak.

1. Automatikus osztályozás alkalmazása adatbázis tervezésben

1.1. Bevezetés

Jól ismert tény, hogy az adatbázisok hatékonysága erősen függ a tulajdonságtípusoknak rekordokba (vagy részrekordokba) történő leképezésétől, illetve a rekordok (szegmenstípusok) nagyobb egységekbe (pl. area) csoportosításától. Ennek megtervezésére szinte természetesen kínálkozik alkalmas, számítógéppel segített csoportképző eljárás alkalmazása még a fizikai tervezést megelőzően. Mivel így tapasztalat szerint jelentősen csökkenteni lehet az adatelérés idejét, az adatbázistervezésben ennek hatékony megoldása nagy gyakorlati jelentőséggel bír [42].

A feladat két fő lépésre tagolt:

— egyedtípusok szegmenstípusokká leképezése (ezt a cikk 3. fejezete tárgyalja),

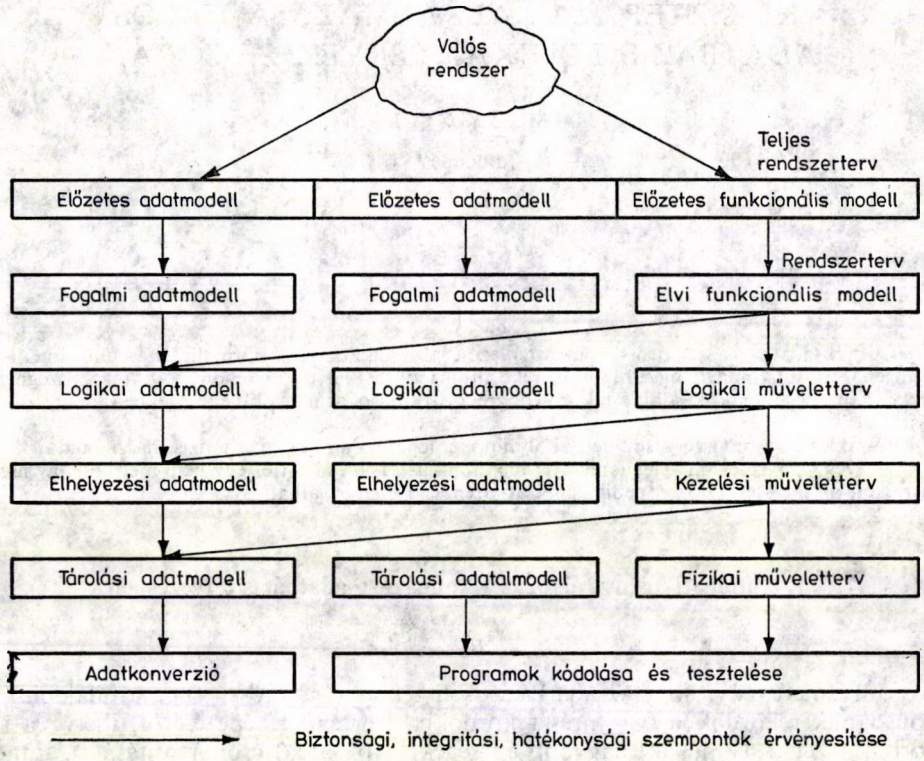
— szegmenstípusok nagyobb egységekké csoportosítása (ezt a 4. fejezet tárgyalja).

A cikkben e két lépés újszerű megoldását írjuk le. A megoldás módszerét mindkét esetben az osztályozásra alkalmazható eljárások egy csoportjából a klaszteranalízis módszereinek köréből választottuk ki. A klaszteranalízis módszereinek sajátosságaival a cikk 2. fejezete foglalkozik.

1.2. A téma jelentősége

A klaszterálás más néven (klaszteranalízis) a többváltozós matematikai-statisztikai módszerek egyike. Igen alkalmas bonyolult rendszerek belső struktúrájának feltárására és tervezésére. Egyre több területen eredményesen alkalmazzák. Tipikus alkalmazások pl. a könyvtári visszakeresés, az alakfelismerési, és a statisztikai problémák területei [18], [44].

Az információs rendszerek tervezésének egyes fázisaiban is már alkalmaznak klaszterálási módszereket. Például az egyes alrendszerek elhatárolására, egy szervezet



1. ábra. Az adatbázistervezés folyamata

információrendszere teljes architektúrájának feltárására alkalmaz heurisztikus kéziklaszterálást az IBM BSP módszere. Bonyolult programrendszerek tervezéséhez is sikeresen használják a számítógépi klaszterálást [40].

Az adatbázis fizikai tervezési menetében — különösen az igen nagy adatbázisok esetén — egyre gyakrabban alkalmazzák az adatszerkezetek tárolószerkezetekre való leképezésének (a rekord elrendezés optimalizálása, az elérési idők minimalizálása) céljából [28], [36], [20].

Tudomásunk szerint azonban eddig csupán elvétve és más módszereket használva alkalmaztak klaszterálást az adatbázis logikai tervezése során. A logikai adatbázistervezés az elvi adatmodellből indul ki. Az elvi adatmodellt pedig egymást átfedő klaszterek (egyedtípusok) halmazaként is felfoghatjuk. Az elvi adatmodell megvalósítása azonban nem célszerű, mert e modell még nem veszi figyelembe a hatékonyság szempontjait. Éppen ezért a logikai adatbázistervezésen belül az elvi modell alkotta *kiinduló klaszterstruktúrát úgy transzformáljuk át, hogy egy egyedtípust több szegmensbe képezünk le, majd e szegmensek közül az adatigény szempontjából közelálló funkciókat kielégítőket egy nagyobb csoportba, un. area-ba vonjuk össze.*

A cikk azzal foglalkozik, hogy e két utóbbi tevékenységre mely klaszterálási módszerek és milyen taxonomikus mértékek alkalmasak,

A cikkben tehát az adatbázistervezésen belüli hozzáféréselemzés és ehhez az eszközként alkalmazott klaszter elemzés kérdéskörével foglalkozunk. *Olyan módszert ismertetünk, amely egy adott elvi adatmodell és adott funkciók alapján lényegében egy újabb — logikai szintű — adatmodellt* hoz létre. A módszer által előállított adatmodellben a tulajdonságtípusok egyrészt az egyedtípusok, másrészt a funkciók csoportosításában szerepelnek.

A fenti megfogalmazásból nyilvánvaló, hogy a javasolt módszer az adatbázisok tervezésének mindhárom szintjét érinti.

Egy elvi adatmodellből kiindulva a módszer segítségével előállított modellnek megfelelően az elvi adatmodell pontosítható és csupán az elvi adatmodell alapján megalkothatjuk a logikai adatmodell első változatát. Alkalmas arra is, hogy segítséget nyújtson az ABKR kiválasztásában. A már választott ABKR korlátait tükröző logikai adatmodellből kiindulva is használható annak hozzáférés-elemzésére. A fizikai tervezés szintjén a módszer az area-tervezés elvégzéséhez döntési szempontokat ad.

A fentiekben leírt körben történő általános alkalmazhatósága mellett a módszer még azzal az előnnyel is rendelkezik, hogy számítógépen történő megvalósítása egyszerű.

Az 1. fejezet további részeiben tömören ismertetjük az 1.1. pontban vázlatosan már említett tervezési folyamat három fő (elvi, logikai, és fizikai) menetének azon tevékenységeit, amelyekkel az általunk javasolt módszer kapcsolatban van. Ez annál is inkább szükséges, mert rá kell mutatni, hogy a javasolt módszer mely tervezési tevékenységet vált ki hatékonyabb, gépesített és elméletileg megalapozott algoritmusokkal, ugyanakkor meg kell mutatni az egész folyamat tevékenységeinek megfelelő illeszkedését is.

1.3. A hozzáférési igények meghatározásának szakaszán belül áttekintett teendők

A logikai tervezés kiindulópontja a már funkcionálisan elemzett elvi adatmodell, ezért indításként áttekintjük most a funkcionális elemzés szakaszának tevékenységeit:

(i) Az elemzés előkészítése:
— a funkciók összegyűjtése,
— adatnevek egyeztetése,
— a tulajdonságtípusok meglétének ellenőrzése,
— adatmodell esetleges kiegészítése,
— mértékadó funkciók kiválasztása elemzésre.

(ii) Az elemzés végrehajtása mértékadó funkciónként:
— belépési pont meghatározása,
— a funkció ellátásához szükséges egyedtípusok sorrendjének megjelölése,
— egyedtípusonként az előfordulások aktuális részhalmazának és az érintett tulajdonságtípusokon végzendő műveleteknek rögzítése.

(iii) Az elvi adatmodell esetleges korrekciója
— elvi adatmodell kiegészítése,
— navigációs utak korrekciója,
— a funkció leírások ellenőrzése.

(iv) Az elemzési eredmények összesítése egyedtípusonként:
— Az elemzett funkciók ellátásához szükséges valamennyi egyedtípus (ezen belül tulajdonságtípusaik) összegyűjtése.

— Egyedtípusonként (ezen belül tulajdonságtípusokra lebontva) a hozzáférési igények összesítése.

A funkcionális elemzés céljai:

— A normalizált adatmodell pontosítása.
E célt az (i) és az (iii) tevékenységcsoportok szolgálják.
— Az adatmodellhez kapcsolódó algoritmusmodellek megalkotása.
Ezt a célt az (ii) tevékenységcsoport szolgálja.
— A logikai adatbázistervezés előkészítése.

Ez utóbbi célt az (i), és az (ii), és az (iv) tevékenység csoport(ok) szolgálják.

Az alábbiakban röviden sorra vesszük az egyes tevékenységek közül azokat, amelyek a logikai adatbázistervezés előkészítését szolgálják.

Az elemzés előkészítését jelentő tevékenységek közül a logikai adatbázistervezés előkészítését közvetve szolgálja a mértékadó funkciók kiválasztása.

1.3.1. Mértékadó funkciók kiválasztása elemzésre

Tapasztalatok szerint egy átlagos szervezet esetében az összes funkció mintegy negyedének elemzése elegendő az adatbázis tervezésekor.

Természetesen a funkciók ellátásához szükséges felhasználói (lekérdező) programok specifikálása a többi funkció elemzését is szükségessé teszi, azonban ezt az adatbázis sémájának véglegesítését követően is el lehet végezni.

Előfordulási gyakoriságuk szerint csoportosítjuk a funkciókat és a továbbiakban egyelőre csak a tömegszerűségük, fontosságuk miatt mértékadókkal foglalkozunk. A mértékadó funkciók kiválasztását a következő szempontok befolyásolják:

— a funkció fontossága, súlya,
— az adatigények tisztázottsága,
— a rendelkezésre álló szervezői kapacitás.

A funkció súlyát, fontosságát több tényező együttesen határozza meg.

Ezek a tényezők általában a:

— gyakoriság,
— mozgatott adattömeg,
— válaszadási idő.

A legfontosabbak azok a funkciók, amelyek nélkül az adatbázis nem tartalmazhat helyes adatokat. Ezek az adatbázis karbantartó funkciók, amelyeket tehát teljes körűen elemeznünk kell.

A további funkciók kiválasztására nézve már csak irányelveket lehet adni.

Természetes, hogy a sűrűn gyakorlandó funkciók végrehajtására nézve az adatbázisnak lehetőleg hatékony működésűnek kell lennie. Válasszuk tehát elemzésre a gyakran ismétlődő funkciókat.

Ha valamely funkció végrehajtása nagy adattömeg feldolgozását igényli az adatbázisból, akkor a hatékonyság még a funkció viszonylag ritkább előfordulása esetében is lényeges. Ezért ezek a funkciók is elemzésre jelölendők.

Az adatbázis párbeszédes üzemmódú használata az egyik olyan helyzet, amikor lényeges a gyors válasz, a várakozási idő csökkentése érdekében. Ha tehát a funkció gyors választ igényel, akkor különös gonddal kell elemezni.

Ettől teljesen eltérő okból, — a feldolgozás hatékonysága érdekében — de szintén lényeges a feldolgozás minél gyorsabb végrehajtása, ha olyan nagy tömegű gyűjtött adat érkezik be, amivel az adatbázist karban kell tartani, majd szoros határidővel

kell sok táblát készítenünk az így karbantartott adatbázis adataiból. Ilyen például a negyedéves, féléves és éves statisztikai kötegelt feldolgozások. Emellett fontosak a navigálás (lásd (ii)) legmegfelelőbb módjának megtalálása érdekében azok a kötegelt feldolgozások is, amelyek az adatmodell túlnyomó (vagy jelentős) részét érintik.

Az eddigiekben még csak szempontokat vettünk fel, melyeket tükröző alkalmas mérték adott körülmények közötti megalkotásával a fenti elvi útmutatás alapján a funkciók ún. ABC — elemzését is el lehet végezni.

A logikai adatbázistervezés szemszögéből összegezve azt kell megállapítanunk, hogy az egyes funkciókat elsősorban fellépésük becsült relatív gyakorisága, a mozgatott adattömeg, és az elvárt, válaszadási idő alapján súlyokkal látjuk el.

A fenti objektíve jól becsülhető súlytényezők mellett még egyéb, szubjektív becslés, alapján adott súlytényezők is elképzelhetőek. Például a funkció súlyába beépíthető az, hogy mennyiben tisztázott annak adatigénye.

Itt két dolgot kell a funkcióra nézve megnézni: az egyik az érintett adattartalom megléte, azaz a kívánt új információ szolgáltatásához szükséges adatoknak szerepelniök kell az adatbázis adatai között. Ez ellenőrizhető az adatmodellen (lásd a 2. ábrát) (Bachman-diagram) majd szükség esetén a kiegészítés megtehető.

Másik az, hogy tisztázott legyen a kívánt információ előállítási algoritmusa, illetve az ehhez szükséges felhasználói munkafolyamat. Ide kell érteni a szükséges adatok adatbázisba kerülésének folyamatát is.

1.3.2. Az elemzés végrehajtása mértékadó funkciónként

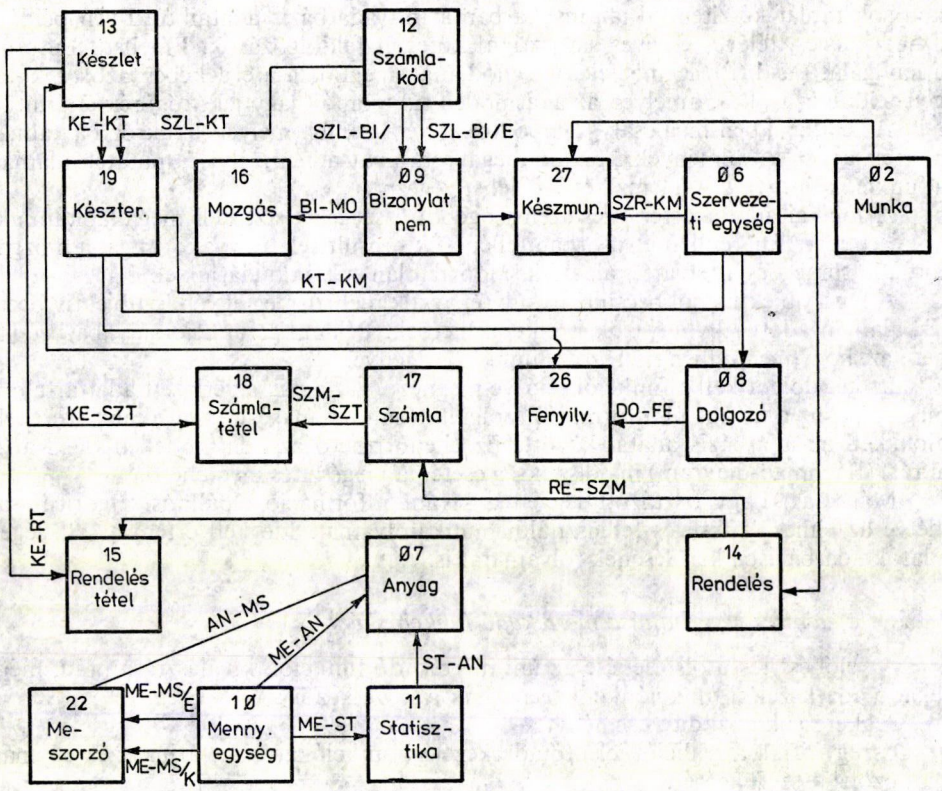
Az előkészítés utolsó lépése az ún. mértékadó funkciók kiválasztása után, mértékadó adatkezelési funkciónként teendőink a következők:

— Megkeressük az indító eseményt.
— Összegyűjtjük a reakciót jelentő adatkezelési funkciókat és feltárjuk ezek egymás közötti összefüggését.
— Megállapítjuk, hogy a feltárt funkciók milyen tulajdonságtípusokat igényelnek, illetve milyen tulajdonságtípusokat állítanak elő.
— Eddigi összes megállapításunkat egy rövid verbális leírásban foglaljuk össze ez kiegészülhet a vonatkozó jogszabályok, szabályzatok stb. szövegrészleteivel, illetve döntési táblákkal, továbbá az input-output bizonylatokkal, kitöltési utasításokkal.
— Megvizsgáljuk, hogy előállítható-e a szükséges információ a már meglevő adatmodellből, illetőleg beépíthetők-e abba a termelődő adatok, átvezethetők-e a módosítások.

Ehhez szükségünk van mind az adatmodell grafikus ábrájára azaz a Bachmann-diagramra, mind pedig az abban szereplő egyed-, illetve tulajdonságtípusok leírására. Ezek segítségével kétféle dokumentumot készítünk:

— navigációs ábrát (lásd a 3. ábrát) és
— a funkció elérési útjának leírását.

A navigációs ábrán azt szemléltetjük, hogy a kívánt adatok elérése, vagy új adat helyének megkeresése érdekében hol kell majd belépnünk az adatbázisba, hogyan kell abban a kapcsolatokon áthaladnunk és hol lépünk ki abból. Az egész folyamatot az adatmodellben történő navigációnak nevezzük. Ehhez kiindulásnak az adatmodell Bachmann-diagramját használhatjuk legcélszerűbben. A navigációs ábrát funkcionális modellnek is szokták nevezni és megszerkesztése a 4. ábra konvencióit követi.



2. ábra. Adatmodell

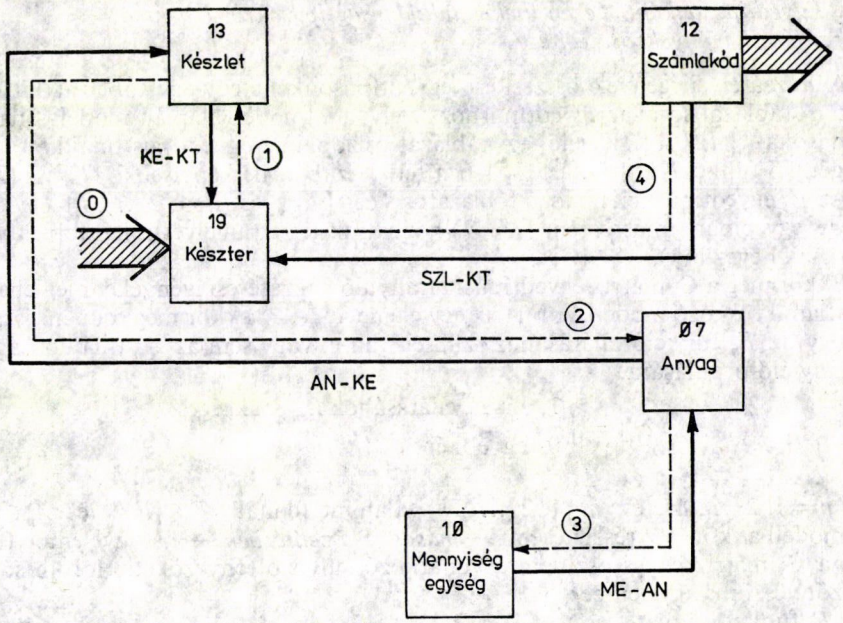
Annak érdekében, hogy a navigációs út jól megkülönböztethető legyen a Bachmann-diagramban feltüntetett egyedtípusok közötti kapcsolatoktól, szaggatott nyilakkal jelöljük azt, külön jelezve a belépés és a kilépés helyét.

A belépési pont meghatározására az ún. vezérlőinformáció szolgál (vö. [24] 125. oldalon). Megjegyezzük, hogy elsősorban az adatbázis kötegelt üzemmódú használatakor tapasztalhatóan — a gyakorlatban ugyanazon adatkezelési funkció vonatkozásában gyakran többféle szóbajöhető vezérlőinformációval is rendelkezünk, s ettől függően viszont többféle belépési pont, illetve navigációs út létezhet, melyek mindegyike ugyanazt az adatvisszakeresési igényt elégíti ki.

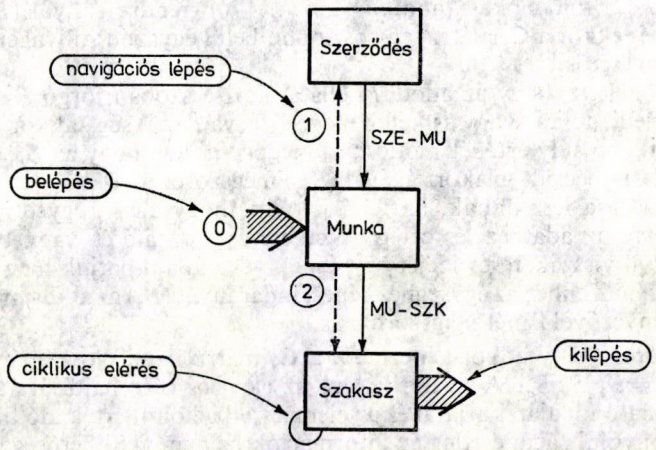
Ilyenkor dönteni kell, hogy melyik tulajdonságtípust tekintjük vezérlőinformációnak.

A navigáció leírásakor meg kell adnunk a kiválasztás másodlagos ismérvét is (ha van ilyen). Sok esetben ugyanis, ha eljutottunk valamely egyedtípushoz, korántsem biztos, hogy annak valamennyi előfordulását el kell érni.

Nem mellékes, hogy egyedi, teljes vagy ciklikus (ismételt) keresésre van-e szükség.



3. ábra. Adatmodell



4. ábra. A navigációs ábra szimbólumai

Végül érdekes még az, hogy az adatkezelés célja az egyes egyedtípusok vonatkozásában éppen visszakeresés, hozzáadás vagy törlés-e.

Az elért egyedelőfordulások kiválasztott tulajdonságtípusain végzendő művelet is lehet visszakeresés vagy módosítás.

1.3.3. Egyedtípusonként (ezen belül tulajdonságtípusokra lebontva) a hozzáférési igények összesítése

A hozzáférési igények összesítése egyedtípusonként egy-egy táblázat kitöltését jelenti. A táblázat sorai az egyedtípushoz az elvi adatmodell alapján rendelt tulajdonságtípusokat $\{A\}^M \subseteq \{A\}^Q$ jelöli. A táblázat oszlopai az egyedtípus tulajdonságtípusait igénylő funkciókat (vö. az 1.4.1-ben tett leszabással) képviselik $\{F\}^N \subseteq \{F\}^R$.

Az E_j egyedtípushoz tartozó táblázatot $(E_j \in \{E\}^J$, és $j=1, 2, \ldots, J)$ *első közelítésben* egy bináris mátrixként (M, N) hozhatjuk létre (amelyet a későbbiekben $\mathbf{C}_j$ mátrixnak nevezünk), $\mathbf{C}_j \in \{\mathbf{C}\}^J$.

Ekkor még a $\mathbf{C}_j$ mátrix egyedtípusra irányuló hozzáférési igényeket meglehetősen pontatlanul tükröző modell. A mátrix egy eleme $c_{mn} \in \mathbf{C}_j$ ekkor még csupán azt fejezi ki, hogy az F_n funkció ellátásához szükség van-e, vagy sem az A_m tulajdonságtípus valahány előfordulására.

$$c_{mn} = \begin{cases} 1, & \text{ha } F_n \text{ ellátásához } A_m \text{ szükséges} \\ 0, & \text{egyébként,} \end{cases}$$

ahol $m=1, 2, \ldots, M$ és $n=1, 2, \ldots, N$ valamint fennáll $F_n \in \{F\}^N$ és $A_m \in \{A\}^M$. Ez a modell akkor lesz részletesebb — *második közelítésben* — ha figyelembe tudjuk venni a $\mathbf{C}_j$ mátrix egyes oszlopaira vonatkozó súlyozó tényezőket. Három súlyozó tényezőt veszünk figyelembe:

— az F_n funkció súlyát,
— az E_j-re irányuló keresések számát (F_n esetében),
— az E_j-re irányuló adatkezelés célját (F_n esetében).

Az $F_n \in \{F\}^N$ funkciókat (ahol $n=1, 2, \ldots, N$) rendre súllyal (S_{F_n}) látjuk el (vö. az (i)-vel) és ekkor a $\mathbf{C}_j$ mátrix egy oszlopon belül egyazon súllyal ellátott bináris mátrixként kezelhető.

Ha az is rögzített, hogy az adott F_n ellátásához a szóban forgó E_j egyedtípusra vonatkozó egyedi, teljes, vagy ismételt keresésre van szükség, akkor ez az n-edik oszlopra vonatkozó súly értékét módosítja, mégpedig úgy, hogy az S_{F_n} súlytényezőt a keresések relatív becsült gyakoriságával (S_{gy}) megszorozzuk.

Hasonló hatása lesz annak, ha feltárjuk azt, hogy az E_j egyedtípusnál az F_n funkciót tekintve az adatkezelés célja visszakeresés, hozzáférés vagy törlés. A legkisebb időigényű visszakeresés idejét egységnek véve állapíthatjuk meg a hozzáadás vagy törlés esetének ehhez az egységhez képest időarányát (S_a) az addig már megkapott súlyt ezzel a tényezővel ismét megszorozzuk.

A fenti három tényezőből képezhető, a $\mathbf{C}_j$ mátrix n-edik oszlopára vonatkozó súly tehát: $S_n = S_{F_n} S_{gy} S_a$. A gyakorlatban az elemzés végrehajtására (vö. az 1.3.2. ponttal) fordítható időalap korlátai és az elemzésre kijelölt mértékadó funkciók nagy száma vagy bonyolultsága, esetleg az információk be nem gyűjthetősége nem mindig teszi lehetővé, hogy a hozzáférési igényeket ennél részletesebben, pontosabban rögzítsék. Természetesen kívánatos, hogy az is legyen rögzített a helyzetfelmérés során, hogy az F_n esetében az E_j tulajdonságtípusain milyen műveleteket kell végezni, s ebből ugyancsak súlyozó tényezők (S_m) képezhetők, amelyek azonban már nem a $\mathbf{C}_j$ mátrix n-edik oszlopára, hanem csak az oszlop megfelelő elemeire vonatkoznak.

Ha a már előbb tárgyalt oszlopokra vonatkozó súlyok mellett ez utóbbi elemekre vonatkozó súlytényezőket is figyelembe vesszük, akkor a $\mathbf{C}_j$ mátrix már nem tekint-

hető egyszerűen egy oszloponként súlyozott bináris mátrixnak, hanem — *harmadik közelítésben* — csak egy eleve nem-bináris mátrixnak, amely már kellő pontossággal modellezi az egyedtípusra vonatkozó hozzáférési igényeket.

Az első közelítésű modell még nem alkalmas a logikai tervezés megalapozásához.

A második és harmadik közelítésű modellek azonban már erre alkalmasak, s ennek megfelelően fogunk majd a 3.2.1. és a 3.2.2. pontban a szegmensképzés módszere szempontjából két alapesetet tárgyalni.

A második közelítésű súlyozott bináris $\mathbf{C}_j$ mátrix modellből kiinduló módszer leírását a 3.2.1. pont míg a harmadik közelítésű (nem-bináris $\mathbf{C}_j$ mátrix) modellből kiinduló módszer ismertetését a 3.2.2. pont tartalmazza.

1.4. A logikai adatbázis tervezés menetén belül áttekintett tevékenységek

A logikai adatbázis tervezés menete során a hatékony működés érdekében az elvi adatmodellen változtatásokat hajtunk végre. Ezek az alábbi szakaszokba tagolhatóak:

(i) Elvi adatmodell ABKR által nem kezelendő részeinek leszabása.
(ii) ABKR kiválasztása.
(iii) A kiválasztott ABKR korlátainak és lehetőségeinek figyelembevételével az elvi adatmodell átalakítása.
(iv) A hatékonyság és biztonság érdekében alkalmazott kompromisszumok figyelembevételével az elvi adatmodell átalakítása.

Eltérő tervezési szituációk és így sorrendek is léteznek, melyek közül gyakori az (i), (iv), (ii), (iii), valamint az (ii), (i), (iv), (iii) sorrend is.

1.4.1. Az elvi adatmodell leszabása

Az elvi adatmodellnek az ABKR által kezelni kívánt részét különválasztjuk a más módon kezelni kívánt részektől. Az értekezés további részeiben a rövidség kedvéért az elvi adatmodell elnevezéssel a már „leszabott" elvi adatmodell-részt fogjuk illetni.

1.4.2. Az ABKR kiválasztása

Az ABKR kiválasztása praktikusan gyakran ezt megelőzően már megtörtént. Elvileg azonban egészen eddig halasztható, s mivel ekkorra az adatmodell hozzáférési viszonyait már felmértük, a funkcionális elemzés során a mértékadó funkciókra részletesen meghatároztuk, teoretikusan megalapozottabb lehet a konkrét helyzetnek legjobban megfelelő ABKR kiválasztása most.

Attól függetlenül, hogy ABKR-t ezúttal választhatunk-e vagy kiválasztását már nem, vagy alig befolyásolhatjuk, szükséges az ABKR működésének hatékonyságát érintő alábbi kérdéseket megvizsgálni.

A funkcióelemzés során készült felmérési anyag átvizsgálásával megállapítjuk, hogy maradt-e olyan egyedtípus az eddig már szűkített adatmodellben, amelyet a többitől elszigetelten használnak, vagyis ahol navigáció gyakorlatilag nincs. Ha ilyet találunk, akkor célszerű ezt is leszabni az ABKR által kezelni kívánt szűkített adatmodellből.

Ezután a felhasználói igényekből kiindulva megvizsgáljuk, hogy valóban indokolt-e ABKR használata, és ha igen, milyen típusú ABKR választása volna előnyös.

A felhasználói igények ugyanis megmutatják, hogy melyek azok a kapcsolattípusok, amelyeket döntően, avagy esetleg kizárólag a kapcsolat irányával ellenkező irányban (tehát alárendelttől fölérendelt egyedtípus felé) bejárva használnak. Az ilyen kapcsolattípust nem célszerű adategyüttessé leképezve megvalósítani. Ehelyett az alárendelt egyedtípusban meghagyjuk a fölérendelt egyedtípus azonosítóját, de adategyüttesként a kapcsolattípust nem deklaráljuk. Ennek az az oka, hogy ahol lehet, ott csökkenteni célszerű a megvalósítandó adategyüttesek számát, mivel azok nagy száma a működés hatékonyságát lerontja.

Ez különösen akkor indokolt, ha az alárendelt egyedtípus adategyüttesben lenne tag. Ahhoz ugyanis, hogy az adategyütteseknek megfelelő táblázatokat új tagelőfordulás hozzáadása, vagy törlése esetén az ABKR karbantartsa, viszonylag sok időre van szükség.

Pl. a CODASYL ajánlásokat követő UDS (Siemens) ABKR esetében (amely sok tekintetben hasonló az IDMS-hez) fennáll az, hogy ha egy rekordtípus pl. 10 adategyüttesben tag, akkor az ennek a rekordtípusnak megfelelő 10 000 új rekordelőfordulás létrehozása elérheti a 100 gépórát is Siemens 7738 típusú számítógépen BS 2000 operációs rendszer feltételezésével.

Annak az oka, hogy egy egyedtípus sok adategyüttesben jöhetne tagként szóba két különböző dolog lehet.

Az egyik gyakori ok az, hogy a felhasználó még az előzetes adatmodell helyzetfelmérése során igen sokféle kapcsolatot sorol fel az egyedtípusok között. Mivel ezek egy része nem fontos kapcsolattípus (és ez a funkcionális elemzés helyzetfelmérése után többé-kevésbé ki is derül), e kapcsolatokat kifejező séma megvalósítása eleve rendkívül gazdaságtalan lenne.

A másik gyakori ok az, hogy a tervezők az esetek nagyrészében a tárterülettel való takarékoskodás és a logikai redundancia csökkentése érdekében inkább hajlanak a komplex adatstruktúrák, adategyüttesek (multiple — set, multiple — member set) megvalósítására, mint a szándékos „helypazarlásra" a gyorsabb karbantartási idők és egyszerűbb programozás érdekében. Kétségtelen viszont, az is, hogy a sok adategyüttes megvalósítása meggyorsítja a tárolt adatok visszakeresését. Lényegében tehát a visszakeresés és a karbantartás szempontjai egymás ellen hatnak: ha gyors visszakeresést tűzünk ki célul, úgy viszonylag lassú karbantartásra kell számítanunk és fordítva.

Ennek megfelelően akkor járunk el logikusan, ha már jóelőre mérlegeljük, hogy adott esetben milyen a felhasználói igény: a minél gyorsabb visszakeresésre, vagy pedig a karbantartásra, azaz a minél frissebb adatokra teszi-e a hangsúlyt?

Így eldönthetjük, hogy megengedjük-e a sok adategyüttest, vagy inkább az adategyüttesek számának csökkentésére törekszünk. Igaz ugyan, hogy elég tipikusnak tekinthető az olyan felhasználói igény, ahol a hangsúly a felhasználó gyors információellátásán van, azonban túlságosan rossz hatékonyságú karbantartás sem engedhető meg. Nos erre, az esetre gyakori ökölszabályként ajánlja [3] a következőket:

— az adatmodell egészére nézve a kapcsolattípusok darabszáma lehetőleg ne haladja meg az egyedtípusok darabszámának kétszeresét:
— ha a kapcsolattípusok darabszáma azonos, vagy kevesebb, mint az egyedtípusok darabszáma, akkor célszerű alaposan felülvizsgálni, hogy valóban indokolt-e ABKR használata (lehetséges, hogy ez igen gazdaságtalan lenne);
— egy rekordtípus legfeljebb 3-4 adategyüttesben legyen tag.

Ha ennél több adategyüttesbeli tagság feltétlenül szükségesnek látszik, akkor megfontolandó, hogy helyes-e CODASYL típusú ABKR használata. Ilyenkor általában indokoltabb olyan adatbázis- vagy file-kezelő rendszer igénybevétele, amely elsősorban invertált szerkezetekkel dolgozik. Egy rekord ugyanis általában azért szokott sok adategyüttesben tag lenni, mert sokféle szempont (többnyire másodlagos kulcs) szerint kell visszakeresni. Erre az esetre pedig az invertált szerkezetek a legmegfelelőbbek.

1.4.3. ABKR figyelembevétele

A logikai adatbázisterv elkészítése során a már leszabott adatmodellt átalakítjuk.

Erre egyrészt azért van szükség, mert az adott esetben alkalmazásra kerülő ABKR általában csak korlátozásokkal képes a normalizált és funkcionálisan elemzett kész elvi modell kapcsolatait kezelni.

1.4.4. Az elvi modell torzítása

Másrészt ettől függetlenül azért is szükséges átalakítás, mert akár az elvi vagy akár az átalakított modell kezelése általában nem biztosítana jó hatásfokot. A hatásfok növelése érdekében az elvi modellen (vagy már az alkalmazásra kerülő ABKR korlátait figyelembevevő módosított változatán) az alábbi főbb szempontok alapján korrekciókat hajtunk végre:

Ha egy egyedtípus tulajdonságtípusainak vannak olyan csoportjai, amelyekre azonos jogosultság vagy sűrű együttes használat jellemző, akkor a szóban forgó egyedtípust nem egyetlen rekordtípusba, hanem e csoportnak megfelelő részrekordokba (szegmenstípusokba) képezzük le.

Ha ugyanazon egyedtípusokra (vagy azok meghatározott szegmenstípusaira) a gyakori együttes használat jellemző, akkor a szóban forgó szegmenstípusok (esetleg az egyedtípusok) összevonását célszerű elvégezni.

A fenti tevékenységeknek háromféle hatása is kitűnik:

— A fizikai tervezés jobban előkészített talajról indul, a file-szervezés eredménye is csak jobb lehet.
— A hatékonyabban kezelhető adatbázis érdekében tudatosan eltérünk az adatmodell normalizált állapotától.
— Nem csupán az egyedtípusok, hanem a kapcsolattípusok megbontása, vagy összevonása is szükségessé válhat.

Egyedtípusok szegmenstípusokba történő leképezése

A leképezés főbb szempontjai a következők. Szegmens egyazon egyedtípusra vonatkozó tulajdonságtípusok halmazának hierarchikusan rendezett részhalmaza, amely lehet csak egyszerű (egyszintű hierarchia), de lehet összetett (compound group) szegmens is (esetleg többszintű hierarchia).

Az összetett szegmens abban különbözik az egyszerű szegmenstől, hogy a szegmensen belül logikailag szorosan összetartozó tulajdonságtípusokat ún. alárendelt csoportokba tagolva — csoportnévvel és ismétlési attributummal lehet jellemezni és az összetett szegmensen belül az alárendelt csoportok többszintű hierarchiát is alkothatnak [11] (90—93 oldalon).

Az ABKR-ek egyrésze számára a szegmensképzés alapvetően szükséges tervezési fázis, mert a szegmens (akár összetett szegmens, akár egyszerű) az a legkisebb adategység, amely az adatbázisban egyszerre kezelhető. Más ABKR-ek azonban egyes tulajdonságtípusok elérését is lehetővé teszik. Ez utóbbiak esetén viszont ugyancsak célszerű lehet szegmensek kialakítása pl. az adatbiztonság, illetve az adatkarbantartás szempontjai miatt, hiszen a szegmens-szintű elérés egyúttal szegmensszintű védelmet is lehetővé tesz.

A szegmens jellemzője, hogy az általa tartalmazott tulajdonságtípusok legalább egy rendezési ismérv segítségével egyértelműen azonosítani tudják azt az egyedelőfordulást, amelyre a szegmens vonatkozik. Az hogy egy szegmenst több rendezési ismérv hívhat meg, a hagyományos adatállományoktól különböző adatbázis lényeges jellemzője. A szegmensek homogén halmazát szegmenstípusnak nevezzük és névattributummal a tartalmazott tulajdonságtípusok és alárendelt csoportok neveivel és ismétlési attributumaival jellemezzük.

A szegmensképzés lényeges jellemzője az, hogy itt az adatbázistervező nem csupán a normalizált adatmodell kapcsolatait, hanem pl. az egyes funkciók adatigényeit és a lekérdezési gyakoriságokat is figyelembe veszi. Az adatbiztonság és karbantartás szempontjait éppen az imént említettük. De egy már működő adatbázis környezetében változik a funkciók köre, ezek adatigénye, fellépésük gyakorisága, feldolgozásuk prioritása is. Mindez az adatbázis átszervezését igényli úgy, hogy eközben a futó programok működtetésében se legyen fennakadás.

A szegmenstervezés számára ez azt jelenti, hogy a későbbiekben várható bővülések lehetőleg új szegmenstípusok legyenek és ne a meglevő szegmenstípusok belső tartalmát és szerkezetét kelljen változtatni. Erre készülve célszerű elkerülni a kevésszámú tulajdonságtípusból összeépített terjedelmes szegmenstípusok gyakorlatát.

Szempont az is, hogy ha kevesebb szegmenstípus van, könnyebb lehet egy program kódolása és a kevesebb elérés miatt rövidebb a program futása, viszont a szegmensek egy konkrét program által nem is használt adatokat cipelnek, melyek mozgatása, lebontása felesleges teher. Egyenletes fizikai tárolás lehetséges, ha a kialakított szegmenstípusok hosszúsága hasonló.

Számos gyakorlati szempont merül fel — mint látjuk —, amelyet az adott helyzetnek megfelelően kisebb vagy nagyobb súllyal kell figyelembe venni, és amelyek abba az irányba hatnak, hogy még az egyes adatok elérését lehetővé tevő ABKR-ek alkalmazása esetén is képezzünk szegmenseket (esetleg itt egyes szegmenstípusokat csupán egy tulajdonságtípus alkot).

A kapcsolattípusok tekintetében fellépő változások: A szegmensképzés eredményeképp egyedtípusonként megkapjuk az egyedtípust jellemző tulajdonságtípusok diszjunkt részhalmazait.

Az adatmodellt alkotó $\{A\}^Q$ halmazt ugyanis (vö. az 1.3. pont (iv)-vel) leképezzük egy, a szegmenstípusok alkotta T-elemű $\{SZ\}^T \neq \emptyset$ halmazzá, mégpedig úgy, hogy $\{A\}^Q$ egyes elemei esetleg több szegmenstípushoz is tartozhatnak, azaz a leképezés nem egyértelmű. Másképp fogalmazva fennáll a szegmenstípusokká leképezett adatmodellre nézve az, hogy

$$\exists_{SZ_t, SZ_z \in \{Sz\}^T} [(SZ_t \cap SZ_z \neq \emptyset) \wedge (SZ_t \nsubseteq SZ_z) \wedge (SZ_z \nsubseteq SZ_t)],$$

ahol $t = 1, 2, \ldots, T$ és $z = 1, 2, \ldots, T$ valamint $t \neq z$,

emellett viszont az egy egyedtípushoz tartozó W -elemű $\{SZ\}^W$ halmazra nézve, ahol $\{SZ\}^W \subseteq \{SZ\}^P$ is fennáll az, hogy

$$\forall_{SZ_u, SZ_v \in \{SZ\}^W} [SZ_u \cap SZ_v = \emptyset) \vee (SZ_u \subseteq SZ_v) \vee (SZ_v \subseteq SZ_u)],$$

ahol $u = 1, 2, \ldots, W$ és $v = 1, 2, \ldots W$ valamint $u \neq v$. A tulajdonságtípusoknak a szegmensképzéssel nyert, egyazon egyedtípushoz tartozó diszjunkt részhalmazai egyikében helyezkedik el az egyedtípus azonosítója. A kapott többi szegmenstípusba ugyancsak be kell vinnünk — az egyedtípushoz tartozást kifejezendő — az egyedtípus azonosítóját. Ez történhet úgy, hogy az azonosítót megismételjük, de úgy is, hogy pointer képviseli.

Ez a gyenge logikai redundancia egy különös esete, ahol a kapcsoló tulajdonságtípus mindkét szegmenstípus azonosítójával megegyezik.

Egy kapcsolattípust tekintve, amelynek vagy a fölérendelt, vagy az alárendelt egyedtípusát szegmensekre bontottuk, e kapcsolattípus megbontása is szükségessé válhat.

Az eredeti (szétbontás előtti) egyedtípust jellemző fölérendeltségi kapcsolattípusok azon a (szétbontás utáni) szegmenstípuson keresztül valósulhatnak meg, amely a kapcsoló tulajdonságtípust ténylegesen tartalmazza. Az eredeti egyedtípust jellemző alárendeltségi kapcsolattípusok a szétbontás után kapott valamennyi szegmenstípus esetében megmaradnak. A szegmensképzés eredményeképp tehát megkapjuk egyrészt a szegmenstípusok halmazát, másrészt a szegmenstípusok közötti kapcsolattípusok halmazát.

Szegmenstípusok összevonása

A szegmensképzés eredményéből kiindulva összevonásokat kell majd végrehajtanunk *több cél* érdekében is:

Az első cél az, hogy a hatékonyság érdekében ezen a módon korlátozhatjuk azt, hogy az adatmodell egészére nézve a kapcsolattípusok darabszáma mennyivel haladja meg a szegmenstípusok számát, továbbá azt, hogy egy szegmenstípus legfeljebb 3-4 adategyüttesben legyen tag (vö. az 1.4.2. pontban kifejtettekkel). Általában cél az, hogy korlátozzuk azt, hogy egy szegmenstípus hány adategyüttesben legyen tag. Hogy ez fontos, vagy mellékes cél-e, az az adott esetben megválasztott ABKR azon tulajdonságától függ, hogy gyorsan, vagy lassan képes-e kezelni nagymennyiségű új tagelőfordulás beillesztését (törlését). Ha az ABKR ilyen tekintetben gyors, akkor ez a cél nem túl lényeges és a szegmenstípusok összevonásának fő célja inkább az area-képzés segítése lesz.

A szegmenstípusok összevonásának következménye az, hogy a kapcsolattípusok tekintetében is változásoknak kell megjelenni. Ezek az alábbiak ([24]):

A) Ha a két szegmenstípus ugyanazon egyedtípus szegmenstípusa volt, vagy ez nem áll ugyan fenn, de kölcsönös függésben vannak, akkor:
   — valamelyik szegmenstípusban levő azonosító megszűnik,
   — így a kapcsolattípus is megszűnik,
   — ha a két szegmenstípus fölérendeltjei között van azonos szegmenstípus, akkor az egyik fölérendeltségi kapcsolattípust elhagyjuk,
   — a két szegmenstípus alárendeltjei egyaránt az új, összevont szegmenstípus alárendeltjei lesznek.

B) Ha a két szegmenstípus eltérő egyedtípusokhoz kötődik és közöttük funkcionális függés áll fenn, akkor két esetet valósíthatunk meg:
   a) Az alárendelt szegmenstípusba vonjuk be a fölérendeltet:
      — mindkét szegmenstípus fölérendelt kapcsolatai továbbra is megmaradnak,
      — az összevont szegmenstípus azonosítója a két szegmenstípus azonosítójától összetett azonosító lesz,
      — és ezért megváltoznak mindkét szegmenstípus alárendelt kapcsolatai is.
   b) A fölérendelt szegmenstípusba vonjuk be az alárendeltet:
      — a fölérendelt szegmenstípus fölötti kapcsolattípusok nem változnak,
      — a fölérendelt szegmenstípus egyéb alárendelt kapcsolatai változatlanok,
      — az alárendelt szegmenstípus alárendelt kapcsolattípusai megváltoznak, mert a kapcsolatot most az új összevont szegmenstípus azonosítója jelenti.

Az alárendelt szegmenstípusnak a fölérendeltbe bevonásával tulajdonképpen a kapcsolatot adategyüttes helyett a fölérendelt szegmenstípusba helyezett ismétlődő csoporttal hoztuk létre. Ez általában a visszakeresési idő csökkentése és a szükséges tárolóigény növelése irányában hat (hacsak nem hozunk létre változó hosszúságú szegmenseket).

A második cél az, hogy az egyes adatkezelési funkciók által kezelt adatalmodellek közötti átfedéseket felismerve, a nagymértékben átfedő adatalmodellekből viszonylag homogén csoportokat képezve meghatározzuk, hogy mely funkciókhoz lehet majd egy közös alsémát definiálni. Ezzel a programozási munka volumenét is csökkenthetjük.

A harmadik cél az, hogy area (vagy realm) felosztást kell végeznünk. Ez azt jelenti, hogy szegmenstípusokat kell nagyobb egységekbe areakba csoportosítanunk.

A negyedik a harmadikkal szorosan összefonódik, s az areak-nak particióba, fileokba történő leképezését jelenti azon ABKR-ek esetén, ahol particiók, ill. fileok definiálhatók.

Mind a négy fenti tevékenység a szegmenstípusok csoportosításának feladata, amelyekben közös még az is, hogy alapvetően annak az alapján végezhető el, hogy a szegmenstípusokat mely adatkezelési funkciók használják. Eltérő jellegzetesség az, hogy amíg az első és másodikként említett teendők a logikai tervezés, addig a harmadik és negyedik teendők a fizikai tervezés menetében kerülnek sorra. Az itt most utalásszerűen, a 3. és a 4. fejezetben pedig bővebben kifejtettek tehát nem csupán az adatbázis logikai, hanem fizikai tervezésének tevékenységeit is segítik.

1.5. A fizikai adatbázis-tervezés menetén belül áttekintett tevékenységek

Kiinduló pontja a logikai adatbázis- terv, és a funkcionális elemzés során feltárt adatok. Ezek közül a leglényegesebbek:
— a hozzáférési igények második (esetleg harmadik) közelítésű modelljét (lásd az 1.3 pontban) alkotó tényezők.

A fizikai tervezés menete két szakaszra bontható:
— hozzáférési módok tervezése szegmenstípusonként,
— tárolási szerkezet tervezése.

A hozzáférési mód tervezése egy szegmenstípus esetében az alábbi tevékenységekre tagolható szakasz:
— meg kell állapítani, hogy a kiválasztott ABKR esetében hányféleképpen és milyen

módon oldható meg a teljes, az egyedi, és az ismétléses keresés, ezen belül azt is meg kell állapítani, hogy melyek lehetnek az ABKR standard rekordtípus elhelyezési módjai,
— szegmenstípus alkalmas elhelyezési módját a hozzáférési igények modelljéből kiindulva meg kell határozni,
— adategyüttesenként külön-külön el kell dönteni a használandó pointer-típust,
— meg kell határozni az adategyüttesen belüli tagrekord-előfordulások logikai rendezettségének fajtáját,
— az adategyüttes tagsági viszonyok változásának kezelési módját.

A tárolási szerkezet meghatározásának szakasza az alábbi főbb tevékenységekre bontható:
— az adatbázis areakra történő felosztása,
— az areak leképezése a particiókba, ill. fileokba (ez nem minden ABKR-esetén kell),
— az areakat alkotó pagek méretének meghatározása,
— adatbázis helyigényének kiszámítása.

Meg kell jegyezni, hogy a részletes fizikai tervezést igen nagy mértékben befolyásolják:
— a logikai adatbázis-terv,
— az adatbázis integritásának biztosítása,
— az adott ABKR lehetőségei,
— a rendelkezésre álló tárolók milyensége,
— az alkalmazandó programnyelv *(host-language)* adatkezelési lehetőségei.

Éppen ezért a logikai adatbázis-tervre építve csupán a tárolási szerkezet és a hozzáférési mód tervezés kezdő tevékenységeihez tudunk jobban algoritmizált segítséget adni.

Az általunk javasolt és a 4. fejezetben leírt eljárás segíti a szegmenstípus elhelyezési módjának és az areaknak a megtervezését.

1.5.1. Elhelyezési mód meghatározása

Példaként tekintsük kiválasztott ABKR-nek az IDMS-t. Itt három standard elhelyezési mód van
(LOCATION MODE):
— CALC célja: a szegmens előfordulások egyenletes eloszlását (random) lehet biztosítani az arean, továbbá szimbolikus kulccsal közvetlen elérés biztosítható,
— VIA SET célja: lehetőleg minél közelebb lehessenek egymáshoz azok a szegmensek, amelyeket együtt kívánnak elérni,
közvetlen elérés nem lehetséges.
— DIRECT célja: adatbáziskulccsal közvetlen elérés biztosítható.

Egyedi keresés megoldható:
— Közvetlenül (DIRECT),
— Random (CALC).

Teljes keresés megoldható:
— egyedi közvetlen elérés ismétlésével,
— az area soros keresésével,
— szingular set létrehozásával majd elérésével.

Ismételt keresés megoldható:

— fölérendelt egyed előfordulásból az alárendelt egyedelőfordulás elérésével (adategyüttes létrehozása szükséges, a tagrekordot általában VIA, a tulajdonos rekordot CALC módon célszerű elhelyezni),

— egy egyedtípus másodlagos ismérv azonos értékével rendelkező előfordulásainak elérésével (indexeléssel az IDMS *Sequential Processing Facility* (SPF)-jével).

A tervezés kiindulópontja az, hogy adatmodellenként meghatározható egy szegmenstípusok alkotta precedenciasorrend, amelyek az adatmodell egészét tekintve egy teljes precedenciahálót alkotnak. A tervezőnek egyrészt a szegmenstípus e precedenciahálóban elfoglalt helyét és azt figyelembevéve, hogy a szegmenstípus elérése iránti igények között van-e közvetlen, elérési igény, másrészt a tulajdonos-tag lánc hosszát és a tagrekord elérése iránti igény gyakoriságát figyelembevéve lehet dönteni.

Megfogalmazható néhány általános irányelv az elhelyezési mód kiválasztására nézve:

— Csak azokat a rekordtípusokat szabad VIA SET módon elhelyezni, amelyek közvetlen elérésére soha sincsen szükség.

— Egy adatmodell hierarchiába rendezhető elérési precedencia-sorrendet is képvisel. A hierarchia legfelső szintjén nem lehet VIA, annál inkább szükséges CALC vagy DIRECT elhelyezés. A legalsó szinteken viszont pont fordított a helyzet.

— Közbenső szinteken csak akkor célszerű VIA alkalmazása, ha az alsó szinteken ez nem jelent majd nagy elérési időt.

De a fenti irányelvekből még általában nem válhat egyértelmű szabály, mert az egy szegmenstípusra irányuló hozzáférési igények sokszor egymást kizáró elhelyezési módokat követelnek.

Ha ez így van, akkor az eligazodás céljából először össze kell gyűjteni az egy szegmenstípusra irányuló hozzáférési igényeket.

És itt szoros kapcsolat van a következő pontban ismertetett area-leképzéssel. Ez utóbbinál ugyanis mintegy az előző inverz feladataként össze kell gyűjteni azokat a szegmenstípusokat, amelyek a rájuk irányuló hozzáférési igények szempontjából egymáshoz közelieknek tekinthetők.

1.5.2. Az adatbázis areakra történő felosztása

Az area-koncepció azon a gondolaton alapul, hogy általában nincs szükség a teljes adatbázisra, hanem mindig csak egy jól körülhatárolt részére, ahhoz, hogy adatait elérjék, feldolgozzák.

Többnyire annak az alapján rendelhetünk szegmenstípusokat egy areaba, hogy a szóban forgó szegmenstípusokat ugyanazon adatkezelési funkciók használják. Emellett vannak speciális az alkalmazott ABKR-től függő okok is.

Így például az IDMS SPF-je alkalmazása esetén az indexeknek és a rekordelőfordulásoknak külön areaban kell lenniük.

Alapvetően azonban abból kell kiindulnunk, hogy az egy areaba kerülő szegmenstípusokat annak alapján tudjuk összeválogatni, hogy az adatkezelési funkciók egy csoportja szemszögéből nézve homogén részhalmazt alkotnak.

Mielőtt ezt az összeválogatást elkezdenénk, végre kell hajtani a szegmenstípusok ún. véglegezését. Mint azt már 1.4.4. szakaszban láttuk, a szegmensképzés eredménye-

képp olyan szegmenstípusokat nyerünk, amelyeket az adatkezelési funkcióknak az egyes adattípusok iránti hozzáférési igényeit tekintve alakítottunk ki.

E szegmenstípusok azonban általában csak javasolt és még nem végleges klaszterszerkezetet alkotnak. Ennek az az oka, hogy létezik egy ettől független az adatok karbantartásából az adat helyességéért való felelősségéből, valamint egy másik, az adatok elérését korlátozó, adatbiztonsági szempontrendszer, amelyek helyzetfelmérés után közvetlenül megadják, hogy szempontjukból mely adattípusokat lenne célszerű vagy éppenséggel kötelező együtt kezelni. Esetleg további szempontrendszer(eke)t is figyelembe kell venni.

Az adott feladattól és környezettől függően más és más e szempontok egymáshoz viszonyított súlya. Minden felvett (lásd fent) szempontrendszer egy feltehetően eltérő klaszterszerkezetet (itt: szegmensfelosztást) tükröz. A rendszer tervezőjének lesz az a feladata, hogy egyedtípusonként áttekintve az — mondjuk három szempont szerinti — egymás melletti klaszterszerkezeteket, véglegezze a klaszterek, (azaz a szegmenstípusok) szerkezetét.

E véglegezéshez támaszt az ad, ha pl. egy képernyőn egymás mellett úgy jelenhet meg a (több szempont szerinti) többféle klaszterszerkezet, a hogy tervező könnyen áttekintse, s viszonylag gyorsan lehessen a véglegezésről dönteni.

Az egy csoportba vagy osztályba kerülő szegmenstípusok meghatározásának tehát kiindulópontja az, hogy az előbb említett véglegezés eredményeképp rögzítettük, hogy mely tulajdonságtípusok vannak egy szegmenstípusban.

Így az egy funkció szempontjából a tulajdonságtípusok egy részhalmazát jelentő adattal modell helyett a szegmenstípusok megfelelő részhalmazát is tekinthetjük adatalmodellnek, de megkülönböztető jelzővel látjuk el és *szegmensszintű adatalmodellnek* nevezzük.

A szegmensszintű adatalmodellek összessége egy táblázatba rendezhető.

A táblázat egyes oszlopai az adatmodellt kezelő funkciókat képviselik, sorai pedig a szegmensképzés eredményeként meghatározott szegmenstípusokat. A táblázat első közelítésben egy bináris mátrixként hozható létre (ezt $\mathbf{B}(T, R)$ mátrixnak nevezzük). A mátrix egy eleme $b_{tr} \in \mathbf{B}$ azt fejezi ki, hogy az F_r, funkció ellátásához szükség van-e, vagy sem az SZ_t szegmenstípusban levő valamelyik tulajdonságtípus valahány előfordulására.

$$b_{tr} = \begin{cases} 1, & \text{ha } F_r \text{ ellátásához } SZ_t \text{ szükséges} \\ 0, & \text{egyébként} \end{cases}$$

ahol $t=1, 2, \ldots, T$ és $r=1, 2, \ldots, R$ valamint fennáll $F_r \in \{F\}^R$ és $SZ_t \in \{SZ\}^T$. A táblázat r-edik oszlopának $b_{tr}=1$ elemei alkotják az F_r-hez tartozó szegmensszintű adatalmodellt. A **B** mátrix sorainak és oszlopainak permutálásával blokkdiagonálishoz közelálló alakú mátrixot képezhetünk, amelyen szemmel is kivehetők az egymástól többé-kevésbé jól elhatárolható blokkok, az ún. *szegmensosztályok*. Általában egy szegmensosztály fog majd egy areat alkotni.

Ezen túlmenően lehetőség lesz arra, hogy a tervező a szegmensosztályok közötti átfedéseket figyelembevéve hajtsa végre az areak fileokba való leképezését abban az esetben, ha az alkalmazott ABKR-ben ez szükséges.

A fileok tartalmának és szerkezetének meghatározását az ún. *mértékadó adatbázisrekordok* (lásd később a 4. fejezetben) kiválasztásával lehet elvégezni. A tulajdonságtípusoknak szegmenstípusokba, azoknak pedig areakba (és file-okba) csoporto-

sításánál figyelembe kell venni azt a célt, hogy az adatbázis egészét tekintetbe véve minimalizáljuk a szegmenselőfordulások elérési idejét.

Adott esetben meg lehet kísérelni egy minimalizálandó célfüggvény definiálását az F_r funkció által használt tulajdonságtípusok halmaza A_{F_r} és az SZ_t szegmenstípusba tartozó tulajdonságtípusok halmaza A_{SZ_t} között [42].

A hozzáférési modellre épülő optimalizálásnál azonban tudomásul kell vennünk azt a tényt, hogy amíg az S_{F_r} és az S_{gy} súlytényezők értékeit a felhasználó adatigényeiből kiindulva meg tudjuk határozni, addig az S_a és a később hivatkozott S_m súlytényezők konkrét értékeit (ezek arányszámok) nagymértékben meghatározzák az implementáció (a géptípus, a háttértárolók, az ABKR, a választott file-szerkezetek) részben már adott, részben pedig a fizikai tervezés későbbi szakaszában még eldöntendő körülményei. Éppen ezért a szegmensképzés és az areaképzés olyan, a fizikai tervezés későbbi menetéből visszacsatoló, többfordulós folyamat, amelyben mindig jelentős tere marad, a visszacsatolások miatt felmerülő korrekciók, s azok hatásai mérlegelésének.

2. Tudományos előzmények

2.1. Az osztályozás módszerei

Az adatmodellben szereplő tulajdonságtípusok szegmenstípusokba, majd a szegmenstípusoknak nagyobb egységekbe való nyalábolása osztályozási feladat. [48] szerint az osztályozás feladatai három csoportba oszthatók:

— adott osztályok és az azokba sorolt egyedek sajátosságaiból az egyes osztályok jellemzőinek meghatározása,
— adott egyedeknek előre megadott osztályokba (nomenklatúrába) történő besorolása (más néven prekordinált osztályozás),
— adott egyedek halmaza és jellemzőik ismeretében azokból alkalmas osztályok (gyakoriak a class, cluster, clump, group, clique elnevezések) meghatározása (másnéven posztkordinált vagy automatikus osztályozás).

A cikkben kizárólag ezen utóbbi problémakörrel, pontosabban az ún. automatikus osztályozásnak ezen belül is a cluster analízisnek (a továbbiakban: klaszterálás) az információs rendszerek logikai tervezésében való alkalmazásával foglalkozunk. A prekordinált osztályozással szemben, a posztkordinált osztályozásnál az osztályozás alapja egy csoportképző algoritmus és nincsen előre megadott nomenklatúra.

Az alkalmas csoportképző algoritmus rendszerint valamelyik többváltozós matematikai-statisztikai módszernek egyik eljárása. Rendszerint több különböző eljárást is ki szoktak próbálni.

A legismertebb módszerek:

A faktoranalízis, valamint a többszörös korreláció- és regresszióelemzés, továbbá a kanonikus korrelációszámítás, a többszempontú varianciaszámítás, illetve ennek és a többszempontú regreszszió- és korrelációelemzésnek a kombinációja a kovarianciaelemzés. Viszonylag új módszerek a többdimenziós mértékskálázás, valamint a log-lineáris elemzés. Osztályozásra, tipizálásra főleg klaszterálást és diszkriminancia-elemzést szoktak alkalmazni, az előbbit gyakran előzetes faktoranalízissel (főkomponensanalízis) kombinálva.

Az automatikus osztályozást végző matematikai-statisztikai módszerek általában egy A ún. adatmátrixból indulunk ki, amely úgy állítható elő, hogy adott G_j

egyedhez (kivett mintához) ahol $j=1, \ldots J,$) éppen I számú tulajdonság $\{H\}^I$, (ahol $i=1, \ldots, I$) tartozik és egy konkrét tulajdonság H_i (ahol $H_i \in \{H\}^I$) meglétét egy G_j (ahol $G_j \in \{G\}^J$) egyed esetén $a_{ij} \neq 0$ jelöli (ahol $a_{ij} \in \mathbf{A}$), illetve $a_{ij}=0$, ha a megfelelő (H_i) tulajdonság nem lép fel a G_j egyed esetében.

Az előbbi meghatározás a matematikai-statisztika szóhasználatával élt.

Ha most ugyanezt a meghatározást az adatmodellezés [24] szakkifejezéseit alkalmazva fogalmazzuk át, akkor azt mondhatjuk, hogy a tulajdonságtípusokat abból kiindulva akarjuk csoportosítani, hogy adott A_q tulajdonságtípushoz (és ez itt nem kivett minta, hanem egy meghatározott terjedelmű halmaz), — ahol $q=1, 2, \ldots, Q$ — éppen R számú funkció F_r (ahol $r=1, 2, \ldots, R$) tartozik és egy konkrét F_r (ahol $F_r \in \{F\}^R$) funkciónak egy adott A_q (ahol $A_q \in \{A\}^Q$) tulajdonságtípus iránt fellépő adatigényét $a_{qr} \neq 0$ jelöli (ahol $a_{qr} \in \mathbf{A}$), illetve $a_{qr}=0$, ha a megfelelő (F_r) funkció nem lép fel adatigénnyel az A_q tulajdonságtípussal szemben. Láthatjuk, hogy ott, ahol a matematikai — statisztika egyedet mond, az adatmodellezés esetünkben megfelelő kifejezése éppen a tulajdonságtípus elnevezés. Ahol pedig a matematikai-statisztika tulajdonságról beszél, ott az adatmodellezés a funkció elnevezést használja. A pontos értelmezés érdekében ezért lerögzítjük, hogy az utóbbi fogalomra vonatkozó két elnevezés közül a *funkció* elnevezést fogjuk használni.

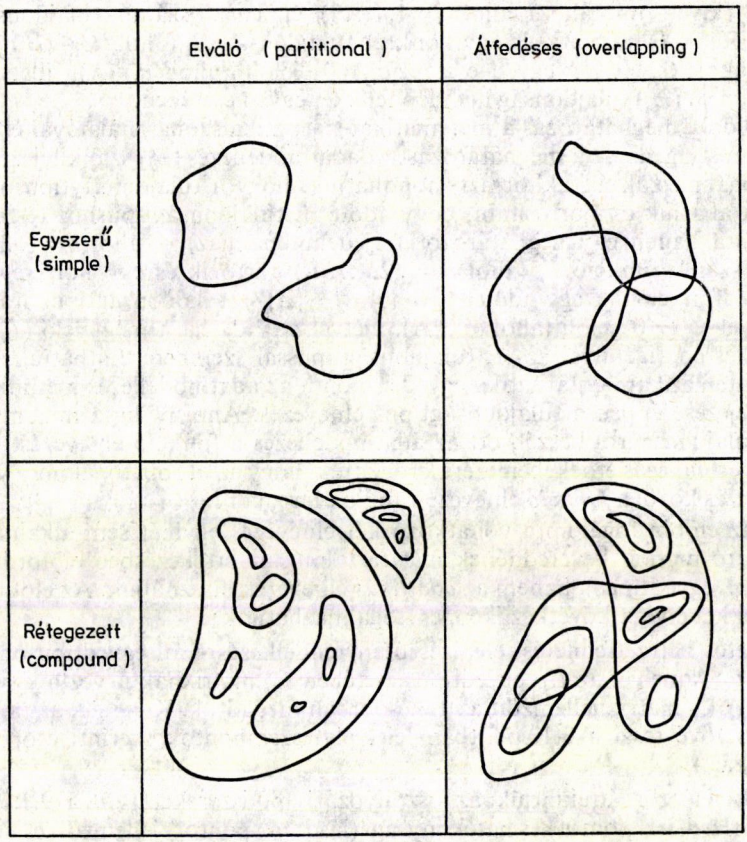
De az előbbi fogalomra vonatkozó két elnevezés egyikét sem alkalmazzuk itt, mert zavaró módon keverednének más tartalmú, az értekezésben előforduló fogalmakkal (pl. egyedtípus), hanem az *adattípus* elnevezést használjuk. Az előbb rögzített kifejezésekkel élve a következőképpen fogalmazhatunk:

Előzetes helyzetfelmérés, elemzés után rendelkezésre áll egyedtípusonként egy-egy táblázat, amelyet az E_j egyedtípus esetében $\mathbf{C}_j$ mátrixnak nevezünk (vö. az 1.3. ponttal). A $\mathbf{C}_j$ mátrix jellemzőit az 1.3. pontban megadtuk. Keressük azt a módszert, amely lehetővé teszi a 3.1. pontban leírt elvi szempontok szerinti csoportképzést (osztályozást).

Ehhez először áttekintjük az osztályozási módszereket (vö. a 2.1. fejezettel) majd elvi (lásd 3.1. pontot) és hatékonysági (lásd 3.2. pontot) jellemzőiket figyelembevéve ki fogjuk közülük választani a szegmenstípusképzéshez (lásd 3.1.1. és 3.1.2. pontban) és a nagyobb egységek képzéséhez (lásd a 4.1. és a 4.2. pontban) alkalmas módszert (illetve e módszerek egy szűk lehatárolt körét). A módszer megválasztásával összhangban kijelöljük az alkalmazható taxonomikus mértéket is (lásd a 3.1.3. és a 3.2.2. pontban).

Az előzőek szerint automatikus osztályozásra elsősorban faktor- és sajátérték elemző módszerek, diszkriminancia elemzés és klaszterálás jöhet szóba. Tekintettel arra, hogy az előbbiek $O(K^3)$ lépésszámú eljárásokra vezetnek, gyakorlati alkalmazásukra [56] (53. oldal) nagy adathalmazok esetén ritkán kerül sor, s általában helyettük klaszteráló módszereket alkalmaznak. Esetünkben az osztályzandó tulajdonságtípusok nagyságrendje $K \sim 10^3$, így sem faktoranalízis, sem sajátérték-elemző módszer használata nem volna hatékony [17] (124. oldal). Előtérbe kerül tehát esetünkben is a klaszteráló módszerrel létrehozott adatbázis struktúra. A megfelelő *klaszterálási módszer kiválasztásánál igen fontos az, hogy meghatározzuk, milyen is az adatmodell tulajdonságtípusai által alkotott klaszterek szerkezete.*

[29] (540. oldal) a klaszterekről megállapítja, hogy praktikusan megkülönböztetünk diszjunkt (elváló vagy partitional, vagy disjoint) és átfedéses (overlapping) klasztereket.



5. ábra. A klaszterek egymáshoz való viszonya, fajtái

De a klaszteréket is lehet klaszterálni, kisebb csoportokat nagyobbakká összefogni (pl. alacsonyabb hasonlósági küszöbértéket bevezetve). Ekkor rétegezett klaszterekről, egyébként pedig egyszerű (vagy egyszintű) klaszterekről beszélünk.

Hierarchikus a klaszterek szerkezete akkor, ha elváló és rétegezett.

A fenti csoportosítást az 5. ábra illusztrálja (átvéve [29]-ből).

[32] a klaszterek szerkezetét, illetve a klaszterálás alapvető lehetőségeit egzaktább módon az alábbiak szerint határozta meg:

$$C = \{C_1, C_2, \ldots, C_k\}$$

a klaszterek halmaza és C_i klaszter — ahol $(i=1, \ldots, k)$.

Egyszerű (simple) klaszterek:

$$\underset{C_i, C_j \in C}{\forall} [C_i \nsubseteq C_j) \wedge (C_j \nsubseteq C_i)], \quad \text{ahol} \quad i \neq j \quad \text{és} \quad j = 1, 2, \ldots, k.$$

Rétegezett (compound) klaszterek:

$$\exists_{C_i, C_j \in C} [(C_i \subseteq C_j) \vee (C_j \subseteq C_i)], \quad \text{ahol} \quad i \neq j.$$

Ha bármely két klaszterre fennáll, hogy
$C_i \cap C_j = \emptyset$, vagy az egyiket a másik teljes egészében tartalmazza, akkor a klaszterálás elváló, egyébként pedig átfedéses.
Elváló (partitional) klaszterálás:

$$\forall_{C_i, C_j \in C} [(C_i \cap C_j = \emptyset) \vee (C_i \subseteq C_j) \vee (C_j \subseteq C_i), \quad \text{ahol} \quad i \neq j.$$

Átfedéses (overlapping) klaszterálás:

$$\exists_{C_i, C_j \in C} [(C_i \cap C_j \neq \emptyset) \wedge (C_i \nsubseteq C_j) \wedge (C_j \nsubseteq C_i)], \quad \text{ahol} \quad i \neq j.$$

A klaszterálások egy sorozatában C^g ahol g egy ordinális — vagy intervallum —, illetve arányskálán mért paramétert (hasonlósági küszöbértéket) jelöl, a sorozat minden egyes lépésében nyert klasztert rétegezett klaszternek nevezzük. Ha a sorozat minden tagja elváló klaszter, akkor a sorozat egészét előállító eljárást hierarchikus klaszterálásnak, egyébként pedig átfedéses (átlapolt) rétegezett (itt egyszerűen csak nemhierarchikus) klaszterálásnak nevezzük. Az ordinális paraméter segítségével végzett hierarchikus klaszterálást taxonomikus osztályozásnak is nevezik.

2.2. Nem-hierarchikus klaszteráló módszerek jellemzése [56]

A nem-hierarchikus módszerek általában [56]/51. oldal) diszjunkt klaszterek meghatározására szolgálnak.
A klaszterek (szegmenstípusok) kívánt minimális száma általában előre is megadható.
Emellett egyes módszereknél megadható minden egyes szegmenstípusba kerülő adattípusok számát tekintve egy alsó, illetve felső korlát.
Amennyiben az egyes klaszterek egymást átfednék, ennek mértékét többnyire szabályozni lehet. Gyakran adott egy optimalizálandó célfüggvény. Szinte valamennyi osztályozó algoritmus iteratív. Az eljárások egy része kiszűri a kivételeket (outliereket).
Többnyire nincs szükség hasonlósági együtthatók meghatározására. Számos ilyen módszer azért gyors, mert csak egyszer kell elérni egy adattípust. A nemhierarchikus módszerek végrehajtásának lépésszáma $O(K \log K)$ — ahol K az adattípusok száma, — vagy ahhoz közelálló érték [17] (124. oldal).
Célszerű azonban figyelembe venni azt is, hogy a szegmensképzés eredménye általában nem független az algoritmus által kezelt adattípusok sorrendjétől és a kapott szegmensek karbantartása miatt esetleg igen gyakori szegmensátszervezésre kerülhet sor [56] (58. oldal).
Ez pedig rögtön módosítja e módszerek időigényére vonatkozó meggondolásainkat is. Ha ugyanis egy klaszterálást követő karbantartás úgy befolyásolhatja a klaszterek elhatárolását, hogy a karbantartás után újra kell képezni a klasztereket, akkor

a helyes időbecslés [56] szerint:

$$O[\sum_{i=1}^{U}(K_i \log K_i)],$$

ahol $i=1, \ldots, U$ a karbantartások számát, és $K_i - K_{i-1}$ az egy alkalommal felvett karbantartott adattípusok számát jelenti.

2.3. Hierarchikus klaszteráló módszerek jellemzése

2.3.1. Elvi alapok

A 2.1. pont végén bevezettük a taxonomikus osztályozás fogalmát. [45] megállapítja (94. oldalon), hogy egy halmaz elemei között értelmezett ekvivalencia (tehát egy reflexív, szimmetrikus és tranzitív) reláció egyértelműen meghatározza az adathalmaz taxonomikus felosztását.

Más szóval az ekvivalencia reláció gráfjának az a sajátossága, hogy összefüggő részgráfokra, egymástól jól elváló komponensekre bontható.

Tekintsük a C_j halmazt, amely itt az E_j egyedtípushoz tartozó adattípusok halmaza $C_j=\{A_1, A_2, \ldots, A_M\}$, majd a $C_j \times C_j = \{(A_1, A_1), (A_1, A_2), \ldots, (A_M, A_M)\}$ szorzathalmazt.

E szorzathalmazt egy célszerűen megválasztott képlet (az ún. taxonomikus mérték) segítségével a való számok egy részhalmazára egy alkalmas klaszteráló algoritmus képezi le. A taxonomikus mérték pedig a $C_j \times C_j$ szorzathalmaz minden egyes elempárja esetén a két elem közötti egyneműséget (homogénitást) méri. Az algoritmus e mértékét figyelembevéve tudja biztosítani, hogy a kapott szegmenstípusok (klaszterek) valóban homogén elemekből álljanak.

A $C_j \times C_j$ szorzathalmaz elemeire páronként vonatkozó taxonomikus mértékeket egy ún. **D** mátrixba rendezhetjük (vö. a 2. táblázattal), amely egy összetartozási relációt (távolságot) vagy hasonlóságot fejez ki a megfelelő C_j halmazban levő adattípusok között. Ha fennáll $d_{ij}=d_{ji}$, ahol $d_{ij} \in \mathbf{D}$ és akkor ez a reláció reflexív és szimmetrikus is, lehet séges azonban, hogy a tranzitivitás hiányzik, így az összetartozási reláció még esetleg nem ekvivalencia. Az ekvivalencia reláció fennállásának kimutatásában segíthet az, hogy e relációhoz rendelt mátrixnak az a sajátossága (lásd [45] 95. oldalon), hogy a sorok és oszlopok permutálása révén blokk-diagonális alakra hozható úgy, hogy a főátlóra illeszkedő blokkon kívüli mátrixelemek értéke nulla.

Ugyancsak segíthet a felismerésben az, hogy egy reláció akkor és csak akkor tranzitív, ha önmagával vett kompozíciója ([45] 91. oldal) nem bővíti a reláció terjedelmét.

Ezen utóbbi összefüggésre épül az ún. tranzitív lezárás módszere (lásd: [45] 155. oldal), amely azonban meglehetősen számításigényes, lassú eljárás. Kimondható az, hogy [45] (161. oldal) az egyszerűlánc módszer a tranzitív lezárás módszerének természetes általánosítása.

Segíthet az ekvivalencia reláció fennállásának felismerésében az is, hogy a relációhoz rendelt gráfnak — mint arra már fentebb utaltunk — az a sajátossága, hogy egymástól elváló részgráfokra bontható. [22], valamint [60] is kimutatta, hogy mind az az információ, amely egy sokdimenziós térben elhelyezkedő szögpontok csoportosításához az egyszerű — lánc módszer végrehajtásakor szükséges, e szögpontokat kifeszítő minimális fa (a továbbiakban röviden: MFF) előállításával nyerhető.

Adott egy összekötött összefüggő, irányítatlan gráf P.

Tekintsük külön szögpontjainak halmazát $V \neq \emptyset$ és éleinek halmazát $E \neq \emptyset$, ahol $E \subseteq V \times V$.

Definíció. Feszítőfa E azon részhalmaza, amelyre fennáll, hogy V bármely két eleme között csupán egyetlen út van. Tehát a feszítőfa alkotta részgráfban nincsen hurok.

Definíció. Minimális feszítőfa (MFF) akkor képezhető, ha feltesszük, hogy E minden eleméhez egy súlyt (távolság, vagy költség, stb.) lehet rendelni. P-nek egy minimális kifeszítőfája olyan feszítőfa, amelynél az élek súlyait összegezve a kapott összeg minimális.

Esetünkben a P szögpontjait a csoportosítandó adattípusok jelentik és a két szögpont közötti él súlya a két adattípus közötti távolság (vagy hasonlóság) értéke.

MFF előállításának idő- és tárigény szempontjából jelentősen eltérő módszerei vannak ([22] 59. oldal). Kimutatták, hogy az MFF előállítását az ún. *Prim-algoritmus* [46] a többi addig kidolgozott módszernél hatékonyabban végzi.

A Prím-algoritmus az alábbi:

Az M számú adattípus (X) alkotta halmaz $P=\{X_1, X_2, \ldots, X_M\}$ két részhalmazra (A és B) bontható. A-ba tartoznak az MFF-be már bevont, B-be pedig az MFF-be még be nem vont adattípusok.

1. lépés: Kezdetben $A \neq \emptyset$ és $B=P$.

2. lépés: Rendeljük A-hoz bármely $X \in B$-t.

3. lépés: Keressük meg A-ban és B-ben azt az adattípuspárt, amelynek a távolsága minimális (legközelebbi szomszédok).

4. lépés: A B-ben így kiválasztott adattípust B-ből elvéve A-hoz rendeljük.

5. lépés: Az algoritmust befejezzük, ha már $B \neq \emptyset$, ha viszont még $B \neq \emptyset$, akkor a 3. lépéstől újra lefolytatjuk.

A *Prím-algoritmus* azért volt más módszerekhez képest előnyösebb [21], mert az eljárás során [22] (59. oldal):

— a szögpontok közötti távolságokra csupán egyszer van szükség,

— agglomeratív módszerrel könnyen lehet hierarchikus klasztereket is előállítani.

Hátránya maradt viszont, hogy a 3. lépés végrehajtása sok összehasonlítást igényel és vagy arra van szükség, hogy igen gyorsan elérhető tárban tartsunk $M(M-1)/2$ előre kiszámított távolságértéket (amire gyakran nincs elegendő tárterület), vagy ezeket ismételten elő kell állítani, ami viszont az algoritmust lelassítja. A *Prím-algoritmus* hatékony számítógépes megvalósítását [12] és [59] adták meg.

Kimutatható [37] (194. oldal), hogy van ennél tár- és időigény tekintetében egyaránt hatékonyabb eljárás (az ún. E^* algoritmus). Kevesebb számítást és operatív tárat igénylő eljárás az, amelynél eleve kevesebb összehasonlításra van szükség, hiszen a 3. lépésben a legközelebbi szomszédok megtalálásához elegendő csupán a legutoljára A-hoz csatolt egyetlen adattípus és a B-ben levő többi adattípus közötti távolságok kiszámítása.

[4] kidolgoztak egy ennél is gyorsabb eljárást, amelyet [7] még továbbjavított.

BENTLEY és FRIEDMAN [4] definiálják az alábbi fogalmakat:

— különálló szögpont: a gráf azon pontja, amely nincs más ponttal összekötve,

— fragment: részfa a gráf egészén belül,

— szögpont és részfa távolsága: egy részfa és egy azon kívüli szögpont távolsága alatt páronként a szögpont és a részfát egyes alkotó szögpontok között vett távolságok közül a minimális távolságot értjük,
— szögpont legközelebbi szomszédja: az a másik szögpont, amelynek a tekintett első szögponttól való távolsága nem nagyobb mint más, egyéb szögpontokhoz való távolságai,
— részfa legközelebbi szomszédja: az a szögpont, amelynek a tekintett részfától való távolsága nem nagyobb, mint más, egyéb részfákhoz való távolságai.

MFF előállításának két fő lépése van:
— Bármely különálló szögpontot legközelebbi szomszédjához kötjük.
— Bármely részfát legközelebbi szomszédjához kötjük.

PRIM kimutatta, hogy M szögpont között $M-1$ összekötést elvégezve megkapjuk az MFF-t.

A fenti két lépést kombinálva különböző algoritmusok kreálhatók, de megvalósításukat tekintve ezeket alapvetően két csoportba oszthatjuk. Ugyanis a részfa legközelebbi szomszédjának megkereséséhez le kell tárolni vagy:

a) a részfa minden egyes szögpontjához a hozzá legközelebbi különálló szögpont azonosítóját és kettejük távolságát, vagy
b) minden egyes különálló szögponthoz a részfában levő legközelebbi szomszédjának azonosítóját és távolságát.

Az így letárolt kapcsolatok az MFF létrehozásához szóbajövő gráféleket adják meg. Ha egy részfához új szögpontot adnak (vagy régit elvesznek) e kapcsolatokat is karban kell tartani. Ezt figyelembevéve kimutatható, hogy (lásd [4]) általában hatékonyabb, ha a fenti b) változatot alkalmazzuk pontosan úgy, ahogy azt PRIM [46], DIJKSTRA [12] és YAO [59] teszik.

Vannak azonban esetek, ahol az a) változat jobb. Ilyen eset az, ha a részfa legközelebbi szomszédjának keresésekor — egy előfeldolgozás eredményeképp — már nem kell a keresést az egész ponthalmazra kiterjeszteni, hanem csak egy kisebb részhalmazon belül folytatjuk le azt. E kisebb részhalmazon belül is magának a legközelebbi szomszédkeresésnek az időigényét csökkenti az, ha részfánként olyan ún. prioritási-sort alkotunk, amelyben a sor elején találjuk majd a részfa legközelebbi szomszédját a részfában levő párjával együtt.

BENTLEY és FRIEDMAN algoritmusa a *Prim-algoritmus* egy olyan változata, amely bármely — még a háromszögegyenlőtlenségnek eleget nem tevő — taxonomikus távolságmértékkel is alkalmazható és kihasználja a koordinátatér geometriájának sajátosságait. Hatékonysága közel $O(M \log M)$, így lényegesen gyorsabb, mint az addigi $O(M^2)$ körüli eljárások, viszont a tárigény nagyjából 10—20%-kal (ez a dimenziók számától függ) meghaladja azokét.

CHEN és LEE változtatása arra épült, hogy a *Bentley—Friedman algoritmus* esetében előfordulhat, hogy másodszorra, vagy többször is kell legközelebbi szomszédot keresin olyan szögponthoz, amelyhez ezt egyszer már megtettük, s nyilvánvaló, hogy szükségtelenül állítunk elő egyszer már kiszámított, (de megnem őrzött) távolságértékeket. Ezért a tárolóterület rovására az eljárás feltétlenül tovább gyorsítható akkor, ha az egyszer már kiszámított távolságokat megőrizzük.

Erre természetesen nincs mindig lehetőség gépi korlátok miatt, BENTLEY és FRIEDMAN éppen ezért mást javasol. A probléma megoldását úgy kívánják elérni, hogy

„jó" kezdőpontból kezdjék el az MFF felépítését. Jó kezdőpont-véleményük szerint — ott lehet, ahol a szögpontok egymástól viszonylag legtávolabb állnak. Ennek kijelöléséhez viszont szükség van a teljes ponthalmazra vonatkozó előzetes, durva elemzésre, amellyel fel kell deríteni a sűrűn egymás mellett álló homogén pontcsoportokat, azaz homogén klasztereket.

Az előállított — erre mint láttuk az előbb, számos hatékony módszer van — MFF-ből elhagyva azokat az éleket, amelyek súlya (hossza) egy paraméterérték (hasonlósági küszöbérték) felett van, megkapjuk a gráf komponenseit, a paraméter értékének monoton lépcsőzetes változtatásával pedig végülis egyszerű lánccal összekötött hierarchikus elrendezésű klasztereket kapunk.

Meg kell jegyeznünk, hogy az MFF-nek egyszerű lánccal összekötött klaszterek hierarchiájába történő transzformálásakor információveszteség is fellép (lásd [22] 59. oldal.) E transzformációra mégis abból a praktikus okból kifolyólag van szükség, mert az egyszerű-lánc hierarchia adatstruktúrájának karbantartása lényegesen egyszerűbb, jobb hatásfokkal oldható meg [56] (60. oldal), mint az MFF-é. [46] kimondja, hogy a hasonlósági küszöb egy diszkrét értékének beállításakor az MFF ezen értékkel egyenlő vagy hosszabb éleit elvágva kapott részgráfok alkotják az eredeti ponthalmaz alkotta gráfból képezhető klasztereket és ezeken kívül az eredeti gráfból más klaszterek már nem képezhetők.

2.3.2. Hierarchikus klaszteráló módszerek csoportosítása

Összegezve [17] szerint a hierarchikus klaszteráló eljárás típusát tekintve lehet:
— felülről-lefelé építkező (egyszerű divizív),
— alulról-felfelé építkező (egyszerű agglomeratív),
— paraméterekkel vezérelt (s így lehet akár divizív, akár agglomeratív jellegű).

Az egyszerű divizív módszer fő lépései (adatbázistervezés esetére konkrétizált kifejezéseket használva):

1. Mind a K adattípus egyetlen kiinduló szegmenstípusba összefogott.
2. Valamennyi éppen létező (és legalább két adattípust tartalmazó) szegmenstípust kétfelé választunk (vágás).
3. A 2. lépést ismételjük, míg valamennyi adattípus külön szegmenstípusba kerül.

Az alkalmazott módszerek között főleg az optimális vágás megítélése (mérése) tekintetében van eltérés. Az összes lehetséges felbontás ($2^{K-1}-1$) áttekintése igen idő- és tárigényes folyamat.

Az egyszerű agglomeratív módszer fő lépései:

1. Mind a K adattípust önálló szegmenstípusnak (klaszternek) tekintjük.
2. Eggyé olvasztjuk a két leginkább hasonló szegmenstípust (klasztert).
3. A 2. lépést ismételjük, míg valamennyi adattípus egy szegmenstípusba kerül.

Az alkalmazott módszerek ([5]) főleg abban különböznek egymástól, hogy a klaszterek (szegmenstípusok) közötti hasonlóságot (különbséget) milyen függvénnyel adják meg.

A paraméterrel vezérelt módszer fő lépései:

1. szakasz: Esetleg nem diszjunkt, de homogén klasztereket előállító eljárás alkalmazása.

2. szakasz: Az 1. szakaszban előállított egyes homogén klasztereken belül (pl. a legközelebbi szomszéd keresésével) minimális feszítő fák előállítása.

3. szakasz: Dendrogram előállítása egy hasonlósági küszöbérték (cut-off vagy threshold paraméter) értékének monoton lépcsőzetes változtatásával (ha a paraméter hasonlósági mérték, — akkor monoton növelése divizív, csökkentése agglomeratív eljárás).

A módszer ilyen összetett szervezését az alábbiak indokolják:

A minimális feszített fák jól elváló (diszjunkt) klasztereket állítanak elő (3. szakasz). A minimális feszített fák létrehozásához szükség van a legközelebbi szomszéd típusú keresések valamelyikére (2. szakasz).

Mivel ilyen típusú keresések (nearest neighbour, illetve best matching) egyrészt igen időigényesek, másrészt tárigényük is nagy, ezért általában nem lehet a teljes sokaságra végrehajtani, de egy részsokaságra igen. E részsokaságokat egy előcsoportosítással hozzák létre (1. szakasz), választják el egymástól.

Egy másik szempontból, az eljárás jellegét tekintve [1] (132. oldal) ez lehet valamilyen:

— lánc módszer,
— centroid-módszer,
— hibanégyzet — vagy variancia-módszer.

A hierarchikus klaszterálást végző eljárásokat a szerzők egy része [56] (56. oldal) a nem-hierarchikus módszerekhez képest előnyösebben ítéli meg az alábbi okok miatt:

— A hierarchikus klaszterek esetén viszonylag hatékonyabb visszakereső eljárások léteznek.
— Ha az adatok közötti természetes szerkezet olyan, hogy egymást részben átlapoló klaszterek vannak, akkor a hierarchikus klaszterálás végrehajtása gyorsabb folyamat.
— A hierarchikus klaszter struktúra tárigénye lényegesen kisebb, mint az egy szintbe (átlapolva) rendezetté.
— Az ún. kivételeket (outlier) kiszűrik az agglomeratív módszerek [19].
— Az agglomeratív klaszteráló eljárások időbeli bonyolultsága $O(K^2)$ körüli (lásd [56] 58. oldal), de létezik néhány ennél valamelyest gyorsabb agglomeratív eljárás is.

A karbantartás következtében gyakori átszervezések időigényét is figyelembevéve a nemhierarchikus módszerek hatékonysága már korántsem múlja felül jelentős mértékben a hierarchikus eljárásokét, mert az ún. egyszerű-lánc módszerekkel képzett klaszterekre fennáll az, hogy:

— a klaszterálás eredménye, független az algoritmus által kezelt adattípusok sorrendjétől,
— viszonylag érzéketlenek a karbantartás során ért változásokra, [32], [56] (56. oldal).

Ugyanakkor más szerzők (lásd [35]) felvetik a hierarchikus klaszteráló eljárások fő hátrányait:

— Az eljárások nem feltétlenül biztosítanak optimális megoldást [1] (190. oldal), mert egy korai lépésben végrehajtott összevonás (vagy vágás) véglegesen rögzíti az adattípus klaszterhez tartozását és ez később sem változtatható.
— Az egyszerű lánc módszerek ún. lánchatást [1] is tartalmaznak.
— A divizív módszerek nem szűrik ki az outliereket [19].

— A számítási eljárás vagy a hasonlósági vagy a távolsági mátrix elemeinek kiszámítását is igényli (így esetünkben a szegmensképzési algoritmus idő-, illetve tárigénye nagyobb) [17] (116. oldal).
— A klaszterek kívánt számát nem lehet előre megadni.
— A divizív módszerek idő- és tárigénye gráf-modell alkalmazása esetén nagy.

Hipergráf-modell esetén a vágások hatékonysága lényegesen jobb lesz [18], továbbá nem kell taxonomikus mértéket kiszámítani, így az algoritmus időbeli bonyolultsága $O(K)^4(\varepsilon)^2$, ahol K a hipergráf szögpontok száma, ε a hipergráfélek száma.

3. Szegmensképzés módszere

3.1. Elvi meggondolások

3.1.1. A szegmensképzéshez alkalmas klaszterálási módszer kiválasztása

A leképezés kiindulópontja az elvi adatmodell, azaz az információs rendszer típusain (egyed-, tulajdonság- és kapcsolattípusok) értelmezett struktúra, amelyben a típusokat és a relációkat tartalmi, jelentéstani viszonyokból eredeztetjük.

Ez a struktúra szintetikus adatmodellezés esetén úgy jön létre, hogy a modellt első közelítésben véges számú, egyébként véges értékkészlettel rendelkező tulajdonságtípusok halmazának fogjuk fel, azaz ún. adatelemkatalógust állítunk elő. A következő lépésben a tulajdonságtípus halmazból (amelyet a rendszer referenciahalmazának tekinthetünk) kijelöljük az egyedtípusok részhalmazát.

Analitikus adatmodellkészítésnél e két lépés sorrendje — durván szólva — felcserélődik.

Ki kell emelni, hogy adott rendszerben minden egyedtípus meghatározott tulajdonságtípusok halmazával jellemezhető. Egy ilyen halmaz homogén klasztert alkot. Ugyanakkor létezik az ún. gyenge logikai redundancia jelensége [24] (40. oldal), ami azt jelenti, hogy egy egyedtípust jellemző tulajdonságtípus(ok) egyben más egyedtípus(okat) azonosítanak. Ezeket a tulajdonságtípusokat kapcsoló tulajdonságtípusoknak nevezzük és létezésük a tulajdonságtípusok halmazai alkotta homogén klaszterek átlapolódását jelenti.

A már előállított elvi adatmodell esetében ismert tehát a kiinduló részhalmazok száma, azok tartalma (a tulajdonságtípusok) külön-külön, és az átfedést jelentő halmazelemek egy része.

Átfedés jelentkezik ugyanis az ún. erős logikai redundancia esetében is, ez azonban az elvi adatmodell szintjén nem jelent kapcsolatot. Addig, tehát amíg az átfedések (redundancia) egyes eseteit (a kapcsolattípusokat képező tulajdonságtípusokat) külön is számontartjuk, addig más eseteit (erős logikai redundancia) a tervezés során kapcsolatként nem kell figyelembe venni.

Ez az egyik oka annak, amiért nem áll módunkban a nem-hierarchikus klaszteráló eljárások alkalmazása. Ezek az eljárások amelyek az átfedések kezelését is lehetővé tennék, ugyanis nem tudnak különbséget tenni erős és gyenge logikai redundancia között, mindkettőt egyképpen átfedésnek fogják fel. Márpedig a tervezés kapcsolatként csak a gyenge logikai redundanciát tartja számon.

De van másik — valamelyest gyengébb — kizáró ok is. Ez az a követelményünk, hogy a szegmensképző algoritmust eredménye legyen független az algoritmus által

kezelt tulajdonságtípusok sorrendjétől. Ez ugyanis kirekeszti a nem-hierarchikus klaszteráló módszerek többségét és néhány hierarchikus módszert is. Ha a nem- hierarchikus módszerek nem alkalmazhatók, maradnak a hierarchikus eljárások. Hálós adatmodell esetében azonban az egyedtípusok közötti átfedések miatt ezek nem alkalmazhatók az adatmodell referenciahalmazának egészére, csupán az egyes egyedtípusokra nézve külön-külön, egymás után.

A hierarchikus módszerek alkalmazása ilyen formában lehetséges is, mert az egy egyedtípushoz tartozó szegmenstípusoknak — az egyedtípus azonosítójának kiemelése után — diszjunkt részhalmazokat kell alkotniuk, és szükséges is mert egy szegmenstípus definíció szerint hierarchikus belső szerkezetű.

[32] arra mutat rá, hogy a hierarchikus klaszterálás alkalmazása akkor előnyös, ha nem homogén klaszterek elkülönítése a célunk, hanem a klaszterek optimális összekapcsolását keressük. De jelen esetben éppen ez az egyik fő cél, hiszen a klasztereknek (szegmenstípusok) a későbbi visszakeresések szempontjából optimális összekapcsolását keressük.

Azzal, hogy a hierarchikus klaszterálást nem alkalmazzuk egyszerre az egész referenciahalmazra, hanem sorban annak homogén részhalmazaira, egyben le is győzzük azt a nehézséget, hogy e módszercsaládra $O(m^2)$ körüli értékek jellemzők (ahol $m \sim 10^3 - 10^4$ a referenciahalmaz elemszáma) és tárigénye is jelentős. A fenti $O(m^2)$ érték az agglomeratív módszereket jellemzi, a divizív módszerek lépésszáma általában lényegesen nagyobb. Épp ezért az utóbbiakat nem is vesszük most számításba.

A létrehozott szegmenstípusok érzéketlensége a környezetből jövő változásokra ugyancsak fontos követelmény és befolyásolja, hogy mely klaszteráló módszereket lenne jó itt alkalmazni. Jelen esetben ugyanis az adatbázis túl gyakori átszervezését célszerű elkerülni. Lehetőleg olyan szegmenstípusokat hozzunk létre, amelyeket túl gyakran átrendezni még akkor sem kell, ha változik az ügytípusok, adatkezelési funkciók vagy a bizonylatok köre, adattartalma, illetve távolabbról a külső vagy a szervezeti belső szabályozás. De ezen túlmenően is célszerű olyan klaszteráló módszert választani, hogy az egyes adatkezelési funkciók adatigényének nem teljesen pontos meghatározása, majd ezt követő pontosítása a szegmenstípusok átrendezése iránti igényt majd lehetőleg csak kismértékben befolyásolja. Ez az előzőhöz hasonló, de más eredetű hatásra vonatkozó zavarérzékenységi-stabilitási követelmény.

[32] kimutatta, hogy nem csupán a hierarchikus agglomeratív eljárások, hanem valamennyi klaszteráló eljárás közül az egyszerű-lánc módszerek tesznek eleget a fenti, valamint az előbb már említett sorrend-függetlenségi követelményeknek is.

De lánc-módszer alkalmazásakor az általában kedvezőtlen láncha tás is fellép. A mi esetünkben azonban ebből hátrány nem következik, mert a lánc-módszert csupán a referenciahalmaz egy homogén részhalmazára (egy egyedtípus tulajdonságtípusaira) alkalmazzuk egyidőben.

Esetünkben továbbá ahelyett, hogy önkényesen vagy tapasztalati alapon a klaszteráláshoz hasonlósági küszöbértéket vennénk fel, célszerűbb volna olyan eljárást alkalmazni, amelynél erre nincs szükség. Annál is inkább ez a helyzet, mert az osztályozást nem az areakra, s azon belül a szegmenstípusokra, adattípusokra hierarchikusan, egyidejűleg végezzük, hanem előbb (lásd: a 3.1.2. pontban) csak adattípusokra, külön menetben (lásd: a 4.1. pontban) a szegmenstípusokra. Az általunk választott algoritmus nem is igényli hasonlósági küszöbértékparaméter meghatározását.

A 3.1.2. pontban röviden ismertetett algoritmus az ún. n. *Wrocławi taxonómia,* [16] mely hasonlóságot (távolságot) figyelembe véve először meghatározza mindegyik

adattípus legközelebbi szomszédját (1. fázis). Így olyan két elemből álló kezdő klasztereket (hipergráf élek) képez, amelyek „esetleg összefűzhető láncszemeknek" tekinthetők. Egyszerű lánccá akkor „fűzi" őket, azaz akkor olvasztja egy klaszterbe (hipergráf komponensbe) ha van közös elemük (3. fázis).

Mivel az algoritmus az összefűzésnél már nem a kezdő klaszterekben levő adattípusok hasonlósági mértéke, hanem egy közös adattípus megléte, vagy hiánya alapján dönt afelől, hogy két kezdő klasztert eggyé olvasszon-e vagy sem, hasonlósági küszöbérték felvételére egyáltalán nincs is szükség. Az eljárás az egyszerű lánc módszer egy speciális esetének is tekinthető, ahol először egy legközelebbi szomszédsági relációt értelmezünk, azaz minden adattípushoz hozzárendeljük a hasonlósági (**D**) mátrix szerint hozzá legközelebb álló másik adattípust, majd ezzel a relációval hajtunk végre műveleteket, amíg végülis ekvivalencia relációt kapunk. Annak bizonyítását, hogy eközben MFF előállítása történik meg, [16] adja meg.

Természetesen ettől az algoritmustól eltérő, más megoldások is alkalmasak a feladat megoldására, de ott hasonlósági küszöbértéket fel kell venni. Ezen az áron alkalmazhatnánk a legtöbb hierarchikus agglomeratív módszert, így pl. az egyszerűlánc módszer különböző változatait is.

3.1.2. Szegmensképző algoritmus ismertetése

A javasolt algoritmus bináris változók esetén történő alkalmazására az 1—4. táblázatok egy számpéldát is végigkövetnek.

Az algoritmus eg már felállított **D** mátrixból indul ki (vö. az 1. táblázattal). A 3.2. pontban kitérünk a **D** mátrixok **C** mátrixokból történő létrehozásának folyamatára is.

1. TÁBLÁZAT

A **D** mátrix (felvett adatokkal)

$1 \xrightarrow{j} M$

Elemi adat / Elemi adat	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8
A_1		0,513	0,221	0,221	0,914	0,971	0,193	0,474
A_2	0,513		0,107	0,000	0,391	0,495	0,111	0,237
A_3	0,221	0,107		0,981	0,391	0,333	0,666	0,513
A_4	0,221	0,000	0,981		0,333	0,221	0,666	0,510
A_5	0,914	0,391	0,391	0,333		0,970	0,301	0,555
A_6	0,971	0,495	0,333	0,221	0,970		0,221	0,474
A_7	0,193	0,111	0,666	0,666	0,301	0,221		0,693
A_8	0,474	0,237	0,513	0,510	0,555	0,474	0,693	

(1 ↓ i ↓ M)

A **C** mátrixok létrehozásának folyamatára viszont az 1.3.3. pont tartalmazott útmutatást. A szegmensképző algoritmus az alábbi négy fő fázisból áll:

1. Fázis: Ennek a fázisnak az a célja, hogy kijelöljük azokat az adattípus-párokat, amelyeket majd a 3. fázis elindításához szükséges kezdő klaszereknek tekintünk. Más szavakkal, létrehozzuk azokat a két-két elemet tartalmazó hipergráf éleket, amelyeknek később az unióját kívánjuk képezni (ahol ez egyáltalán lehetséges). Ezt a célt úgy érjük el, hogy a **D** mátrix minden egyes oszlopában kiválasztunk egy-egy maximális értékű elemet (max. d_{ij}=konstans, ahol $i=1, \ldots I$) első lépésként.

Természetesen a **D** szimmetriája miatt akár soronként is haladhatnánk, de az értekezésben nem ezt a konvenciót követjük.

Az oszlop maximális elemének kiválasztásakor esetleg több azonos maximális értéket is találhatunk, ezért soronként képezzük a sor elemeinek összegét és a keresett oszlop maximális elemének majd azt tekintjük, amelyikhez tartozó sorösszeg a legnagyobb. Ennek a választásnak az az oka, hogy az oszlop (indexe) által jelölt adattípussal való szegmensképzésbe a szóba jövő néhány adattípus közül először azt kívánjuk bevonni, amelynek a többi adattípusokhoz való kapcsolata, hasonlósága összességében viszonylag kisebb. Ha az oszlop több azonos maximális értékű eleme esetében még a hozzájuk tartozó sorösszegek is egyenlő értékek lennének, akkor közülük egyet a többi adattípusokhoz való hasonlósági együtthatóinak megoszlását figyelembevéve (vagy akár önkényesen) választunk ki.

A számpélda 1. fázisa végének állapotát mutatja be a 2. táblázat.

2. Fázis: Ennek a fázisnak az a célja, hogy kijelöljük azt a kezdő klasztert (hipergráf élt), másszóval a szegmens két elemből álló „magját", amely a 3. fázisban sorrakerülő folyamat jó kiindulópontja lesz.

Ezért meg akarjuk találni a **D** mátrixban azt a (hipergráf élt alkotó) két adattípust, amelyek egymáshoz a mátrixon belül a legnagyobb mértékben hasonlóak. Ezt úgy

2. TÁBLÁZAT

A D mátrix valamennyi oszlopában: oszloponkénti maximális elem(ek) kijelölése (és az oszloponkénti abszolút maximális elem kiválasztása)

Elemi adat / Elemi adat	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	$\sum_{i=j}^{M} d_{ij}$
A_1		0,513	0,221	0,221	0,914	0,971	0,193	0,474	3,507
A_2	0,513		0,107	0,000	0,391	0,495	0,111	0,237	1,854
A_3	0,221	0,107		0,981	0,391	0,333	0,666	0,513	3,212
A_4	0,221	0,000	0,981		0,333	0,221	0,666	0,510	2,932
A_5	0,914	0,391	0,391	0,333		0,970	0,301	0,555	3,855
A_6	0,971	0,495	0,333	0,221	0,970		0,221	0,474	3,685
A_7	0,193	0,111	0,666	0,666	0,301	0,221		0,693	2,851
A_8	0,474	0,237	0,513	0,510	0,555	0,474	0,693		3,456

érjük el, hogy az 1. fázisban az oszloponként kiválasztott maximális értékű elemek közül a mátrix abszolút maximális elemét (azaz a hipergráf élek közül a legnagyobb súlyút) választjuk ki. Mivel a **D** mátrix szimmetriája miatt legalább két ilyen azonos értéket találunk, ezért közülük ugyanúgy a megfelelő sorösszegek alapján választjuk ki a keresett mátrixelemet, mint azt az 1. fázisnál fentebb már leírtunk. E mátrixelem sor- és oszlopindexe közvetlenül megmutatja a szegmens magját képező két adattípust. A számpélda a 2. fázis végének állapotát a 3. táblázaton mutatja be.

3. TÁBLÁZAT

A **D** mátrix abszolút maximális elemének kijelölése: d_{34} a sorösszegértékek összehasonlításával 3,212 > 2,932 történt

Elemi adat / Elemi adat	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	$\sum_{i=j}^{M} d_{ij}$
A_1		0,513	0,221	0,221	0,914	0,971	0,193	0,474	3,507
A_2	0,513		0,107	0,000	0,391	0,495	0,111	0,237	1,854
A_3	0,221	0,107		0,981	0,391	0,333	0,666	0,513	3,212
A_4	0,221	0,000	0,981		0,333	0,221	0,666	0,510	2,932
A_5	0,914	0,391	0,391	0,333		0,970	0,301	0,555	3,855
A_6	0,971	0,495	0,333	0,221	0,970		0,221	0,474	3,685
A_7	0,193	0,111	0,666	0,666	0,301	0,221		0,693	2,851
A_8	0,474	0,237	0,513	0,510	0,555	0,474	0,693		3,456

3. Fázis: Ennek a fázisnak az a célja, hogy az előbb (a 2. fázisban) kijelölt klasztertől (hipergráf éltől) kiindulva sorra eggyé olvasszuk azokat a kezdő klasztereket (vö. az 1. fázissal), amelyekben van közös rész. Ennek végrehajtása érdekében ezért a szegmens magjának kijelölése után meg kell vizsgálni, hogy vannak-e még **D**-ben további adattípusok, amelyeket e mag köré csoportosítunk.

Eddig (vö. 1. fázis) azt vizsgáltuk, hogy az első adattípus mely másikkal van a legszorosabb kapcsolatban (hasonlóság), ezután azt vesszük szemügyre, hogy e második mely továbbiakhoz kapcsolódik a legszorosabban. Ehhez a második adattípusnak megfelelő sorban keresünk az első adattípuson kívüli, az előzőek során (vö. 1. fázis) az oszlopokban már megjelölt maximális elem(ek)et. Ha ilyent nem találunk, akkor a 3. fázist azaz a szegmens kialakítását befejezzük és a 4. fázisra ugrunk.

Ha viszont ilyen(ek)et találunk, akkor az ezek oszlopindexe által jelölt adattípusokat is ebbe a szegmensbe tartozónak tekintjük és oszlopukat a **D** mátrixból töröljük. Az így kiválasztott adattípusokra vonatkozóan ugyancsak sorra meg kell vizsgálni, — a második adattípus kapcsolódásaira nézve előbb leírt eljárás megismétlésével — hogy mely továbbiakhoz kötődnek a legszorosabban, és ha a megfelelő sorban találunk még megjelölt (maximális) oszlopeleme(ke)t, akkor az ez(ek)nek megfelelő adattípus(ok)at is felvesszük a szegmensbe.

Felvételüket az indokolja, hogy az így kijelölt adattípus(ok mutatnak) legnagyobbfokú hasonlóságot a szegmensbe legutóbb felvett adattípussal, míg a szegmensen kívüli adattípusokhoz való hasonlóság(uk) viszonylag kismértékű. Az eljárást ciklikusan folytatjuk mindaddig, amíg olyan adattípushoz (sorhoz) jutunk, amelyben (az 1. fázisban) oszloponkénti maximális elem kijelölése alkalmával megjelölt újabb elemet már találunk.

A számpélda a 3. fázis második szegmenstípusa képzése végének állapotát a 4. táblázaton mutatja be.

4, TÁBLÁZAT

Az első szegmenstípus SZ_1: $A_3 - A_4$ képzése után a megmaradt $\mathbf{D}^{1,2,3,4,5,6,7,8}$ minormátrixból a második szegmenstípus elemeinek SZ_2: $A_2 - A_1 - A_6 - A_5$ összegyűjtése

Elemi adat / Elemi adat	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	
A_1		0,513 (7. lépés)		0,914	0,971	0,193	0,474	3,507	
A_2	0,513			0,391	0,495	0,111	0,237		
A_3									
A_4									
A_5	0,914	0,391			0,970	0,301	0,555	(5. lépés)	
A_6	0,971	0,495 (3. lépés)		0,970		0,221	0,474	3,685	(1. lépés)
A_7	0,193	0,111		0,301	0,221		0,693		
A_8	0,474	0,237		0,555	0,474	0,693			
	(2. lépés)	(8. lépés)		(4. lépés)	(6. lépés)				

4. Fázis: Miután így kijelöltük a **D** mátrix, (illetve az ahhoz tartozó **C** mátrix) által reprezentált egyedtípus egyik szegmenstípusának tulajdonságtípusait, az ezeknek megfelelő adattípusok sorait és oszlopait a **D** mátrixból elhagyva, annak redukált mátrixát képezzük. Amennyiben e redukált mátrix mérete legalább (3×3), arra nézve újabb szegmenstípus képzéséhez a fent leírt eljárást kell ismét végezni, a 2. fázistól eltekintve. Ha viszont a redukált mátrix mérete kisebb, mint (3×3), akkor arra az egyedtípusra nézve, amelyre a kiinduló **D** mátrix is vonatkozik, a szegmenskéző algoritmust már nem működtetjük tovább, hanem áttérünk a következő egyedtípushoz (ha van még ilyen) tartozó **D** mátrix feldolgozására az 1. fázistól kezdve.

3.1.3. Hasonlósági együttható választása

A kiválasztás szempontjai:

— Milyen skálán mérhetőek az adattípust jellemző ismérvek értékei.
— Szimmetriatulajdonsággal rendelkező mértékre van-e szükség vagy sem.
— A klaszterálási algoritmus végrehajtása szempontjából mely mérték választása jelent nagyobb hatékonyságot.

Tekintsük most a klaszterálandó **C** mátrixban rendezett kiinduló adatainkat.

Ha $c_{mn} \in \mathbf{C}$ csakis F_n-nek A_m iránti becsült relatív elérési gyakoriságát fejezi ki, akkor esetünkben valamennyi ismérv értékeit egyaránt intervallumskálán mérhetjük.

Ezt figyelembe véve esetleg alkalmazható lenne valamely metrikus (pl. az euklideszi) távolságmérték. Az előbbi feltétel azonban esetleg nem mindig fog teljesülni. Emellett tekintetbe kell venni azt a már erősebb megkötést is, hogy két adattípus (A_i és A_j) számunkra most fontos jellemzői — azaz hogy együtt ($P11_{ij}$), illetve egymástól külön-külön ($P01_{ij}$ és $P10_{ij}$) hány funkció igényli, illetve együtt hány funkció egyáltalán nem ($P00_{ij}$) igényli — szempontjából itt az a döntő, hogy mikor jelenik meg a két adattípus egy funkcióban együtt ($P11_{ij}$) és sokkal kevésbé érdekes az, hogy mikor nem. A fentiek figyelembevételével csak olyan taxonomikus mértéket alkalmazhatunk, amely $P11_{ij}$ és $P00_{ij}$ súlyozása tekintetében nem szimmetrikus, és nem csupán intervallum, hanem pl. ordinális változók esetére is (lásd: 2.4.3.) alkalmas mérték. A számunkra célszerű hasonlósági együttható olyan legyen, amely kifejezi annak feltételes valószínűségét, hogy két adattípust egy véletlenszerű kiválasztott funkció egyaránt igényel ($P11_{ij}$), míg az „együtt nem igényli" eseteket ($P00_{ij}$) figyelmen kívül hagyjuk. Ilyen normalizált asszimetrikus hasonlósági együttható pl. az ún. *Jaccard együttható:*

$$d_{ij} = \frac{P11_{ij}}{P11_{ij} + P01_{ij} + P10_{ij}}.$$

Az asszimetriára vonatkozott előző utalásunk (hangsúlyozzuk, hogy ez csupán $P00_{ij}$ figyelembe nem vételére vonatkozik) mellett ez azt jelenti, hogy d_{ij} a $P11_{ij}$ monoton növekvő függvénye és fennáll, hogy itt $d_{ij} = d_{ji}$, továbbá $0 \leqq d \leqq 1$. [1] (90. oldal) kifejti, hogy akár egyszerű, akár teljes lánc (*single-linkage* és *complete-linkage*) módszer végrehajtása esetében a *Jaccard* és a *Dice-féle együttható*

$$d_{ij} = \frac{2P11_{ij}}{2P11_{ij} + P01_{ij} + P10_{ij}}$$

alkalmazása teljesen azonos hatású.

A *Czekanowski* vagy más néven *Dice-féle együtthatót* úgy tekintjük, mint a *Jaccard együttható* kiterjesztését.

[22], valamint [60] ugyanakkor kimutatták az egyszerű-lánc módszer és a minimális feszítőfa előállítása közötti szerves összefüggést. Tekintettel arra, hogy a 3.1.2. pontban leírt algoritmus ugyancsak minimális feszítőfát állít elő, a mi esetünkben a *Jaccard* és a *Dice együttható* egyaránt alkalmazható lenne, így a két fenti hasonlósági együttható közül azt fogjuk használni, amelynek kiszámítása az egész feldolgozási folyamat hatékonyságát jobban növeli.

Ennek eldöntéséhez azonban kissé részletesebben elemezni kell a 3.1.2. pontban már tömören ismertetett algoritmus tervezett végrehajtási folyamatának néhány részletét.

3.2. A hatékonyság és a gyakorlati megvalósítás szempontjai

Kezdetben a **C** mátrixból kiindulva d_{ij} értékeket határozunk meg, amelyek majd egy **D** mátrixot alkotnak. Ez a **D** mátrix főátlóra szimmetrikus ($d_{ij} = d_{ji}$), a főátló elemeinek ($i = j$) értéket ($d_{ij} = 1$ lenne) nem adtunk, míg a többi mátrixelem ($i \neq j$) értéke $0 \leqq d_{ij} \leqq 1$.

Az algoritmus legidőigényesebb része éppen a **D** mátrix előállítása, azaz a $d_{ij} \in \mathbf{D}$ hasonlósági együtthatók kiszámítása a $c_{mn} \in \mathbf{C}$ értékekből, mert ehhez például M számú sor között kell $M(M-1)/2$ számú — ha **D** szimmetrikus (itt ez a helyzet), akkor csak $M(M-1)/4$-páronként sorelemről — sorelemre haladó összehasonlítást, majd az előállított $P11$, $P01$, $P10$ hasonlósági paraméterekkel aritmetikai műveleteket végeznünk.

A **D** mátrix előállítására azonban kidolgoztak [6], [58] már gyorsabb módszereket is.

Ezek helyett egy még hatékonyabb algoritmust alkalmazunk, amellyel **D** egy sorát egyszerre előállíthatjuk.

5. TÁBLÁZAT

A **C** mátrix sorösszegeinek képzése:
$\{P_m\}$ létrehozása a *Dice-együtthatók* nevezőjében szereplő érték két tagját

$1 \longrightarrow n \longrightarrow N$

		F_1	F_2	F_3	F_4	F_5	F_6	F_7	F_8	F_9	F_{10}	F_{11}	F_{12}	F_{13}	F_{14}	F_{15}	P
1	A_1	1	0	0	0	1	1	1	1	1	0	1	1	1	0	0	9
	A_2	0	1	1	0	0	1	0	0	0	1	1	1	0	1	0	7
m	A_3	1	1	0	1	1	0	1	1	0	1	1	0	0	0	1	9
	A_4	1	1	0	1	0	1	0	1	0	1	0	1	1	1	1	10
M	A_5	0	1	1	1	1	0	0	1	1	0	1	1	0	1	1	10
	A_6	0	0	0	0	1	1	0	0	1	1	0	0	0	1	1	6
	A_7	1	0	1	1	0	1	0	1	0	0	0	0	1	0	0	6
	A_8	1	0	0	0	1	0	0	0	1	1	1	1	0	0	1	7

A gyorsításra az ad módot, hogy ahol $d_{ij}=0$, ott fennáll ugyancsak $P11_{ij}=0$ is, így elegendő csupán a $P11_{ij}$ értékeket megállapítani. Habár valamennyi (M) adattípushoz a hasonlósági együtthatók $d_{ij}=0$, vagy $d_{ij}\neq 0$ értékének megállapításához ugyan változatlanul $M(M-1)/2$ — szimmetrikus **D** mátrix esetén csupán ennek a fele mennyiségű összehasonlítást kell elvégezni, de az mégis gyorsabb, mint a hasonlósági együtthatók értékének teljes meghatározása. Ez utóbbi kiszámítására csak akkor kerül sor, ha $P11_{ij}\neq 0$. Az időmegtakarítás annál nagyobb, minél többször fordul elő $d_{ij}=0$ érték.

Az általunk jelenleg implementált algoritmus öt lépésből áll.

Az algoritmus használható mind hasonlósági együttható, mind távolságmérték alkalmazása esetében. Az előbbi esetet a 3.2.1. pont, míg az utóbbi esetet a 3.2.2. pont írja le.

3.2.1. Hasonlósági együttható alkalmazása

1. lépés: Az 1.3. pontban leírt $\mathbf{C}(M \times N)$ mátrixból kiindulva képezzük először az M-elemű **P** vektort, amelynek egy elemét (p_m) úgy képezzük, hogy összeadjuk a **C** mátrix egy sorában álló elemeket:

$$p_m = \sum_{n=1}^{N} c_{mn}, \quad \text{ahol} \quad m = 1, 2, \ldots, M.$$

A **P** vektor előállítása a **C** mátrix ellenőrzését szolgálja. A $p_m = 0$ eset ugyanis azt jelenti, hogy a **C** mátrixba bevont A_m adattípus ott tulajdonképpen felesleges, hiszen az E_j egyedtípusnak egy olyan tulajdonságtípusáról van szó, amelyet az adatmodellre irányuló egyetlen adatkezelési funkció sem igényel. A **P** vektor bármely két elemét (p_i és p_j, ahol $i \neq j$ és $i = 1, \ldots, M$ és $j = 1, \ldots, M$) összeadva a (lásd az 5. táblázatot) $p_i + p_j = 2P11_{ij} + P01_{ij} + P10_{ij}$ értéket kapjuk. A $p_i + p_j$ előállítása két fő célt szolgál:

Egyrészt segítségével a *Dice-együttható* viszonylag gyorsan előállítható lesz.

Másrészt bizonyos mértékig ellenőrizhető az is, hogy a **C** mátrix helyesen volt-e kitöltve. Hiszen amennyiben ugyanis $p_i + p_j = 0$, az csak úgy állhat elő, hogy $P11_{ij} = P01_{ij} = P10_{ij} = 0$, amiből következik $d_{ij} = 0$ is. Ez az eset azt jelenti, hogy a szóban forgó **C** mátrix helyesbítését kell elvégeznünk, hiszen a tekintett két adattípus (A_i és A_j) egyikét sem igényli a **C** mátrixban szereplő egyetlen funkció sem. A **C** mátrix létrehozásánál tehát hiba történt, amelyet korrigálni kell. Vagy az volt a hiba, hogy a **C** egyszerűen felesleges adattípust (A_i vagy A_j-t) tartalmaz, vagy az, hogy a funkciók tényleges adatigényét pontatlanul tükrözi a **C** mátrix, elírás történhetett.

Az is előfordulhatott, hogy a szóban forgó két adattípus valamelyikét nem ehhez a **C** mátrixhoz, hanem egy másik egyedtípushoz tartozó **C** mátrixhoz kellett volna rendelni.

$p_i + p_j$ közvetlen előállíthatósága amellett szól, hogy a *Jaccard helyett a Dice-együtthatót alkalmazzuk.*

6. TÁBLÁZAT

A $\mathbf{J}_1$ részmátrix, ahol $l = 1$, $m = 1, 2, \ldots, 8$, $k = 1, 2, \ldots, 13$

$1 \longrightarrow^{k} K$

	F_1	F_5	F_6	F_7	F_8	F_9	F_{11}	F_{12}	F_{13}
A_1	1	1	1	1	1	1	1	1	1
A_2	0	0	1	0	0	0	1	1	0
A_3	1	1	0	1	1	0	1	0	0
A_4	1	0	1	0	1	0	0	1	1
A_5	0	1	0	0	1	1	1	1	0
A_6	0	1	1	0	0	1	0	0	0
A_7	1	0	1	0	1	0	0	0	1
A_8	1	1	0	0	0	1	1	1	0

2. lépés: Ezután előállítjuk a **C**-ből annak l-edik (kezdetben $l=1$) redukált mátrixát $\mathbf{J}_l$-et, $\mathbf{J}_l=[\mathbf{C}_{k_1}, \ldots, \mathbf{C}_{k_r},]$ ahol $r \leqq N$ és $1 \leqq k_1 < k_2 < \ldots < k_r \leqq N$ úgy, hogy ha $\mathbf{C}_j$ oszlopvektor l-edik eleme 1, akkor (fennáll $k_i=j$) bekerül $\mathbf{J}_l$-be annak egy oszlopaként, egyébként pedig nem.

Tegyük fel, hogy $\mathbf{J}_l(M \times K)$ méretű mátrix lesz (6. táblázat).

Ezáltal a $\mathbf{J}_l$ redukált mátrixba vontuk mindazon változók (itt funkciók) körét, amelyek az A_l és a többi adattípus közös) együtt előforduló) jellemzőiként potenciálisan felléphetnek a **C** mátrixon belül.

7. TÁBLÁZAT

A $\mathbf{J}_1$ részmátrix sorösszegeinek képzése: $\{P_{11}\}_1$, létrehozza a *Dice-együtthatók* számlálójában álló érték felét ($l=1$ és $m=1, 2, \ldots, 8$, $k=1, 2, \ldots, 13$)

	F_1	F_5	F_6	F_7	F_8	F_9	F_{11}	F_{12}	F_{13}	P_{11}	
A_1	1	1	1	1	1	1	1	1	1	9	P_{11}
A_2	0	0	1	0	0	0	1	1	0	3	A_1-A_2 P_{11}
A_3	1	1	0	1	1	0	1	0	0	5	A_1-A_3 P_{11}
A_4	1	0	1	0	1	0	0	1	1	5	A_1-A_4 P_{11}
A_5	0	1	0	0	1	1	1	1	0	5	A_1-A_5 P_{11}
A_6	0	1	1	0	0	1	0	0	0	3	A_1-A_6 P_{11}
A_7	1	0	1	0	1	0	0	0	1	4	A_1-A_5 P_{11}
A_8	1	1	0	0	0	1	1	1	0	5	A_1-A_8

3. lépés: Képezzük először a $\mathbf{P11}_l$ vektort, amelynek elemei a $p11_{ml}$ skalárok, ahol a

$p11_{ml}=\sum_{k=1}^{K} j_{mk}$ sorra megadják az A_l és az A_m (ahol $m \neq l$) közötti hasonlósági együtthatók számlálójában levő értékeket, azaz a $P11_{ml}$ értékeket (kivéve $l=m$ esetében, ahol $d_{ml}:=0$). $j_{mk} \in \mathbf{J}_l$, ahol $k=1, 2, \ldots K$ és $l=$konst. (vö. 7. táblázattal).
— Ahol $p11_{ml}=0$ (de $P01_{ml} \neq 0$ és $P10_{ml} \neq 0$) ott $d_{ml}=0$ és mivel A_l-t és A_m-t nem vonjuk majd össze egy szegmenstípusba, így vele e célból további műveleteket már nem is fogunk végezni, tehát letárolására sem lesz szükségünk (ha a **D** mátrixot nem tömbként, hanem listaszerkezetben tároljuk le).

— Ahol azonban $p11_{ml} \neq 0$, ott fennáll $0 < d_{ij} \leqq 1$ és a már előbb előállított p_m felhasználásával kiszámítjuk a d_{ml} értékeket, azaz a **D** mátrix egy sorát (vagy oszlopát).

— Amennyiben $d_{ij}=1$, akkor az A_i és A_j-nek megfelelő adattípusokat feltétlenül egy szegmenstípusba vonjuk össze.

Ez ugyanis azt jelenti, hogy a két adattípust a **C** mátrixon belül az jellemzi, hogy őket kizárólag együtt igénylik valamennyi funkció esetén. Praktikusan emellett még figyelembe kell venni a következőket:

Természetesen az sincs kizárva, hogy itt nem két külön, hanem egy és ugyanazon — de valamilyen hiba miatt kétszer és eltérő indexszel azonosított — adattípusról van szó.

Ugyanazon adattípusnak az információrendszer majdani adatbázisában történő fizikailag többszöri (redundáns) tárolását általában célszerű elkerülni.

Persze különbséget teszünk az egy szegmenstípuson belüli és a szegmenstípusok között fennálló redundancia esete között. Az utóbbit a következő (4.) lépésnél tárgyaljuk, az előbbi esetet itt.

Egy szegmenstípuson belül mindenképpen el akarunk kerülni bármiféle redundanciát. Ezt egyrészt a szegmensképzés kiindulásához használt normalizált (harmadik normál alakú) adatmodell elkészítése, másrészt egy, a **D** mátrix előállítása közben elvégzendő logikai ellenőrzés biztosítja. Ez utóbbi ellenőrzés azt jelenti, hogy ha két (A_i és A_j) adattípus (ahol $i\neq j$) valamennyi taxonomikus jellemzője egybeesik (azaz $d_{ij}=1$), akkor a két adattípust külön meg kell vizsgálni és tényleges tartalmi egybeesés esetén az egyiket törölni kell a szóban forgó **C** mátrixból. Kiköthetünk olyan feltételt is, hogy már egymáshoz igen közelinek jellemzett (pl. $0{,}95\leqq d_{ij}\leqq 1$) adattípusokat is külön megvizsgálunk. — Ha $0<d_{ij}<1$, akkor összevonásra sor kerülhet.

Eltérően azoktól a módszerektől (pl. SALTON (1968), amelyek egy *H hasonlósági küszöbértéket* rögzítenek és $H<d_{ij}<1$ esetén (ahol $0<H<1$) végzik el az összevonást egy klaszterbe, a 3.1.2. pontban leírt algoritmushoz ilyen paraméter *rögzítésére nincs szükség*.

4. lépés: Kiválasztjuk soronként a max (d_{ml}) elemeket, egyben mindjárt képezzük a sorösszeget is (lásd: a 3.1.2. pontban az 1. és 2. fázist). Azonban jó előre el kell határoznunk, hogy csupán ezeket tároljuk-e el, vagy valamennyi $d_{ml}\in\mathbf{D}$ most kiszámított hasonlósági együttható értékét megőrizzük.

Ezzel kapcsolatban az alábbi szempontokat kell figyelembe vennünk:

A 3.1.2. pontban kifejtett eljárás gépi megvalósítása tekintetében *alapvetően három alternatíva közül választhatunk* [1]

— A teljes hasonlósági (**D**) mátrixot tároljuk. Ekkor a tárigény viszonylag nagy és előállítása a **C** mátrixból $2M^2-9M/2$ számú összehasonlítást igényel. Így ez az alternatíva akkor kerül előtérbe, ha a **C** mátrixok többségére fennáll az, hogy a klaszterálandó adattípusok számát (M) a funkciók száma (N) meghaladja, azaz $M<N$.

— A teljes hasonlósági (**D**) mátrix tárolása helyett csupán az egyes oszlopainak (vagy soroknak) maximális elemét tároljuk. Erre akkór nyílik lehetőség, ha a **C** mátrixhoz rendelt funkciókat arány vagy intervallumskálán mérhetjük [1]. Mivel a mi esetünkben $c_{mn}\in\mathbf{C}$ becsült relatív gyakoriságot fejez ki, ezen előfeltétel fennáll. Ha a **C** mátrixok döntő többségére fennáll az, hogy $M>N$, akkor ennek az alternatívának a tárigénye az előbbihez képest lényegesen kisebb lesz, ugyanakkor azonban $(2M^2-11M)2$ összehasonlítást, majd $M(M-1)$ hasonlósági együttható kiszámítását igényli [1].

— Háttértárolón helyezzük el a szekvenciális sorrendbe rendezett hasonlósági (**D**) mátrixot. Célszerűen akkor tehetjük meg, ha a klaszteráló algoritmus hatékonyságát ez nem korlátozza. A feszítőfák előállítására épülő algoritmusok hatékonyságát a fenti megoldás általában nem korlátozza, így ez számos esetben alkalmazható [1].

Azt, hogy a fenti három alapvető alternatíva közül melyiket alkalmazzuk most, azt az alábbi meggondolások befolyásolják:

Tekintettel arra, hogy az adatbázis újraszerkesztésére annak életciklusa alatt sokszor szükség lesz, olyan klaszteráló módszert célszerű alkalmazni, amely mindig támaszkodhat az előző szerkesztést meghatározó taxonomikus mérték (vagy annak tényezői) eltárolt értékeire. Ez arra utal, hogy *jelen esetben az eltárolt hasonlósági mátrixon alapuló implementációt helyesebb alkalmazni.* Ezt az állítást még két további érv támasztja alá:

> Az egyik az, hogy az adattípusok szegmenstípusokká csoportosításának alapját képező változók egy része intervallum-skálán, más része ordinális skálán mért változó lehet. Változók alatt a funkciókat értjük, szubjektív súlyozásuk eredményezhet ordinális skálán mért változókat (vö. az 1.3.1. ponttal). Ez esetben pedig az eltárolt hasonlósági mátrixon alapuló implementáció a korrekt megoldás. A másik érv az, hogy az esetek elsöprő többségében a funkciók száma kisebb, mint az adattípusoké.

[1] szerint ez utóbbi körülmény fennállása esetén az eltárolt hasonlósági mátrixszal megvalósított agglomeratív klaszterálás hatékonyabb, mint az egyéb implementációs lehetőségek.

Tekintettel arra, hogy nem csupán a — 3.1.2. pontban szereplő algoritmus segítségével képzett — szegmenstípusok (klaszterek) későbbi karbantartásához, hanem az oszloponként kiválasztásra kerülő max (d_{ml}) elem egyértelmű meghatározása, majd a redundancia eltüntetése, vagy megtartása gyors eldöntéséhez mindjárt a szegmensképzés utáni tervezési szakaszban is szükség van több d_{ij} mátrixelem ($d_{ij} \in \mathbf{D}$) gyors elérésére, helyes itt a hasonlósági együtthatók alkotta **D** mátrix teljes letárolása. Ezt mind a **C**, mind a szóban forgó **D** mátrixok száma, valamint átlagos terjedelme, továbbá az általában rendelkezésre álló közvetlen elérésű tárolók kapacitása is megengedi. *A mátrixok számára és terjedelmére vonatkozó durva becslésünk az alábbi:*

> Annyi **C**, illetve **D** mátrix van, ahány egyedtípus. Ezek száma egy átlagos információs rendszernél 20—100 között van. Egyedtípusonként tapasztalat szerint 5—30 adattípus fordul elő. Az adatkezelési funkciók száma egyedtípusonként elég eltérő lehet, de a tervezéskor figyelembevett kisebb hányaduk miatt ez a szám hozzávetőleg 15—20. Ebből adódik, hogy egy **C** mátrix mérete legfeljebb 20×30, egy **D** mátrixé pedig ugyancsak legfeljebb 30×30, mindkét fajta mátrixból legfeljebb 100—100 van (egy átlagosnak tekintett információs rendszernél).

5. lépés: Mivel **D** szimmetrikus, elegendő az alsó háromszögmátrix előállítása, tehát l értékét eggyel megnövelve megvizsgáljuk, hogy $l < M$ fennáll-e. Ha igen, visszatérünk a 2. lépés elejére és új redukáltmátrixot képezünk.

Amennyiben $l < M$ már nem áll fenn, akkor a **D** mátrix előállítása befejeződik.

3.2.2. Távolságmérték alkalmazása

Az implementáció megoldását befolyásoló tényező az elvileg megfelelő taxonomikus mérték kiválasztása. Újabb indokot hozunk fel amellett, hogy miért volt szükség az 1.3. pontban a második és a harmadik közelítésű hozzáférési modell éles elhatárolására.

Mivel a harmadik közelítésű hozzáférési modell már semmiképpen sem tekinthető oszloponként súlyozott bináris mátrixnak, nem érvényesülhetnek azok a klaszteráló eljárás hatékonyabb végrehajtását lehetővé tevő sajátosságok, amelyeket a második közelítésű modellnél kihasználhattunk.

Az algoritmus hatékonysága érdekében ugyanis bináris mátrixot lenne célszerű kezelni. Nem elsősorban azért, mert bináris változók esetén a taxonomikus mértékek gazdagabb készletéből [1] válogathatunk, hanem főleg azért, mert a tárterülettel takarékoskodhatunk (egy mátrix-elem tárolásához tömb társzerkezet esetén — nem kell egy byte, hanem egy bit is elegendő), s így nagyobb méretű mátrixok is kezelhetők.

A bináris változók esetére kiválasztott hasonlósági együttható sem lehet már számunkra megfelelő *(Dice-együttható)*.

Hasonlósági együtthatókat ugyanis csak bináris változók esetében képezhetünk. Mivel változóink most többértékűek, ezért át kell térni a taxonomikus távolságmértékre, továbbá hasonlósági mátrixról távolsági mátrixra.

Itt jelentkezik azonban egy másik sajátosság. Akár a második, akár a harmadik közelítésű modellből indulunk ki, a taxonomikus mértékkel szembeni elvárásunk az, hogy a mérték fejezze ki, hogy két tulajdonságtípusnak egy szegmenstípusba összevonását mérlegelve fontosabb közös jellemzőjük az, hogy egy adott adatkezelési funkció esetén együtt előfordulnak, mint az, ha nem fordul ott elő egyikük sem. Így azt várjuk el, hogy a taxonomikus mérték tükrözze értékelésünk e súlyozásbeli aszszimmetriáját. A hasonlósági együtthatók jó része rendelkezik is ilyen tulajdonsággal, a távolságmértékek közül azonban igen kevés. A kívánt jellemzőkkel rendelkező távolságmértéket e szűk körben kell keresnünk.

Tekintettel arra, hogy hasonlósági együtthatóként legalkalmasabbnak a *Dice* (vagy *Czekanowsky-féle*) *koefficiens* látszik, az ennek kiterjesztése inverzeként meghatározható *Lance—Williams-féle nem-metrikus távolságmérték* alkalmazása a következetes választás.

Kétségtelen az is, hogy ez esetben nem tudjuk kiaknázni a hasonlósági együttható alkalmazásából adódó előnyöket és el kell tekinteni a 3.2.1. pontban vázolt eljárás alkalmazásától.

([1] 113. oldal) rámutat, hogy a LANCE és WILLIAMS [34] által használt nem-metrikus távolságmérték

$$\mathscr{L}_{ij} = \frac{\sum_{n=1}^{N} |c_{in} - c_{jn}|}{\sum_{n=1}^{N} (c_{in} + c_{jn})} = \frac{1 - 2P11_{ij}}{2P11_{ij} + P01_{ij} + P10_{ij}}$$

éppen a *Dice-együttható* valós többértékű változók esetére való általánosítása, azaz belőle bináris változók esetén a *Dice-együttható* 1-es komplemensét (vö. 3.1.2.) kapjuk.

Mivel azonban ez nem tesz eleget a metrika valamennyi követelményének,

(ugyanis $c_{mn} < 0$, ahol $c_{mn} \in \mathbf{C}$ esetében $L_{ij} < 0$ is előfordulhat), csak $c_{mn} \geqq 0$-nál alkalmazható.

A mi esetünkben éppen fennáll az, hogy $c_{mn} \geqq 0$. Igy nem pl. az euklideszi távolságmértéket hanem a nem-bináris válltozók esetére alkalmazható *Lance—Williams féle távolságot* fogjuk alkalmazni, mivel nem fordulhat elő negatív súlytényező.

Ennek megfelelően nem hasonlósági, hanem távolsági mátrixot állítunk majd elő és a 3.2.1, pont algoritmusa is több ponton értelemszerűen módosulni fog.

4. Szegmenstípusok összevonása, areak kialakításának folyamata

4.1. Szegmenstípusok osztályozása

A szegmensképzés során olyan csoportosítást végeztünk el, amely funkciók meghatározott adattípusok iránti hozzáférési igényéből és az egyes adattípusoknak egymással való — a fenti tekintetben vett — hasonlóságából indult ki. Ráadásul a hatékonyabban végrehajtható csoportképző eljárás érdekében az egyedtípusokra vonatkozó adattípusokat előzetes csoportképzéssel már összevontuk (ezek a **C** mátrixok). Egy ilyen csoportból ($\mathbf{C}_j \in \{\mathbf{C}\}^J$) indul ki a szegmenstípusokra való még részletesebb széttagolás folyamata. Az így előállított szegmenstípusokból lesznek majd nagyobb összevont egységek építőkövei. Ezek tervezésére azért van szükség, mert önmagában (általában) egyetlen, a fenti csoportosítási szempontok alapján képzett szegmenstípus sem képes biztosítani egy szempontból teljes lekérdezést (tehát pl. egy egyedtípusra, vagy egy meghatározott funkció által igényelt valamennyi adattípus lehívását), hanem ez gyakran csak több szegmenstípusban található adattípusokra támaszkodva lehetséges. Az egy funkció elintézéséhez (vagy egy egyedtípus jellemzéséhez) szükséges adatok lekérdezéséhez célszerűen ezért a megfelelő egyedtípusok között még a funkcionális elemzés során létre kellett hoznunk a (közvetlen vagy közvetett) navigációs kapcsolatot.

Az adatbázis tervezésének most a következő szakaszában kijelöljük, hogy egy areaba azon szegmenstípusok kerüljenek, amelyeket ugyanazon adatkezelési funkciók használnak, így a kapott eredményeket az areafelosztásnál hasznosítjuk.

Ezen az alapon a szegmenstípusok olyan osztályozását fogjuk elvégezni, amely teljes és biztosítja, hogy a szegmenstípusok egy osztálya (röviden: egy szegmensosztály) alkalmas legyen egyszerre (minél) több funkció adatigényének kielégítésére (tehát ebből a szempontból homogénitást mutasson).

A szegmensosztályok képzése közvetlenül az areafelosztás megoldását célozza, közvetve pedig előkészíti az elhelyezési módok meghatározását, szegmenstípusok (esetleg egyedtípusok) összevonását, kapcsolattípusok szétbontását esetleges összevonását.

8. TÁBLÁZAT

Szegmensképzés végeredménye (az 1. táblázat **D** mátrixa adataiból kiindulva)

Egy egyedtípushoz tartozó szegmenstípusok	Kiinduló mátrix	Elemi adatok csoportja
SZ_1	$\mathbf{D}$	$A_3 - A_4$
SZ_2	$\mathbf{D}^{1,2,5,6,7,8}_{1,2,5,6,7,8}$	$A_1 - A_6 - A_5 - A_2$
SZ_3	$\mathbf{D}^{7,8}_{7,8}$	$A_7 - A_8$

A matematikai—statisztika nyelvezetéhez igazodva csoportosítandó egyedeknek most így a szegmenstípusokat, változóknak pedig a funkciókat tekintve az előbbi mondatban megfogalmazott cél voltaképpen a változók és az egyedek „harmonizált" osztályozásának végrehajtását igényli [1]. ANDERBERG kifejti, hogy ennek egy lehetséges megoldása az, hogy felváltva előbb az egyedeket, majd a változókat ciklikusan klaszteráljuk úgy, hogy végeredményben egy kölcsönösen „harmonizált" eredményt kapjunk. Elméleti kutatásokat e téren többen végeztek [15], [39], [13]. Ilyen jellegű feladatokra például alkalmas a PROMENADE [25] vagy OLPARS [49] programcsomag, illetve azok on-line interaktív üzemmódra továbbfejlesztett változata. De ezek az eszközök számunkra nem hozzáférhetőek, és drágák is.

A szegmensosztályok képzésére az alábbi módszer javasolható:

Az 1.3. pontban ismertetett funkcionális felmérés és a 3. fejezetben áttekintett szegmensképző eljárás végeredményéből (szegmenstípusok) kiindulva hozzunk létre egy olyan **B** bináris mátrixot, amelynek oszlopai az egyes funkciókhoz tartozó szegmensszintű adat-almodelleket, sorai az előzőleg már meghatározott szegmenstípusokat képviselik (lásd a 9. táblán).

Sem az egy szegmenstípussal leírt szegmensek számát, sem az egy-egy almodell előfordulásának gyakoriságát jellemző adatokat itt most nem kell figyelembe venni, mivel ezt a szegmensképzésnél — vö. 1.3.2. ponttal — már megtettük).

Az így létrehozott **B** mátrixot blokk-diagonális mátrix előállítására (**B***) törekedve rendezzük át, úgy, hogy a $b_{ir} \neq 0$ értékű elemei minél nagyobb mértékben a főátló körül sűrűsödjenek. Ezáltal az egyes szegmenstípusok osztályba tartozása szemlélet alapján is könnyen eldönthető lesz. $\mathbf{B}^* \sim \mathbf{B}$, mert az (vö. 10. táblázattal) **B** átrendezését elemi átalakításokkal (sor-, illetve oszlopcserék) végezzük el.

Feltehető azonban eleve, hogy automatikus osztályozásunk nem lehet taxonomikus és a kapott osztályok átfedésben lesznek. Ennek az az előrelátható oka, hogy gyakorlatilag mindig van több olyan funkció, amely adatigényének kielégítését biz-

9. TÁBLÁZAT

A **B** mátrix
Valamennyi egyedtípushoz tartozó szegmenstípusok felsorolása

	F_1	F_2	F_3	F_4	F_5	F_6	F_7	F_8	F_9	F_{10}	F_{11}	F_{12}
SZ_1	1	0	1	1	0	0	0	0	0	0	0	0
SZ_2	0	0	0	0	0	1	0	0	0	0	0	1
SZ_3	0	1	0	0	0	0	0	0	1	1	1	1
SZ_4	0	0	0	1	1	0	1	1	0	0	0	0
SZ_5	0	0	0	1	1	0	1	1	0	0	0	0
SZ_6	0	0	0	0	0	1	0	0	0	0	0	1
SZ_7	0	0	0	1	1	0	1	1	0	0	0	0
SZ_8	0	1	0	0	0	0	0	0	1	1	1	0
SZ_9	1	0	1	1	0	0	0	0	0	0	0	0

10. TÁBLÁZAT

B* mátrixsza átrendezett táblázat (vö. 9. táblázat)

	F_1	F_3	F_4	F_5	F_7	F_8	F_6	F_{12}	F_9	F_{10}	F_{11}	F_2
SZ_1	1	1	1	0	0	0	0	0	0	0	0	0
SZ_9	1	1	1	0	0	0	0	0	0	0	0	0
SZ_4	0	0	1	1	1	1	0	0	0	0	0	0
SZ_5	0	0	1	1	1	1	0	0	0	0	0	0
SZ_7	0	0	1	1	1	1	0	0	0	0	0	0
SZ_6	0	0	0	0	0	0	1	1	0	0	0	0
SZ_2	0	0	0	0	0	0	1	1	0	0	0	0
SZ_3	0	0	0	0	0	0	0	1	1	1	1	1
SZ_8	0	0	0	0	0	0	0	0	1	1	1	1

11. TÁBLÁZAT

Blokkdiagonális mátrix

1	1	1	1	0	0	0	0	0	0	0
1	1	1	1	0	0	0	0	0	0	0
0	0	0	0	1	1	1	1	1	0	0
0	0	0	0	1	1	1	1	1	0	0
0	0	0	0	1	1	1	1	1	0	0
0	0	0	0	0	0	0	0	0	1	1
0	0	0	0	0	0	0	0	0	1	1

tosító szegmensszintű adatalmodell csak több szegmensosztályból gyűjthető össze. Ez egyben azt is jelenti, hogy emiatt szegmenskapcsolat nem csupán egy-egy szegmens-osztályon belül, hanem szegmensosztályok között is fennáll.

A fenti körülményeket figyelembe véve az elméletileg lehetséges legkedvezőbb osztályképzés, amelyet a blokkdiagonális alakúra átrendezett **B** mátrix (vö. 11. táblázattal) tükröz — igen ritkán fog a gyakorlatban előállni. A praktikusan remélhető legkedvezőbb osztályképzés az az eset, amelyet a 12. táblázaton láthatunk, mivel:

a) Egy szegmenstípus egyértelműen a funkciók egy jól körülhatárolható részhalmazához rendelhető, mivel ezek belőle egyaránt merítenek.

Ha még ez a praktikusan remélhető legkedvezőbb eset sem jöhet létre, mert számos olyan szegmenstípus adódik, amely a fenti a) feltételnek nem tesz eleget, akkor ezt az fejezi ki, hogy az átrendezett **B*** mátrixban a fődiagonálistól oldalra találunk

12. TÁBLÁZAT

Praktikusan kedvező szegmensosztályképzés

1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
0	0	0	1	1	1	1	1	0	0	0	0	0	0	0
0	0	0	0	0	0	1	1	1	1	1	1	1	0	0
0	0	0	0	0	0	1	1	1	1	1	1	1	0	0
0	0	0	0	0	0	0	0	0	0	0	0	1	1	1
0	0	0	0	0	0	0	0	0	0	0	0	1	1	1

13. TÁBLÁZAT

Gyakorlatban várható K mátrix alakja

1	1	1	1	1	0	0	0	0	0	1	0	0	0	0
1	1	1	1	1	0	0	0	0	0	1	0	0	0	0
0	0	0	1	1	1	1	1	0	0	0	0	0	1	0
0	0	0	0	0	0	1	1	1	1	1	1	1	0	0
0	1	1	0	0	0	1	1	1	1	1	1	1	0	0
0	1	0	0	0	0	0	0	0	0	0	0	1	1	1
0	0	0	1	1	0	0	0	0	0	0	0	1	1	1

olyan homogén blokkokat, amelyek a fődiagonálisban megjelenő szegmensosztályok közötti kapcsolatokat fejezik ki (vö. 13. táblázat).

A szegmensosztályok képzéséhez olyan módszerre van szükségünk, amely figyelembe veszi azt a körülményt, hogy a tervezés következő lépése érdekében (lásd: 4.2. pont) elő kell segíteni egy olyan interaktív ember—gép kapcsolatot, amely a modell sokdimenziós terét a display kétdimenziós síkjára úgy vetíti ki, hogy az ilyenkor szükségképpen fellépő torzulások (a leképezés nem távolságtartó) ellenére is gyorsan el lehessen dönteni, hogy mely szegmenstípusok tartoznak egy szegmensosztályba (klaszterba) és melyek nem.

Az erre a célra a gyakorlatban alkalmazott módszerek közül gyakoriak:

— Táblázat — átrendezés (*Data-rearranging*, vagy kváziklaszterálás).
— Főkomponens analízis.
— Több-dimenziós skálázás (multidimensional scaling).

A mi esetünkben a táblázat — átrendezés kiválasztása látszik célszerűnek, hiszen kiinduló adataink táblázatba rendezetten már rendelkezésre állnak. Ugyanakkor a BMDP, vagy más hazánkban elérhető klaszterálási programcsomagok (MTA) általában rendelkeznek valamilyen táblázat — átrendező batch programmal.

A táblázat-átrendezésre több heurisztikus algoritmus közül választhatunk [14], [27], [50], [33]. A táblázat-átrendezés egyes algoritmusai gráfelméleti meggondolásokon alapulnak. Az alábbi heurisztikus módszer alkalmazását azért javasoljuk, mert alapját ilyen elméleti meggondolások képezik.

A módszer alapgondolata az, hogy mivel a sokdimenziós térben meghatározott minimális feszítőfa egészét nehézkesen lehet két dimenzióban megjeleníteni úgy, hogy a leképezésből adódó torzítások hatását kiküszöböljük, nem lenne célszerű itt a csoportképzéshez kiindulásul minimális feszítőfát létrehozni. Viszont az előbbi létrehozásához képest az ún. minimális feszítőút előállítása egyszerűbb, ugyanakkor a létező csoportok vizuálisan jól megkülönböztethetők.

Definíció. Feszítőútnak nevezzük azt az összefüggő részgráfot, amely eleget tesz az alábbi feltételeknek:

1. Minden szögpontja a gráfon belül van.
2. Ezek között van kettő olyan szögpont, amelyekhez csak egyetlen él csatlakozik.
3. A részgráf összes többi szögpontja esetében pedig egy szögponthoz mindig csakis kettő-kettő él csatlakozik.

A feszítőút a feszítőfának egy speciális esete.

Definíció. Minimális feszítőútnak nevezzük azt a feszítőutat, amelynek teljes hossza a gráfban lehetséges valamennyi feszítőutaké között a legrövidebb. [38] kimutatta, hogy egy minimális feszítőút megtalálása az „utazó ügynök probléma" megoldásával egyenértékű. [51]-ben azonban nem a fenti probléma megoldására gyakran alkalmazott algoritmusok valamelyikét használják, mert megállapítják, hogy ha sok (legalább néhányszáz) tétel klaszterálásáról van szó, akkor ezen algoritmusok túl lassúak.

A gyorsabb eljárás érdekében megelégszenek azzal is, hogy ne a minimális, hanem egy ahhoz eléggé közelálló hosszúságú azaz egy ún. rövid feszítőutat kapjanak meg.

Erre a célra egy olyan iteratív heurisztikus algoritmust alkalmaznak, amely a kiinduló **B** mátrix oszlopait, majd sorait permutálja.

Minden egyes permutáció alkalmával arra törekszünk, hogy a **B** mátrixra vonatkozó alábbi célfüggvény értékét maximalizáljuk.

$$f(\mathbf{B}) = \frac{1}{2} \sum_{t=1}^{T} \sum_{r=1}^{R} b_{tr} \cdot (b_{t,r+1} + b_{t,r-1} + b_{t-1,r} + b_{t+1,r}),$$

ahol $b_{tr} \geqq 0$ továbbá $b_{0r} = b_{T+1,r} = b_{t,R+1} = b_{t0}$. A $f(\mathbf{B})$ bármely méretű és formájú mátrixra számítható.

A célfüggvényt Mc. Cormick—Schweitzer—White [41] javasolta.

$$f(\mathbf{B}) = \frac{1}{2} \sum_{t=1}^{T} \sum_{r=1}^{R} [b_{tr} \cdot (b_{t,r+1} + b_{t,r-1}) + b_{tr} \cdot (b_{t-1,r} + b_{t+1,r})] =$$

$$= \frac{1}{2} \sum_{t=1}^{T} \sum_{r=1}^{R} (f_t(\mathbf{B}) + f_r(\mathbf{B})).$$

Minthogy $f(\mathbf{B})$ kiszámításánál, ha a **B** sorain haladunk végig, akkor csak a belső összeg második tagja $f_r(\mathbf{B})$ változhat, továbbá, ha **B** oszlopait vesszük sorra, akkor csak

a belső összeg első tagja $f_t(\mathbf{B})$ változhat, az algoritmus is két menetre, azaz oszlopok és a sorok permutálására tagolható.

Az algoritmus két meneténck lényege az alábbi:

I. menet: Megkeres az oszlopvektorok között egy rövid feszítőutat. Átrendezi az oszlopokat, úgy, hogy azok sorrendjét a rövid feszítőút jelöli ki.

II. menet: Az oszlopcserék végrehajtása után megkeres a sorvektorok között egy rövid feszítőutat. Átrendezi a sorokat úgy, hogy azok sorrendjét e rövid feszítőút jelöli ki.

Az algoritmus első meneténck lépései:

1. lépés: Válasszunk ki egy tetszőleges oszlopot a $(T \times R)$ méretű **B** mátrixból. $I=1$.

2. lépés: Közvetlenül a kiválasztott oszlop mellé jobbra és balra is (ha ez lehetséges) egyenként sorra odapróbáljuk a megmaradt $R-I$ számú oszlopot és minden egyes ilyen belepróbálás alkalmával megnézzük (kiszámítjuk), hogy hogyan nőne meg, vagy csökkene a $f(\mathbf{B})$ értéke.
Amely oszlopcserénél a $f(\mathbf{B})$ értéke maximális lenne, azt az oszlopcserét ténylegesen végrehajtjuk. [51] változtatása az, hogy a $f(\mathbf{B})$ helyett bármely távolságmérték használható. Ez esetben az oszlopcserét a minimális távolságérték esetén kell végrehajtani.

3. lépés: $I=I+1$ és ismételjük meg a 2. lépést, amíg csak $I+R$ be nem következik. Az algoritmus időigénye (T^2R+TR^2) 2-vel arányos, azaz az időigény a **B** mátrix méretétől és az alkalmazott távolságmértéktől függ, [41] és csak kismértékben befolyásolja a kezdő oszlop illetve sor megválasztása.

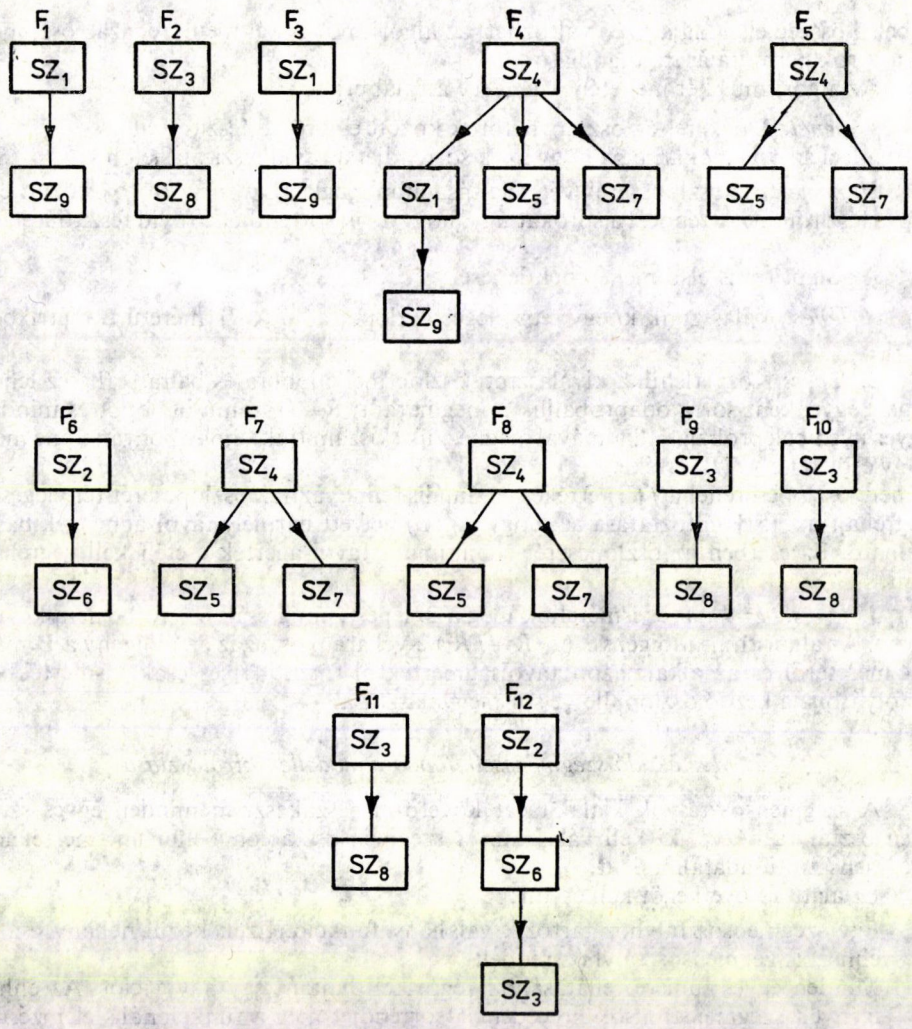
4.2. Mértékadó szegmensszintű adatalmodellek kiválasztása

A szegmensosztályok kialakítását követő első szakaszban minden egyes szegmensosztályra nézve, ki kell választani a szegmensosztályon belüli ún. mértékadó szegmensszintű adatalmodellt.

Ez alatt a következőt kell érteni:

Egy szegmensosztályhoz tartozik valahány funkció. Ezek közül néhány csakis ehhez a szegmensosztályhoz rendelt.
Minden egyes funkció ellátásához végre kell hajtani egy navigációt. Az ehhez igényelt szegmenstípusok precedencia-sorrendjét még a funkcionális elemzésnél fel kellett tárni. Ebből funkciónként egy szegmenselérési fa konstruálható.
A fa gyökere az a szegmenstípus, amely tartalmazza a [24] által vezérlőinformációnak nevezett tulajdonságtípust. A gyökérszegmenst nevezi a CODASYL [11] 120. oldal *entry-defining group*-nak.
A szegmensosztályhoz egyértelműen rendelt funkciók navigációi által meghatározott szegmens-elérési fák csomópontként és levélként jobbára ugyanazon szegmenstípusokat tartalmazzák, de a szegmenstípusok fában elfoglalt szintje, helye általában eltérő.

Az a célunk, hogy az egy szegmensosztályhoz egyértelműen rendelt szegmenselérési fák közül kiválasszunk egyet, — a szegmensosztály mértékadó szegmensszintű adatalmodelljét — mert általában az ennek megfelelő tartalmú és szerkezetű areat fogjuk majd a tárolón megvalósítani.



6. ábra. Összevont szegmensszintű adatmodellek

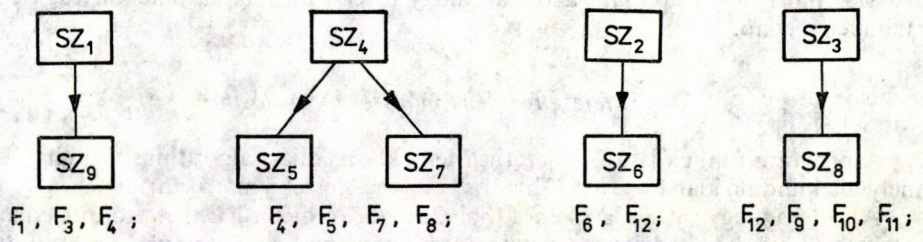
A fenti cél eléréséhez először bejelöljük — a **B** mátrixot oszloponként kiegészítve, $b_{tr} \neq 0$ elemeihez ($b_{tr} \in \mathbf{B}$) további jelölést téve — azt, hogy egy-egy szegmenselérési fában az érintett szegmenstípusok mely szintszámmal jellemezhetők. Az így kiegészített **B** mátrixot **F** mátrixnak nevezzük.

Ezután az **F** mátrix egy-egy oszlopa által képviselt relációkat grafikus formában jelenítjük meg. A szegmensosztályhoz egyértelműen tartozó funkciókra egyenként is megállapítva az érintett szegmenstípusok navigációs kapcsolatait, ezután a kapott fastruktúrákat összevonással eggyé olvasztjuk, és az így összevont struktúrához a szegmensosztály egésze szempontjából érvényes hierarchiaszinteket állapítunk meg. Az összevonás alapja természetesen egyrészt azt, hogy funkciónként a szegmensszintű

14. TÁBLÁZAT

F mátrixszá átjelölt táblázat (vö. 10. táblázat)
(Az átjelölés végrehajtását a 6. ábra és 10. táblázat birtokában lehet elvégezni).
Egy mátrixelem felső részében a meghívó szegmenstípus, alatta hirearchiaszint jele

	F_1	F_3	F_4	F_5	F_7	F_8	F_6	F_{12}	F_9	F_{10}	F_{11}	F_2
SZ_1	$\bar{1}$	$\bar{1}$	SZ_4 2									
SZ_9	SZ_1 2	SZ_1 2	SZ_1 3									
SZ_4			$\bar{1}$	$\bar{1}$	$\bar{1}$	$\bar{1}$						
SZ_5			SZ_4 2	SZ_4 2	SZ_4 2	SZ_4 2						
SZ_7			SZ_4 2	SZ_4 2	SZ_4 2	SZ_4 2						
SZ_6							SZ_2 2	SZ_2 2				
SZ_2							$\bar{1}$	$\bar{1}$				
SZ_3								SZ_6 3	$\bar{1}$	$\bar{1}$	$\bar{1}$	$\bar{1}$
SZ_8									SZ_3 2	SZ_3 2	SZ_3 2	SZ_3 2



7. ábra. Mértékadó szegmensszintű adatmodellek szegmensosztályonként

adatalmodellek mely elemei esnek egybe, másrészt az, hogy ez utóbbi egybeesők között fennálló relációk mennyiben esnek egybe.

Ezeket a lépéseket elvégezhetjük számítógép segítsége nélkül is, hiszen kisebb feladatok esetén a gyakorlatban feltehetően csupán néhány tucat gráf közül kell kiválasztani a mértékadó szegmensszintű adatalmodellnek megfelelő struktúrát.

Ha ezt minden szegmensosztályra nézve már végrehajtottuk, akkor előállítottuk valamennyi mértékadó szegmensszintű adatalmodellt.

Az egy szegmensosztályon belüli mértékadó szegmensszintű adatalmodellek kiválasztása és ezzel az area e belső szerkezetének kijelölése után kezdünk foglalkozni a több szegmensosztályhoz is kötődő, funkciókkal. Ennek az a célja, hogy az előbb

meghatározott mértékadó szegmensszintű adatalmodellek közötti kapcsolatokat úgy alakítsuk ki, hogy a több szegmensosztályból is merítő funkciók adatigényeire legyenek messzemenően tekintettel. Itt két eset van:

a) Több szegmenscsoport közti kapcsolatot képviselő funkciók szóba jöhető szegmensszintű adatalmodelljei úgy kapcsolják össze a szegmensosztályok mértékadó adatalmodelljeit, hogy az egyszerűen azok egyesítését jelenti, külön-külön tekintett belső szerkezetük legcsekélyebb változtatása nélkül.

b) Több szegmensosztály mértékadó szegmensszintű almodelljeinek összekötését a szóba jöhető szegmensszintű adatalmodellek úgy valósítják meg, hogy lesz olyan mértékadó szegmensszintű adatalmodell, melynek belső szerkezete is megváltozik.

Ezt lehetőleg kerüljük el.

5. Gyakorlat, megvalósítás

5.1. Általános tapasztalatok

A 3. és 4. fejezetben leírt eljárás softwaretermékké fejlesztve és az adatbázistervezési folyamatba illesztve 1983-ban elkészült. TPA 1140, TPA 1148 és SZM—4 típusú számítógépekre FORTRAN nyelvű programok állnak rendelkezésre.

1986-ban hajtjuk végre az eljárás IBM/370 és az ESZR II. sorozatának néhány modelljére történő adaptálását.

Két vállalat éles adataival próbáltuk ki a módszert. A nyert tapasztalatok azt bizonyítják, hogy a logikai tervezés érintett tevékenységei lényegesen gyorsabban és megbízhatóbban végezhetők el. Atekintetben, hogy ez az eljárás mennyivel ad „jobb" leképezést és így hatékonyabb adatbázistervezést, mint a más tervezési módszerrel nyert leképezés, csak hosszabb idő elteltével lesznek adataink.

Már most megállapítható azonban, hogy a tervezői munka hatékonysága egyértelműen megnő.

5.2. A 3. fejezetben leírt módszer megvalósítása

A program funkciója a 3. fejezetben leírt klaszterálási algoritmus végrehajtása, amelynek kiinduló alapadatait a **C** bináris vagy egész értékű mátrix tartalmazza.

A program egy futás alatt képes több különböző méretű **C** mátrixot, illetve egy mátrixnak különböző variánsait is feldolgozni, valamint megadhatók korábbi futtatásokból vagy más módon kialakított klaszterjavaslatok is, amelyek arra használhatók egy későbbi fázisban, hogy a különböző szempontok alapján nyerhető klaszterálások egymással összevethetők legyenek a végleges változat meghatározása előtt.

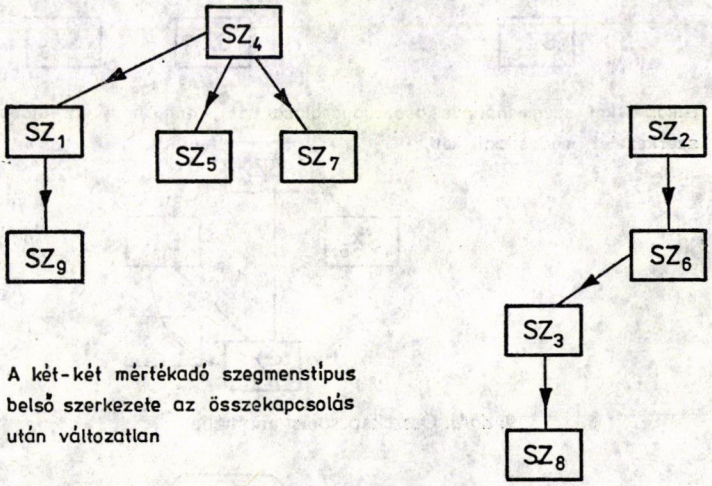
A klaszterálási algoritmus eredményét a program a következő formában állítja elő.

Ha a klaszterálás során az $A(1), \ldots, A(M)$ elemeket $K(\leqq M)$ klaszterba sorolja, akkor a klasztereket keletkezésük sorrendjében 1-től K-ig sorszámozza, valamint ha az I. klaszterba $(1 \leqq I \leqq K)$ az $A(I_1), \ldots A(I_{n_I})$ elemek kerülnek (n_I az I. klaszter elemszáma), akkor az eredményt tartalmazó $Q(1), \ldots, Q(M)$ mennyiségekre $Q(I_1) = \ldots = Q(I_{n_I}) = I$, azaz tetszőleges $J(1 \leqq J \leqq M)$ esetén $Q(J)$ azon klaszternek a sorszáma, amelybe a J-edik elem $A(J)$ tartozik. Amennyiben az inputban már meglevő klaszterálásokat is meg kívánunk adni, annak is ilyen formájúnak kell lennie.

A feldolgozható adatok korlátai:
a) legfeljebb 32 adattípus és
b) legfeljebb 128 funkciótípus egy **C** mátrixon belül,
c) a **C** mátrixok számát a háttértároló kapacitása szabja meg.

A program FORTRAN nyelven készült, ezért az input adatok leírásánál a FORTRAN FORMAT tételeit adjuk meg az egyes adattételek mellett.

Bináris változók esetén a *Dice-együtthatókból* képzett hasonlósági mátrix, egész változók esetén pedig a *Lance—Williams-féle nem-metrikus távolságmátrix* a klaszterálási algoritmus alapja.



8. ábra. Mértékadó szegmensszintű adatmodellek összekapcsolása

A program három fázisban működik:
a) adatellenőrző fázis,
b) hasonlóság, ill. távolság mátrixot előállító fázis,
c) klaszterálás.

Mivel az inputban szerepelnek a mátrixok darabszámára és méretére vonatkozó információk is, ezért az adatellenőrző fázis gondosan ellenőrzi, hogy az input megfelel-e az előírt formátumnak.

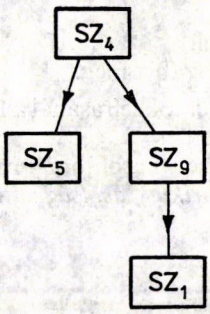
Bármilyen előforduló hiba a futás befejezését eredményezi, mert lehetetlenné válik az input és a program megfelelő szinkronizációja és a további fázisok eredménye értelmetlenné válna.

A b) fázis futása során a program kihasználja a PDP—FORTRAN előnyös bitkezelési tulajdonságait, ha a **C** mátrix bináris értékekből áll.

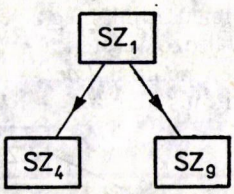
A c) fázis a már leírt formában állítja elő a klaszterálás eredményét a tartalmazó **Q** vektort. A további feldolgozások (area tervezés) előkészítéseképpen a program igényel, de nem használ két azonosító vektort az inputban:
a) adattípus azonosító,
b) funkciótípus azonosító.

Ezeket egy **C** mátrixra vonatkozóan egyszer kell megadni, értékük egész szám lehet,

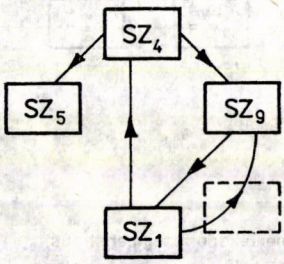
Tételezzük fel, hogy F_4 -re nem a 6. ábrán látott reláció, hanem



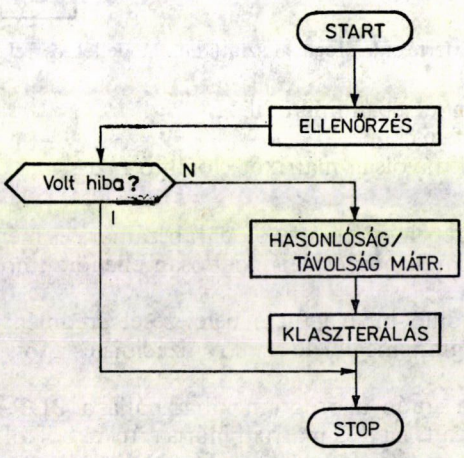
állna fenn, továbbá F_3 -ra és F_1 -re sem a 6. ábrán látható reláció igaz, hanem



Ekkor a két szegmenstípus összekapcsolása azt jelenti, hogy az utóbbi szerkezetét módosítani kell:



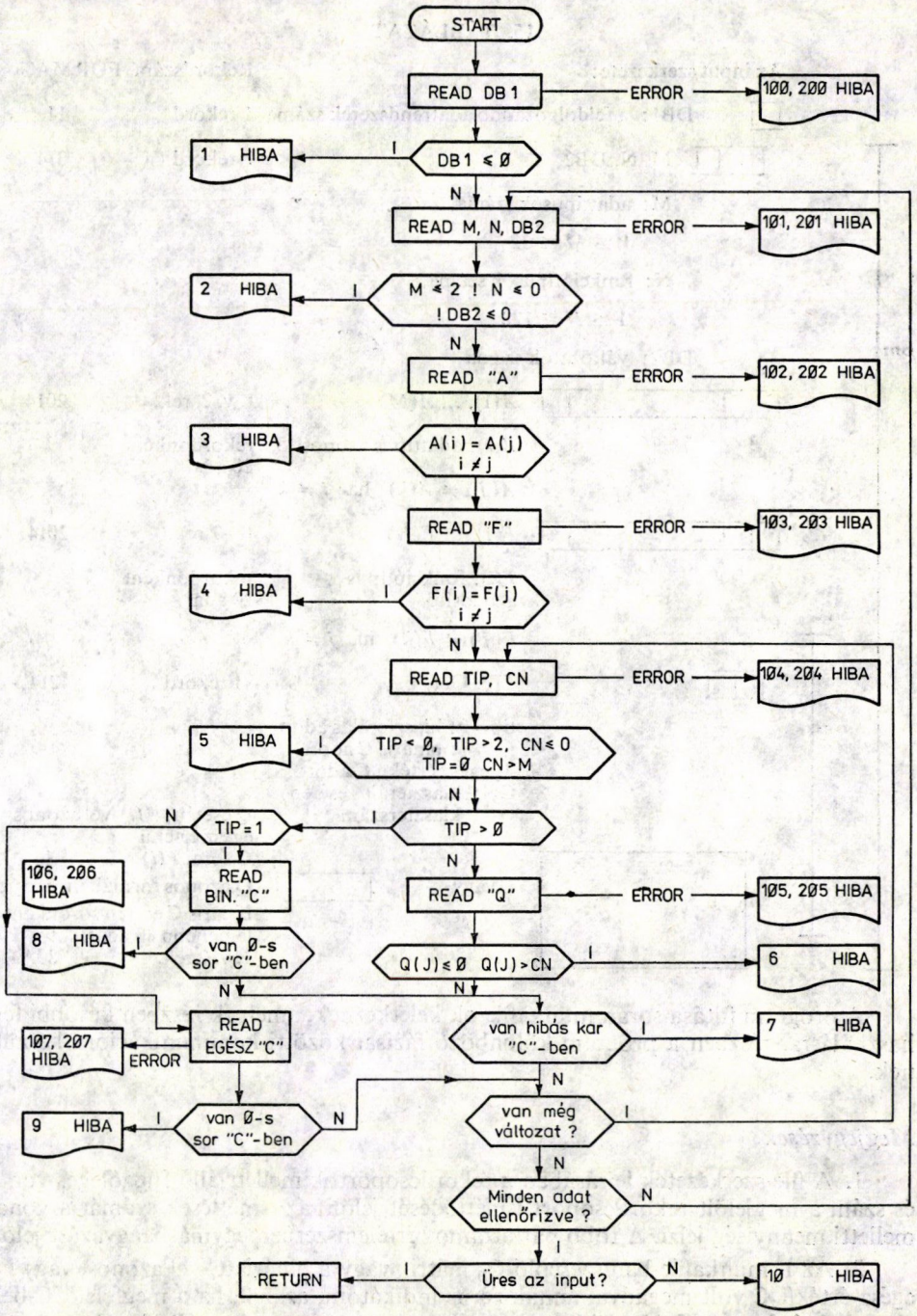
9. ábra. Összekapcsolási anomáliák



10. ábra. A feldolgozás fázisainak nagyvonalú áttekintése

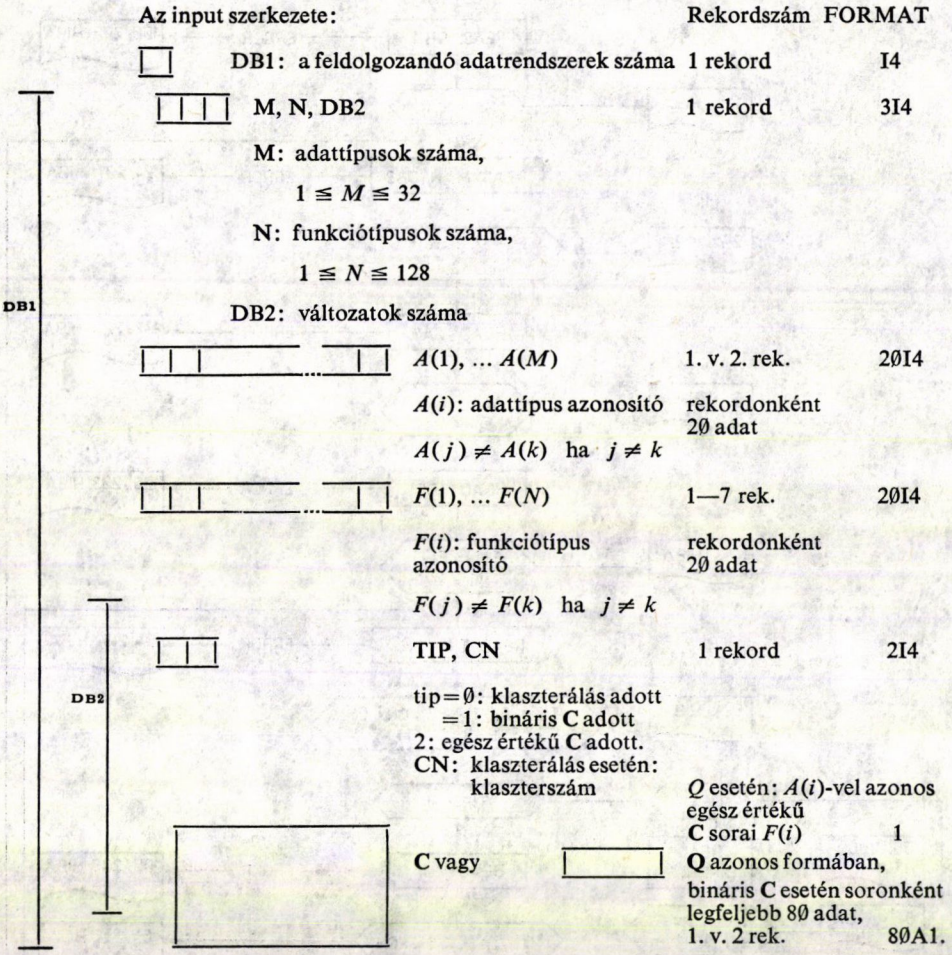
az adattípus azonosítókkal is és a funkciótípus azonosítóknak is egyedieknek kell lennie egy **C** mátrixon belül.

Az input legfeljebb 80 karakteres rekordokból áll, így jól alkalmazkodik mind a kártyás, mind a képernyős környezethez.



11. ábra. Ellenőrzés

15. TÁBLÁZAT



A program futása során munkafile-ok keletkeznek, amelyek részben későbbi felhasználásra, részben a program különböző fázisai közötti kommunikációra készülnek.

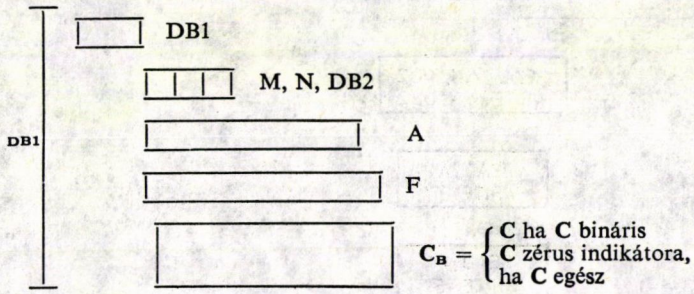
Megjegyzések.

1. A file-szerkezetek leírásában a rekordcsoportok mellett álló függőleges vonal és szám a megjelölt rekordcsoport ismétlődését jelöli, az ismétlések számát a vonal melletti mennyiség jelzi. A több párhuzamos értelemszerűen egymásbaágyazást jelöl.

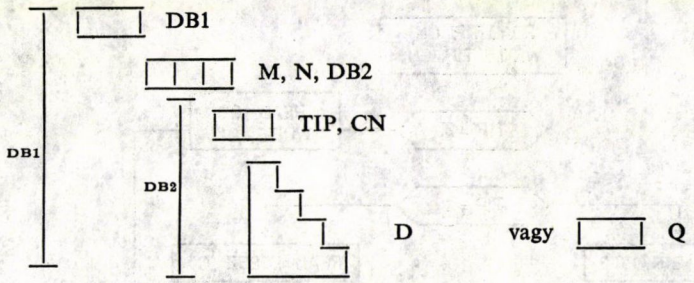
2. Az 1. munkafile-ban szereplő $\mathbf{C}_B$ mátrix vagy a bináris **C**-vel azonos, vagy ha egész értékű **C** volt megadva, annak zérus indikátora, azaz Ø, ha a megfelelő **C**-beli elem Ø, és 1, ha nem Ø. Ha több változat is szerepel az inputban, ($DB2 > 1$), ez a $\mathbf{C}_B$

16. TÁBLÁZAT

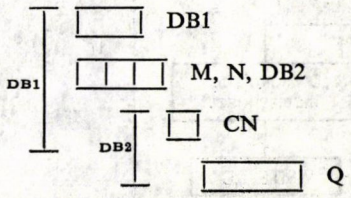
1. Munkafile:



2. Munkafile:



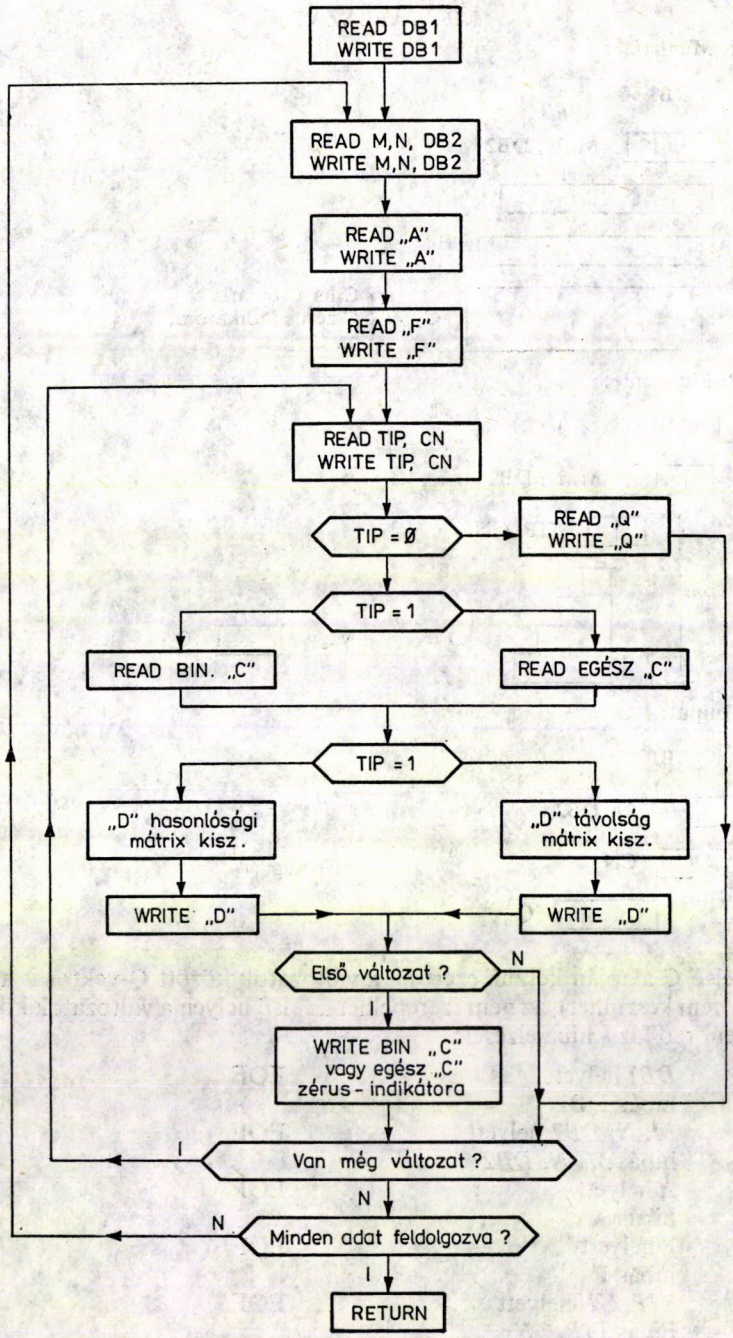
Output:



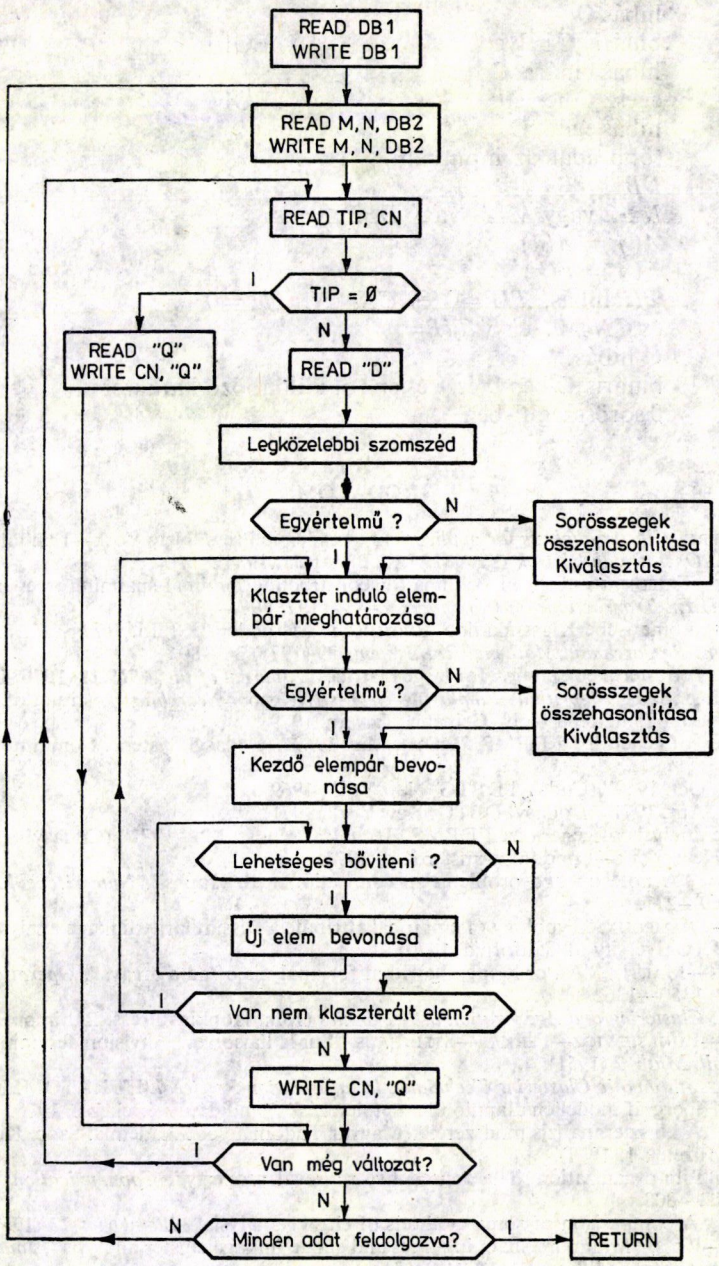
mindig az első **C** alapján készül, ezért ha a változatok között **Q**-vektor is adott (ami alapján C_B nem készülhet), az nem szerepelhet az első helyen a változatok között.

Az ellenőrző fázis hibajelzései:

1ØØ	*DB*1 helyett	EOF
2ØØ	hibás DB1	
1Ø1	*M*, *N*, *DB*2 helyett	EOF
2Ø1	hibás *M*, *N*, *DB*2	
1Ø2	**A** helyett	EOF
2Ø2	hibás **A**	
1Ø3	**F** helyett	EOF
2Ø3	hibás **F**	
1Ø4	*TIP*, *CN* helyett	EOF
2Ø4	hibás *TIP*, *CN*	
1Ø5	**Q** helyett	EOF



12. ábra. Hasonlóság/távolság mátrix létrehozása



13. ábra. Klaszterálás

2Ø5	hibás **Q**	
1Ø6	bináris **C** helyett	EOF
2Ø6	hibás bináris **C**	
1Ø9	egész **C** helyett	EOF
2Ø9	hibás egész **C**	
22	több adat az inputban	
1	$DB1 \leqq Ø$	
2	$N \leqq 2$ vagy $M \leqq 1$ vagy $DB2 \leqq Ø$	
3	$A(j)=A(k)$	
4	$F(j)=F(k)$	
5	TIP hibás, $TIP=0$ és $CN>N$, $TIP=0$ és $CN \leqq Ø$, első $TIP=Ø$	
6	**Q** hibás	
7	bináris **C**-ben Ø és 1 és sp-től különböző karakter	
8	Ø sorösszeg **C**-ben	

IRODALOM

[1] ANDERBERG, *Cluster analysis for applications* (Academic Press, New York—London, 1973).
[2] BANA, I., *Osztott adatbázisok* (SZÁMALK, Budapest, 1984).
[3] BENTLEY—FRIEDMAN, "Fast algorithms for constructing minimal spanning trees in coordinate space", *IEEE Transactions on Computers* **C—27** (1978) 97—105.
[4] BOCK, "Automatische Klassifikation. Statistische Methoden II", Ed. *Walter, Lecture Notes on Operations Research and Mathematical Systems* **39** (1970) 36—80.
[5] CAGAN, "Term-term correlation for large matrices", *Journal of the ASIS* **21** (1970) 163.
[6] CHEN—LEE, *Some Algorithms Employing Nearest Neighbor Searching* (Institute of Comp. Dec. Sci., National Tsing Hua Univ, Hsinchu, Taiwan, 1979).
[7] "A survey of generalized DBMS. Report, May 1969" (Codasyl Systems Committe, New York, 1969).
[8] "Report, Oct. 1969" (Codasyl DBTG, New York, 1969).
[9] "Report, Apr. 1971" (Codasyl DBTG, New York, 1971).
[10] "Feature analysis of generalized DBMS. Technical report, May 1971" (Codasyl Systems Committee, New York—London—Amsterdam, 1971).
[11] DIJKSTRA, "A note on two problems in connection with graphs", *Numerische Mathematik* **1** (1959) 269—271.
[12] DUBIN—CHAMPOUX, "Typology of empirical attributes. Dissimilarity linkage analysis. Technical report 3" (University of California, 1970).
[13] DEUTSCH—MARTIN, "An ordering algorithm for analysis of data arrays", *Operations Research* **19** (1971) 1350—1362.
[14] FISCHER, *Clustering and Aggregation in Economics* (John Hopkins Press, Baltimore, Md., 1698).
[15] FLOREK—LUKASIEWICZ—PERKAL—STEINHAUS, "Sur la liaison et la division des point d'ensemble fini", *Coll. Math.* **2** (1951).
[16] FRITSCHE, *Automatic Clustering Techniques in Information Retrieval* (EURATOM, 1974).
[17] FUTÓ, „Hipergráf modellen alapuló klaszterelemzés", kandidátusi értekezés, 1978.
[18] FÜSTÖS, „A klaszteranalízis módszerei" (Magyar Tudományos Akadémia. Szoc. Kut. Int. módszertani füzetek 1. 1977).
[19] GHOSH, "File organization. The consecutive retrieval property", *Communications of ACM* **15** (1972) 802—808.
[20] GOWER, „A comparison of some methods of cluster analysis", *Biometrics* **23** (1967) 623—637.
[21] GOWER—ROSS, "Minimum spanning trees and single linkage cluster analysis", *Journal of Royal Stat. Ass. C. Applied Statistics* **18** (1969) 54—64.
[22] GYORSOKNÉ—VERŐ, *SÁMÁN adatbázistervezési gyakorlatok* (SZÁMALK, Budapest, 1982).
[23] HALASSY, *Adatmodellezés, adatbázis-tervezés* (SZÁMOK, Budapest, 1980).
[24] HALL, "Avoiding informational distortions in automatic grouping programs", *Systematic Zoology* **18** (1969) 328—329.

[25] HALL—BALL—WOLF—EUSEBIO, "PROMENADE. An improved interactive graphics man/machine system for pattern recognition" (Stanford Research Institute, Menlo Park, Calif., Rep. No. RADC—TR—68—572, 1969).
[26] HARTIGAN, "Direct clustering of data matrices", *Journal of the American Stat. Ass.* **67** (1970) 123—129.
[27] HOFFER—SEVERANCE, "Use of cluster analysis in physical data bases" (Farmingham, Mass., 1975) 69—86.
[28] HORVÁTH, „Automatikus osztályozás", *Könyvtári Figyelő* **5** (1978).
[29] HORVÁTH—LEITNER—MEZEY—PONGRÁCZ, *A tanácsi információrendszer fejlesztése* (SZÁMALK, Budapest, 1983).
[30] "Business system planning", kézikönyv (IBM, 1972).
[31] JARDINE—SIBSON, *Mathematical Taxonomy* (Wiley, New York, 1971).
[32] KOVÁCS LÁSZLÓ BÉLA, Combinational Methods of Discrete Programming (Akadémiai Kiadó, Budapest, 1980).
[33] LANCE—WILLIAMS, "Computer programs for hierarchial polythetic classification. Similarity analysis" (1966).
[34] LANCE—WILLIAMS, "A general theory of classificatory sorting strategies. Clustering systems", *Computer Journal* **10** (1967) 271—277.
[35] LEE—TSENG, "Multikey sorting", *International Journal of Policy Analysis and Information Systems* **3** (1979) 1—20.
[36] LEE, "Clustering analysis and its applications", *Advances in Information Systems Science* **8** (1981) 153—158.
[37] LENSTRA, "Clustering a data array and the travelling salesman problem", *Operations Research* **20** (1970) 993—1009.
[38] LITOFSKY, "Utility of automatic classification for information storage and retrieval", ph. d. diss. (1969).
[39] MASUDA, "Optimization of program organization by cluster analysis", *Proceedings of IFIP* **74** (1979) 261—265.
[40] MCCORMICK—SCHWEITZER—WHITE, "Problem decomposition and data reorganization by a clustering technique", *Operations Research* **22** (1974) 413—414.
[41] NISHIGAKI—NOGI—MIYAMOTO, "Segments organization by cluster analysis", *Information Processing in Japan* **16** (1976) 153—158.
[42] ORBÁN, *Programozás Warnier-módszerrel* (SZÁMALK, Budapest, 1982).
[43] PÁRNICZKY, *A statisztikai informatika alapjai* (SKV, Budapest, 1976).
[44] PÁRNICZKY, „Doktori disszertáció", MTA TMB, Budapest, 1980.
[45] PRIM, "Shortest connection matrix network and some generalizations", *Bell System Technical Journal* **36** (1957) 3189—4401.
[46] RÓÓB, NIMANAL (NIMIGÜSZI, 1980).
[47] SALTON, *Automatic Information Storage and Retrieval* (McGraw—Hill, New York, 1968).
[48] SAMMON, "On-line pattern analysis and recognition system" (Rome Air Development Center. Griffiss Air-Force Basis, New York, Rep. No. RADGTR—68—263, 1968).
[49] SLAGLE—CHANG—HELLER, "Experiments with some clustering analysis algorithms", *Pattern Recognition* **6** (1974) 181—187.
[50] SLAGLE—CHANG—HELLER, "A clustering and datareorganization algorithm", *IEEE Transactions on Systems, Man and Cybernetics* **15** (1975).
[51] SÖRGEL, "Mathematical analysis of documentation system", *Information Storage Retrieval* **3** (1967) 129—173.
[52] SZIAM *Szintetikus adatmodellező, kézikönyv* (SZÁMOK, Budapest, 1981).
[53] SZIRA—MEZEY, „A megyei tanácsi számítógépes információs mintarendszer részletes koncepciója és intézkedési terve" (ÁSZI, 1981).
[54] TORGERSON, *Theory and Methods of Scaling* (John Wiley, New York, 1960).
[55] VAN RIJSBERGEN, *Information Retrieval* (Butterworths, London, 1979).
[56] WILLETT, "A fast procedure for the classification of similarity coefficients in automatic classification", *Information Processing and Management* **17** (1981) 53—60.
[57] WRIGHT, "An algorithm for computing correlation matrices", *Journal of the ASIS* **23** (1972) 130.

[58] YAO, "An O/E log log V) algorithm for finding minimal spanning trees", *Information Processing Letters* 4 (1975) 21—23.
[59] ZAHN, "Graph — theoretical methods for detecting and describing gestalt clusters", *IEEE Transactions on Computers* 20 (1971) 68—86.

(Beérkezett: 1983. szeptember 23.)

(Átdolgozva beérkezett: 1984. május 8.)

MEZEY GYULA
ORSZÁGOS MŰSZAKI INFORMÁCIÓS KÖZPONT ÉS KÖNYVTÁR
1428 BUDAPEST, PF. 12.

ПРОЕКТИРОВАНИЕ СЕТЕВОЙ БАЗЫ ДАННЫХ

Д. Мезеи

В статье рассматриваются вопросы проектирования сегментов и область (арэа) базы данных. Отдельные типы нормализованной и функциально проанализированной принципиальной модели данных преобразуются в рекорды либо в части рекордов (сегменты), затем рекорды (части рекордов) объединяются в большие группы (области арэа). Статья описывает также такие алгоритмы, которые используют методы кластерного анализа. Для построения сегментов использовались алгометративное иерархическое кластеризацие и неметрическое измерение расстояний. При проектировании областей использовался метод перегриннировки таблиц.

Alkalmazott Matematikai Lapok **11** (1985) 297—309

KONJUGÁLT IRÁNYOK ELŐÁLLÍTÁSA — A KONJUGÁLT PÁROK MÓDSZERE

HEGEDŰS CSABA J., BODÓCS LÁSZLÓ

Budapest

A dolgozat általános algoritmusokat tárgyal, amelyek tetszőleges téglalap alakú, valós vagy komplex elemű mátrixra nézve konjugált irányokat készítenek. A származtatás vetítések sorozatával történik, a konjugált irányok rekurzióval készülnek.

A kapott sémák lényegében tartalmazzák az összes korábbi rekurzív sémát, köztük *Lánczos módszerét*, amellyel kvadratikus mátrixok kontinuáns alakra hozhatók.

Az általánosítás előnyei: nagyobb a szabadsági fokok száma és lehetőség van a konjugált-irány módszerek egységes tárgyalására.

1. Bevezetés

Célunk olyan általános algoritmusok leírása, amelyek egy tetszőleges A mátrixra nézve konjugált irányokat állítanak elő. Az első ilyen típusú algoritmusokat HESTENES és STIEFEL [10], illetve HESTENES [9] készítették pozitív definit, hermitikus mátrixokra. Módszereik egy pozitív definit kvadratikus alak minimalizálásán alapultak. Az e dolgozatban leírt algoritmusok egymást követő projekciók alkalmazásával származtathatók. Ez a származtatási mód a konjugált-irány rekurziók általánosabb osztályához vezet, mely tartalmazza a korábbi módszereket, köztük *Lánczos módszerét*.

A 2. fejezetben definiáljuk a konjugált párok fogalmát, majd bevezetünk olyan projektorokat, amelyek segítségével a konjugált párok rekurzív módon állíthatók elő. A 3. fejezetben rövid történeti áttekintést adunk olyan ortogonalizációs eljárásokról, melyek vetítéssel származtathatók. A 4. fejezetben ismertetjük a rekurziós algoritmusok kapcsolatát a lineáris egyenletek, a szinguláris érték feladatok és a sajátérték feladatok megoldásával.

A mátrixokat nagybetűkkel, a vektorokat kisbetűkkel, a skalárokat a dimenziók és indexek kivételével görög kisbetűkkel jelöljük. Az n-dimenziós komplex vektorteret C^n jelöli, az $m \times n$ típusú komplex elemű mátrixok halmaza $C^{m,n}$. Az $\mathbf{A}$ mátrix transzponáltja $\mathbf{A}^T$, transzponált konjugáltja $\mathbf{A}^H$. A *Kronecker szimbólumot* δ_{ij} jelöli. A $\mathbf{v}_j$, $j=1, 2, \ldots, i$ vektorokat $\{\mathbf{v}_j\}_{j=1}^{i}$, s a $\mathbf{v}_j$, $\mathbf{u}_j$ vektorpárok rendszerét hasonlóan $\{\mathbf{v}_j, \mathbf{u}_j\}_{j=1}^{i}$ jelöléssel fogjuk megadni. Az n-dimenziós egységmátrix $\mathbf{I}_n$. A μ_j skalárokból készített diagonálmátrix $\langle \mu_j \rangle$.

2. A konjugált párok előállítása vetítéssel

Az $\mathbf{A}$ mátrixra nézve konjugált vektorpárok, röviden: az $\mathbf{A}$-konjugált párok fogalma az ortogonalitási tulajdonság általánosításaként keletkezett.

2.1. Definíció. Legyen $\mathbf{A} \in C^{m,n}$, $\mathbf{v}_j \in C^m$, $\mathbf{u}_j \in C^n$ és a $\{\mathbf{v}_j, \mathbf{u}_j\}_{j=1}^{i}$ rendszer elégítse ki az alábbi feltételt:

$$\mathbf{v}_j^H \mathbf{A} \mathbf{u}_k = \alpha_j \delta_{jk}, \quad \alpha_j \neq 0, \quad j, k = 1, 2, \ldots, i. \tag{2.1}$$

Ekkor a $\{\mathbf{v}_j, \mathbf{u}_j\}_{j=1}^{i}$ rendszert az A mátrixra nézve A-konjugált párok rendszerének fogjuk nevezni.

Megjegyezzük, hogy STEWART [17] egy kevésbé szigorú definíciót adott az A-konjugált párok fogalmára, nevezetesen az ő feltétele (2.1) helyett a következő volt:

$$\mathbf{v}_j^H \mathbf{A}\mathbf{u}_k = 0, \quad j < k. \tag{2.2}$$

E feltétel mellett az ortogonalitási relációknak csak mintegy fele teljesül. Jelen dolgozat szerzői úgy vélik, hogy a (2.2) feltételt kielégítő vektorrendszert inkább félig A-konjugált párok rendszerének kéne nevezni, mivel a két definíció között minőségi különbség van a következményeit tekintve. A (2.1) feltétel mellett — mint ezt később látni fogjuk — lehetőség van az A-konjugált párok rekurzív előállítására úgy, hogy legfeljebb négy vektor ismerete szükséges egy következő A-konjugált pár előállításához. Ugyanakkor a (2.2) feltételt kielégítő vektorok esetén valamennyi korábban előállított A-konjugált párra szükség van. Minthogy a konjugált-irány algoritmusok alkalmazása csak nagyméretű ritka mátrixok esetén kerülhet szóba, emiatt a módszerek közötti különbség jelentős.

A 2.1. definíció speciális A mátrixok esetén speciális vektorrendszerekhez vezet. Amennyiben A pozitív definit mátrix, és $\mathbf{v}_j = \mathbf{u}_j$ minden j-re, akkor HESTENES és STIEFEL [10] definíciója szerint A-ortogonális rendszert kapunk. Ha A egységmátrix, akkor biortogonális rendszert, ha pedig a $\mathbf{v}_j$, $\mathbf{u}_j$ vektorok is megegyeznek, akkor ortogonális rendszert kapunk.

Feltéve, hogy a $\{\mathbf{v}_j, \mathbf{u}_j\}_{j=1}^{i}$ rendszer A-konjugált párok rendszere, definiálhatjuk az alábbi két projektort:

$$\mathbf{P}_i^l = I_m - \sum_{j=1}^{i} \mathbf{A}\mathbf{u}_j \mathbf{v}_j^H / \mathbf{v}_j^H \mathbf{A}\mathbf{u}_j \tag{2.3}$$

és

$$\mathbf{P}_i^r = \mathbf{I}_n - \sum_{j=1}^{i} \mathbf{u}_j \mathbf{v}_j^H \mathbf{A} / \mathbf{v}_j^H \mathbf{A}\mathbf{u}_j, \tag{2.4}$$

ahol $\mathbf{I}_n$ és $\mathbf{I}_m$ egységmátrixok.

A projektorok alábbi tulajdonságai könnyen ellenőrizhetők:

$$\mathbf{P}_i^l = \prod_{j=1}^{i} (\mathbf{I}_m - \mathbf{A}\mathbf{u}_j \mathbf{v}_j^H / \mathbf{v}_j^H \mathbf{A}\mathbf{u}_j), \tag{2.5}$$

$$\mathbf{P}_i^r = \prod_{j=1}^{i} (\mathbf{I}_n - \mathbf{u}_j \mathbf{v}_j^H \mathbf{A} / \mathbf{v}_j^H \mathbf{A}\mathbf{u}_j), \tag{2.6}$$

ahol a tényezők felcserélhetők, továbbá

$$\mathbf{v}_i^H \mathbf{P}_k^l = 0, \quad \mathbf{v}_i^H \mathbf{A}\mathbf{P}_k^r = 0, \quad i \leqq k, \tag{2.7}$$

$$\mathbf{P}_k^r \mathbf{u}_i = 0, \quad \mathbf{P}_k^l \mathbf{A}\mathbf{u}_i = 0, \quad i \leqq k. \tag{2.8}$$

A (2.3) és a (2.4) projektorok felhasználhatók a $\{\mathbf{v}_j, \mathbf{u}_j\}_{j=1}^{i}$ rendszer bővítésére.

2.2. TÉTEL. ([7], [8]). Legyen $\{\mathbf{v}_j, \mathbf{u}_j\}_{j=1}^{i}$ A-konjugált párok rendszere és legyen $\mathbf{r}_{i+1} \in C^m$ és $\mathbf{q}_{i+1} \in C^n$ olyan, hogy

$$\mathbf{r}_{i+1}^H \mathbf{P}_i^l \mathbf{A}\mathbf{P}_i^r \mathbf{q}_{i+1} \neq 0. \tag{2.9}$$

Ekkor a

$$\mathbf{v}_{i+1}^H = \mathbf{r}_{i+1}^H \mathbf{P}_i^l \tag{2.10}$$

és

$$\mathbf{u}_{i+1} = \mathbf{P}_i^r \mathbf{q}_{i+1} \tag{2.11}$$

vektorok egy újabb **A**-konjugált párt képeznek.

Bizonyítás. A (2.7), (2.8) és (2.9) összefüggések felhasználásával elegendő a 2.1. definíciót ellenőrizni.

A következőkben megvizsgáljuk, hogy hány **A**-konjugált pár állítható elő. Szükségünk lesz az alábbi lemmára:

2.3. LEMMA. Legyen $\{\mathbf{v}_j, \mathbf{u}_j\}_{j=1}^i$ **A**-konjugált párok rendszere. Ekkor a $\{\mathbf{v}_j^H\}$, $\{\mathbf{u}_j\}$, $\{\mathbf{v}_j^H\mathbf{A}\}$ és az $\{\mathbf{A}\mathbf{u}_j\}$ rendszerek vektorai lineárisan függetlenek.

Bizonyítás. Indirekt módon tegyük fel, hogy például a $\{\mathbf{v}_j^H\}_{j=1}^i$ rendszer valamely $\mathbf{v}_k$ eleme az állítással ellentétben előállítható a többi elem lineáris kombinációjaként. Ekkor

$$\mathbf{v}_k^H \mathbf{A}\mathbf{u}_k = \sum_{j=1, j\neq k}^{i} \alpha_j \mathbf{v}_j^H \mathbf{A}\mathbf{u}_k = 0$$

ellentmond a 2.1. definíciónak, mert $\mathbf{v}_k^H\mathbf{A}\mathbf{u}_k$ nem lehet nulla. Hasonló módon látható be a lineáris függetlenség a többi rendszernél is.

2.4. Definíció. Az **A**-konjugált párok rendszerét teljesnek nevezzük, ha a maximális számú vektort tartalmazza.

2.5. TÉTEL. Egy adott **A** mátrix esetén legfeljebb annyi **A**-konjugált pár állítható elő, amennyi a mátrix rangja. Ebben az esetben teljes **A**-konjugált rendszert és az **A** mátrix egy minimális diadikus felbontását kapjuk.

Bizonyítás. A (2.9) feltételből következik, hogy nem állítható elő új **A**-konjugált pár, ha valamely ϱ egész számra teljesül, hogy $\mathbf{P}_\varrho^l \mathbf{A} \mathbf{P}_\varrho^r = 0$, ahonnan átrendezéssel

$$\mathbf{A} = \sum_{j=1}^{\varrho} \mathbf{A}\mathbf{u}_j \mathbf{v}_j^H \mathbf{A} / \mathbf{v}_j^H \mathbf{A}\mathbf{u}_j. \tag{2.13}$$

Minthogy a 2.3. lemma szerint az $\{\mathbf{A}\mathbf{u}_j\}$ és $\{\mathbf{v}_j^H\mathbf{A}\}$ vektorok lineárisan függetlenek, így az **A** mátrix egy minimális diadikus előállításához jutottunk, amiből következik, hogy az **A** mátrix rangja $\varrho(\mathbf{A}) = \varrho$.

Bevezetve a

$$\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_\varrho], \quad \mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_\varrho] \tag{2.14}$$

oszlopvektorokból készített mátrixokat, (2.13) az alábbi alakban is felírható:

$$\mathbf{A} = \mathbf{A}\mathbf{U}(\mathbf{V}^H\mathbf{A}\mathbf{U})^{-1}\mathbf{V}^H\mathbf{A}, \tag{2.15}$$

ahol $\mathbf{V}^H\mathbf{A}\mathbf{U}$ diagonális mátrix. Az $\mathbf{X} = \mathbf{U}(\mathbf{V}^H\mathbf{A}\mathbf{U})^{-1}\mathbf{V}^H$ mátrix az **A** mátrix (1,2)-es (reflexív) általánosított inverze, mivel $\mathbf{A}\mathbf{X}\mathbf{A} = \mathbf{A}$ és $\mathbf{X}\mathbf{A}\mathbf{X} = \mathbf{X}$ egyaránt teljesülnek.

A következőkben rátérünk olyan algoritmusok ismertetésére, amelyeknél egyszerű rekurzióval készíthetők **A**-konjugált párok. Az algoritmusokat olyan általános

formában adjuk meg, ahol az egyes paraméterek speciális választásával különféle variánsok készíthetők.

2.6. Algoritmus. Tetszőleges nemzérus $\mathbf{r}_0 \in C^m$ és $\mathbf{q}_0 \in C^n$ kezdővektorokkal készítsük el $i=0, 1, \ldots$ értékeire az

$$\mathbf{r}_{i+1} = \mathbf{P}_i^l \mathbf{r}_i, \quad \mathbf{q}_{i+1}^H = \mathbf{q}_i^H \mathbf{P}_i^r, \tag{2.16}$$

$$\mathbf{v}_{i+1}^H = \mu_{i+1} \mathbf{r}_{i+1}^H \mathbf{C} \mathbf{P}_i^l, \quad \mathbf{u}_{i+1} = \lambda_{i+1} \mathbf{P}_i^r \mathbf{K} \mathbf{q}_{i+1} \tag{2.17}$$

vektorokat, ahol $\mathbf{A} \in C^{m,n}$, $\mathbf{C} \in C^{m,m}$, $\mathbf{K} \in C^{n,n}$, $\mathbf{K}$ és $\mathbf{C}$ hermitikus, pozitív definit mátrixok, μ_i és λ_i nemzérus skalárok, $\mathbf{P}_0^l = \mathbf{I}_m$ és $\mathbf{P}_0^r = \mathbf{I}_n$.

Az algoritmusban a $\mathbf{K}$ és $\mathbf{C}$ mátrixok, valamint a μ_i, λ_i skalárok egyébként szabadon választhatók. Az előbbiek kondícionáló mátrixokként, az utóbbiak pedig a konjugált vektorok hosszának skálázására használhatók. Ha a (2.17) összefüggéseket összehasonlítjuk (2.10) és (2.11)-gyel, akkor látható, hogy ebben az előállításban a $\mu_i \mathbf{r}_i^H \mathbf{C}$ és $\lambda_i \mathbf{K} \mathbf{q}_i$ alapvektorok szerepelnek, ahol $\mathbf{r}_i$ és $\mathbf{q}_i$ szintén vetítéssel álltak elő.

A következőkben megmutatjuk, hogy a (2.16) és (2.17) összefüggések jelentősen egyszerűsödnek. A levezetést párhuzamosan készítjük a jobb és bal oldali vektorokra azzal a feltételezéssel, hogy (2.9) teljesül.

Minthogy elegendő a projektorok (2.5) és (2.6) előállításában az utolsó tényezőt venni, így (2.16) a következő alakra egyszerűsíthető:

$$\mathbf{r}_{i+1} = \mathbf{r}_i - \mathbf{A}\mathbf{u}_i \mathbf{v}_i^H \mathbf{r}_i / \mathbf{v}_i^H \mathbf{A}\mathbf{u}_i, \quad \mathbf{q}_{i+1}^H = \mathbf{q}_i^H - \mathbf{v}_i^H \mathbf{A} \mathbf{q}_i^H \mathbf{u}_i / \mathbf{v}_i^H \mathbf{A}\mathbf{u}_i. \tag{2.18}$$

Ugyanakkor (2.17) felhasználásával valamely β_{ij} és γ_{ij} skalárokkal

$$\mathbf{r}_i^H \mathbf{C} = \sum_{j=1}^{i} \beta_{ij} \mathbf{v}_j^H, \quad \mathbf{K}\mathbf{q}_i = \sum_{j=1}^{i} \gamma_{ij} \mathbf{u}_j \tag{2.19}$$

illetve (2.7) és (2.8) alkalmazásával

$$\mathbf{v}_i^H \mathbf{r}_k = \mathbf{v}_i^H \mathbf{P}_{k-1}^l \mathbf{r}_{k-1} = 0, \quad \mathbf{q}_k^H \mathbf{u}_i = \mathbf{q}_{k-1}^H \mathbf{P}_{k-1}^r \mathbf{u}_i = 0, \quad i < k \tag{2.20}$$

adódik. A (2.19) és (2.20) összefüggésekből

$$\mathbf{r}_i^H \mathbf{C} \mathbf{r}_k = \sum_{j=1}^{i} \beta_{ij} \mathbf{v}_j^H \mathbf{r}_k = 0, \quad \mathbf{q}_k^H \mathbf{K} \mathbf{q}_i = \sum_{j=1}^{i} \gamma_{ij} \mathbf{q}_k^H \mathbf{u}_j = 0 \tag{2.21}$$

következik $i<k$-ra és a szimmetria miatt ezek az összefüggések $k<i$-re is igazak maradnak, kaptuk tehát, hogy a $\mathbf{q}_j$ vektorok K-ortogonális, az $\mathbf{r}_j$ vektorok pedig $\mathbf{C}$-ortogonális rendszert képeznek.

A (2.19) összefüggésekkel (2.18) megfelelő formuláit beszorozva az ortogonalitási tulajdonságok alkalmazásával nyerhetők az alábbi összefüggések, figyelembe véve, hogy $\beta_{ii} = 1/\mu_i$ és $\gamma_{ii} = 1/\lambda_i$:

$$\begin{aligned} \mu_i \mathbf{r}_i^H \mathbf{C} \mathbf{r}_i &= \mu_i \mathbf{r}_i^H \mathbf{C} \mathbf{A}\mathbf{u}_i \mathbf{v}_i^H \mathbf{r}_i / \mathbf{v}_i^H \mathbf{A}\mathbf{u}_i = \mathbf{v}_i^H \mathbf{r}_i, \\ \lambda_i \mathbf{q}_i^H \mathbf{K} \mathbf{q}_i &= \lambda_i \mathbf{v}_i^H \mathbf{A} \mathbf{K} \mathbf{q}_i \mathbf{q}_i^H \mathbf{u}_i / \mathbf{v}_i^H \mathbf{A}\mathbf{u}_i = \mathbf{q}_i^H \mathbf{u}_i, \end{aligned} \tag{2.22}$$

valamint (2.18)-ból helyettesítéssel adódik:

$$\mathbf{r}_k^H \mathbf{C}\mathbf{A}\mathbf{u}_j/\mathbf{v}_j^H \mathbf{A}\mathbf{u}_j = \mathbf{r}_k^H \mathbf{C}(\mathbf{r}_j - \mathbf{r}_{j+1})/\mathbf{v}_j^H \mathbf{r}_j,$$

(2.23)

$$\mathbf{v}_j^H \mathbf{A}\mathbf{K}\mathbf{q}_k/\mathbf{v}_j^H \mathbf{A}\mathbf{u}_j = (\mathbf{q}_j^H - \mathbf{q}_{j+1}^H)\mathbf{K}\mathbf{q}_k/\mathbf{q}_j^H \mathbf{u}_j.$$

Így a (2.17) formulák az alábbi alakra redukálhatók az ortogonalitási tulajdonságok segítségével:

$$\mathbf{v}_{i+1}^H = \mu_{i+1}\mathbf{r}_{i+1}^H \mathbf{C}\Big(\mathbf{I}_m - \sum_{j=1}^{i} \mathbf{A}\mathbf{u}_j \mathbf{v}_j^H/\mathbf{v}_j^H \mathbf{A}\mathbf{u}_j\Big) = \mu_{i+1}\mathbf{r}_{i+1}^H \mathbf{C} + \mathbf{v}_i^H \mu_{i+1}\mathbf{r}_{i+1}^H \mathbf{C}\mathbf{r}_{i+1}/(\mu_i \mathbf{r}_i^H \mathbf{C}\mathbf{r}_i),$$

(2.24)

$$\mathbf{u}_{i+1} = \lambda_{i+1}\Big(\mathbf{I}_n - \sum_{j=1}^{i} \mathbf{u}_j \mathbf{v}_j^H \mathbf{A}/\mathbf{v}_j^H \mathbf{A}\mathbf{u}_j\Big)\mathbf{K}\mathbf{q}_{i+1} = \lambda_{i+1}\mathbf{K}\mathbf{q}_{i+1} + \mathbf{u}_i \lambda_{i+1}\mathbf{q}_{i+1}^H \mathbf{K}\mathbf{q}_{i+1}/(\lambda_i \mathbf{q}_i^H \mathbf{K}\mathbf{q}_i).$$

Vezessük be a $\|\mathbf{q}_i\|_{\mathbf{K}}^2 = \mathbf{q}_i^H \mathbf{K}\mathbf{q}_i$ és $\|\mathbf{r}_i\|_{\mathbf{C}}^2 = \mathbf{r}_i^H \mathbf{C}\mathbf{r}_i$ energetikai normákat, így eredményünket az alábbi tételben fogalmazhatjuk meg:

2.7. TÉTEL. Feltéve, hogy a nevezők nem nullák, a 2.6 algoritmusban adott rekurziók $i>0$ esetben a következő alakra egyszerűsíthetők:

$$\mathbf{r}_{i+1} = \mathbf{r}_i - \mathbf{A}\mathbf{u}_i \mu_i \|\mathbf{r}_i\|_{\mathbf{C}}^2/\mathbf{v}_i^H \mathbf{A}\mathbf{u}_i,$$

$$\mathbf{q}_{i+1}^H = \mathbf{q}_i^H - \mathbf{v}_i^H \mathbf{A}\lambda_i \|\mathbf{q}_i\|_{\mathbf{K}}^2/\mathbf{v}_i^H \mathbf{A}\mathbf{u}_i,$$

(2.25)

$$\mathbf{v}_{i+1} = \bar{\mu}_{i+1}\mathbf{C}\mathbf{r}_{i+1} + \mathbf{v}_i \bar{\mu}_{i+1}\|\mathbf{r}_{i+1}\|_{\mathbf{C}}^2/(\bar{\mu}_i \|\mathbf{r}_i\|_{\mathbf{C}}^2),$$

$$\mathbf{u}_{i+1} = \lambda_{i+1}\mathbf{K}\mathbf{q}_{i+1} + \mathbf{u}_i \lambda_{i+1}\|\mathbf{q}_{i+1}\|_{\mathbf{K}}^2/(\lambda_i \|\mathbf{q}_i\|_{\mathbf{K}}^2),$$

ahol az $\mathbf{r}_i$ vektorok $\mathbf{C}$-ortogonálisak, a $\mathbf{q}_i$ vektorok $\mathbf{K}$-ortogonálisak és a $\mathbf{v}_i$, $\mathbf{u}_i$ vektorok $\mathbf{A}$-konjugált párokat képeznek. Néhány választás λ_i és μ_i-re:

$$\text{(I)} \quad \lambda_i = 1, \qquad \mu_i = 1,$$

(2.26)

$$\text{(II)} \quad \lambda_i = 1/\|\mathbf{q}_i\|_{\mathbf{K}}, \quad \mu_i = 1/\|\mathbf{r}_i\|_{\mathbf{C}},$$

$$\text{(III)} \quad \lambda_i = 1/\|\mathbf{q}_i\|_{\mathbf{K}}^2, \quad \mu_i = 1/\|\mathbf{r}_i\|_{\mathbf{C}}^2.$$

Amennyiben az $\mathbf{A}$ mátrix hermitikus, $\lambda_i = \mu_i$, $\mathbf{C}=\mathbf{K}$ és az $\mathbf{r}_0$, valamint a $\mathbf{q}_0$ kezdővektorok megegyeznek, akkor (2.25) első és második, illetve harmadik és negyedik formulája megegyezik, és az összefüggések $\lambda_i = \mu_i = 1$ mellett ugyanazok lesznek, mint HESTENES [9] általánosított konjugált gradiens módszerénél. Ha $\mathbf{K}=\mathbf{C}=\mathbf{I}$ és $\lambda_i = \mu_i = 1$ minden i indexre, akkor a jólismert konjugált gradiens rekurzióhoz jutunk, ezért a (2.25) rekurziót *Hestenes—Stiefel féle rekurziónak* fogjuk nevezni. A skálázó paramétereket (III) szerint választva kapjuk a legegyszerűbb alakot. Ezt a konjugált gradiens módszerre már HESTENES és STIEFEL is észrevették [10, 9. fejezet].

A következőkben még röviden ismertetünk egy rekurziós sémát, amely ugyancsak konjugált irányokat állít elő.

2.8. Algoritmus. Legyen $\mathbf{A}\in C^{m,n}$, $\mathbf{B}\in C^{n,m}$, $\mathbf{0}\neq\mathbf{r}_0\in C^m$, és $\mathbf{0}\neq\mathbf{q}_0\in C^n$; $\mathbf{r}_0$, $\mathbf{q}_0$ egyébként tetszőleges induló vektorok, valamint λ_i, μ_i tetszőleges nemzérus számok.

Készítsük el $i=0, 1, \ldots$ értékeire az alábbi vektorokat:

$$\begin{aligned} &\mathbf{r}_{i+1} = \mathbf{P}_i^l \mathbf{r}_i, \quad \mathbf{q}_{i+1}^H = \mathbf{q}_i^H \mathbf{P}_i^r, \\ &\mathbf{v}_{i+1}^H = \mu_{i+1} q_{i+1}^H \mathbf{B}\mathbf{P}_i^l, \quad \mathbf{u}_{i+1} = \lambda_{i+1} \mathbf{P}_i^r \mathbf{B}\mathbf{r}_{i+1}, \end{aligned} \tag{2.27}$$

ahol $\mathbf{P}_i^l$ és $\mathbf{P}_i^r$ a (2.3) és a (2.4) által definiált projektorokat jelölik.

Ekkor hasonló fogásokkal, mint a 2.6. algoritmusnál, bebizonyítható a

2.9. TÉTEL. A 2.8. algoritmus rekurziói nullánál nagyobb indexekre az alábbi összefüggésekre egyszerűsíthetők, feltéve, hogy a nevezők nem nullák:

$$\begin{aligned} \mathbf{r}_{i+1} &= \mathbf{r}_i - \mathbf{A}\mathbf{u}_i \mu_i \mathbf{q}_i^H \mathbf{B}\mathbf{r}_i / \mathbf{v}_i^H \mathbf{A}\mathbf{u}_i, \\ \mathbf{q}_{i+1}^H &= \mathbf{q}_i^H - \mathbf{v}_i^H \mathbf{A} \lambda_i \mathbf{q}_i^H \mathbf{B}\mathbf{r}_i / \mathbf{v}_i^H \mathbf{A}\mathbf{u}_i, \\ \mathbf{v}_{i+1}^H &= \mu_{i+1} \mathbf{q}_{i+1}^H \mathbf{B} + \mathbf{v}_i^H \mu_{i+1} \mathbf{q}_{i+1}^H \mathbf{B}\mathbf{r}_{i+1} / (\mu_i \mathbf{q}_i^H \mathbf{B}\mathbf{r}_i), \\ \mathbf{u}_{i+1} &= \lambda_{i+1} \mathbf{B}\mathbf{r}_{i+1} + \mathbf{u}_i \lambda_{i+1} \mathbf{q}_{i+1}^H \mathbf{B}\mathbf{r}_{i+1} / (\lambda_i \mathbf{q}_i^H \mathbf{B}\mathbf{r}_i). \end{aligned} \tag{2.28}$$

A $\mathbf{q}_i$, $\mathbf{r}_i$ vektorok $\mathbf{B}$-konjugált párok, a $\mathbf{v}_i$, $\mathbf{u}_i$ vektorok pedig $\mathbf{A}$-konjugált párok.

Speciális esetben, ha $\mathbf{B}=\mathbf{A}^H$, akkor ez a rekurzió egyszerre két $\mathbf{A}$-konjugált rendszert készít. Ha a mátrixok kvadratikusak és $\mathbf{B}$ egységmátrix, akkor megmutatható, hogy a rekurziós formulák ekvivalensek *Lánczos módszerével,* emiatt ezt a rekurziót *Lánczos-féle rekurziónak* fogjuk nevezni. A bizonyítást a 4. fejezetben tárgyaljuk. Mint látható, a *Lánczos-féle rekurzió* legegyszerűbb alakját akkor veszi fel, ha a skálázó paramétereket $\lambda_i=\mu_i=1/\mathbf{q}_i^H\mathbf{B}\mathbf{r}_i$-nek választjuk.

Megjegyezzük még, hogy a tárgyalt sémák tovább általánosíthatók az $\mathbf{A}$, $\mathbf{A}^H\mathbf{A}$ és $\mathbf{A}\mathbf{A}^H$ mátrixok hatványainak bevezetésével.

A kapott rekurziókkal kapcsolatban több általános kérdés is felvethető. Az egyik ilyen, hogy biztosítható-e, hogy a nevező ne váljon idő előtt nullává? A válasz a *Hestenes—Stiefel séma* esetén igenlő, s egy további cikkben megmutatjuk, hogy a *Hestenes—Stiefel rekurziók* mindig módosíthatók úgy, hogy az utolsónak kapott $\mathbf{v}_i$ és $\mathbf{u}_i$ vektorok egyszerre érkezzenek az $\mathbf{A}^H$, illetve az $\mathbf{A}$ mátrix nullterébe. A *Lánczos-féle rekurzió* esetében ilyen általános választ jelenleg nem ismerünk, de megadhatók olyan speciális esetek, amikor a nullosztó elkerülhető.

Mindkét módszer felhasználható lineáris egyenletrendszer megoldására. Kimutatható továbbá, hogy a *Hestenes—Stiefel típusú rekurzió* a szinguláris érték dekompozíció feladatával rokon, míg a *Lánczos-séma* a mátrixok sajátértékfeladatával van összefüggésben. A 4. fejezetben ezeket meg fogjuk mutatni.

A legfontosabb kérdés a stabilitás kérdése, nevezetesen: a rekurzív generálás során az egyes ortogonális és konjugált vektorrendszerek elemei mennyire romlanak el. Az közismert, hogy a *Lánczos-módszer* vektorainál a hibák elég gyorsan felszaporodnak, s emiatt csak speciális esetben, szimmetrikus mátrixok esetén szokták a módszert használni. Numerikus vizsgálataink során azt tapasztaltuk, hogy a *Hestenes—Stiefel típusú rekurzió* is kevéssé stabil, emiatt jelen formájukban egyik módszert sem ajánljuk, mint megbízható és garantáltan működő algoritmust. Azonban léteznek olyan indítóvektorok és rekurzió variánsok, amelyek mellett a stabilitás fokozható. Ezeket az eredményeket külön cikkben fogjuk közölni.

3. A vetítéseken alapuló ortogonalizációs algoritmusok fejlődése

Az előző fejezetben bemutatott konjugált-irány rekurziók speciális esetekben egyszerűbb ortogonalizációs eljárásokat adnak. A projekciós elv tükrében érdekes és tanulságos áttekinteni az ortogonalizációs módszerek fejlődését.

A *Gram—Schmidt ortogonalizáció* e század elején született. Eszerint a $\{\mathbf{q}_j\}$ lineárisan független vektorrendszerből vetítések sorozatával készítünk egy $\{\mathbf{u}_j\}$ ortogonális rendszert:

$$\mathbf{u}_1 = \mathbf{q}_1, \quad \mathbf{u}_{i+1} = \Big(\mathbf{I} - \sum_{j=1}^{i} \mathbf{u}_j \mathbf{u}_j^H / \mathbf{u}_j^H \mathbf{u}_j\Big)\mathbf{q}_{i+1}. \tag{3.1}$$

A (3.1) formula $\mathbf{A}=\mathbf{I}$ esetére (2.10) és (2.11) speciális alakja, amikor a két oldal egybeesik.

Ha $\mathbf{A}$ továbbra is egységmátrix, de adott két lineárisan független rendszer $\{\mathbf{r}_j\}$ és $\{\mathbf{q}_j\}$, akkor (2.10) és (2.11) ezekből biortogonális rendszert készít:

$$\begin{gathered} \mathbf{v}_1 = \mathbf{r}_1, \quad \mathbf{v}_{i+1}^H = \mathbf{r}_{i+1}^H\Big(\mathbf{I} - \sum_{j=1}^{i} \mathbf{u}_j \mathbf{v}_j^H / \mathbf{v}_j^H \mathbf{u}_j\Big), \\ \mathbf{u}_1 = \mathbf{q}_1, \quad \mathbf{u}_{i+1} = \Big(\mathbf{I} - \sum_{j=1}^{i} \mathbf{u}_j \mathbf{v}_j^H / \mathbf{v}_j^H \mathbf{u}_j\Big)\mathbf{q}_{i+1}. \end{gathered} \tag{3.2}$$

Ennek a formulának azonban az alkalmazások során nem volt különösebb jelentősége.

A fejlődésben a következő lépést FOX—HUSKEY—WILKINSON [5] munkája jelentette, akik pozitív definit, hermitikus $\mathbf{A}$ mátrixra nézve készítettek $\mathbf{A}$-ortogonális vektorokat egy $\{\mathbf{q}_j\}$ lineárisan független rendszer elemeiből. Ehhez (3.1)-ben annyi módosítást kell csinálni, hogy az $\mathbf{A}$ mátrixot be kell írni a projektor kifejezésébe:

$$\mathbf{u}_1 = \mathbf{q}_1, \quad \mathbf{u}_{i+1} = \Big(\mathbf{I} - \sum_{j=1}^{i} \mathbf{u}_j \mathbf{u}_j^H \mathbf{A} / \mathbf{u}_j^H \mathbf{A}\mathbf{u}_j\Big)\mathbf{q}_{i+1}. \tag{3.3}$$

A kapott formula ismét (2.10) és (2.11) speciális alakja.

A módszerek fejlődésében a következő lépés LÁNCZOS [12], [13] módszerének megjelenése, amely egy általános mátrixot kontinuáns alakra hoz. A módszer első megjelenési alakjából nem ismerhető fel a (2.27) előállítás. Azonban FLETCHER [4] megmutatta, hogy a *Lánczos módszer* formulái átírhatók a (2.28) alakra, ahol $\mathbf{A}$ kvadratikus mátrix, $\mathbf{B}=\mathbf{I}$ és $\lambda_i=\mu_i=1$. A hasonlósági transzformációval kapott kontinuáns mátrix nemzérus elemei a fellépő belső szorzatok értékeiből állíthatók össze, amint azt a 4. fejezetben mi is megmutatjuk. A *Lánczos módszer* mögött így

a következő vetítések fedezhetők fel:

$$\mathbf{v}_1 = \mathbf{q}_1,$$

$$\mathbf{q}_{i+1}^H = \mathbf{q}_i^H\Big(\mathbf{I}_n - \sum_{j=1}^{i} \mathbf{u}_j \mathbf{v}_j^H \mathbf{A}/\mathbf{v}_j^H \mathbf{A}\mathbf{u}_j\Big),$$

$$\mathbf{v}_{i+1}^H = \mathbf{q}_{i+1}^H\Big(\mathbf{I}_n - \sum_{j=1}^{i} \mathbf{A}\mathbf{u}_j \mathbf{v}_j^H/\mathbf{v}_j^H \mathbf{A}\mathbf{u}_j\Big),$$

(3.4)
$$\mathbf{u}_1 = \mathbf{r}_1,$$

$$\mathbf{r}_{i+1} = \Big(I_n - \sum_{j=1}^{i} \mathbf{A}\mathbf{u}_j \mathbf{v}_j^H/\mathbf{v}_j^H \mathbf{A}\mathbf{u}_j\Big)\mathbf{r}_i,$$

$$\mathbf{u}_{i+1} = \Big(\mathbf{I}_n - \sum_{j=1}^{i} \mathbf{u}_j \mathbf{v}_j^H \mathbf{A}/\mathbf{v}_j^H \mathbf{A}\mathbf{u}_j\Big)\mathbf{r}_{i+1},$$

ahol **A** kvadratikus n-edrendű mátrix. Összehasonlítva (3.2)-vel látjuk, hogy a projektorokban megjelenik az **A** mátrix, és az alapvektorok is vetítéssel készülnek. Itt a $\mathbf{v}_j$, $\mathbf{u}_j$ vektorok **A**-konjugáltak, a $\mathbf{q}_j$, $\mathbf{r}_j$ vektorok pedig biortogonális rendszert alkotnak. A (2.28) rekurziókat *O'Leary* [14] is megadta arra az esetre, amikor **A** és **B** kvadratikus mátrixok.

A jólismert konjugált-gradiens (illetve konjugált-irány) módszert HESTENES és STIEFEL [10] közölték egy átfogó dolgozatban. Az ő formulájuk a következő alakkal ekvivalens:

$$\mathbf{u}_1 = \mathbf{q}_1,$$

(3.5)
$$\mathbf{q}_{i+1} = \Big(\mathbf{I}_n - \sum_{j=1}^{i} \mathbf{A}\mathbf{u}_j \mathbf{u}_j^H/\mathbf{u}_j^H \mathbf{A}\mathbf{u}_j\Big)\mathbf{q}_i,$$

$$\mathbf{u}_{i+1} = \Big(\mathbf{I}_n - \sum_{j=1}^{i} \mathbf{u}_j \mathbf{u}_j^H \mathbf{A}/\mathbf{u}_j^H \mathbf{A}\mathbf{u}_j\Big)\mathbf{q}_{i+1},$$

ahol a különbség (3.3)-hoz képest annyi, hogy a $\{\mathbf{q}_i\}$ bázisvektorokat vetítésekkel generáljuk. Ez az előállítás (2.16) és (2.17) speciális esete, amikor **A** hermitikus, pozitív definit, n-edrendű mátrix, $\mathbf{K}=\mathbf{C}=\mathbf{I}_n$, $\lambda_i=\mu_i=1$ minden i-re és a jobb és a bal oldal egybeesik. A kapott $\{\mathbf{q}_j\}$ rendszer ortogonális, az $\{\mathbf{u}_j\}$ rendszer pedig **A**-ortogonális, és a rekurziók (2.25) szerint egyszerűsödnek. Mint látható, a (3.5) összefüggések (3.4) egy speciális alakjának tekinthetők.

A módszer következő bővítése HESTENES [9] nevéhez fűződik. HESTENES (3.5)-öt egy hermitikus, pozitív definit **K** mátrix hozzávételével kibővítette:

$$\mathbf{u}_1 = \mathbf{K}\mathbf{q}_1,$$

$$\mathbf{q}_{i+1} = \Big(\mathbf{I}_n - \sum_{j=1}^{i} \mathbf{A}\mathbf{u}_j \mathbf{u}_j^H / \mathbf{u}_j^H \mathbf{A}\mathbf{u}_j\Big)\mathbf{q}_i, \tag{3.6}$$

$$\mathbf{u}_{i+1} = \Big(\mathbf{I}_n - \sum_{j=1}^{i} \mathbf{u}_j \mathbf{u}_j^H \mathbf{A} / \mathbf{u}_j^H \mathbf{A}\mathbf{u}_j\Big)\mathbf{K}\mathbf{q}_{i+1}.$$

Így a $\{\mathbf{q}_j\}$ rendszer **K**-ortogonális, az $\{\mathbf{u}_j\}$ rendszer pedig **A**-ortogonális lesz.

A *Hestenes—Stiefel típusú konjugált párok módszerét* HEGEDŰS [7] publikálta arra az esetre, amikor **K** és **C** egységmátrixok és $\lambda_i = \mu_i = 1$ minden i-re. A **K** és **C** mátrixokkal való bővítést BODÓCS [3] diplomamunkája tartalmazza. Az itt ismertetett eredményeket a szerzők egy angol nyelvű közleményben is közreadták [8]. Végezetül megemlítjük, hogy ABAFFY, BROYDEN és SPEDICATO lineáris egyenletrendszerre egy újabb módszerosztályt dolgoztak ki, amelyhez tartozó algoritmust *ABS-algoritmusnak* nevezik [1]. Ezt ABAFFY és SPEDICATO egy újabb dolgozatukban általánosították [2], amelyben kimutatják az összefüggést az *ABSg-algoritmus* és az itt közölt algoritmusok között.

4. Néhány további összefüggés

Ebben a szakaszban megmutatjuk, hogy a lineáris egyenletrendszerek megoldása, a mátrixok sajátértékproblémája és a szinguláris érték dekompozíció feladata kapcsolatban állnak a konjugált irány algoritmusokkal. Elsőként a lineáris egyenletrendszerek megoldásával foglalkozunk.

4.1. TÉTEL. ([7]). Legyen $\mathbf{A} \in C^{m,n}$, $\varrho(\mathbf{A}) = \varrho$, $\{\mathbf{v}_j, \mathbf{u}_j\}_{j=1}^{\varrho}$ teljes **A**-konjugált rendszer, $\mathbf{x}_1 \in C^n$ és $\mathbf{r}_1 = \mathbf{b} - \mathbf{A}\mathbf{x}_1$, ahol $\mathbf{b} \in C^m$. Akkor az

$$\mathbf{A}\mathbf{x} = \mathbf{b} \tag{4.1}$$

lineáris egyenletrendszer akkor és csak akkor konzisztens, ha $\mathbf{P}_\varrho^l \mathbf{r}_1 = \mathbf{0}$. Ekkor a megoldások

$$\mathbf{x} = \mathbf{x}_1 + \sum_{j=1}^{\varrho} \mathbf{u}_j \mathbf{v}_j^H \mathbf{r}_1 / \mathbf{v}_j^H \mathbf{A}\mathbf{u}_j + \mathbf{P}_\varrho^r \mathbf{t} \tag{4.2}$$

alakban állíthatók elő, ahol $\mathbf{t} \in C^n$ tetszőleges vektor.

Bizonyítás. Teljes **A**-konjugált rendszer esetén (2.13)-ból következik, hogy $\mathbf{P}_\varrho^l \mathbf{A} = \mathbf{0}$, innen $\mathbf{P}_\varrho^l \mathbf{A}\mathbf{x} = \mathbf{P}_\varrho^l \mathbf{b} = \mathbf{0}$, és hasonlóan $\mathbf{P}_\varrho^l \mathbf{r}_1 = \mathbf{P}_\varrho^l (\mathbf{b} - \mathbf{A}\mathbf{x}_1) = \mathbf{0}$ is szükséges következmény. Másrészt $\mathbf{P}_\varrho^l \mathbf{r}_1 = \mathbf{0}$ elégséges, mert (2.3) alapján

$$\mathbf{r}_1 = \mathbf{b} - \mathbf{A}\mathbf{x}_1 = \mathbf{A} \sum_{j=1}^{\varrho} \mathbf{u}_j \mathbf{v}_j^H \mathbf{r}_1 / \mathbf{v}_j^H \mathbf{A}\mathbf{u}_j, \tag{4.3}$$

ahonnan

$$\mathbf{b} = \mathbf{A}\Big[\mathbf{x}_1 + \sum_{j=1}^{\varrho} \mathbf{u}_j \mathbf{v}_j^H \mathbf{r}_1 / \mathbf{v}_j^H \mathbf{A}\mathbf{u}_j\Big]. \tag{4.4}$$

Létezik tehát egy partikuláris megoldás. Az általános megoldást (4.2) adja, mivel $\mathbf{P}_Q^r t$ az $\mathbf{Ax}=\mathbf{0}$ homogén egyenlet általános megoldása, l. [16, 1.10.2. tétel].

Megjegyzések. A (4.1) egyenletrendszer egy partikuláris megoldásának előállításához az is elegendő, ha valamely $i<\varrho(\mathbf{A})$ indexre $\mathbf{P}_i^l\mathbf{r}_1=\mathbf{0}$, minthogy már ennek teljesüléséből is következik (4.3), illetve (4.4), ahol ϱ helyett most i írható.

Definiáljuk a további $\mathbf{x}_i$-ket az

$$\mathbf{x}_i = \mathbf{x}_{i-1}+\mathbf{u}_{i-1}(\mathbf{v}_{i-1}^H\mathbf{r}_1/\mathbf{v}_{i-1}^H\mathbf{A}\mathbf{u}_{i-1}) \tag{4.5}$$

összefüggéssel, ekkor a további reziduum-vektorokra kapjuk:

$$\mathbf{r}_i = \mathbf{b}-\mathbf{A}\mathbf{x}_i = \mathbf{b}-\mathbf{A}\Big(\mathbf{x}_1+\sum_{j=1}^{i-1}\mathbf{u}_j\mathbf{v}_j^H\mathbf{r}_1/\mathbf{v}_j^H\mathbf{A}\mathbf{u}_j\Big) = \mathbf{P}_{i-1}^l\mathbf{r}_1 = \mathbf{P}_{i-1}^l\mathbf{r}_{i-1}. \tag{4.6}$$

Ebből látható, hogy a (2.16)—(2.17) vagy a (2.27) rekurziókat alkalmazva a reziduum-vektorok és a rekurziók $\mathbf{r}_i$ vektorai mind a *Hestenes—Stiefel típusú,* mind a *Lánczos típusú módszernél* egybeejthetők.

A rekurzió során az $\mathbf{r}_1$ vektort nem szükséges megőrizni, mert a $\mathbf{v}_i^H\mathbf{r}_1$ skalárszorzat átírható, felhasználva $\mathbf{v}_i$ definícióját. A *Hestenes—Stiefel típusú módszernél* az eredmény:

$$\mathbf{v}_i^H\mathbf{r}_1 = \mu_i\mathbf{r}_i^H\mathbf{C}\mathbf{P}_{i-1}^l\mathbf{r}_1 = \mu_i\|\mathbf{r}_i\|_{\mathbf{C}}^2, \tag{4.7}$$

a *Lánczos típusú módszernél* pedig

$$\mathbf{v}_i^H\mathbf{r}_1 = \mu_i\mathbf{q}_i^H\mathbf{B}\mathbf{P}_{i-1}^l\mathbf{r}_1 = \mu_i\mathbf{q}_i^H\mathbf{B}\mathbf{r}_i. \tag{4.8}$$

Ezek olyan értékek, amelyek a rekurzió készítésekor amúgy is előfordulnak.

A továbbiakban megmutatjuk, hogy a 2. szakaszban mutatott rekurziók mátrixegyenlettel is megadhatók. Ehhez bevezetjük a következő jelöléseket. Jelölje $\mathbf{R}$ azt a mátrixot, melynek j-edik oszlopvektora $\mathbf{r}_j$, s hasonlóan a $\mathbf{v}_j$, $\mathbf{q}_j$, $\mathbf{u}_j$ vektorokból is készítsük el a $\mathbf{V}$, $\mathbf{Q}$ és $\mathbf{U}$ mátrixokat. $\mathbf{M}=\langle\mu_j\rangle$ és $\Lambda=\langle\lambda_j\rangle$ nemszinguláris diagonálmátrixokat jelölnek.

Bevezetünk még egy speciális alsó háromszögmátrixot, melynek minden nemzérus eleme 1 és az inverze olyan alsó bidiagonális mátrix, amelynek főátlójában 1-ek, a főátló alatt -1-ek állnak. Az $\mathbf{S}$ mátrix analógja a diszkrét integráloperátornak: $(\mathbf{S}\mathbf{y})_i=\sum_{j=1}^{i}y_j$, míg az inverze a diszkrét differenciáloperátornak: $(\mathbf{S}^{-1}\mathbf{y})_i=y_i-y_{i-1}$, $i\neq 1$.

Feltételezve, hogy a *Hestenes—Stiefel típusú rekurzió* az i-edik lépésig jutott el, az előállított vektorok segítségével felírhatók az alábbi összefüggések:

$$\mathbf{A}\mathbf{U}\mathbf{D}_a^{-1}\mathbf{M}\mathbf{D}_r = \mathbf{R}\mathbf{S}^{-1}-\mathbf{r}_{i+1}\mathbf{e}_i^T, \tag{4.9}$$

$$\mathbf{D}_q\Lambda\mathbf{D}_a^{-1}\mathbf{V}^H\mathbf{A} = \mathbf{S}^{-T}\mathbf{Q}^H-\mathbf{e}_i\mathbf{q}_{i+1}^H, \tag{4.10}$$

$$\mathbf{D}_r\mathbf{S}^{-1}\mathbf{D}_r^{-1}\mathbf{M}^{-1}\mathbf{V}^H = \mathbf{R}^H\mathbf{C}, \tag{4.11}$$

$$\mathbf{U}\Lambda^{-1}\mathbf{D}_q^{-1}\mathbf{S}^{-T}\mathbf{D}_q = \mathbf{K}\mathbf{Q}, \tag{4.12}$$

ahol

$$\mathbf{D}_r = \mathbf{R}^H\mathbf{C}\mathbf{R}, \quad \mathbf{D}_q = \mathbf{Q}^H\mathbf{K}\mathbf{Q}, \quad \mathbf{D}_a = \mathbf{V}^H\mathbf{A}\mathbf{U}, \tag{4.13}$$

és $-T$ a transzponált inverzet jelöli. Balról $\mathbf{e}_j^T$, ill. jobbról $\mathbf{e}_j$ — a j-edik egységvektorokkal — szorozva a (2.25) rekurzió formulái $i=j$-re adódnak.

A (2.28) *Lánczos-féle rekurzió* mátrixegyenletei a következők:

$$\mathbf{D}\Lambda\mathbf{D}_a^{-1}\mathbf{V}^H\mathbf{A} = \mathbf{S}^{-T}\mathbf{Q}^H - \mathbf{e}_i\mathbf{q}_{i+1}^H, \tag{4.14}$$

$$\mathbf{AUD}_a^{-1}\mathbf{MD} = \mathbf{RS}^{-1} - \mathbf{r}_{i+1}\mathbf{e}_i^T, \tag{4.15}$$

$$\mathbf{DS}^{-1}\mathbf{D}^{-1}\mathbf{M}^{-1}\mathbf{V}^H = \mathbf{Q}^H\mathbf{B}, \tag{4.16}$$

$$\mathbf{U}\Lambda^{-1}\mathbf{D}^{-1}\mathbf{S}^{-T}\mathbf{D} = \mathbf{BR}, \tag{4.17}$$

ahol

$$\mathbf{D} = \mathbf{Q}^H\mathbf{BR} \quad \text{és} \quad \mathbf{D}_a = \mathbf{V}^H\mathbf{AU}. \tag{4.18}$$

A mátrixegyenletekkel történő reprezentációnak több előnye is van. Így például egy módszer elemzésekor számos kis összefüggést kell levezetni, amelyek mátrixegyenletekben tömör formában megtalálhatók. Más rekurziók is egyszerűen előállíthatók, például a $\mathbf{Q}$ és $\mathbf{R}$ mátrixok kiküszöbölésével olyan rekurzió készíthető, amely csak a $\mathbf{V}$ és $\mathbf{U}$ mátrix elemeire támaszkodik. Az $\mathbf{A}$ mátrix konjugált irányokkal kapott faktorizációja is könnyen vizsgálható. Most a mátrixegyenletes alakot arra fogjuk felhasználni, hogy megmutassuk a szinguláris érték feladattal és a sajátérték-feladattal való összefüggéseket.

Helyettesítve (4.11) és (4.12)-ből kapjuk

$$\mathbf{R}^H\mathbf{CAKQ} = \mathbf{D}_r\mathbf{S}^{-1}\mathbf{D}_r^{-1}\mathbf{M}^{-1}\mathbf{D}_a\Lambda^{-1}\mathbf{D}_q^{-1}\mathbf{S}^{-T}\mathbf{D}_q, \tag{4.19}$$

ami kontinuáns mátrix, mivel diagonális mátrixok és egy alsó és felső bidiagonális mátrix szorzata. Amennyiben a rekurzió teljes rendszert állított elő, és a $\mathbf{C}$ és $\mathbf{K}$ mátrixok $\mathbf{C}=\mathbf{C}_1\mathbf{C}_1^H$, $\mathbf{K}=\mathbf{K}_1\mathbf{K}_1^H$ alakúak, akkor $\mathbf{C}_1\mathbf{AK}_1$ és $\mathbf{D}_r^{-1/2}\mathbf{R}^H\mathbf{CAKQD}_r^{-1/2}$ szinguláris értékei ugyanazok lesznek, mert az előbbit balról, illetve jobbról egy-egy unitér mátrixszal, vagy olyan parciális izometriával szoroztuk, amelyek teljesen kifeszítik a nemzérus szinguláris értékekhez tartozó altereket. Innen látható, hogy a $\mathbf{K}=\mathbf{I}$, $\mathbf{C}=\mathbf{I}$ speciális esetben a $\mathbf{D}_r$, $\mathbf{D}_q$ és $\mathbf{D}_a$ diagonálmátrixokból olyan kontinuáns mátrix készíthető, amelyből az $\mathbf{A}$ mátrix szinguláris értékei meghatározhatók.

Megmutatjuk, hogy a *Hestenes—Stiefel féle rekurzióból* származtatható olyan algoritmus, mely két ortonormált vektorrendszer segítségével rekurzív módon egy téglalap alakú mátrixot bidiagonális alakra hoz.

Legyen $\mathbf{A}\in C^{m,n}$ és készítsük el az $\mathbf{AA}^H$ mátrixra a (4.9)—(4.13) rekurziókat, amikor $\mathbf{U}=\mathbf{V}$, $\mathbf{R}=\mathbf{Q}$, $\Lambda=\mathbf{M}$, $\mathbf{C}=\mathbf{K}$:

$$\mathbf{AA}^H\mathbf{VD}_a^{-1}\mathbf{MD}_r = \mathbf{RS}^{-1} - \mathbf{r}_{i+1}\mathbf{e}_i^T, \tag{4.20}$$

$$\mathbf{D}_r\mathbf{S}^{-1}\mathbf{D}_r^{-1}\mathbf{M}^{-1}\mathbf{V}^H = \mathbf{R}^H\mathbf{C}, \tag{4.21}$$

$$\mathbf{D}_a = \mathbf{V}^H\mathbf{AA}^H\mathbf{V}, \tag{4.22}$$

ahol $\mathbf{C}=\mathbf{I}$ esetben

$$\mathbf{R}^H\mathbf{AA}^H\mathbf{V} = \mathbf{D}_r\mathbf{S}^{-1}\mathbf{D}_r^{-1}\mathbf{M}^{-1}\mathbf{D}_a$$

alsó bidiagonális mátrix. Itt a baloldali $\mathbf{R}^H$ mátrix sorvektorai normáltak lesznek, ha $\mathbf{D}_r^{-1/2}$-vel szorzunk balról. Az $\mathbf{A}^H\mathbf{V}$ mátrix oszlopvektorai ortogonálisak (4.22) sze-

rint és jobbról $\mathbf{D}_a^{-1/2}$-vel szorozva normáltak, így a

$$(\mathbf{D}_r^{-1/2}\mathbf{R}^H)\mathbf{A}(\mathbf{A}^H\mathbf{V}\mathbf{D}_a^{-1/2}) = \mathbf{D}_r^{1/2}\mathbf{S}^{-1}\mathbf{D}_r^{-1}\mathbf{M}^{-1}\mathbf{D}_a^{1/2} \tag{4.23}$$

alsó bidiagonális mátrix szinguláris értékei megegyeznek $\mathbf{A}$ szinguláris értékeivel, ha a rekurziót teljesen végigcsináltuk, azaz, ha $\mathbf{v}_{i+1}^H\mathbf{A}=0$. A konjugált gradiens rekurzióknak ezt az összefüggését GOLUB és KAHAN publikálták [6]. Összehasonlítva (4.19)-cel, (4.23) egyszerűbb alakra vezet, azonban a kondíciószám négyzetelődése miatt a (4.20)—(4.21) iterációnál gyengébb stabilitás várható.

A következőkben rátérünk a *Lánczos-féle rekurzió* és a *sajátérték-feladat* összefüggéseinek megmutatására. Először is vegyük észre, hogy (4.16)—(4.18)-ból helyettesítéssel a

$$\mathbf{Q}^H\mathbf{B}\mathbf{A}\mathbf{B}\mathbf{R} = \mathbf{D}\mathbf{S}^{-1}\mathbf{D}^{-1}\mathbf{M}^{-1}\mathbf{D}_a\Lambda^{-1}\mathbf{D}^{-1}\mathbf{S}^{-T}\mathbf{D} \tag{4.24}$$

szimmetrikus kontinuáns mátrix adódik. A $\mathbf{B}=\mathbf{I}$, $\mathbf{A}\in C^{n,n}$ speciális esetben (4.18)-ból látható, hogy $\mathbf{Q}$ és $\mathbf{R}$ biortogonális vektorokat tartalmaznak és a belső szorzatuk eggyé tehető, ha a $\mathbf{D}=\mathbf{Q}^H\mathbf{R}$ összefüggésben a $\mathbf{D}$ diagonálmátrixot átvisszük a jobb oldalra. Ezek alapján a $\mathbf{D}^{-1}\mathbf{Q}^H\mathbf{A}\mathbf{R}$ vagy $\mathbf{Q}^H\mathbf{A}\mathbf{R}\mathbf{D}^{-1}$ kontinuáns mátrixok az $\mathbf{A}$ mátrix egy hasonlósági transzformáltját adják, ha a $\mathbf{Q}$ és $\mathbf{R}$ mátrixok kvadratikusak. Ha pedig a rekurzió úgy áll meg, hogy egyszerre $\mathbf{r}_{i+1}=\mathbf{q}_{i+1}=0$, akkor a kontinuáns mátrix sajátértékei megegyeznek $\mathbf{A}$ sajátértékeivel, de nem biztos, hogy az összessel. A *Lánczos-módszer* befejeződési feltételeit bővebben l. a [11], [15] munkákban.

Megmutatjuk, hogy a (4.14)—(4.18) rekurzió a $\mathbf{C}=\mathbf{I}$, $\mathbf{A}\in C^{n,n}$ speciális esetben ekvivalens *Lánczos módszerével.* Tegyük fel, hogy $\mathbf{r}_{i+1}=\mathbf{q}_{i+1}=0$, és helyettesítsük $\mathbf{V}^H$-t és $\mathbf{U}$-t (4.16) és (4.17)-ből (4.14) és (4.15)-be, használjuk fel továbbá (4.24)-et. Az eredmény:

$$\mathbf{Q}^H\mathbf{A} = (\mathbf{Q}^H\mathbf{A}\mathbf{R})\mathbf{D}^{-1}\mathbf{Q}^H, \tag{4.25}$$

$$\mathbf{A}\mathbf{R} = \mathbf{R}\mathbf{D}^{-1}(\mathbf{Q}^H\mathbf{A}\mathbf{R}). \tag{4.26}$$

A j-edik vektorokat kifejezve kapjuk, hogy

$$\mathbf{q}_j^H\mathbf{A} = \alpha_{j-1}\mathbf{q}_{j-1}^H + \beta_j\mathbf{q}_j^H + \gamma_{j+1}\mathbf{q}_{j+1}^H, \tag{4.27}$$

$$\mathbf{A}\mathbf{r}_j = \alpha_{j-1}\mathbf{r}_{j-1} + \beta_j\mathbf{r}_j + \gamma_{j+1}\mathbf{r}_{j+1}, \tag{4.28}$$

ahol

$$\alpha_{j-1} = \mathbf{q}_j^H\mathbf{A}\mathbf{r}_{j-1}/\mathbf{q}_{j-1}^H\mathbf{r}_{j-1} = \mathbf{q}_{j-1}^H\mathbf{A}\mathbf{r}_j/\mathbf{q}_{j-1}^H\mathbf{r}_{j-1},$$
$$\beta_j = \mathbf{q}_j^H\mathbf{A}\mathbf{r}_j/\mathbf{q}_j^H\mathbf{r}_j \quad \text{és} \quad \gamma_{j+1} = \mathbf{q}_j^H\mathbf{A}\mathbf{r}_{j+1}/\mathbf{q}_{j+1}^H\mathbf{r}_{j+1}. \tag{4.29}$$

A $\mathbf{q}_j$ és $\mathbf{q}_{j+1}$ vektorok relatív hossza megválasztható úgy, hogy $\gamma_{j+1}=1$ legyen minden j-re. Az eredeti rekurzióban ez például a $\lambda_j=-\mathbf{v}_j^H\mathbf{A}\mathbf{u}_j$, $\mu_j=1/\mathbf{q}_j^H\mathbf{r}_j$ választással tehető meg. Ezáltal $\alpha_{j-1}=\mathbf{q}_j^H\mathbf{r}_j/\mathbf{q}_{j-1}^H\mathbf{r}_{j-1}$ lesz, amivel pontosan visszakapjuk a *Lánczos-módszer* formuláit, vö. [15, 496. o.].

Végül még megjegyezzük, hogy a (4.9)—(4.18) összefüggések felhasználásával származtathatók blokk konjugált irány algoritmusok, ha az $\mathbf{S}$ mátrixban szereplő 1-ek helyére egységmátrixokat írunk. Ilyen algoritmusokat származtatott O'LEARY [14].

IRODALOM

[1] ABAFFY, J., BROYDEN, C. G., SPEDICATO, E., "A class of direct methods for linear systems", *Numer. Math.* **45** (1984) 361—376.
[2] ABAFFY, J., SPEDICATO, E., "A generalization of the ABS algorithm for linear systems", Quaderni de Dipartimento di Matematica, Statistica e Informatica e Applicazioni, Report 1985 N. 4, Instituto Universitario di Bergamo, Bergamo, Italy, 1985.
[3] BODÓCS, L., „A konjugált irányok módszere lineáris egyenletrendszerek megoldására", szakdolgozat, Eötvös Loránd Tudományegyetem, Budapest, 1980.
[4] FLETCHER, R., "Conjugate gradient methods for indefinite systems", in: *Proc. of the Dundee Biennial Conference on Numerical Analysis Ed. G. A. Watson* (Lecture Notes in Mathematics 506, Springer, Berlin, 1976) 73—89.
[5] FOX, L., HUSKEY, H. D., WILKINSON, J. H., "Notes on the solution of algebraic simultaneous equations", *Quart. J. Mech. Appl. Math.* **1** (1948) 149—173.
[6] GOLUB, G., KAHAN, W., "Calculating the singular values and pseudo-inverse of a matrix", *SIAM J. Numer. Anal.* **2** (1965) 205—224.
[7] HEGEDŰS, Cs. J., "Generalization of the method of conjugate gradiets: The method of conjugate pairs", in: *Collection of Scientific Papers in Collaboration with the Joint Institute for Nuclear Research,* Dubna, USSR and the Central Research Institute for Physics, Budapest, Hungary. Algorithms and Programs for Solution of Some Problems in Physics (Report KFKI 1979—82, MTA KFKI, Budapest, 1979) 199—209.
[8] HEGEDŰS, Cs. J., BODÓCS, L., "General recursions for A-conjugate vector pairs", Report KFKI 1982—56, MTA KFKI, Budapest, 1982.
[9] HESTENES, M. R., "The conjugate gradient method for solving linear systems", in: *Proc. Symposia on Applied Mathematics, Vol. VI. Numerical Analysis* (McGraw—Hill, New York, 1956) 83—102.
[10] HESTENES, M. R., STIEFEL, E., "Methods of conjugate gradients for solving linear systems", *J. Res. Nat. Bur. Standards Sect. B* **49** (1952) 409—436.
[11] HOUSEHOLDER, A. S., *The Theory of Matrices in Numerical Analysis* (Blaisdell Publishing Company, New York, Toronto and London, 1964).
[12] LÁNCZOS, C., "An iteration method for the solution of the eigenvalue problem of linear differential and integral operators", *J. Res. Nat. Bur. Standards Sect B* **45** (1950) 255—283.
[13] LÁNCZOS, C., "Solution of systems of linear equations by minimized iterations", *J. Res. Nat. Bur. Standards Sect. B* **49** (1952) 33—53.
[14] O'LEARY, D. P., "The block conjugate gradient algorithm and related methods", *Linear Algebra and Appl.* **29** (1980) 293—322.
[15] RALSTON, A., *Bevezetés a Numerikus Analízisbe* (Műszaki Könyvkiadó, Budapest, 1969).
[16] RÓZSA, P., *Lineáris Algebra és Alkalmazásai* (Műszaki Könyvkiadó, Budapest, 1974).
[17] STEWART, G. W., "Conjugate direction methods for solving systems of linear equations", *Numer. Math.* **21** (1973) 285—297.

(Beérkezett: 1985. március 29.)

HEGEDŰS CSABA J. ÉS BODÓCS LÁSZLÓ
MTA KÖZPONTI FIZIKAI KUTATÓ INTÉZET
1525 BUDAPEST, PF. 49.

GENERATION OF CONJUGATE DIRECTIONS: THE METHOD OF CONJUGATE PAIRS

Cs. J. HEGEDŰS and L. BODÓCS

A general class of algorithms for generating conjugate directions with respect to an arbitrary rectangular matrix is described. The method is based on successive projections. The obtained schemes essentially contain most of the previously known recursive generation schemes, including Lánczos' tridiagonalization. The possible advantages of the generalization are that the degrees of freedom are increased and it allows unified description and analysis of conjugate direction algorithms.

Alkalmazott Matematikai Lapok **11** (1985) 311—318

SZORZATFÜGGVÉNYEK KONKÁVITÁSI TARTOMÁNYÁRÓL

RAPCSÁK TAMÁS és BORZSÁK PÉTER

Budapest

A dolgozatban a $G(t)=\prod_{i=1}^{n} F_i(t_i)$ alakú függvények konkávitási tartományát vizsgáljuk. Először szükséges és elegendő feltételeket adunk arra, hogy a $G(t)$ függvény *negatív Hesse-mátrixa* pozitív szemidefinit legyen. A kapott feltételek alapján a konkávitási tartomány vizsgálatát egy szeparábilis függvény kvázikonvexitására vezetjük vissza, amelyre elegendő feltételeket adunk. Végül normális eloszlás esetén nézzük meg, hogy a konkávitási tartomány hogyan változik a dimenzió-szám növelésével.

1. Bevezetés

A dolgozatban a $G(\mathbf{t})=\prod_{i=1}^{n} F_i(t_i)$ alakú függvények konkávitási tartományát vizsgáljuk. Egy függvény konkávitási tartományán azt a legbővebb konvex halmazt értjük, amelyen a függvény konkáv. Ilyen jellegű problémákkal elsősorban a sztochasztikus programozási [1], [4], [5] és a statisztikai feladatok [8] megoldásakor találkozunk. Számítástechnikailag lényegesen kedvezőbb ugyanis az az eset, mikor a valószínűségi függvény a vizsgált tartományon már konkáv.

MILLER és WAGNER [4] vizsgálta először független valószínűségi változók együttes eloszlásfüggvényét, amely ilyen típusú szorzatfüggvényre vezetett. Normális eloszlású, nem független valószínűségi változók együttes eloszlásfüggvényének konkávitási kérdéseivel foglalkozott PRÉKOPA [5]. Ezeknél az eredményeknél lényegesen általánosabb és alapvetőbb eredményeket ért el később PRÉKOPA [6], [7] mikor együttes log-konkáv sűrűségfüggvénnyel rendelkező valószínűségi változók együttes eloszlásfüggvényének konkávitási kérdéseivel foglalkozott.

Független valószínűségi változók együttes eloszlásfüggvényének konkávitási tartományát vizsgálta BAWA [1]. Az itt szereplő eredmények ez utóbbi cikk [1] eredményeinek általánosításai.

2. A probléma megfogalmazása

Tegyük fel, hogy az $F_i(t_i)$, $t_i \in R_i$, $i=1, \ldots, n$ kétszer folytonosan differenciálható függvények, ahol R_i az R^n n-dimenziós euklideszi tér i-edik koordináta tengelyét jelöli azaz $R^n = R_1 \oplus R_2 \oplus \ldots \oplus R_n$. Tekintsük a

$$G(\mathbf{t}) = \prod_{i=1}^{n} F_i(t_i), \quad \mathbf{t} \in R^n \tag{2.1}$$

alakú függvényeket és nézzük meg mi annak a szükséges és elegendő feltétele, hogy $G(\mathbf{t})$ valamilyen tartományon konkáv legyen. Először egy szükséges feltételt bizonyítunk, amellyel az $F_i(t_i)$ függvények osztályát tudjuk leszűkíteni.

2.1. LEMMA. Ha $G(\mathbf{t})$ nem az azonosan nulla függvény, akkor a $G(\mathbf{t})$ függvény konkávitásának szükséges feltétele, hogy a vizsgált tartományon minden i, $i=1, \ldots, n$ esetén

$$F_i(t_i) > 0 \quad \text{vagy} \quad F_i(t_i) < 0 \tag{2.2}$$

teljesüljön.

Bizonyítás. Először tegyük fel, hogy a $G(\mathbf{t})$ függvény értéke valamilyen $\hat{\mathbf{t}}$ pontban nulla, úgy hogy a szorzatban csak egy tényező nulla. Az általánosság megszorítása nélkül feltehetjük, hogy $F_1(\hat{t}_1)=0$. Az ismert, hogy a $G(\mathbf{t})$ függvény konkávitásának szükséges és elegendő feltétele, hogy az értelmezési tartomány egy konvex részhalmazán a $G(\mathbf{t})$ függvény *Hesse-mátrixának* negatívja, a $-\mathbf{H}(\mathbf{t})$ mátrix minden pontban pozitív szemidefinit legyen. Kiszámolva a $-\mathbf{H}(\hat{\mathbf{t}})$ mátrixot, azt kapjuk, hogy csak az első oszlop és az első sor különbözhet nullától. Az itt szereplő elemek a következők:

$$\begin{aligned} h_{11} &= -f_1'(\hat{t}_1) \prod_{i=2}^{n} F_i(\hat{t}_i), \\ h_{r1} = h_{1r} &= -f_1(\hat{t}_1) f_r(\hat{t}_r) \prod_{\substack{i=2 \\ i \neq r}}^{n} F_i(\hat{t}_i). \end{aligned} \tag{2.3}$$

Ebből viszont következik, hogy tetszőleges $\mathbf{x}$ vektor mellett

$$-\mathbf{x}^T \mathbf{H}(\hat{\mathbf{t}})\,\mathbf{x} = -2x_1 \sum_{i=2}^{n} h_{1i} x_i - x_1^2 h_{11}, \tag{2.4}$$

amely érték lehet pozitív is, negatív is, tehát a $-\mathbf{H}(\hat{\mathbf{t}})$ mátrix nem pozitív szemidefinit.

Tegyük fel most, hogy a szorzatban két tényező nulla azaz

$$F_1(\hat{t}_1) = 0, \quad F_2(\hat{t}_2) = 0. \tag{2.5}$$

Így a *Hesse-mátrixban* csak h_{12} különbözhet nullától, amiből következik, hogy $-\mathbf{H}(\hat{\mathbf{t}})$ nem pozitív szemidefinit.

Ha a szorzat függvényben kettőnél több tényező nulla valamelyik pontban, akkor ebben a pontban a $G(\mathbf{t})$ gradiense nulla, amiből következik, hogy ez globális maximum pont, ami szintén ellentmondáshoz vezet.

Tegyük fel tehát a továbbiakban, hogy a vizsgált tartományon

$$F_i(t_i) > 0, \quad i = 1, \ldots, n \tag{2.6}$$

és

$$F_i'(t_i) = f_i(t_i) > 0, \quad i = 1, \ldots, n. \tag{2.7}$$

2.2. LEMMA. Annak szükséges és elegendő feltétele, hogy a vizsgált tartomány

egy $\mathbf{t} \in R^n$ pontjában a $-\mathbf{H}(\mathbf{t})$ mátrix pozitív szemidefinit legyen, az hogy az

$$f_i'(t_i) \leqq 0, \quad i = 1, \ldots, n, \tag{2.8}$$

$$\sum_{i=1}^{n} \frac{1}{\left(\frac{F_i(t_i)}{f_i(t_i)}\right)'} \leqq 1 \tag{2.9}$$

feltételek teljesüljenek.

Bizonyítás. Ismert az az állítás, hogy egy mátrix akkor és csak akkor pozitív szemidefinit, ha az összes minorjának a determinánsa nem negatív. Mivel

$$-\mathbf{H}(\mathbf{t}) = G(\mathbf{t}) \begin{pmatrix} -\frac{f_1'(t_1)}{F_1(t_1)} & -\frac{f_1(t_1)f_2(t_2)}{F_1(t_1)F_2(t_2)} & \ldots & -\frac{f_1(t_1)f_n(t_n)}{F_1(t_1)F_n(t_n)} \\ -\frac{f_2(t_2)f_1(t_1)}{F_2(t_2)F_1(t_1)} & -\frac{f_2'(t_2)}{F_2(t_2)} & \ldots & -\frac{f_2(t_2)f_n(t_n)}{F_2(t_2)F_n(t_n)} \\ \vdots & \vdots & & \vdots \\ -\frac{f_n(t_n)f_1(t_1)}{F_n(t_n)F_1(t_1)} & -\frac{f_n(t_n)f_2(t_2)}{F_n(t_n)F_2(t_2)} & \ldots & -\frac{f_n'(t_n)}{F_n(t_n)} \end{pmatrix} \tag{2.10}$$

ezért a (2.8) feltétel valóban szükséges.

Először a főminorok determinánsát vizsgáljuk. Jelölje a p-edik főminor determináns értékét $D_p(\mathbf{t})$, $p = 1, \ldots, n$. Vezessük be az alábbi jelöléseket.

$$\frac{f_i(t_i)}{F_i(t_i)} = a_i, \quad \frac{f_i'(t_i)}{f_i(t_i)} = b_i, \quad i = 1, \ldots, n. \tag{2.11}$$

Ha oszloponként a_i-t, $i = 1, \ldots, p$ kiemelünk, akkor

$$D_p(\mathbf{t}) = \prod_{i=1}^{p} a_i \begin{vmatrix} -b_1 & -a_1 & \ldots & -a_1 \\ -a_2 & -b_2 & \ldots & -a_2 \\ \vdots & & & \\ -a_p & -a_p & \ldots & -b_p \end{vmatrix} \tag{2.12}$$

majd minden sorból a_i-t, $i = 1, \ldots, p$ kiemelve kapjuk, hogy

$$D_p(\mathbf{t}) = \prod_{i=1}^{p} a_i^2 \begin{vmatrix} -\frac{b_1}{a_1} & -1 & \ldots & -1 \\ -1 & -\frac{b_2}{a_2} & \ldots & -1 \\ \vdots & & & \\ -1 & -1 & \ldots & -\frac{b_p}{a_p} \end{vmatrix}. \tag{2.13}$$

Ezután minden sorból vonjuk le az első sort. Így

$$D_p(\mathbf{t}) = \prod_{i=1}^{p} a_i^2 \begin{vmatrix} -\frac{b_1}{a_1} & -1 & \dots -1 \\ -1+\frac{b_1}{a_1} & -\frac{b_2}{a_2}+1 & \dots 0 \\ \vdots & & \\ -1+\frac{b_1}{a_1} & 0 & \dots -\frac{b_p}{a_p}+1 \end{vmatrix}. \tag{2.14}$$

Fejtsük ki a determinánst az első oszlop szerint és jelöljön a zárójelen belül az i sorindexeket. Akkor

$$D_p(\mathbf{t}) = \prod_{i=1}^{p} a_i^2 \times$$

$$\times \left[-\frac{b_1}{a_1} \prod_{i=2}^{p} \left(-\frac{b_i}{a_i}+1\right) + \left(-1+\frac{b_1}{a_1}\right) \sum_{i=2}^{p} (-1)^{i+1}(-1)^{1+i-1}(-1) \frac{\prod_{l=2}^{p}\left(-\frac{b_l}{a_l}+1\right)}{\left(-\frac{b_i}{a_i}+1\right)} \right] =$$

$$= \prod_{i=1}^{p} a_i^2 \prod_{i=2}^{p} \left(-\frac{b_i}{a_i}+1\right) \left[-\frac{b_1}{a_1} + \left(-1+\frac{b_1}{a_1}\right) \sum_{i=2}^{p} \frac{1}{\left(-\frac{b_i}{a_i}+1\right)} \right] = \tag{2.15}$$

$$= \prod_{i=1}^{p} a_i^2 \prod_{i=2}^{p} \left(-\frac{b_i}{a_i}+1\right) \left[\left(-\frac{b_1}{a_1}+1\right) - \left(-\frac{b_1}{a_1}+1\right) \sum_{i=2}^{p} \frac{1}{\left(-\frac{b_i}{a_i}+1\right)} - \frac{-\frac{b_1}{a_1}+1}{-\frac{b_1}{a_1}+1} \right] =$$

$$= \prod_{i=1}^{p} a_i^2 \prod_{i=1}^{p} \left(-\frac{b_i}{a_i}+1\right) \left[1 - \sum_{i=1}^{p} \frac{1}{-\frac{b_i}{a_i}+1} \right].$$

Mivel

$$-\frac{b_i}{a_i}+1 = \frac{f_i^2(t) - f_i'(t)F_i(t)}{f_i^2(t)}, \quad i = 1, \dots, n \tag{2.16}$$

és mivel az $f_i'(t_i) \leqq 0$, $i=1, \dots, n$ feltételek a $-\mathbf{H}(\mathbf{t})$ pozitív szemidefinitségének szükséges feltételei, ezért a $\prod_{i=1}^{p}\left(-\frac{b_i}{a_i}+1\right)$ érték bármely $p=1, \dots, n$ esetén nem negatív. Ebből következik, hogy a $D_p(\mathbf{t})$ nemnegativitásának szükséges feltétele a

$$\sum_{i=1}^{p} \frac{1}{-\frac{b_i}{a_i}+1} \leqq 1, \quad p = 1, \dots, n \tag{2.17}$$

egyenlőtlenségek teljesülése. Mivel a bal oldalon csupa nemnegatív érték áll, ezért ha

az egyenlőtlenség $n=p$ esetén igaz, akkor az összes többi p értékre is igaz. Ez viszont (2.16) alapján éppen a (2.9) egyenlőtlenséget jelenti.

A számításokból következik, hogy ha a (2.8) és a (2.9) feltételek teljesülnek, akkor a $D_p(\mathbf{t})$, $p=1, \ldots, n$ értékek nem negatívak. Tehát a $D_p(\mathbf{t})$, $p=1, \ldots, n$ főminorok nemnegativitásának a (2.8) és (2.9) szükséges és elegendő feltételei. Mivel a $G(\mathbf{t})$ függvény invariáns a benne szereplő $F_i(t_i)$, $i=1, \ldots, n$ függvények sorrendjére, ezért a minorokat ugyanilyen módon lehet vizsgálni és a minorok nemnegativitásának szükséges és elegendő feltételei ugyancsak a (2.8), (2.9) feltételek lesznek. Ezzel bebizonyítottuk az állítást.

A 2.2. lemmából következik, hogy a $G(\mathbf{t})$ függvény konkávitási tartományát a

$$\sum_{i=1}^{n} \frac{1}{\left(\frac{F_i(t_i)}{f_i(t_i)}\right)'} \leqq 1 \tag{2.18}$$

egyenlőtlenség által meghatározott halmaz tartalmazza és ott az

$$f_i'(t_i) \leqq 0, \quad i = 1, \ldots, n \tag{2.19}$$

egyenlőtlenségek teljesülését is meg kell követelni. Mivel a konkávitási tartománynak konvex halmaznak kell lenni, ezért a (2.18) egyenlőtlenség által meghatározott halmaz akkor a legbővebb, ha az konvex, azaz ha a

$$\hat{G}(\mathbf{t}) = \sum_{i=1}^{n} \frac{1}{\left(\frac{F_i(t_i)}{f_i(t_i)}\right)'} \tag{2.20}$$

függvény kvázikonvex. Azonban $\hat{G}(\mathbf{t})$ szeparábilis függvény, amelynek kvázikonvexitásához elegendő az összegben szereplő függvényekről azt megmutatni, hogy vagy mindegyik monoton csökkenő vagy mindegyik monoton növekedő.

2.3. LEMMA. Ha van olyan $\mathbf{t}^* \in R^n$ pont, hogy a $t_i \geqq t_i^*$, $i=1, \ldots, n$ tartományokban az $f_i(t_i)$, $i=1, \ldots, n$ függvények nem növekedőek és logaritmikusan konkávak, akkor ott a $\hat{G}(\mathbf{t})$ függvény minden tagja monoton csökkenő.

Bizonyítás. Mivel az $F_i(t_i)$, $i=1, \ldots, n$ függvények pozitívak és monoton növekedőek, az $f_i(t_i)$, $i=1, \ldots, n$ függvények pozitívak és a megfelelő $t_i \geqq t_i^*$, $i=1, \ldots, n$ tartományokban monoton csökkenőek és logaritmikusan konkávak ezért az

$$\left(\frac{F_i(t_i)}{f_i(t_i)}\right)' = \frac{f_i^2(t_i) - f_i'(t_i) F_i(t_i)}{f_i^2(t_i)}, \quad i = 1, \ldots, n \tag{2.21}$$

függvények pozitívak és monoton növekedőek. Ebből következik, hogy a $\hat{G}(\mathbf{t})$ függvényben szereplő mindegyik függvény monoton csökkenő a megfelelő $t_i \geqq t_i^*$, $i= =1, \ldots, n$ tartományokban, tehát a $\hat{G}(\mathbf{t})$ függvény kvázikonvex a $\mathbf{t} \geqq \mathbf{t}^*$ tartományon.

3. Néhány megjegyzés az eredményekkel kapcsolatban

1. *Megjegyzés.* A 2.2., 2.3. lemmákból látható, hogy a $G(\mathbf{t})$ szorzatfüggvény konkávitásának a vizsgálatát a (2.18) egyenlőtlenség teljesülésére lehet visszavezetni. Mivel a $G(\mathbf{t})$ függvény *negatív Hesse-mátrixa* pozitív szemidefinitségének szükséges és elegendő feltételei között szerepelnek az $f_i'(t_i) \leqq 0$, $i=1, \ldots, n$ feltételek, ezért a 2.3. lemmában az új feltevés csak az, hogy az $f_i(t_i)$, $i=1, \ldots, n$ függvények az értelmezési tartományuk egy részén logaritmikusan konkávak. Ez azt jelenti, hogy ha a 2.2., 2.3. lemma feltételei teljesülnek, akkor elegendő csak egy olyan $\mathbf{t}^* \in R^n$ pontot megkeresni, amelyre a (2.18) egyenlőtlenség igaz, mert ebből következik, hogy minden $\mathbf{t} \geqq \mathbf{t}^*$ értékre is igaz, ami egyben azt is mutatja, hogy ez a $G(\mathbf{t})$ függvény konkávitási tartományának a része lesz. A legkisebb t^* érték adja a konkávitási tartományt.

2. *Megjegyzés.* A 2.3. lemma feltételeiből és a (2.21) egyenlőtlenségből következik, hogy az $F_i(t_i)$, $i=1, \ldots, n$ pozitív, monoton növekedő függvények és az $f_i(t_i)$, $i=1, \ldots, n$ pozitív, monoton nem növekedő logaritmikusan konkáv függvények hányadosai, az $\dfrac{F_i(t_i)}{f_i(t_i)}$, $i=1, \ldots, n$ függvények konvexek a $t_i \geqq t_i^*$, $i=1, \ldots, n$ tartományokban.

3. *Megjegyzés.* BAWA [1] cikkében független, egyforma eloszlású valószínűségi változókat tekint, azaz feltételezi, hogy az $F_i(t_i)$, $i=1, \ldots, n$ függvények eloszlásfüggvények és az $F_i = F_j$, $i, j=1, \ldots, n$ egyenlőségek teljesülnek. Így a 2.2. lemmában szereplő minorok determinánsainak a vizsgálata egyszerűbbé válik. A 2.3. lemmában szereplő $f_i(t_i)$, $i=1, \ldots, n$ függvényekre vonatkozó feltételek megegyeznek a BAWA [1] cikkben szereplő feltételekkel. Ott azonban e feltételek segítségével a szigorúan unimodális befejeződésű sűrűségfüggvénnyel "strongly unimodular upper tails" rendelkező eloszlásfüggvények osztálya van definiálva. HAJEK és SIDAK hasonlóan definiálta a szigorúan unimodális sűrűségfüggvénnyel rendelkező eloszlásfüggvények osztályát [3].

4. *Megjegyzés.* PRÉKOPA [6], [7] és BAWA [1] cikkeiben szerepel az, hogy a normális, a gamma, a Weibull, a lognormális és a kettős exponenciális eloszlásfüggvények sűrűségfüggvénye teljesíti a 2.3. lemma feltételeit.

5. *Megjegyzés.* A (2.18) egyenlőtlenség speciális esete az [1] cikkben szereplő egyenlőtlenség. Tekintsük például a

$$G(t) = \prod_{i=1}^{n} F(t_i) \tag{3.1}$$

függvényt, ahol $F(t_i)$, $i=1, \ldots, n$ valamelyik az előző megjegyzésben szereplő eloszlás eloszlásfüggvényét jelenti. Ebben az esetben a (2.18) egyenlőtlenség alapján elegendő olyan t^* értéket találni $(\mathbf{t}^* = (t^*, t^*, \ldots, t^*))$ amelyre

$$\frac{nf^2(t^*)}{f^2(t^*) - f'(t^*)F(t^*)} \leqq 1. \tag{3.2}$$

Ebből átrendezéssel kapjuk, hogy a

$$\frac{-f'(t^*)F(t^*)}{f(t^*)} - (n-1)f(t^*) \geqq 0 \tag{3.3}$$

egyenlőtlenségnek kell teljesülni, ami megegyezik az [1] cikkben szereplő egyenlőtlenséggel.

6. *Megjegyzés.* Egyforma eloszlású, független valószínűségi változók esetén a (3.3) egyenlőtlenség segítségével vizsgálni tudjuk, hogy a (3.1) képletben szereplő $G(\mathbf{t})$ függvény esetén a konkávitási tartomány hogyan változik a dimenziószám azaz az n értékének a növekedésével. Ha a normális eloszlást tekintjük, akkor mivel $f(t)=\frac{1}{\sqrt{2\pi}} e^{-t^2/2}$, $f'(t)=-tf(t)$, így a (3.3) egyenlőtlenségből az alábbit kapjuk.

$$\frac{tF(t)}{f(t)}+1 \geqq n. \tag{3.4}$$

Kiszámolva az 1. táblázatban szereplő értékeket, a (3.4) egyenlőtlenségből következik, hogy melyek azok a legnagyobb n értékek (2. táblázat), amelyek mellett a 2. táblázatban szereplő tartományok még a megfelelő $G(\mathbf{t})$ függvények konkávitási tartományai.

1. TÁBLÁZAT

t	0	0,5	0,6	1	1,5	2
$\frac{tF(t)}{f(t)}$	0	0,9820	1,3068	3,4764	10,8093	36,1926

A 2. táblázatból látszik, hogy a dimenziószám növekedésével a konkávitási tartomány csökken.

2. TÁBLÁZAT

$\mathbf{t} \geqq 0$	$n=1$
$\mathbf{t} \geqq 0{,}5$	$n=2$
$\mathbf{t} \geqq 0{,}6$	$n=3$
$\mathbf{t} \geqq 1$	$n=4$
$\mathbf{t} \geqq 1{,}5$	$n=11$
$\mathbf{t} \geqq 2$	$n=37$

7. *Megjegyzés.* GERENCSÉR adott szükséges és elegendő feltételeket kétszer folytonosan differenciálható konvex és konkáv tagokból álló szeparábilis függvények szigorú pszeudokonvexitására [2].

IRODALOM

[1] BAWA, V. S., "On chance constrained programming problems with joint contraints", *Management Science* **19** (1973) 1326—1331.
[2] GERENCSÉR, L., "On a close relation between quasiconvex and convex functions and related investigations", *Math. Operationsforschung und Statistik* **4** (1973) 201—211.
[3] HAJEK, J. and SIDAK, Z., *Theory of Rank Tests* (Academic Press, N. Y., 1967).
[4] MILLER, B. L. and WAGNER, H. M., "Chance constrained programming with joint constraints", *Operations Research* **13** (1965) 930—945.

[5] PRÉKOPA, A., "On probabilistic constrained programming", in: *Proc. of Princeton Symposium on Math. Programming, ed. Kuhn, H. W.* (Princeton University Press, Princeton, N. Y., 1970) 113—118.
[6] PRÉKOPA, A., Sztochasztikus rendszerek optimalizálási problémáiról, Akadémiai doktori disszertáció, Bp. 1970.
[7] PRÉKOPA, A., "Logarithmic concave measures with application to stochastic programming", *Acta Sci. Math.* **32** (1971) 301—316.
[8] SOMMERVILLE, P. N., "Some problems of optimum sampling", *Biometrika* **41** (1954) 420—429.

(Beérkezett: 1984. október 12.)

RAPCSÁK TAMÁS ÉS BORZSÁK PÉTER
MTA SZÁMÍTÁSTECHNIKAI ÉS AUTOMATIZÁLÁSI KUTATÓ INTÉZET
1111 BUDAPEST, KENDE U. 8—10.

ON THE CONCAVITY SET OF THE PRODUCT FUNCTIONS

T. RAPCSÁK AND P. BORZSÁK

In this paper the question is the following: where is concave the $G(t)=\prod_{i=1}^{n} F_i(t_i)$ function. First necessary and sufficient conditions are obtained for the positive semidefinitness of $-G(\mathbf{t})$. It turns out that it is sufficient to give conditions for the quasiconvexity of a separable function. Finally the relation of the concavity set of $G(\mathbf{t})$ and of the number of dimension is investigated in the case of the normal distribution functions.

Alkalmazott Matematikai Lapok 11 (1985) 319—334

A LEHMER—SCHUR MÓDSZER OPTIMALIZÁLÁSÁRÓL

GALÁNTAI AURÉL

Gödöllő

A dolgozatban a *Lehmer—Schur típusú módszerek* optimalitását és realizálásának néhány kérdését vizsgáljuk. Néhány fontos részosztályban megadjuk az optimális módszereket és éles becslést adunk meg az elérhető abszolút optimumra vonatkozóan.

1. Bevezetés

LEHMER ([8]) 1961-ben publikálta az első olyan iterációs módszert valós vagy komplex együtthatós polinomegyenletek közelítő megoldására, amely globálisan, tehát minden esetben konvergens. Módszere azon alapul, hogy a *Schur—Cohn teszt* segítségével tetszőleges polinomról el tudjuk dönteni, hogy van-e gyöke egy tetszőlegesen megadott nyílt egységkörlemezen, vagy nincs. Ennek alapján a *Lehmer—Schur módszer* a következőképpen írható le.

Tekintsük a komplex sík egy olyan nyílt körlapját, amely tartalmazza a szóbanforgó polinom legalább egy gyökét. Fedjük le ezt a körlemezt kisebb sugarú nyílt körlemezekkel és a *Schur—Cohn teszt* segítségével válasszunk ki ebből a fedésből egy olyan kisebb sugarú körlemezt, amely tartalmazza a polinom valamelyik gyökét. Az így kapott körlemezt újra lefedjük nálánál kisebb sugarú nyílt körlemezekkel és ezek közül a *Schur—Cohn teszt* segítségével ismét kiválasztunk egy olyat, amelyik tartalmazza a polinom valamelyik gyökét. Ha az eljárást így folytatjuk, akkor a komplex számsík nyílt körlemezeinek egy olyan sorozatát kapjuk, amelyek sugara szigorúan monoton csökkenő és amelyek mindegyike tartalmazza a polinom egy gyökét. Ha a körök sugarai zérushoz tartanak, akkor az eljárás konvergens és a körök középpontjai a polinom valamelyik gyökéhez tartanak.

Az eljárás különböző kérdéseit vizsgálja LEHMER [9] dolgozata. A számítógépes realizálás kérdéseit és az eljárás stabilitását részletesen STEWART [13] vizsgálta. Az eljárás optimalitását HENRICI [7] és FRIEDLI [3] tanulmányozta. HENRICI az egységkör kongruens fedései között talált egy olyan nyolc körös fedést, amelyet az irodalomban optimálisnak tekintenek a kongruens körfedéseken alapuló *Lehmer—Schur módszerek* osztályában. HENRICI eredményeit továbbfejlesztve FRIEDLI bevezette az ún. q-fedések fogalmát. Igazolta, hogy minden $l \geqq 3$ rögzített körszám esetén létezik optimális q-fedés és nincs ennél jobb nem q-fedés. Heurisztikus módon meghatározott továbbá két olyan 11, ill. 22 körös q-fedést, amelyekhez tartozó *Lehmer—Schur módszerek* komplexitási mutatója mintegy 60%-kal kisebb az eredeti *Lehmer módszerénél.*

A dolgozatban a következő kérdéseket vizsgáljuk. A második szakaszban ismertetjük a *Schur—Cohn tesztet* és igazolunk egy az alkalmazására vonatkozó állítást.

A harmadik szakaszban a dolgozatban vizsgált *Lehmer—Schur módszerosztályt* definiáljuk és az elemzések alapjául szolgáló komplexitási mutatókat. A negyedik szakaszban meghatározzuk az optimális 3, ill. 4 körös *Lehmer—Schur módszereket* és élesnek tekinthető becslést adunk meg az elérhető optimumértékre vonatkozóan. Megadunk továbbá egy olyan q-fedés szerkesztési eljárást is, amellyel kis körszám ($5 \leqq l \leqq 10$) esetén közel optimális q-fedés konstruálható. Az ötödik szakaszban a körgyűrűszerű fedések és a kongruens körfedések osztályában optimális *Lehmer—Schur módszereket* is meghatározzuk.

A szakasz befejezéseként megemlítjük, hogy LEHMER módszerének publikálása óta több globálisan konvergens polinom gyökközelítő eljárást konstruáltak. HENRICI [7] dolgozatában a *Schur—Cohn tesztet* általánosabb ún. *proximitási tesztekkel* helyettesítette. A *Graeffe-módszer* globálisan konvergens változatát dolgozta ki TURÁN PÁL [16]. A *Newton-módszer* globálisan konvergens változatát MUROTA [11] publikálta.

2. A Schur-kritérium

Tekintsük a

$$p(z) = p_0(z) = \sum_{i=0}^{n} a_{i0} z^i = 0 \quad (a_{i0} \in \mathbf{C},\ a_{00} a_{n0} \neq 0) \tag{2.1}$$

polinomot és legyen

$$p_1(z) = T[p_0(z)] = \bar{a}_{00} p_0(z) - a_{n0} z^n \bar{p}_0(1/\bar{z}) = \sum_{i=0}^{n-1} a_{i1} z^i. \tag{2.2}$$

Világos, hogy $p_1(z)$ legfeljebb $(n-1)$-ed fokú. Legyen továbbá

$$p_j(z) = T\{T^{j-1}[p_0(z)]\} \quad (j \geqq 2,\ T^1 = T) \tag{2.3}$$

és legyen

$$p_j(z) = \sum_{i=0}^{n-j} a_{ij} z^i. \tag{2.4}$$

A képzési szabály alapján világos, hogy $p_j(z)$ foka mindaddig kisebb mint $p_{j-1}(z)$ foka, amíg $p_{j-1}(z)$ legalább elsőfokú. Ugyancsak triviális, hogy $p_{n+1}(z) \equiv 0$. Legyen

$$k = \min \{j | p_j(0) = 0\} \quad (k \leqq n+1). \tag{2.5}$$

Ekkor az $N[p_0(z)]$ karakterisztikus függvényt az

$$N[p_0(z)] = \begin{cases} 1, & \text{ha} \quad \exists j \in \{1, \ldots, k-1\}: a_{0j} < 0 \\ 0, & \text{ha} \quad a_{0j} > 0 \quad (j = 1, \ldots, k-1) \quad \text{és} \quad \operatorname{gr}(p_{k-1}(z)) = 0 \\ -1, & \text{egyébként} \end{cases}$$

előírással definiálhatjuk. Igazolható ([12]), hogy $N[p_0(z)] = 1$ esetén a polinomnak legalább egy gyöke van a $\{z \in \mathbf{C} \,|\, |z| < 1\}$ nyílt egységkörben. Ha $N[p_0(z)] = 0$, akkor $p_0(z)$-nek nincs gyöke a nyílt egységkörben.

Minthogy a $p_0(z)$ polinomnak akkor és csak akkor van gyöke a nyílt $D(c, r) = \{z \in \mathbf{C} \,|\, |z-c| < r\}$ körlemezen, ha a $g(z) = p_0(rz+c)$ polinomnak gyöke van a $D(0, 1)$ nyílt egységkörlapon, a *Schur-kritérium* segítségével a $p_0(z) \to g(z) = p_0(rz+c)$ transzformáció végrehajtása után $N[g(z)]$ meghatározásával eldönthetjük, hogy az

adott $D(c, r)$ nyílt körlemez tartalmazza-e a $p_0(z)$ polinom valamelyik gyökét, vagy nem.

Az $N[p_0(z)]=-1$ kivételes esetre vonatkozik a

2.1. TÉTEL. Ha $N[p_0(z)]=-1$, akkor létezik olyan $\delta=\delta(p_0)>0$ szám, hogy $N[p_0(\gamma z)]\geqq 0$ teljesül minden $\gamma\in[1-\delta, 1+\delta]\setminus\{1\}$ esetén.

Bizonyítás. Nyilvánvalóan elég az állítást az $(1, 1+\delta]$ intervallumra igazolni. Jelölje $a_{ij}(\gamma)$ a $p_j(\gamma z)$ polinom együtthatóit $(0\leqq i\leqq n-j, 0\leqq j\leqq n+1)$. Az $a_{ij}(\gamma)$ együtthatók γ polinomjai és $a_{i0}(\gamma)=a_{i0}\gamma^i (0\leqq i\leqq n)$. A képzési szabály alapján világos, hogy $a_{0,j+1}(\gamma)\equiv 0$ akkor és csak akkor teljesülhet, ha $\operatorname{gr}(p_j(\gamma z))\equiv 0$. Jelölje most $k(\gamma)$ a $p_0(\gamma z)$ polinomnak megfelelő (2.5)-beli k értéket és tegyük fel, hogy $N[p_0(z)]=-1$. Ekkor fennáll, hogy

$$a_{0j}(1)>0 \quad (1\leqq j<k(1)), \quad \operatorname{gr}(p_{k(1)-1}(z))>0.$$

Az $a_{ij}(\gamma)$ polinomok folytonossága miatt van olyan $I_1=(1, 1+\delta_1]$ intervallum $(\delta_1>0)$, hogy minden $\gamma\in I_1$ esetén

$$a_{0j}(\gamma)>0 \quad (1\leqq j<k(1)), \quad \operatorname{gr}(p_{k(1)-1}(\gamma z))>0.$$

Minthogy $a_{0,k(1)}(\gamma)\not\equiv 0$, van olyan $I_2=(1, 1+\delta_2]\subset I_1$ intervallum, hogy vagy $a_{0,k(1)}(\gamma)<0 (\gamma\in I_2)$, vagy $a_{0,k(1)}(\gamma)>0$ $(\gamma\in I_2)$ teljesül. Az első esetben nyilvánvalóan fennáll, hogy $N[p_0(\gamma z)]=1$ $(\gamma\in I_2)$. A második esetben $\operatorname{gr}(p_{k(1)}(\gamma z))\equiv 0$ esetén kapjuk, hogy $N[p_0(\gamma z)]=0$ teljesül minden $\gamma\in I_2$ értékre. Ha $\operatorname{gr}(p_{k(1)}(\gamma z))\not\equiv 0$, akkor van olyan $I_3=(1, 1+\delta_3]\subset I_2$ intervallum amelyen teljesül, hogy $\operatorname{gr}(p_{k(1)}(\gamma z))>0$ minden $\gamma\in I_3$ esetén. Ekkor ismételten alkalmazhatjuk az előbbi okfejtéseket. Az eljárás a $p_j(\gamma z)$ polinomok fokszámának csökkenése miatt véges lépésben megadja a tétel állításában szereplő intervallumot. Ezzel állításunkat igazoltuk.

A *Schur-kritérium* alkalmazásához, adott $p_0(z)$ polinom és $D(c, r)$ körlemez esetén, szükségünk van a $p_0(z)\to p_0(rz+c)$ transzformáció végrehajtására. Ezt két lépésben tehetjük meg úgy, hogy először végrehajtjuk a $p_0(z)\to g(z)=p_0(z+c)$ transzformációt, majd a $g(z)\to h(z)=g(rz)$ transzformációt. Az első transzformáció végrehajtására STEWART [13] az *iterált Horner-sémát* javasolja. Célszerűbb azonban a SHAW és TRAUB [15] által konstruált alábbi algoritmust alkalmazni, mert a transzformáció közel optimális műveletszámmal ($3n-2$ multiplikatív művelet, $0{,}5n(n+1)$ additív művelet (hajtható végre és a *Shaw—Traub eljárás* WOZNIAKOWSKI [17] eredménye szerint az fl, fl_2 és $\overline{fl}$ aritmetikákban numerikusan stabilis.

1. Algoritmus

Számítsuk ki a

$$T_i^{-1}=a_{n-i-1,0}x^{n-i-1}, \quad (i=0, 1, \ldots, n-1)$$

$$T_j^j=a_{n0}x^n, \quad (j=0, 1, \ldots, n)$$

$$T_i^j=T_{i-1}^{j-1}+T_{i-1}^j, \quad (j=0, 1, \ldots, n, \quad i=j+1, j+2, \ldots, n)$$

mennyiségeket.

Kimutatható ([15]), hogy $T_n^j=(p_0^{(j)}(x)/j!)x^j$ $(j=0, 1, \ldots, n-1)$. Minthogy

$$p_0(t+x) = \sum_{i=0}^{n} (p_0^{(i)}(x)/i!)t^i, \quad p_0^{(n)}(x)/n! = a_{n0}$$

az 1. algoritmussal a $p_0(z)\to p_0(z+c)$ transzformáció végrehajtható.

A $g(z)\to h(z)=g(rz)$ transzformációt részletesen STEWART [13] elemezte. A transzformáció stabilitásának növelésére a következő skálázást javasolta. Legyen $h(z)=\sum_{i=0}^{n} c_i z^i$ és Ω a számítógépen ábrázolható legnagyobb pozitív szám.

2. Algoritmus.

(i) Határozzuk meg a

$$\sigma = \max \{\lambda | 0 < \lambda \leqq \Omega,\ \lambda|b_i| \leqq \Omega \quad (i = 0, 1, \ldots, n)\}$$

számot, ahol b_i-t a $g(z)=\sum_{i=0}^{n} b_i z^i$ egyenlőség definiálja.

(ii) Ha $r<1$, akkor legyen

$$c_i = (\sigma r^i) b_i \quad (i = 0, 1, \ldots, n).$$

(iii) Ha $r>1$, akkor legyen

$$c_i = (\sigma r^{i-n}) b_i \quad (i = n, n-1, \ldots, 0).$$

Az algoritmussal kapcsolatos további stabilitási vizsgálatokat tartalmaz [13].

3. A Lehmer—Schur módszerek osztálya

Jelölje $\Phi(0, 1)=\{D_i\}_{i=1}^{l}$ a $D_0=D(0, 1)$ nyílt egységkörlap olyan fedését, amely kielégíti a $D_i=D(c_i, r_i)$, $c_i\in D_0$, $r_i<1$ $(1\leqq i\leqq l)$ és $l\geqq 3$ feltételeket. A tetszőleges $D(c, r)$ körlemez ehhez hasonló $\Phi(c, r)$ fedését a $\Phi(c, r)=\{D(c+rc_i, rr_i)\}_{i=1}^{l}$ előírással definiáljuk. A körök sorrendje tetszőleges lehet, azonban rögzített.

Legyen $D(z_0, R_0)$ olyan nyílt körlemez, amely tartalmazza a $p_0(z)$ polinom legalább egy gyökét. Például $z_0=0$ és $R_0=1+\max_i |a_{i0}/a_{n0}|$ megfelelő választás. Ezekután a *Lehmer—Schur algoritmus* $(d+1)$-edik lépését a következőképpen definiálhatjuk $(d\geqq 0)$.

Az $L(\Phi)$ algoritmus

(i) Az $i=1, 2, \ldots, l-1$ értékekre rendre ellenőrizzük, hogy van-e gyöke a $p_0(z)$ polinomnak a $D(z_d+R_d c_i, R_d r_i)$ nyílt körlapon. Ha igen, akkor legyen

$$z_{d+1} = z_d + R_d c_i, \quad R_{d+1} = R_d r_i$$

és a $(d+1)$-edik lépést befejeztük.

(ii) Ha nem találtunk gyököt a $\Phi(z_d, R_d)$ fedés első $l-1$ körében, akkor legyen

$$z_{d+1} = z_d + R_d c_l, \quad R_{d+1} = R_d r_l.$$

Az algoritmus $O((\max_i r_i)^d)$ sebességgel konvergál. Az esetleg előforduló $N[h(z)]=-1$ kivételes esetben az $R_d r_i$ értéket olyan $\gamma R_d r_i (\gamma>1)$ értékkel helyettesítjük, amelyre $N[h(\gamma z)]\geqq 0$ teljesül (lásd még [12]).

LEHMER eredeti algoritmusa a

$$\Phi^L(0,1) = \{D(0,1/2),\ D(c_i, 2/5) | i = 1, \ldots, 8\}$$

fedést használja, ahol $c_i = 0{,}75 \exp(2\pi j(i-1)/8)/\cos(\pi/8) (i=1, \ldots, 8)$. A j az imaginárius egységet jelöli. HENRICI módszerét a $\Phi^H(0,1) = \{D(0,r), D(c_i, r) | i=1, \ldots, 7\}$ fedés definiálja, ahol $r=(1+2\cos(2\pi/7))^{-1}$, $c_i = 2r\cos(\pi/7)\exp(2\pi j(i-1)/7)$ $(i=1, \ldots, 7)$.

A következőkben bevezetjük a vizsgálataink alapját képező komplexitási mutatókat. Válasszuk a számítási költség egy egységének a *Schur-kritérium* egy tetszőleges körlapra való alkalmazását. Az általánosság megszorítása nélkül feltehetjük, hogy $R_0=1$ és tekintsük mindazon (2.1) alakú polinomok halmazát amelyeknek van gyöke a $D(0,1)$ egységkörlemezen. Legyen továbbá

$$X_1(\Phi) = \max\left\{\frac{1}{|\log r_1|}, \ldots, \frac{l-1}{|\log r_{l-1}|}, \frac{l-1}{|\log r_l|}\right\}. \tag{3.1}$$

Ha a polinomok gyökét $\varepsilon>0$ pontossággal kívánjuk meghatározni $(\varepsilon>0)$, akkor a vizsgált polinomosztályban az $L(\Phi)$ algoritmus legnagyobb számítási összköltsége aszimptotikusan $X_1(\Phi)|\log\varepsilon|$ $(\varepsilon\to 0)$. Eszerint az $L(\Phi)$ algoritmus *Traub-féle komplexitási mutatóját* ([14], [7], [3]) a (3.1) mennyiséggel definiálhatjuk. Az átlagos számítási összköltséget jellemző komplexitási mutatót FRIEDLI [3] alapján azon feltevés mellett definiálhatjuk, hogy a vizsgált polinomosztály minden tagjának pontosan egy gyöke van az egységkörben, az egységkörbe nem eső gyökei pedig olyanok, hogy az iterációs eljárás nem vezethet ki az egységkörből. Tegyük fel továbbá, hogy a gyökök egyenletes eloszlásúak az egységkörlapon. Legyen p_i $(i=1, \ldots, l)$ annak a valószínűsége, hogy a $\Phi(0,1)$ fedés i-edik körében találunk gyököt, azaz p_i az i-edik körlemez bevonása után lefedett $D_0 \cap (D_i \setminus \bigcup_{j=1}^{i-1} D_j)$ új terület és az egységkörlap területének hányadosa. A tett feltevések mellett igazolható, hogy az $\varepsilon>0$ pontosságú közelítések $(\varepsilon\to 0)$ számítási összköltségének várható értéke aszimptotikusan $Z_1(\Phi)|\log\varepsilon|$ $(\varepsilon\to 0)$, ahol

$$Z_1(\Phi) = \Big(\sum_{i=1}^{l-1} i p_i + (l-1)\, p_l\Big)\Big/\Big(\sum_{i=1}^{l} p_i |\log r_i|\Big). \tag{3.2}$$

Az $L(\Phi)$ algoritmus átlagos komplexitási mutatóját tehát a (3.2) mennyiséggel definiálhatjuk. A $Z_1(\Phi)$ komplexitási mutató esetén a körök sorrendje lényeges, ui. nyilvánvalóan függ a körök sorrendjétől. A két mutató között a triviális $Z_1(\Phi)\leqq X_1(\Phi)$ egyenlőtlenség áll fenn.

A dolgozat fő célja a *Lehmer—Schur típusú módszerek* X_1 mutató szerinti optimalizálása. Ez alatt a következőket értjük.

Legyen $\mathcal{F}$ az eljárás definíciójában megengedett $\Phi(0,1)$ fedések halmaza. Az $\mathcal{F}^*\subset\mathcal{F}$ részosztályban az $L(\Phi^*)$ $(\Phi^*\in\mathcal{F}^*)$ módszert optimálisnak nevezzük, ha $\inf\{X_1(\Phi) | \Phi\in\mathcal{F}^*\} = X_1(\Phi^*)$. Minthogy adott $\mathcal{F}^*$esetén nem mindig létezik az osztályhoz tartozó optimális Φ^* fedés, a következőképpen kell eljárnunk. Legyen

$\tilde{\mathcal{F}}$ a $\bar{D}_0$ zárt egységkörlemez összes olyan zárt körlemezekből álló $\psi(0,1)=\{\bar{D}_i\}_{i=1}^{l}$ fedésének halmaza, amelyre $D_i=D(c_i,r_i)$, $c_i\in\bar{D}_0$, $r_i<1$ $(i=1,\ldots,l)$, $\bigcup_{i=1}^{l}\bar{D}_i\supset$ $\supset\bar{D}_0$ és $l\geqq 3$ teljesül. Világos, hogy minden $\psi\in\tilde{\mathcal{F}}$ esetén a $\{D_i|\bar{D}_i\in\psi\}$ halmazrendszer vagy $\mathcal{F}$-hez tartozik, vagy véges sok pont kivételével lefedi a D_0 nyílt egységkörlemezt. Minthogy ezen véges sok pont gyök volta könnyen ellenőrizhető, tekinthetjük a *Lehmer—Schur módszereket* az $\tilde{\mathcal{F}}$ fedésosztály felett megadottnak. Eszerint egy $\mathcal{F}^*\subset\tilde{\mathcal{F}}$ fedésosztályban az $L(\Phi^*)$ módszert optimálisnak tekintjük, ha $\Phi^*\in\mathcal{F}^*$ és $X_1(\Phi^*)=\inf\{X_1(\Phi)||\Phi\in\mathcal{F}^*\}$. STEWART [13] elemzései szerint stabilitási okokból célszerű bizonyos mértékű túlfedést biztosítani. Ezért $\tilde{\mathcal{F}}\setminus\mathcal{F}$-beli optimum esetén úgy is eljárhatunk, hogy az optimális fedésben levő körlemezek sugarának kismértékű növelésével és a növelés utáni körlemez határok elhagyásával olyan $\mathcal{F}$-beli fedést definiálunk, amelynek komplexitási mutatója adott pontossággal közelíti meg az optimumértéket.

FRIEDLI és HENRICI vizsgálataikban a (3.1), (3.2) mutatók helyett a technikailag könnyebben kezelhető

$$X_2(\Phi)=\max\{i/|\log r_i||i=1,\ldots,l\}, \tag{3.3}$$

$$Z_2(\Phi)=\Big(\sum_{i=1}^{l} ip_i\Big)/\Big(\sum_{i=1}^{l} p_i|\log r_i|\Big) \tag{3.4}$$

mutatókat használják, amelyek annak felelnek meg, hogy az utolsó l-edik körre is alkalmazzuk a *Schur-kritériumot*. Ez a skatulya elv miatt nyilvánvalóan szükségtelen, másrészt $X_1(\Phi)\leqq X_2(\Phi)$, $Z_1(\Phi)\leqq Z_2(\Phi)$ $(Z_2(\Phi)\leqq X_2(\Phi))$ miatt az optimumok is különbözhetnek. A logaritmus megválasztása az optimalitást nyilvánvalóan nem befolyásolja. Tetszőleges $d\in\mathbf{N}(d>1)$ alapú logaritmus használata esetén a tekintett komplexitási mutatók azzal a szemléletes tartalommal rendelkeznek, hogy mennyi maximális, ill. átlagos költség kell a gyökközelítés egy jeggyel történő javításához a d alapú számrendszerben.

Vizsgálatainkban FRIEDLI [3] alapján a 10-es alapú logaritmust használjuk. Az optimalizálást az X_2 komplexitási mutató szerint is elvégezzük. FRIEDLI és HENRICI módszereivel történő összehasonlítások esetén a (3.2), (3.4) komplexitási mutatókat is megadjuk.

4. A q-fedések és az optimális Lehmer—Schur módszerek

FRIEDLI [3] alapján a $\Phi(0,1)=\{D(c_i,r_i)|i=1,\ldots,l\}\in\mathcal{F}$ fedést *q-fedésnek* nevezzük, ha létezik $0<q<1$, hogy $r_i=q^i$ $(i=1,\ldots,l)$. A q-fedésekre fennáll, hogy $X_2(\Phi)=Z_2(\Phi)=1/|\log q|$. FRIEDLI megmutatta, hogy minden $\Phi(0,1)\in\mathcal{F}$ fedéshez megadható olyan $\hat{\Phi}(0,1)$ q-fedés, amelyre $X_2(\Phi)=X_2(\hat{\Phi})$. Igaz ugyanis, hogy a $q=\max_{1\leqq i\leqq l} r_i^{1/i}$ választás mellett a $\hat{\Phi}(0,1)=\{D(c_i,q_i)|i=1,\ldots,l\}$ körhalmaz $\mathcal{F}$-beli fedés és $X_2(\Phi)=X_2(\hat{\Phi})$. Tehát

$$\inf_{\Phi\in\mathcal{F}} X_2(\Phi)=\inf_{\Phi\in\mathcal{F}_2} X_2(\Phi), \tag{4.1}$$

ahol $\mathcal{F}_2$ a q-fedések halmazát jelöli. FRIEDLI [3] igazolta továbbá, hogy minden rögzí-

tett $l \geqq 3$ körszám esetén van megoldása a zárt egységkörlemez zárt q-fedéseire vonatko-

$$Q_2(l) = \min \left\{ q \middle| \bigcup_{i=1}^{l} \overline{D(c_i, q^i)} \supset \bar{D}_0 \right\} \tag{4.2}$$

zó extremum feladatnak. Összevetve ezt a rögzített körszám esetén is fennálló (4.1) egyenlőséggel kapjuk, hogy

$$\inf_{\Phi \in \mathscr{F}^l} X_2(\Phi) = \min_{\Phi \in \mathscr{F}_2^l} X_2(\Phi) = 1/|\log Q_2(l)|, \tag{4.3}$$

ahol $\mathscr{F}^l \subset \mathscr{F}$ az l-körös fedések, $\mathscr{F}_2^l \subset \tilde{\mathscr{F}}$ pedig az l-körös zárt q-fedések halmazát jelöli. Az egyenlőség a (4.2) feladat megoldására áll fenn.

FRIEDLI heurisztikus úton, véletlenszerű kereséseket használva meghatározott egy 11 körös q-fedést a $q=0{,}7698$ értékkel és egy 22 körös q-fedést a $q=0{,}7663$ értékkel. Alsó becsléseket adott meg továbbá $Q_2(l)$ értékére a $\sum_{i=1}^{l} q^{2i} = 1$ egyenlet pozitív valós gyökének segítségével. A következőkben FRIEDLI eredményeit kiegészítjük, ill. élesítjük, amennyiben megadjuk az optimális 3 és 4 körös *Lehmer—Schur módszereket* és az övénél élesebb alsó becslést adunk meg $Q_2(l)$ értékére. Ez utóbbi becslést közel optimális q-fedések konstruálására is felhasználjuk. Az előzőekhez hasonló eredményeket adunk meg az X_1 mutató esetére is. Ehhez szükségünk van a kvázi q-fedések fogalmának bevezetésére is. Egy $\Phi(0, 1) = \{D(c_i, r_i) | i = 1, \ldots, l\} \in \mathscr{F}$ fedést *kvázi q-fedésnek* nevezzünk, ha létezik $0 < q < 1$, hogy $r_i = q_i$ $(i = 1, \ldots, l-1)$ és $r_l = q^{l-1}$. Minden $\Phi(0, 1) \in \mathscr{F}$ fedéshez van olyan $\hat{\Phi}(0, 1)$ kvázi q-fedés, amelyre $X_1(\Phi) = X_1(\hat{\Phi})$. Legyen ugyanis

$$q = \max \left\{ \max_{1 \leqq i < l} r_i^{1/i}, r_l^{1/(l-1)} \right\}$$

és

$$\hat{\Phi}(0, 1) = \{D(c_i, q^i),\ D(c_l, q^{l-1}) | i = 1, \ldots, l-1\}.$$

Ekkor $\hat{\Phi}(0, 1) \in \mathscr{F}$ és $X_1(\Phi) = X_1(\hat{\Phi})$. Ha a kvázi q-fedések halmazát $\mathscr{F}_1$ jelöli, akkor (4.1)-hez hasonlóan igaz, hogy

$$\inf_{\Phi \in \mathscr{F}} X_1(\Phi) = \inf_{\Phi \in \mathscr{F}_1} X_1(\Phi). \tag{4.4}$$

FRIEDLI q-fedésekre vonatkozó bizonyítását felhasználva analóg módon igazolhatjuk, hogy minden rögzített $l \geqq 3$ körszám esetén van megoldása a zárt kvázi q-fedésekre vonatkozó

$$Q_1(l) = \min \left\{ q \middle| \bigcup_{i=1}^{l} \overline{D(c_i, r_i)} \supset \bar{D}_0,\ r_i = q^i (1 \leqq i < l),\ r_l = q^{l-1} \right\} \tag{4.5}$$

extremum feladatnak. Ha az l-körös zárt kvázi q-fedések halmazát $\mathscr{F}_1^l \subset \tilde{\mathscr{F}}$ jelöli, akkor igaz a (4.3)-nak megfelelő

$$\inf_{\Phi \in \mathscr{F}^l} X_1(\Phi) = \min_{\Phi \in \mathscr{F}_1^l} X_1(\Phi) = 1/|\log Q_1(l)| \tag{4.6}$$

egyenlőség. Világos, hogy (4.5) megoldása l-körös X_1 optimális *Lehmer—Schur módszert* szolgáltat.

A definíciók alapján könnyen látható, hogy $Q_1(l)$, $Q_2(l)$ monoton csökkenő és $Q_1(l) \leqq Q_2(l)$ $(l \geqq 3)$. Ezért igaz a

4.1. ÁLLÍTÁS. $\lim_{l \to +\infty} Q_1(l) = \lim_{l \to +\infty} Q_2(l)$ és

$$\inf_{\Phi \in \mathcal{F}} X_i(\Phi) = \lim_{l \to +\infty} 1/|\log Q_2(l)| \quad (i = 1, 2). \tag{4.7}$$

Bizonyítás. A későbbiekben igazolni fogjuk, hogy $Q_i(l) > 0{,}7489$ fennáll $l \geqq 3$, $i = 1, 2$ esetén. Emiatt $\inf_{\Phi \in \mathcal{F}} X_i(\Phi) = \lim_{l \to +\infty} 1/|\log Q_i(l)|$ $(i = 1, 2)$. Minthogy $Q_1(l) \leqq Q_2(l)$ $(l \geqq 3)$, azért fennáll, hogy $\inf_{\Phi \in \mathcal{F}} X_1(\Phi) \leqq \inf_{\Phi \in \mathcal{F}} X_2(\Phi)$. Jelölje $\Phi_1^l \in \mathcal{F}_1^l$ az l-körös extremális kvázi q-fedést, $\Phi_2^l \in \mathcal{F}_2^l$ pedig az l-körös extremális q-fedést. Ekkor fennáll, hogy

$$X_2(\Phi_2^l) \leqq X_2(\Phi_1^l) \leqq \frac{l}{l-1} X_1(\Phi_1^l) \quad (l \geqq 3),$$

ahonnan $\inf_{\Phi \in \mathcal{F}} X_2(\Phi) \leqq \inf_{\Phi \in \mathcal{F}} X_1(\Phi)$ következik. Emiatt $\inf_{\Phi \in \mathcal{F}} X_1(\Phi) = \inf_{\Phi \in \mathcal{F}} X_2(\Phi)$, ahonnan az állítás többi része már közvetlenül adódik.

Jelölje $q_1(l)$ a

$$\sum_{i=1}^{l-1} \arcsin q^i + \arcsin q^{l-1} = \pi \quad (l \geqq 3) \tag{4.8}$$

egyenlet, $q_2(l)$ pedig a

$$\sum_{i=1}^{l} \arcsin q^i = \pi \quad (l \geqq 3) \tag{4.9}$$

egyenlet egyetlen, a (0, 1) intervallumba eső valós gyökét.

Igaz a következő

4.1. TÉTEL. Az extremális sugárértékre fennállnak a

$$Q_i(l) \geqq q_i(l) \quad (l \geqq 3;\ i = 1, 2) \tag{4.10}$$

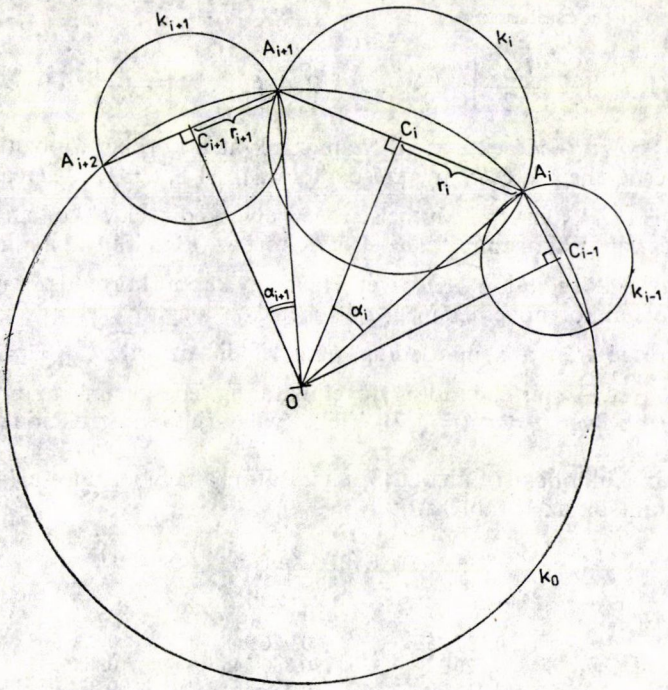
egyenlőtlenségek. Az $l = 3, 4$ $(i = 1, 2)$ esetekben egyenlőség áll fenn.

Bizonyítás. Csak az $i = 2$ esetre vonatkozó egyenlőtlenséget igazoljuk, mert a másik eset bizonyítása hasonló. Először azt igazoljuk, hogy a $q = q_2(l)$ értéknél kisebb paraméterű q-fedés nem fedheti le a D_0 egységkörlemezt. Jelölje k_i a D_i körlemez határát $(i = 0, 1, \ldots, l)$. Ha az adott q értékhez van olyan $\{\overline{D(c_i, q^i)} | i = 1, \ldots, l\}$ körhalmaz, amely lefedi $\overline{D}_0$-t, akkor a $k_1, \ldots, k_l$ köröknek le kell fedniük a k_0 egységkörvonalat. Ez csak akkor lehetséges, ha minden körnek van olyan átmérője, amelynek végpontjai az egységkörön vannak (azaz a k_i körök „átmérővel" fedik le k_0 egy ívét) és a körök ezeken az átmérő végpontokon csatlakoznak egymáshoz az 1. ábra szerint.

Eszerint a k_0 körvonal akkor van lefedve, ha

$$\sum_{i=1}^{l} \alpha_i = \sum_{i=1}^{l} \arcsin r_i = \sum_{i=1}^{l} \arcsin q^i \geqq \pi.$$

A $q = q_2(l)$ érték kielégíti ezt az egyenlőtlenséget és világos, hogy ennél kisebb q érték



1. ábra

esetén k_0, ill. D_0 nem fedhető le semmilyen q-fedéssel. Az $l=3, 4$ esetekben $A_1A_2A_3$, ill. $A_1A_2A_3A_4$ az egységkörbe beírt húrhárom-, ill. húrnégyszögek és így MOLNÁR JÓZSEF [10] egy tétele szerint az egységkörvonalat átmérővel fedő körlemezek az egységkörlemezt is lefedik. Ezzel tételünket igazoltuk.

A becslésben szereplő $q_i(l)$ $(i=1, 2)$ értékek szigorúan monoton csökkenőek és $q_1(l)<q_2(l)$ $(l\geqq 3)$. A definíciók alapján könnyen látható, hogy $\lim_{l\to+\infty} q_i(l)=$ $=q_\infty$ $(i=1, 2)$, ahol q_∞ a

$$\sum_{i=1}^{\infty} \arcsin q^i = \pi \tag{4.11}$$

egyenlet egyetlen, a (0, 1) intervallumba eső valós gyöke.

Kis l értékek $(l=5, 6, 7)$ esetén közvetlenül is ellenőrizhető, hogy nem lehet D_0-t $q=q_i(l)$ $(i=1, 2)$ paraméterű q-fedéssel, ill. kvázi q-fedéssel lefedni. Ugyanakkor az alsó becslés szerkezete lehetővé teszi közel optimális fedések konstruálását kis l értékekre. A $q_i(l)$ egy felfelé kerekített értékéből kiindulva az 1. ábra szerint lefedjük az egységkörvonalat oly módon, hogy a körök átmérővel fedjenek. A körök sorrendje pozitív irányú körüljárás esetén feleljen meg a sugarak, ill. a megfelelő hatványkitevők

$$1, 2\left[\frac{l-1}{2}\right]+1, \ldots, 5, 3, 2, 4, 6, \ldots, 2\left[\frac{l}{2}\right] \tag{4.12}$$

sorrendjének a q-fedések és a

$$1,\ l-1,\ 2\left[\frac{l-3}{2}\right]+1,\ \ldots,\ 5,\ 3,\ 2,\ 4,\ 6,\ \ldots,\ 2\left[\frac{l-2}{2}\right],\quad l-1 \tag{4.13}$$

sorrendnek a kvázi q-fedések esetén. Számozzuk át a köröket a pozitív körüljárási irány szerint és jelölje a k_i és k_0 metszéspontjait A_i, A_{i+1} $(i=1, \ldots, l)$ az 1. ábrának megfelelően. Ha A_{l+1} az $\widehat{A_1A_2}$ ív belsejében helyezkedik el, akkor az $\widehat{A_1A_{l+1}}$ íven kétszeres fedés van. Ekkor toljuk el az első, k_1 kört az origó irányába mindaddig amíg k_1 lefedi az egységkörvonal $\widehat{A_2A_{l+1}}$ ívét. Ha az így kapott körrendszer nem fedi le az egységkörlapot, akkor növeljük meg q értékét és a megnövekedett sugarú köröket toljuk el az origó irányába mindaddig amíg lefedik az $\widehat{A_iA_{i+1}}$ $(i=2, \ldots, l)$, ill. $\widehat{A_2A_{l+1}}$ $(i=1)$ ívet. A q értékét addig növeljük, amíg fedést nem kapunk. Természetesen a legkisebb ilyen $q=q_l^i$ $(i=1,2)$ értéket választjuk konstrukciónk paramétereként.

Eljárásunk különbözik FRIEDLI [3] eljárásától, amelyből csak a (4.12) feltételt vettük át. Tekintsük az 1. táblázatot.

1. TÁBLÁZAT

l	FRIEDLI [3]	$q_2(l)$	q_l^2
3	0,737 35	0,926 968	0,926 968
4	0,720 27	0,860 545	0,860 545
5	0,713 20	0,820 865	0,821
6	0,710 03	0,796 856	0,799 2
7	0,708 53	0,781 736	0,789 5
8	0,707 80	0,771 866	0,784
9	0,707 45	0,765 234	0,779 6
10	0,707 28	0,760 677	0,778 1
...	.....	.....	.....
∞	$\sqrt{2}/2$	0,748987*	0,7663**

Az 1. táblázat FRIEDLI [3] alsó becsléseit, a $q_2(l)$ értéket és a konstruált q_l^2 értékeket tartalmazza hat tizedesre kerekített formában. A *-gal jelölt szám 3×10^{-4} pontosságú alsó becslése q_∞-nek. A másik, **-gal jelzett szám FRIEDLI 22 körös q-fedésének paramétere.

A táblázat alapján több fontos következtetést vonhatunk le. A $q_i(l)\leqq Q_i(l)\leqq q_l^i$ $(i=1, 2)$ egyenlőtlenség alapján nyilvánvaló, hogy

$$0{,}748\,987 \leqq \lim_{l\to+\infty} Q_1(l) \leqq 0{,}7663, \tag{4.14}$$

amiből az abszolút optimumra vonatkozó

$$7{,}966\,49 \leqq \inf_{\Phi\in\mathcal{F}} X_1(\Phi) \leqq 8{,}650\,43 \tag{4.15}$$

becslést nyerjük. Ezek és a $q_2(l)$ alsó becslések jobbak mint FRIEDLI becslései. Ugyanakkor az is következik belőlük, hogy a FRIEDLI által talált $q_{22}^2=0{,}7663$ érték a $\lim_{l\to+\infty} Q_1(l)$ abszolút optimális sugárérték legfeljebb 2×10^{-2} hibájú, tehát igen jó, közelítése.

Az általunk konstruált fedések közel optimálisnak tekinthetők az $l=5, 6, 7$ esetben. Egyrészt fennáll $\delta Q_2(l) \leqq 0{,}008$ $(l=5, 6, 7)$, másrészt az extremális q-fedésnek hasonlónak kell lennie az általunk megadotthoz, amennyiben a $\bar{D}_0 \setminus D_1$ félhold kell a (4.12) szerint elrendezett $k_2, \ldots, k_l$ köröknek lefednie. Ez abból a tényből következik, hogy az extremális q-fedésben a k_i $(i=1, \ldots, l)$ körök mindegyikének fednie kell a k_0 körvonal egy pozitív hosszúságú ívét. Ha ugyanis elhagyjuk bármelyik, mondjuk a legkisebb sugarú kört, akkor $q_l^2 < q_2(l-1)$ $(4 \leqq l \leqq 7)$ miatt a megmaradó körök nem tudják k_0-t, ill. D_0-t lefedni. Hasonló módon lehet belátni, hogy $l \geqq 8$ esetén a q-sugarú $(q \leqq q_8^2)$ körnek (k_1) fednie kell a k_0 egységkörvonal egy pozitív hosszúságú ívét stb.

Az átlagos komplexitási mutatóra $Z_i(\Phi) \leqq X_i(\Phi)$ $(i=1, 2)$ miatt

$$\inf_{\Phi \in \mathcal{F}} Z_i(\Phi) \leqq \inf_{\Phi \in \mathcal{F}} X_1(\Phi) \leqq 8{,}650\,43 \tag{4.16}$$

fennáll. Az azonban nem ismeretes, hogy $\inf_{\Phi \in \mathcal{F}} X_1(\Phi)$-nél van-e kisebb átlagos komplexitási mutatójú fedés.

Tekintsük most a kvázi q-fedésekre vonatkozó 2. táblázatot.

2. TÁBLÁZAT

l	$q_1(l)$	q_l^1
3	0,915 186	0,915 186
4	0,849 557	0,849 557
5	0,812 875	0,813 2
6	0,791 289	0,795 4
7	0,777 869	0,788 5
8	0,769 160	0,783 5
...	.....	.....
∞	0,748 987	0,766 3

A táblázatból hasonló következtetések vonhatók le mint az 1. táblázatból. A két táblázat adatai alapján igazolható, hogy $l=5, 6, 7$ esetén az extremális kvázi q-fedéseknek az általunk konstruáltakhoz hasonlóaknak kell lenniük. Ha ugyanis bármelyik kört, mondjuk a legkisebb q^{l-1} sugarút elhagynánk, akkor $q_l^1 < q_2(l-1)$ $(l=5, 6, 7)$ miatt a megmaradó $l-1$ körvonal nem tudja a k_0 körvonalat lefedni.

Lehmer és Henrici módszerei az

$$X_1(\Phi^L) = 20{,}135, \quad X_1(\Phi^H) = 19{,}9089, \quad Z_2(\Phi^L) = 11{,}143, \quad Z_2(\Phi^H) = 11{,}168$$

komplexitási mutatókkal rendelkeznek ([3]). Az eddigi eredmények alapján könnyen igazolható, hogy a legkisebb körszámú fedés, amelyre mindkét mutatóban fennáll az $X_i(\Phi) \leqq Z_2(\Phi^L)$ $(i=1, 2)$ egyenlőtlenség, legalább 6 körös q-fedés. Az általunk konstruált hat körös q-fedés is kielégíti ezt a feltételt, ezért a fedés adatait a 3. táblázatban megadjuk.

3. TÁBLÁZAT

i	Re c_i	Im c_i	r_i
1	0,543 001 636	0	0,799 2
2	−0,691 665 726	−0,128 510 216	0,638 720 64
3	−0,406 039 69	0,683 577 178	0,510 465 536
4	−0,239 074 846	−0,818 753 509	0,407 964 057
5	0,281 818 842	0,845 796 974	0,326 044 874
6	0,348 926 122	−0,849 297 389	0,260 575 064

A most megadott fedés az X_i komplexitási mutatókban kb. 50%-os javítást jelent LEHMER, ill. HENRICI módszereihez képest. Stabilitási okokból célszerű q értékét 0,8-ra megnövelni (vö. [13]).

5. Optimális kongruens és körgyűrűszerű fedések

Az előző szakaszban beláttuk, hogy rögzített körszám esetén az optimális *Lehmer—Schur módszerek* csak q-fedéseken, ill. kvázi q-fedéseken alapulhatnak, vagy pedig olyan fedéseken, amelyek sugaraira $r_j \leqq (Q_i(l))^j$ $(j=1, \ldots, l)$ teljesül legalább egy esetben egyenlőséggel. Ugyanakkor a (4.1) abszolút optimum érték nem egy kitüntetett véges körszámú fedésre érhető el. Ezért érdekes megvizsgálnunk olyan fedésosztályokat is, amelyek egyrészt az eredeti *Lehmer-módszernek* természetes általánosításai ([5], [7]), másrészt az optimum véges körszámra áll fenn.

Jelölje $\mathscr{F}_3$ a

$$\Phi(0, 1) = \{D(0, 1-\delta),\ D(c_i, r) \mid i = 2, \ldots, l\} \tag{5.1}$$

alakú fedések osztályát, ahol $0<\delta<1$ és $l \geqq 6$.

A kongruens

$$\Phi(0, 1) = \{D(c_i, r) \mid i = 1, \ldots, l\} \tag{5.2}$$

körfedések $(l \geqq 3)$ osztályát pedig jelölje $\mathscr{F}_4$.

Tekintsük először az $\mathscr{F}_3$ fedésosztályra vonatkozó optimum feladatot, azaz az $X_i(\Phi) \to \min$ $(\Phi \in \mathscr{F}_3)$ problémát $(i=1, 2)$. Az (5.1) fedésosztályhoz az

$$R(l-1, \delta) = \min \left\{ \max_{2 \leqq i \leqq l} r_i \,\middle|\, \bigcup_{i=2}^{l} \overline{D(c_i, r_i)} \supset \bar{D}_0 \setminus D_1 \right\} \tag{5.3}$$

geometriai extremum feladatot rendelhetjük hozzá, ahol D_1 a $D(0, 1-\delta)$ körlemezt jelöli. FRIEDLI [3] q-fedésekre vonatkozó bizonyításának egyszerű módosításával könnyen belátható, hogy az (5.3) feladatnak mindig van megoldása. A feladat jellegéből következik, hogy rögzített l körszám esetén $R(l-1, \delta)$ monoton nő, ha a δ körgyűrűvastagság nő. Rögzített δ esetén pedig $R(l-1, \delta)$ monoton csökkenő, ha az l körszám nő. Legyen a továbbiakban $k=l-1$ és tekintsük a 2. ábrát!

A 2. ábra szerint megadott sugarú és helyzetű körökkel az ábrázolt körgyűrű nyilván lefedhető, míg kisebb sugarú körökkel már k_0 sem fedhető le. Eszerint vékony körgyűrűk esetén

$$R(k, \delta) = \sin(\pi/k) \quad (0 < \delta \leqq 2 \sin^2(\pi/k),\ k \geqq 3), \tag{5.4}$$

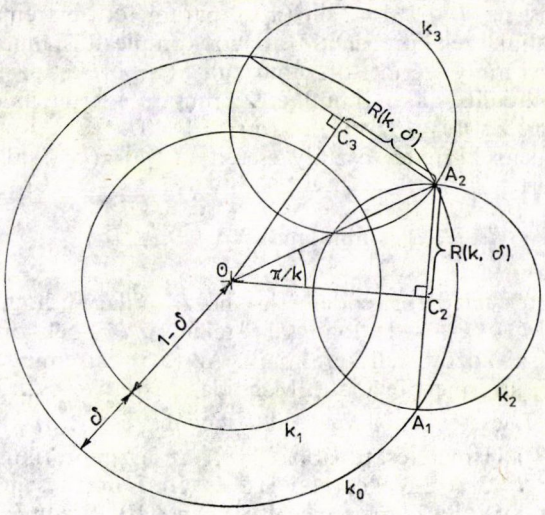
és az extremális fedés a 2. ábra szerinti.

Az $l=4, 5$ $(k=3,4)$ esetekben az (5.4) extremális körfedések a teljes egységkörlapot lefedik (vö. [10], [18]). Ezért kell az $l \geqq 6$ kikötést tenni a körgyűrűszerű fedések vizsgálatában.

A körgyűrűszerű, azaz a $\Phi \in \mathscr{F}_3$ fedések esetén

$$X_i(\Phi) = \max \left\{ \frac{1}{|\log(1-\delta)|},\ \frac{k+i-1}{|\log r|} \right\} \quad (i = 1, 2).$$

Az $R(l-1, \delta)$ extremális sugárérték monotonitása miatt rögzített l és $\Phi \in \mathscr{F}_3$ esetén



2. ábra

fennáll az éles

$$X_i(\Phi) \geqq (k+i-1)/|\log R(k,\delta)| \geqq (k+i-1)/\left|\log \sin\frac{\pi}{k}\right| \quad (0<\delta<1)$$

egyenlőtlenség ($i=1, 2$). Az egyenlőség akkor áll fenn, ha $1/|\log(1-\delta)| \leqq (k+i-1)/|\log \sin(\pi/k)|$. Eszerint rögzített l körszám esetén az $\mathcal{F}_3$-beli X_i optimális *Lehmer—Schur módszereket* a 2. ábra szerinti,

$$\delta \in I_l^i = \left[1-(\sin(\pi/k))^{1/(k+i-1)},\ 2\sin^2(\pi/k)\right] \quad (i=1, 2)$$

paraméterű extremális fedések szolgáltatják. Minthogy $(k+i-1)/|\log \sin(\pi/k)|$ minimumát $i=1$ esetén a $k=8$, $i=2$ esetén pedig a $k=9$ értékre éri el, igaz a következő

5.1. ÁLLÍTÁS. Az $\mathcal{F}_3$ fedésosztályban az optimális *Lehmer—Schur módszereket* $i=1$ esetén az $l=9$, $\delta \in I_9^1$ paraméterű, $i=2$ esetén pedig az $l=10$, $\delta \in I_{10}^2$ paraméterű, 2. ábra szerinti extremális fedések szolgáltatják.

Eszerint $\inf_{\Phi \in \mathcal{F}_3} X_1(\Phi) = 8/|\log \sin(\pi/8)| = 19{,}1773$, illetve $\inf_{\Phi \in \mathcal{F}_3} X_2(\Phi) = 10/|\log \sin(\pi/9)| = 21{,}4616$. Ha az I_l^i intervallumokat jobbról nyilttá tesszük, akkor $\inf X_i$ helyett $\min X_i$-t írhatunk.

Az $R(l-1, \delta)$ extremális sugárérték monotonitási tulajdonságai miatt nincs szükség az (5.3) extremum feladat teljes megoldására, csak a 2. ábra szerinti triviális esetben. Az (5.3) extremum feladatot NAGY DÉNES tanulmányozta és a körgyűrű vastagságára, valamint a k_i ($i=2, \ldots, l$) körök elhelyezkedésére vonatkozó technikai feltételek mellett általánosan megoldotta. Megmutatható, hogy LEHMER (pontosabban G. R. HAGEN, [12] 391. o., 50. feladat) és HENRICI fedései extremálisak a megfelelő paraméterekkel definiált szűkített fedésosztályokban.

Tekintsük most az $l \geqq 6$, $\delta = 2\sin^2(\pi/k)$ paraméterű extremális körgyűrűfedéseket és minimalizáljuk ezek körében a Z_i átlagos komplexitási mutatót. Ha a köröket a pozitív körüljárási irány szerint sorszámozzuk és a köröket az 1, 2, 4, ..., 3, 5, ... sorrend szerint teszteljük, akkor mindkét Z_i mutató szerint az $l = 7$ körös fedés a minimális. Ekkor $Z_1(\Phi) = 10{,}5164$, $Z_2(\Phi) = 10{,}8407$.

Az $\mathscr{F}_4$ kongruens körfedés osztály esetén $X_i(\Phi) = (k+i-1)/|\log r|$. Az (5.2) fedésekhez az ismert

$$R(l) = \min\left\{\max_{1 \leqq i \leqq l} r_i \,\middle|\, \bigcup_{i=1}^{l} \bar{D}_i \supset \bar{D}_0\right\} \tag{5.5}$$

geometriai szélsőérték feladatot rendelhetjük hozzá. Világos, hogy rögzített $l \geqq 3$ és $\Phi \in \mathscr{F}_4$ esetén $X_i(\Phi) \geqq (k+i-1)/|\log R(l)|$, tehát az $\mathscr{F}_4$-beli optimumot az (5.5) probléma megoldásai között kell keresnünk. Az (5.5) extremum feladat a diszkrét geometria egy régi, jól ismert feladata. Megoldása azonban csak néhány esetben ismert. Az $l = 3, 4, 7$ esetekben a megoldás triviális és $R(3) = \sqrt{3}/2$, $R(4) = \sqrt{2}/2$, $R(7) = 1/2$ ([6], [18]). Ismeretes továbbá ([18]) az aszimptotikus $R(l) \sim 1{,}0996/\sqrt{l}$ $(l \to \infty)$ becslés. Az $l = 5, 6$ eseteket NEVILLE és GRÜNBAUM sejtéseinek ([18]) pontosításával BEZDEK KÁROLY oldotta meg 1980-ban ([1]). Eredménye szerint $R(5) = 0{,}6098$ és $R(6) = 0{,}5559$ négy tizedes pontossággal.

Igaz az

5.1. TÉTEL. Az $\mathscr{F}_4$ kongruens körfedések osztályában

$$\inf_{\Phi \in \mathscr{F}_4} X_1(\Phi) = 4/|\log R(5)|, \tag{5.6}$$

és az X_1 mutató szerint optimális *Lehmer—Schur módszert* az ötkörös extremális kongruens körfedés definiálja.

Bizonyítás. Szükségünk van a triviális $R(l) > 1/\sqrt{l}$ $(l \geqq 3)$ egyenlőtlenségre és a $8 \leqq l \leqq 11$ esetén ennél élesebb

$$R(l) > \sin\left(\pi/(l-1)\right) \quad (l > 7) \tag{5.7}$$

egyenlőtlenségre. Tegyük fel, hogy l számú $\sin\frac{\pi}{l-1}$ sugarú zárt körlemez lefedi a $\bar{D}_0$-t. Minthogy $l > 7$ esetén $\sin\frac{\pi}{l-1} < 1/2$, nem lehet az összes körlapnak közös pontja k_0-val. Ugyanakkor k_0 lefedéséhez legalább $l-1$, átmérővel fedő kör kell. Legyenek a k_0-t lefedő körök a $k_1, k_2, \ldots, k_{l-1}$ körök. Jelölje A_i és B_i a k_i és a k_{i+1} körök $(i = 1, \ldots, l-1)$ metszéspontjait úgy, hogy fennálljon $\overline{|OB_i|} < \overline{|OA_i|} = 1$. Ekkor a $B_1, B_2, \ldots, B_{l-1}$ pontok az $\overline{|OB_i|} = 1 - 2\sin^2(\pi/k)$ sugarú, origó középpontú körbe beírt szabályos konvex $l-1$ szöget alkotnak. Minthogy $\overline{|OB_i|} > \frac{1}{2} > \sin(\pi/k)$ $(l > 7)$, az l-edik kör nem fedheti le a $\overset{\frown}{B_1B_2}, \ldots, \overset{\frown}{B_{l-2}B_{l-1}}, \overset{\frown}{B_{l-1}B_1}$ körívekkel határolt belső tartományt. Ez ellentmond a kiinduló feltevésnek, tehát (5.7) igaz. Ha $l > 7$, akkor (5.7)-ből következik, hogy $\Phi \in \mathscr{F}_4$ esetén

$$X_1(\Phi) \geqq 8/|\log \sin(\pi/8)|, \tag{5.8}$$

illetve

$$X_1(\Phi) \geqq (l-1)/|\log 1/\sqrt{l}| \quad (l > 12). \tag{5.9}$$

Az ismert $R(l)$ értékek és az alsó becslések figyelembevételével adódik, hogy

$$\min_{l \geqq 3} (l-1)/|\log R(l)| = 4/|\log R(5)|.$$

Ezzel a tételt igazoltuk.

Az optimum értéke $\inf_{\Phi \in \mathcal{F}_4} X_1(\Phi) = 18{,}6209$. Az X_2 komplexitási mutatóra vonatkozó optimum feladatot az eddig ismert eredmények alapján megoldani nem tudjuk, ui. ehhez az (5.7)-nél jóval élesebb alsó becslés és $8 \leqq l \leqq 12$ esetén az extremum feladat megoldása kellene. Az azonban könnyen igazolható, hogy

$$\inf_{\Phi \in \mathcal{F}_3} X_2(\Phi) < \inf_{\Phi \in \mathcal{F}_4} X_2(\Phi). \tag{5.10}$$

Megjegyezzük, hogy a négy körös optimális *Lehmer—Schur módszerek* X_i komplexitási mutatói lényegesen jobbak mint a szakaszban nyert módszerekéi. A Z_i átlagos komplexitási mutatóval kapcsolatban elegendő az előző szakasz végére utalni.

Végül köszönetemet fejezem ki NAGY DÉNESnek a körgyűrű fedésekkel kapcsolatban nyújtott segítségéért.

IRODALOM

[1] BEZDEK, K., „Über einige Kreisüberdeckungen", *Beiträge zur Algebra und Geometrie* **14** (1983) 7—13.
[2] FEJES TÓTH, L., *Regular figures* (Pergamon Press, Oxford, 1964).
[3] FRIEDLI, A., "Optimal covering algorithms in methods of search for solving polynomial equations", *JACM* **20** (1973) 290—300.
[4] GALÁNTAI, A., „Megjegyzések algebrai egyenletek közelítő megoldásához", *Alkalmazott Matematikai Lapok* 2 (1976) 115—122.
[5] GALÁNTAI, A., "On the optimization of Lehmer—Schur type methods", *Numerikus Módszerek* 6 (1976, ELTE TTK Numerikus és Gépi Matematikai Tanszék).
[6] HAJÓS, GY., NEUKOMM, GY., SURÁNYI, J., *Matematikai versenytételek II.* (Tankönyvkiadó, Budapest, 1964).
[7] HENRICI, P., "Methods of search for solving polynomial equations", *JACM* **17** (1970) 273—283.
[8] LEHMER, D. H., "A machine method for solving polynomial eguations", *JACM* **2** (1961) 151—163.
[9] LEHMER, D. H., "Search procedures for polynomial equation solving", in: *Constructive aspects of the fundamental theorem of algebra"*, (eds.: *Dejon, B., Henrici, P.*), Wiley, London, 1969.
[10] MOLNÁR, J., „Egy elemi geometriai szélsőértékfeladat", *Matematikai és Fizikai Lapok* XLIX (1942) 249—253.
[11] MUROTA, K., "Global convergence of a modified Newton iteration for algebraic equations", *SIAM J. Num. Anal.* **19** (1982) 793—799.
[12] RALSTON, A., *Bevezetés a numerikus analízisbe* (Műszaki Könyvkiadó, Budapest, 1960).
[13] STEWART III., G. W., "On Lehmer's method for finding zeros of a polynomial", *Mathematics of Computation* **23** (1969) 829—835.
[14] TRAUB, J. F., *Iterative Methods for the Solution of Equations* (Prentice—Hall, Englewood Cliffs, N. J., 1964).
[15] SHAW, M., TRAUB, J. F., "On the number of multiplications for the evaluation of a polynomial and some of its derivatives", *JACM* **21** (1974) 161—167).
[16] TURÁN, P., "Power sum method and the approximative solution of algebraic equations", *Mathematics of Computation* **29** (1975) 311—318.

[17] WOZNIAKOWSKI, H., "Rounding error analysis for the evaluation of a polynomial and some of its derivatives", *SIAM J. Num. Anal.* 11 (1974) 780—787.
[18] Шклярский, Д. О., Ченцов, Н. Н., Яилом, И. М., *Геометрические оценки и загачи из комбинаторной геометрии* (Наука, Москва, 1974).

(Beérkezett: 1984. október 15.)

DR. GALÁNTAI AURÉL
AGRÁRTUDOMÁNYI EGYETEM MATEMATIKAI ÉS SZÁMÍTÁSTECHNIKAI INTÉZET
2103 GÖDÖLLŐ

ON THE OPTIMIZATION OF THE LEHMER—SCHUR METHOD

A. GALÁNTAI

In this paper we construct the *optimal Lehmer—Schur method* for coverings using 3 and 4 disks. Near optimal covering algorithms are also given for 5, 6 and 7 disks. The lower estimation of FRIEDLI [3] for the complexity measure of the optimal methods is improved. In addition the optimal methods are established for the annulus like as well as congruent coverings.

Alkalmazott Matematikai Lapok **11** (1985) 335—341

RUNGE—KUTTA MÓDSZEREK ANALITIKUS HIBABECSLÉSEIRŐL

GALÁNTAI AURÉL

Gödöllő

A dolgozatban analitikus hibabecsléseket adunk meg a *Runge—Kutta módszerek* lokális hibájára és megvizsgáljuk a becslések néhány következményét.

1. Bevezetés

Konkrét explicit *Runge—Kutta módszerek* képlethibájára számos a módszerek rendjével azonos nagyságrendű analitikus hibabecslés ismeretes (BIEBERBACH [1], CARR [3], GALLER—ROZENBERG [14], LOTKIN [16]). Ezek a becslések a differenciálegyenletek jobb oldalának (lásd (2.1)) különböző rendű vegyes parciális deriváltjaira vonatkozó a priori korlátoktól függenek, ezért alkalmazhatóságuk erősen korlátozott ([9]).

COOPER [6] olyan, a módszer rendjétől független hibabecsléseket vizsgált, amelyek csak a differenciálegyenlet jobb oldalának *Lipschitz-konstansától* és az elméleti megoldás adott (q-adik, $q \geqq 2$) deriváltjától függenek. Megjegyezzük, hogy hasonló becslések a lineáris többlépéses módszerekre is ismeretesek ([15]).

COOPER vizsgálatainak eredményeként megadta néhány alacsonyabbrendű explicit *Runge—Kutta módszer* és az implicit trapézformula olyan hibabecslését, amely a megoldás második deriváltjától függ. COOPER nem adott meg explicit feltételt ilyen típusú hibabecslések létezésére, ugyanakkor konstrukciói meglehetősen konyolultak.

A dolgozatban CHARTRES és STEPLEMAN [4] más célokra kifejlesztett, de COOPERÉhez hasonló technikáját felhasználva megmutatjuk, hogy meglehetősen természetes feltételek teljesülése esetén mindig létezik a megoldás q-adik deriváltjától függő analitikus hibabecslés. Speciális esetként kapjuk, hogy minden *Runge—Kutta módszernek* van a megoldás második deriváltjától függő képlethibabecslése.

A hibabecslések fennállásának több fontos következménye van. Ezek közül az aktuális konvergencia rendre, a *Peano-féle hibareprezentáció* létezésére és a *B-konvergenciára* vonatkozóakat vizsgáljuk meg.

2. Alapfogalmak, jelölések

Legyen $D \subset R^m$ nyílt tartomány és tekintsük az

$$y' = f(t, y); \quad y(t_0) = y_0 \quad (t_0 \in [a, b],\ y_0 \in D) \tag{2.1}$$

Cauchy-problémát, amelyről feltesszük, hogy $f \in C([a, b] \times D, R^m)$ és létezik $L=$

$=L(f)\geqq 0$ konstans úgy, hogy

$$\|f(t, y)-f(t, z)\| \leqq L\|y-z\| \quad (t\in[a, b];\ y, z\in D). \tag{2.2}$$

Feltesszük továbbá, hogy minden $t_0\in[a, b]$, $y_0\in D$ esetén létezik $y: R\to R^m$ megoldás úgy, hogy $D(y)=[a, b]$, $R(y)\subset D$.

Legyen $t_0<x\leqq b$ tetszőleges pont és legyen π_x a $[t_0, x]$ intervallum összes felosztásainak halmaza. Jelöljön $\Delta_N\in\pi_x$ olyan $\Delta_N=\{t_n\}_{n=0}^N$ felosztást, amelyre

$$t_0 < t_1 <\ldots< t_N = x \tag{2.3}$$

fennáll. Legyen $\|\Delta_N\| := \max\limits_{0\leqq i\leqq n-1} h_i$ a Δ_N felosztás normája, ahol $h_i=t_{i+1}-t_i$ az i-edik lépéshossz. Jelölje továbbá a pontos megoldás $t_i\in\Delta_N$ pontbeli közelítését y_i. Ekkor az s-pontos *Runge—Kutta módszereket* a következő formában definiálhatjuk

$$\begin{aligned} y_{n+1}-y_n &= h_n \sum_{i=1}^{s} c_i k_i(y_n), \\ k_i(y_n) &= f\Big(t_n+a_i h_n,\ y_n+h_n \sum_{j=1}^{s} b_{ij} k_j(y_n)\Big) \quad (i=1, \ldots, s), \end{aligned} \tag{2.4}$$

ahol $a_i, b_{ij}, c_i\in R$ $(i, j=1, \ldots, s)$ adott konstansok, amelyek kielégítik a

$$\sum_{i=1}^{s} c_i = 1, \quad \sum_{j=1}^{s} b_{ij} = a_i \quad (i=1, \ldots, s) \tag{2.5}$$

feltételeket. Az általánosság megszorítása nélkül feltehetjük, hogy $a_i\geqq 0$ $(i=1, \ldots, s)$.

Legyen $c=(c_1, \ldots, c_s)^T\in R^s$, $B=(b_{ij})_{i, j=1}^s$, $a=(a_1, \ldots, a_s)^T\in R^s$ és $D=\operatorname{diag}(a_1, \ldots, a_s)$. Legyen továbbá $e=(1, \ldots, 1)^T\in R^s$.

A (2.4) *Runge—Kutta módszer* $t_n\in\Delta_N$ pontbeli *képlethibáját* az

$$L(y(t_n);\ h_n) = y(t_n)+h_n \sum_{i=1}^{s} c_i k_i(y(t_n))-y(t_{n+1}) \tag{2.6}$$

kifejezés definiálja (vö. pl. [10], [12]).

Szükségünk lesz még a BUTCHERtől [2] származó következő feltételekre

$$\begin{aligned} B(\xi)&: c^T D^{k-1} e = 1/k \quad (1\leqq k\leqq \xi) \\ C(\xi)&: BD^{k-1}e = \frac{1}{k} D^k e \quad (1\leqq k\leqq \xi). \end{aligned} \tag{2.7}$$

Vegyük észre, hogy a (2.5) feltételek a $B(1)$, ill. a $C(1)$ feltételekkel azonosak, tehát minden *Runge—Kutta módszer* kielégíti a (2.7) feltételt a $\xi=1$ értékre.

3. A lokális hiba becslése

Igaz a következő

3.1. TÉTEL. Ha a (2.4) *Runge—Kutta módszer* kielégíti a $B(q)$ és $C(q)$ $(q \geqq 1)$ feltételeket és a (2.1) differenciálegyenlet megoldásának $q+1$-edik deriváltja integrálható, akkor minden $1 \leqq r \leqq q$ indexhez van olyan $d_r > 0$ konstans, hogy $\|\Delta_N\| \leqq 1/(2L\|B\|_\infty)$ esetén

$$\|L(y(t_n);\ h_n)\| \leqq d_r h_n^r \int_{t_n}^{t_n+Ah_n} \|y^{(r+1)}(t)\|\, dt \quad (t_n \in \Delta_N) \tag{3.1}$$

teljesül, ahol $A = \max\{1, \|a\|_\infty\}$.

Megjegyezzük, hogy a (2.5) feltétel miatt minden *Runge—Kutta módszernek* van

$$\|L(y(t_n);\ h_n)\| \leqq d_1 h_n \int_{t_n}^{t_n+Ah_n} \|y''(t)\|\, dt \quad (t_n \in \Delta_N) \tag{3.2}$$

alakú hibabecslése, feltéve hogy y'' létezik és integrálható.

A (3.1) becslésből azonnal következik a

$$\|L(y(t_n);\ h_n)\| \leqq d_r' h_n^{r+1} \sup_{t_n \leqq t \leqq \tilde{t}_{n+1}} \|y^{(r+1)}(t)\| \quad (t_n \in \Delta_N) \tag{3.3}$$

becslés, ahol $d_r' = Ad_r$, $\tilde{t}_{n+1} = t_n + Ah_n$.

A 3.1. tétel bizonyítása. Az egyszerűség kedvéért legyen $h_n = h$, illetve $t_{n+1} = t_n + h$. Definíció szerint

$$L(y(t_n);\ h) = \tag{3.4}$$

$$= -\left[y(t_{n+1}) - y(t_n) - h\sum_{i=1}^{s} c_i y'(t_n + a_i h)\right] + h\sum_{i=1}^{s} c_i\left[k_i(y(t_n)) - y'(t_n + a_i h)\right].$$

Jelölje E az első szögletes zárójelpárban levő kifejezést. A t_n pont körül h szerint sorbafejtve kapjuk, hogy

$$E = \sum_{j=1}^{r} \frac{h^j}{(j-1)!} y^{(j)}(t_n)\left(\frac{1}{j} - \sum_{i=1}^{s} c_i a_i^{j-1}\right) +$$

$$+ \frac{1}{r!}\int_{t_n}^{t_n+Ah} \left[(t_{n+1}-t)_+^r - hr\sum_{i=1}^{s} c_i (t_n + a_i h - t)_+^{r-1}\right] y^{(r+1)}(t)\, dt,$$

ahol a $B(r)$ $(r \leqq q)$ feltétel fennállása miatt az első szumma zérussal egyenlő és

$$x_+^r = \begin{cases} x^r, & \text{ha } x \geqq 0 \\ 0, & \text{ha } x < 0 \end{cases} \quad (r \geqq 0).$$

Tehát

$$\|E\| \leqq \frac{h^r}{r!}(1 + r\|c^T D^{r-1}\|_1) \int_{t_n}^{t_n+Ah} \|y^{(r+1)}(t)\|\, dt.$$

A (3.4) kifejezés többi tagjának becsléséhez tekintsük a következő átalakításokat

$$\|k_i(y(t_n))-y'(t_n+a_ih)\| \leqq L\|y(t_n+a_ih)-y(t_n)-h\sum_{j=1}^{s} b_{ij}y'(t_n+a_jh)\|+$$

$$+hL\sum_{j=1}^{s}|b_{ij}|\|k_j(y(t_n))-y'(t_n+a_jh)\| \leqq L\|y(t_n+a_ih)-y(t_n)-$$

$$-h\sum_{j=1}^{s} b_{ij}y'(t_n+a_jh)\|+hL\|B\|_\infty \max_{1\leqq j\leqq s}\|k_j(y(t_n))-y'(t_n+a_jh)\|.$$

A $\|\Delta_N\|L\|B\|_\infty \leqq 1/2$ feltétel és az $(1-x)^{-1}\leqq 2 (0\leqq x\leqq 1/2)$ egyenlőtlenség alapján átrendezéssel kapjuk, hogy

$$\max_{1\leqq i\leqq s}\|k_i(y(t_n))-y'(t_n+a_ih)\| \leqq 2L \max_{1\leqq i\leqq s}\|y(t_n+a_ih)-y(t_n)-h\sum_{j=1}^{s} b_{ij}y'(t_n+a_jh)\|.$$

Ismét *Taylor-sorfejtést* alkalmazva kapjuk, hogy

$$y(t_n+a_ih)-y(t_n)-h\sum_{j=1}^{s} b_{ij}y'(t_n+a_jh)=\sum_{k=1}^{r}\frac{h^k}{(k-1)!}y^{(k)}(t_n)\left(\frac{a_i^k}{k}-\sum_{j=1}^{s} b_{ij}a_j^{k-1}\right)+$$

$$+\frac{1}{r!}\int_{t_n}^{t_n+A^*h}\left[(t_n+a_ih-t)_+^r-hr\sum_{j=1}^{s} b_{ij}(t_n+a_jh-t)_+^{r-1}\right]y^{(r+1)}(t)\,dt,$$

ahol az első szumma a $C(r)$ $(r\leqq q)$ feltétel miatt zérussal egyenlő és $A^*=\|a\|_\infty$. Tehát

$$\max_{1\leqq i\leqq s}\|k_i(y(t_n))-y'(t_n+a_ih)\| \leqq \frac{2L}{r!}h^r\|D^re+r|B|D^{r-1}e\|_\infty\int_{t_n}^{t_n+Ah}\|y^{(r+1)}(t)\|\,dt,$$

ahol $|B|=\{|b_{ij}|\}_{i,j=1}^{s}$. Az eddigi becslések összegzésével kapjuk, a bizonyítandó

$$\|L(y(t_n);\,h)\| \leqq d_r h^r\int_{t_n}^{t_n+Ah}\|y^{(r+1)}(t)\|\,dt$$

becslést, ahol a korábbiak alapján hibakonstansként a

$$d_r=\frac{1}{r!}(1+r\|c^TD^{r-1}\|_1+2hL\|c\|_1\|D^re+r|B|D^{r-1}e\|_\infty) \tag{3.5}$$

szám választható. Ha felhasználjuk a $\|\Delta_N\|L\|B\|_\infty\leqq 1/2$ feltételt, akkor a formálisan *Lipschitz-konstans* mentes

$$d_r=\frac{1}{r!}(1+r\|c^TD^{r-1}\|_1+\|B\|_\infty^{-1}\|c\|_1\|D^re+r|B|D^{r-1}e\|_\infty) \tag{3.6}$$

hibakonstanst választhatjuk. Ezzel a tétel állítását beláttuk.

Végül megjegyezzük, hogy a *Runge—Kutta formulák* közül az explicitekre $C(2)$ nem teljesül. Az *implicit Runge—Kutta formulák* közül a legalább $2s-2$ rendűekre $B(2s-2)$ és $C(s-2)$ teljesül (lásd még [13]).

4. Következmények, megjegyzések

A (3.1), (3.3) hibabecslések három fontos következményét vizsgáljuk meg. Elsőként a *Runge—Kutta módszerek* aktuális konvergencia rendjével foglalkozunk. Az aktuális konvergencia rend alatt azt a konvergencia sebességet értjük, amelyet egy p-edrendű módszer $p+1$-nél kevesebbszer differenciálható megoldású differenciálegyenleten elér. A lineáris többlépéses módszerekre ismert [5], hogy ha a (2.1) differenciálegyenlet megoldása $q+1$-szer deriválható ($1 \leqq q < p$) és a $q+1$-edik derivált integrálható, akkor minden p-edrendű lineáris többlépéses módszer q-adrendben konvergens. Ugyanakkor CHARTRES és STEPLEMAN [4] megmutatták, hogy a *Runge—Kutta módszerek* nem mindig rendelkeznek az előbbi aktuális konvergencia tulajdonsággal. Megadtak egy elégséges feltételt a fenti konvergencia tulajdonság fennállására és több *explicit Runge—Kutta formula* esetén kimutatták a megfelelő aktuális konvergenciát.

A következőkben megmutatjuk, hogy a (3.1) becslés fennállásából azonnal következik a becslés nagyságrendjének megfelelő aktuális konvergencia.

4.1. TÉTEL. Tegyük fel, hogy a (2.4) *Runge—Kutta módszer* együtthatói kielégítik a $0 \leqq a_i \leqq 1$ $(i=1, \ldots, s)$, a $B(q)$ és a $C(q)$ feltételeket $(q \geqq 1)$. Ha integrálható a megoldás $q+1$-edik deriváltja, akkor a *Runge—Kutta módszer* $O(\|\Delta_N\|^q)$ sebességgel konvergál a megoldáshoz.

Bizonyítás. Könnyen kimutatható ([10], [12]), hogy $\Delta_N \in \pi_x$ és $\|\Delta_N\| L \|B\|_\infty \leqq 1/2$ esetén létezik $D \geqq 0$ konstans úgy, hogy

$$\max_{1 \leqq i \leqq N} \|y_i - y(t_i)\| \leqq D \sum_{i=0}^{N-1} \|L(y(t_i);\, h_i)\|.$$

Az egyenlőtlenség jobb oldalát a (3.1) egyenlőtlenség felhasználásával tovább becsülve kapjuk, hogy

$$\max_{1 \leqq i \leqq N} \|y_i - y(t_i)\| \leqq D d_q \|\Delta_N\|^q \int_{t_0}^{x} \|y^{(q+1)}(t)\|\, dt$$

$(x = t_N)$, ahonnan az állítás közvetlenül adódik.

A tételbelinél gyengébb feltételek mellett igazol elsőrendű aktuális konvergenciát [4] és [11].

A *Runge—Kutta módszerek* $A[\alpha]$-stabilitásának vizsgálatánál fontos szerepe van ([13]) a képlethiba

$$L(y(t_n);\, h_n) = \int_{t_n}^{t_{n+1}} R_k(h_n, t - t_n)\, y^{(k)}(t)\, dt \quad (k \leqq q+1) \tag{4.1}$$

Peano-féle alakban való előállíthatóságának az

$$y' = Ay + g(t); \quad y(t_0) = y_0 \quad (A \in R^{m \times m},\ y_0 \in R^m) \tag{4.2}$$

alakú lineáris differenciálegyenletrendszerek osztályán. A [13] dolgozatbeli eredmények (2.3 lemma) alapján könnyen belátható, hogy $\|\Delta_N\| \leqq 1/(2L\|B\|_\infty)$ esetén a (3.1) becslés fennállásából a (4.1) hibareprezentáció fennállása következik a (4.2) problémák osztályán. Minthogy a (3.1) becslés a $B(q)$ és $C(q)$ feltételek együttes fennállásának következménye [13] alapján indokolt kimondani azt a sejtést, hogy (3.1) alakú hibabecslés akkor és csak akkor létezik, ha az adott módszerre $B(q)$ és $C(q)$ teljesül.

Végül a B-konvergenciával kapcsolatban teszünk néhány megjegyzést. A B-konvergencia fogalmát FRANK—SCHNEID—UEBERHUBER [7] vezették be és azt jelenti, hogy az

$$\langle f(t,u)-f(t,v), u-v\rangle \leqq m\|u-v\|^2 \quad (t\in R^+,\ u,v\in R^m) \tag{4.3}$$

feltételt ($m<0$) kielégítő disszipatív differenciálegyenleteken a *Runge—Kutta módszer* globális hibájára teljesül a

$$\|y_n-y(t_n)\| \leqq C(t_n)h^p \quad (t_n=nh,\ h>0,\ h\leqq h^*) \tag{4.4}$$

egyenlőtlenség, ahol $C(t)$ csak az m konstanstól, az f függvény f_y'-től különböző deriváltjaitól és az $y^{(i)}$ deriváltaktól függhet. A h^* lépéshossz korlát csak az m-től és az f fügvény f_y'-től különböző deriváltjaitól függhet.

A B-konvergencia fogalmának egyik fontos motivációja a konvergencia sebesség jellemzése volt $h\to 0$ esetén ([7]). Ha ugyanis a $C(t)$, illetve a $C(t)$-ben szereplő deriváltak kicsik és a h^* lépéshosszkorlát viszonylag nagy (esetleg nincs), akkor a konvergencia sebessége gyors. Tehát adott $\varepsilon>0$ pontosságú közelítő megoldás a B-konvergens módszerekkel hamarabb elérhető. FRANK—SCHNEID—UEBERHUBER ([7], [8]) számos B-stabilis *Runge—Kutta módszerosztályra* kimutatták a B-konvergenciát olyan p értékekre, amelyekre $B(p)$ és $C(p)$ egyidejűleg fennáll. (Ezt a p értéket *„stage-order"* -nek nevezték).

Ha eltekintünk a (4.3) egyoldali *Lipschitz-feltételtől*, akkor a 3.1, illetve 4.1. tételek alapján minden (2.1) alakú differenciálegyenletekre fennáll a

$$\max_{1\leqq i\leqq N}\|y_i-y(t_i)\| \leqq D'\|y^{(q+1)}\|_{C[t_0,x]}\|\Delta_N\|^q \tag{4.5}$$

egyenlőtlenség feltéve, hogy $y^{(q+1)}\in C[t_0,b]$, $\Delta_N\in\pi_x$, $x\leqq b$ és $\|\Delta_N\|\leqq 1/(2L\|B\|_\infty)$. Ekkor tehát

$$C(t)=D'\|y^{(q+1)}\|_{C[t_0,t]} \quad (t_0\leqq t\leqq b) \tag{4.6}$$

vehető, amely csak az f_y' és az $y^{(q+1)}$ függvényektől függ. Eszerint a B-konvergencia koncepcióban megkövetelt tulajdonságú globális hibabecslést valamilyen $q\geqq 1$ értékre minden *Runge—Kutta módszer* kielégíti és a B-konvergencia csak annyiban jelent többet, amennyiben a (4.3) tulajdonságú differenciálegyenleteken a B-stabilis *Runge—Kutta módszerekre* a h^* lépéshosszkorlát és a D' hibakonstans kedvezőbb.

IRODALOM

[1] BIEBERBACH, L., "On the remainder of the Runge—Kutta formula in the theory of ordinary differential equations", *ZAMP* **2** (1951) 233—248.

[2] BUTCHER, J. C., "Implicit Runge—Kutta processes", *Mathematics of Computation* **18** (1964) 50—64.

[3] CARR, J. W., "Error bounds for the Runge—Kutta single-step integration processes," *JACM* **5** (1958) 39—44.

[4] CHARTRES, B. A., STEPLEMAN, R. S., "Actual order of convergence of Runge—Kutta methods on differential equations with discontinuities", *SIAM J. Num. Anal.* **9** (1972) 476—499.

[5] CHARTRES, B. A., STEPLEMAN, R. S., "Covergence of linear multistep methods for differential equations with discontinuities", *Num. Math.* **27** (1976) 1—10.

[6] COOPER, G. J., "Error bounds for some singlestep methods", *Conference on the numerical solution of differential equations*, (ed.: *Morris, J. L. I.*), *Lecture Notes in Math.*, No. 109, Springer, Berlin, 1969, 140—147.

[7] FRANK, R., SCHNEID, J., UEBERHUBER, C. W., "The concept of B-convergence", *SIAM J. Num. Anal.* **18** (1981) 753—780.
[8] FRANK, R., SCHNEID, J., UEBERHUBER, C. W., "B-convergence of Runge—Kutta methods", Report NR. 48/81. Institut für Numerische Mathematik, Technische Universität, Wien, 1981.
[9] GALÁNTAI, A., „Egylépéses módszerek lokális hibabecslései", *MTA SZTAKI Tanulmányok* 46/1976.
[10] GALÁNTAI, A., "Convergence theorems and error analysis for one-step methods", *Annales Univ. Sci. Budapest, Sect. Math.* **19** (1976) 69—78.
[11] GALÁNTAI, A., "Discrete convergence to generalized solution of Cauchy-problems", Colloquia Mathematica Soc. János Bolyai 30. *"Qualitative theory of differential equations"* Szeged (Hungary) 1979. (ed.: Farkas M.) v. I., II., North-Holland, Amsterdam, 1981, 257—265.
[12] GALÁNTAI, A., „Egylépéses módszercsaládok konvergencia- és hibaanalízise", *Alkalmazott Matematikai Lapok* 9 (1983) 29—42.
[13] GALÁNTAI, A., „Lineáris differenciálegyenletek numerikus módszereinek stabilitása", *Alkalmazott Matematikai Lapok* (megjelenés alatt).
[14] GALLER, B. A., ROZENBERG, D. P., "A generalization of a theorem of Carr on error bounds for Runge—Kutta procedures", *JACM* **7** (1960) 57—60.
[15] JELTSCH, R., NEVANLINNA, O., "Stability and accuracy of time discretizations for initial value problems", *Num. Math.* **40** (1982) 245—296.
[16] LOTKIN, M., "On the accuracy of Runge—Kutta's method", *Math. Tables Aids. Comput.* **5** (1951) 128—133.

(Beérkezett: 1984. június 19.)

DR. GALÁNTAI AURÉL
AGRÁRTUDOMÁNYI EGYETEM MATEMATIKAI ÉS SZÁMÍTÁSTECHNIKAI INTÉZET
2103 GÖDÖLLŐ

ON THE ANALYTICAL ERROR ESTIMATIONS OF RUNGE—KUTTA METHODS

A. GALÁNTAI

In this paper we prove the estimation (3.1) for the local error of *Runge—Kutta methods* under the sufficient conditions $B(q)$ and $C(q)$ (see (2.7)). These conditions are conjectured to be also necessary. Three of the consequences of the estimation are discussed.

Alkalmazott Matematikai Lapok 11 (1985) 343—347

EGY HŐVEZETÉSI PROBLÉMA: A LOKÁLIS POTENCIÁL IDŐFÜGGÉSE

FARKAS HENRIK

Budapest

GLANSDORFF és PRIGOGINE felírtak egy egyenlőtlenséget, az úgynevezett evolúciós kritériumot. Kimutatjuk, hogy a levezetésükben szereplő érvelés matematikailag nem tekinthető szigorúan megalapozottnak, és megfogalmazunk egy tisztán matematikai problémát: meghatározandók az evolúciós kritérium érvényességének pontos feltételei.

1. Bevezetés

Adiabatikusan zárt (izolált) rendszer entrópiája az időnek monoton függvénye [10, 5]. Hallgatólagosan fel szokás tételezni, hogy a környezetétől minden szempontból (nemcsak adiabatikusan) izolált redszernek egyetlen stacionárius állapota van, ami egyben egyensúlyi állapot is. A véges izolált rendszerek entrópiája monoton növekedve tart az S_0 egyensúlyi értékhez, amint $t \to \infty$:

$$\lim_{t \to \infty} S(t) = S_0.$$

A közismert modelleknél ezt a határértéket az entrópia sohasem éri el, azaz

$$S(t) < S_0, \quad t < \infty.$$

Az entrópia tehát *Ljapunov függvény*, és az egyensúlyi állapot aszimptótikusan stabilis [11]. Nem izolált rendszerekre hasonló, általános jellegű állítás nem volt ismert. GLANSDORFF és PRIGOGINE arra törekedtek, hogy analóg állítást fogalmazzanak meg nyitott rendszerek stacionárius állapotra való lecsengésére.

Tekintsünk egy V térfogatot elfoglaló testet, amelyben „tiszta" hővezetés zajlik, hőtágulás és egyéb zavaró folyamatok nélkül. Ekkor a hővezetés differenciálegyenlete [10, 2, 3, 4]:

$$\varrho(\mathbf{r})c(\mathbf{r}, T)\frac{\partial T(\mathbf{r}, t)}{\partial t} = \nabla\big(\lambda(\mathbf{r}, t)\nabla T\big) \tag{1.1}$$

$\mathbf{r}$: helyvektor
t: idő
T: hőmérséklet; $T = T(\mathbf{r}, t)$
ϱ: sűrűség; $\varrho = \varrho(\mathbf{r})$
c: fajhő; $c = c(\mathbf{r}, T)$
λ: hővezetőképesség; $\lambda = \lambda(\mathbf{r}, T)$
∇: a Nabla operátor; $\nabla = \dfrac{\partial}{\partial \mathbf{r}}$

Itt egy olyan általános modellt vettünk alapul (*inhomogén izotróp Fourier test* [3, 4]), ami a legtöbbször jól megfelel a gyakorlat követelményeinek. A $\varrho(\mathbf{r})$, $c(\mathbf{r}, T)$ és $\lambda(\mathbf{r}, T)$ függvényekre kirótt matematikai jellegű követelményeket (folytonosság, differenciálhatóság) nem részletezzük: feltételezzük, hogy az előforduló operációk értelmezettek rájuk. Fizikai jellegű követelmény viszont, hogy ezek értéke pozitív:

$$\varrho > 0, \quad c > 0, \quad \lambda > 0. \tag{1.2}$$

Feltételezzük továbbá, hogy a test ∂V határfelületén teljesül a

$$T(\mathbf{r}, t) = f(\mathbf{r}) \quad \mathbf{r} \in \partial V \tag{1.3}$$

időfüggetlen határfeltétel. Tekintsük most az ehhez a határfeltételhez tartozó $T_0(\mathbf{r})$ stacionárius hőmérsékleteloszlást. Tehát $T_0(\mathbf{r})$ a

$$\begin{aligned} \nabla(\lambda(\mathbf{r}, T_0)\nabla T_0) &= 0 \quad \mathbf{r} \in V, \\ T_0(\mathbf{r}) &= f(\mathbf{r}) \cdot \mathbf{r} \in \partial V \end{aligned} \tag{1.4}$$

peremértékprobléma megoldása. Feltételezzük, hogy ez a megoldás létezik és egyértelmű.

GLANSDORFF és PRIGOGINE az onsageri irreverzibilis termodinamika formalizmusát alkalmazták. E szerint az úgynevezett „termodinamikai erő" az

$$\mathbf{X} = \nabla T^{-1}$$

módon van definiálva. Ezzel a GLANSDORFF és PRIGOGINE által megadott „evolúciós kritérium" [7, 11]:

$$\int_V \lambda(\mathbf{r}, T) T^2 \mathbf{X} \cdot \frac{\partial \mathbf{X}}{\partial t} \, dV \leqq 0. \tag{1.5}$$

Ez a tétel matematikailag kifogástalanul bizonyított: a *Gauss—Osztrogradszkij átalakítás* és az (1.1), (1.3) formulák felhasználásával könnyen látható, hogy esetünkben

$$\int_V \lambda(\mathbf{r}, T) T^2 \mathbf{X} \cdot \frac{\partial \mathbf{X}}{\partial t} \, dV = -\int_V \frac{\varrho(\mathbf{r}) c(\mathbf{r}, T)}{T^2} \left(\frac{\partial T}{\partial t}\right)^2 dV \tag{1.6}$$

amiből (1.2) miatt (1.5) már következik.

Figyeljük meg, hogy ha az

$$L = \lambda T^2 \tag{1.7}$$

mennyiség nem függ a hőmérséklettől, akkor (1.5) átírható a

$$\frac{dP}{dt} \leqq 0 \tag{1.8}$$

formába, ahol a

$$P = \int_V \lambda(\mathbf{r}, T) T^2 (\mathbf{X} \cdot \mathbf{X}) \, dV \tag{1.9}$$

mennyiség az entrópiaprodukció.

2. A probléma felvetése

GLANSDORFF és PRIGOGINE olyan mennyiséget kerestek, ami az entrópiaprodukció szerepét átveszi (1.8)-ban akkor, ha L függ a hőmérséklettől. A $T_0(\mathbf{r})$ stacionárius eloszlás környezetében vizsgálták az (1.5) egyenlőtlenséget: *„Neglecting higher order terms, we may then replace* $\lambda(T)T^2$ *by* $\lambda(T_0)T_0^2$ *in the neighborhood of this stationary state"* ([8]). Ezzel az indoklással jutnak el a

$$\frac{d\Phi}{dt} \leqq 0 \tag{2.1}$$

egyenlőtlenséghez, ahol

$$\Phi = \frac{1}{2}\int_V \lambda(\mathbf{r}, T_0)T_0^2(\mathbf{X}\cdot\mathbf{X})\,dV \tag{2.2}$$

az általuk bevezetett lokális potenciál [8]. Az érvelés logikája azonban nem meggyőző, mert itt a tagok „rendje" pontosan meg nem határozott, laza fogalom [1, 3]. Ezért indokolt az alábbi matematikai probléma felvetése:

1. Probléma. A $T(r, t)$ függvény elégítse ki az (11.) differenciálegyenletet és az (1.3) határfeltételt. $T_0(\mathbf{r})$ legyen az (1.4) peremértékprobléma megoldása. Milyen feltételek mellett érvényes a

$$\frac{d}{dt}\frac{1}{2}\int_V \lambda(\mathbf{r}, T_0)T_0^2(\nabla T^{-1})^2\,dV \leqq 0 \tag{2.3}$$

egyenlőtlenség?

Megjegyzendő, hogy GLANSDORFF és PRIGOGINE megfogalmazása megengedi, hogy (2.3) érvényességét csak a $T_0(\mathbf{r})$ stacionárius eloszlás közelében vizsgáljuk. Ennek megfelelően az 1. probléma enyhített változata:

1a. Probléma. A $T(\mathbf{r}, t)$ függvény elégítse ki az (1.1) differenciálegyenletet és az (1.3) határfeltételt, $T_0(\mathbf{r})$ pedig legyen az (1.4) peremértékprobléma megoldása. Milyen feltételek mellett igaz, hogy létezik olyan $t_1 > t_0$ időpont, hogy a (2.3) egyenlőtlenség teljesül minden $t > t_1$ esetén?

A probléma elképzelhető megoldásai közül különösen kettő lenne érdekes:

a) ha sikerülne bebizonyítani, hogy (2.3) általában (azaz minden „ésszerű" $\lambda(\mathbf{r}, T)$ esetén) érvényes (a fizikai szakirodalomban ez az általánosan elfogadott álláspont);

b) ha sikerülne egy fizikailag reális ellenpéldát találni: olyan $\lambda(\mathbf{r}, T)$ függvényt, amire (2.3) még az 1a. probléma enyhébb megfogalmazása szerint sem érvényes.

3. Megjegyzések

1. A fizikai modelltől eltekintve, és pusztán az idézett érvelés logikai indokoltságát vizsgálva felvethetjük az alábbi problémát:

2. *Probléma.* Legyen $a(\mathbf{r}, t)$ és $b(\mathbf{r}, t)$ két olyan függvény, amelyre teljesül, hogy

$$\int_V a(\mathbf{r}, t) b(\mathbf{r}, t) dV \leqq 0, \tag{3.1}$$

$$\lim_{t \to \infty} a(\mathbf{r}, t) = a_0(\mathbf{r}), \tag{3.2}$$

$$\lim_{t \to \infty} b(\mathbf{r}, t) = 0. \tag{3.3}$$

Milyen feltételek mellett következik ezekből, hogy

$$\int_V a_0(\mathbf{r}) b(\mathbf{r}, t) dV \leqq 0 \tag{3.4}$$

legalábbis elég nagy t-re?

Egy egyszerű ellenpéldát adhatunk (3.4)-re a szakaszonként folytonos függvények köréből. Feleljen meg V-nek a (0, 2) intervallum, és legyen

$$a(x, t) = \begin{cases} 1, & \text{ha } 0 < x \leqq 1 \\ 1 + e^{1-t}, & \text{ha } 1 < x < 2, \end{cases}$$

$$b(x, t) = \begin{cases} e^{-t}(1 + e^{-t}), & \text{ha } 0 < x \leqq 1 \\ -e^{1-t}, & \text{ha } 1 < x < 2. \end{cases}$$

Ekkor (3.1)—(3.3) teljesül, de (3.4) nem.

Az ellenpéldában szereplő a, b függvények az eredeti hővezetési probléma szempontjából több hátrányos tulajdonsággal bírnak:

a) Ésszerű fizikai interpretációjuk nem képzelhető el.

b) Nem folytonosak: $x = 1$-nél szakadásuk van.

c) $x = 2$-nél nemstacionárius határfeltételeknek tesznek eleget.

Habár a b) és a c) hátrányok feltehetően kiküszöbölhetők egy — az integrálokban elhanyagolható járulékú — módosítással az $x = 1$ és az $x = 2$ pont kis környezetében, egy ilyen módosítás az eredeti hővezetési problémához nem sokkal vinne közelebb. Az ellenpélda pusztán a [8]-ban szereplő érvelés hiányos voltát illusztrálja.

2. A lokális potenciálnak, mint T funkcionáljának minimuma van $T = T_0$-nál ([8]). A kérdés csak az, hogy ezt a minimumot időben monoton csökkenően közelíti-e meg, ahogy az evolúciós kritérium állítja, vagy nem.

3. [8]-ban GLANSDORFF és PRIGOGINE az ittenihez képest egy speciális esetet vizsgáltak: a homogén test esetét (ϱ: konstans, $c = c(T)$, $\lambda = \lambda(T)$), a közölt idézet is erre az esetre vonatkozik.

4. A tárgyalt probléma a GYARMATI [10] által bevezetett terminológia szerint az ún. entrópiaképben van megfogalmazva. A (2.3) egyenlőtlenségnek *Fourier-képben* egy egyszerűbb összefüggés felel meg:

$$\frac{d}{dt} \int_V \frac{1}{2} \lambda(\mathbf{r}, T_0)(\nabla T)^2 dV \leqq 0. \tag{3.5}$$

5. A lokális potenciállal megfogalmazott evolúciós kritérium biztosan érvényes, ha a vezetőképesség nem függ az állapottól. Így pl. (2.3) érvényes akkor, ha L nem függ T-től, (3.5) pedig akkor, ha λ nem függ T-től.

6. A tárgyalt evolúciós kritérium jelentőségét szemlélteti, hogy az szerepet játszott PRIGOGINEnak a disszipatív struktúrákkal kapcsolatos későbbi munkáiban [9, 6, 11], amely tevékenységéért 1977-ben *Nobel-díjjal* tüntették ki.

IRODALOM

[1] BÖRÖCZ, Sz., szóbeli közlés.
[2] CARSLAW, H. S. and JAEGER, J. C., *Conduction of Heat in Solids* (Clarendon, Oxford, 1959).
[3] FARKAS, H., „A hővezetés fenomenologikus elméletéről", Kandidátusi értekezés, Magyar Tudományos Akadémia, Budapest, 1974.
[4] FARKAS, H., "On the phenomenological theory of heat conduction", *Int. J. Engng. Sci.* **13** (1975) 1035—1053.
[5] FÉNYES, I., *Termosztatika és termodinamika* (Műszaki Könyvkiadó, Budapest, 1968).
[6] GLANDSDORFF, P., NICOLIS, G. and PRIGOGINE, I., "The Thermodynamic Stability Theory of Non-Equilibrium States", *Proc. Nat. Acad. Sci.* **71** (1974) 197—199.
[7] GLANSDORFF, P. et PRIGOGINE, I., "Sur les propriétés différentielles de la production d'entropie", *Physica* **20** (1954) 773—780.
[8] GLANSDORFF, P. and PRIGOGINE, I., "On a general evolution criterion in macroscopic physics", *Physica* **30** (1964) 351—374.
[9] GLANSDORFF, P. and PRIGOGINE, I., *Thermodynamic Theory of Structure, Stability and Fluctuations*, (Wiley, New York, 1971).
[10] GYARMATI, I., *Nemegyensúlyi termodinamika* (Műszaki Könyvkiadó, Budapest, 1967).
[11] NICOLIS, G. and PRIGOGINE, I., *Self-organization in Nonequilibrium Systems* (Wiley, New York, 1977).

(Beérkezett: 1984. december 2.)

(Átdolgozva beérkezett: 1985. március 15.)

FARKAS HENRIK
BUDAPESTI MŰSZAKI EGYETEM, FIZIKAI INTÉZET
1521 BUDAPEST, MŰEGYETEM RKP. 3.

A PROBLEM OF HEAT CONDUCTION: TIME DEPENDENCE OF THE LOCAL POTENTIAL

H. FARKAS

GLANSDORFF and PRIGOGINE had formulated an inequality, the so-called *"Evolution Criterion"*. It is shown that their reasoning in the deduction of that criterion is not rigorous from a mathematical point of view. A mathematical problem is posed: what are the limits of validity of that evolution criterion?

Alkalmazott Matematikai Lapok **11** (1985) 349—362

DIREKT LOGLINEÁRIS MODELLEK MAXIMUM LIKELIHOOD BECSLÉSE

RUDAS TAMÁS

Budapest

Többdimenziós diszkrét valószínűségeloszlások statisztikai vizsgálatának egyik legelterjedtebb módszere a loglineáris elemzés. Ennek során egy loglineáris eloszláscsalád egy elemét illesztjük a megfigyelésekhez. Az úgynevezett direkt modelleknek megfelelő loglineáris eloszláscsaládok rendelkeznek azzal a tulajdonsággal, hogy kiválasztott marginálisaik (a kompatibilitástól eltekintve) tetszőlegesen megadhatók. A maximum likelihood becslésre ezen modellek esetén zárt képlet létezik. Ennek kiszámítására könnyen programozható algoritmust adunk, majd a loglineáris modellek néhány olyan tulajdonságát tárgyaljuk, amelyek a gyakorlati alkalmazásokban fontosak lehetnek.

1. Bevezetés

Mindenféle többdimenziós statisztikai problémánál fontos kérdés a koordináták egymáshoz való viszonyának tisztázása. A matematikai szempontból legkézenfekvőbb modell ezzel kapcsolatban a koordináták függetlensége, ez azonban a valóban többdimenziós problémáknál nem teljesül.

Sokszor a kérdés fordítva vetődik fel, azaz úgy, hogy ismerve egy valószínűségi vektorváltozó alacsonyabb dimenziós marginális eloszlásait, hogyan tudjuk ezekből előállítani a teljes eloszlást. Ha valamennyi alacsonyabb dimenziós peremeloszlást ismerjük, akkor bizonyos feltételek teljesülése esetén a Kolmogorov-féle alaptétel biztosítja ezek kiterjeszthetőségét, még végtelen dimenziós szorzatterek esetében is.

Kérdés azonban, hogy milyenek azok az eloszlások, amelyek rendelkeznek azzal a tulajdonsággal, hogy a (véges dimenziós) szorzattéren értelmezett eloszlás előállítható pusztán bizonyos marginális eloszlásainak ismeretében.

Erre a kérdésre kézenfekvő az a válasz, hogy a többdimenziós normális eloszlások családja éppen ilyen tulajdonságú, hiszen egy ilyen oszlást a kétdimenziós peremeloszlások már egyértelműen meghatároznak. Ez a válasz csak korlátozott értelemben oldja meg a problémát, hiszen az már nem igaz, hogy mérhető tereknek egy halmazából a kétdimenziós szorzattereken tetszőlegesen megadott normális sűrűségekhez létezne normális sűrűség az összes tér szorzatán úgy, hogy ennek kétdimenziós marginálisai éppen az előre megadottak legyenek (ti. ez csak akkor van így, ha a megadott kovarianciákból képzett mátrix pozitív definit).

A felvetett kérdésekre diszkrét eloszlások körében fogunk választ adni.

A loglineáris eloszláscsalád fogalmának definiálása után a maximum likelihood becslés problémájával fogunk foglalkozni. Ismeretes, hogy bizonyos loglineáris eloszláscsaládok esetén a ML becslés meghatározására zárt képlet létezik. Egy könnyen programozható algoritmust ismertetünk ezekben az esetekben a ML becslés kiszámítására. Az utolsó részben bizonyos loglineáris eloszláscsaládok olyan tulajdonságait bizonyítjuk, amelyek a loglineáris elemzés gyakorlati alkalmazásaiban lehetnek fontosak.

2. A loglineáris eloszláscsalád

Az alábbiakban olyan diszkrét valószínűségeloszlásokkal fogunk foglalkozni, amelyeknél a szóba jövő események halmaza véges. Ilyenkor feltehetjük, hogy a mérhető terek maguk is végesek, és valamennyi részhalmazuk mérhető. Jelölje a vizsgált mérhető tereket

$$(\Omega_i, \mathbf{P}(\Omega_i)), \quad i = 1, \ldots, p$$

ahol Ω_i tetszőleges véges halmaz. Ezen mérhető terek szorzata

$$(\Omega, \mathbf{P}(\Omega)) = (\bigtimes_{i=1}^{p} \Omega_i, \bigotimes_{i=1}^{p} \mathbf{P}(\Omega_i)).$$

A gyakorlatban Ω_i-t legtöbbször egy Z_i diszkrét változó határozza meg. Ω-t kontingencia táblának, $\omega \in \Omega$-t a táblázat cellájának nevezik. A továbbiakban a $\{Z_1, \ldots, Z_p\}$ tetszőleges részhalmazának együttes eloszlását tekintjük a megfelelő mérhető terek szorzatán vizsgált mértéknek. Ha R_γ jelöli a $\{Z_1, \ldots, Z_p\}$ változók egy részhalmazát, akkor a megfelelő szorzatteret Ω_γ-val jelöljük. Ez a kontingencia tábla az Ω táblázat Ω_γ marginálisa, $\omega_\gamma \in \Omega_\gamma$ marginális cella. Ha $(P(\omega), \omega \in \Omega)$ valószínűségeloszlás Ω-n, akkor ennek vetülete Ω_γ-n a $(P(\omega_\gamma), \omega_\gamma \in \Omega_\gamma)$ eloszlás, amit peremeloszlásnak nevezünk. Nyilvánvaló, hogy

$$P(\omega_\gamma) = \sum\nolimits^{*} P(\omega), \quad \gamma \in \Gamma, \quad \omega_\gamma \in \Omega_\gamma,$$

ahol $\sum^{*}$ azokra az ω cellákra való összegzést jelöl, amelyeknek az Ω_γ térre való vetülete éppen ω_γ.

A bevezetésben vázolt problémánál kissé általánosabban azon p dimenziós diszkrét eloszlások családját fogjuk vizsgálni, amelyek bizonyos marginálisaikon értelmezett függvények segítségével előállíthatók. (A későbbiekben még visszatérünk annak vizsgálatára, hogy a peremeloszlások helyett tetszőleges, a marginálisokon értelmezett függvények megengedése mennyiben bővíti az eloszláscsaládot.)

Legyen Γ indexhalmaz és $\{R_\gamma, \gamma \in \Gamma\}$ olyan halmazrendszer, hogy $\gamma \in \Gamma$ esetén $R_\gamma \subset \{Z_1, \ldots, Z_p\}$, továbbá nincs Γ-nak olyan γ_1 és γ_2 eleme, hogy $R_{\gamma_1} \subsetneqq R_{\gamma_2}$.

Adott $\{R_\gamma, \gamma \in \Gamma\}$ mellett a vizsgált valószínűségeloszlások családját azon eloszlások alkotják, amelyek alkalmasan választott f_γ: $\Omega_\gamma \to \mathbf{R}$ függvényekkel a

$$P(\omega) = \prod_{\gamma \in \Gamma} f_\gamma(\omega_\gamma), \quad \omega \in \Omega \tag{2.1}$$

alakba írhatók.

A (2.1) alakú valószínűségeloszlások alkotják (adott $\{R_\gamma, \gamma \in \Gamma\}$ mellett) a loglineáris eloszláscsaládot.

Ha $\gamma \in \Gamma$, az R_γ-beli változókról azt mondják, hogy interakcióban vannak egymással. Maguk az R_γ részhalmazok interakciók. Ha R_γ s változót tartalmaz, akkor $s-1$-ed rendű interakciónak nevezik. A 0 rendű interakciók szokásos neve: főhatás. Feltesszük, hogy $\bigcup_{\gamma \in \Gamma} R_\gamma = \{Z_1, \ldots, Z_p\}$. Két egyszerű példát adunk olyan eloszlásra, amely loglineáris eloszláscsaládba tartozik.

1. Példa. Legyen $p=2$, $\Gamma = \{1, 2\}$, $R_1 = \{Z_1\}$, $R_2 = \{Z_2\}$. Ekkor az az eloszlás, amelynek peremeloszlásai függetlenek, az $\{R_\gamma, \gamma \in \Gamma\}$ által meghatározott loglineáris

eloszláscsaládba tartozik.

$$f_\gamma(\omega_\gamma) = P(\omega_\gamma), \quad \gamma \in \Gamma, \quad \omega_\gamma \in \Omega_\gamma.$$

2. Példa. Legyen $p=3$, $\Gamma=\{(1,3),(2,3)\}$, $R_{(1,3)}=\{Z_1, Z_3\}$, $R_{(2,3)}=\{Z_2, Z_3\}$. Ekkor az az eloszlás melynek első és második marginálisa feltételesen független a harmadikra nézve az $\{R_\gamma, \gamma\in\Gamma\}$ által meghatározott loglineáris eloszláscsaládba tartozik.

$$f_{(1,2)}(\omega_{(1,2)}) = \frac{P(\omega_{(1,2)})}{\sqrt{P(\omega_3)}},$$

$$f_{(2,3)}(\omega_{(2,3)}) = \frac{P(\omega_{(2,3)})}{\sqrt{P(\omega_3)}}.$$

Azt a feltevést, hogy az ismeretlen együttes eloszlás egy loglineáris eloszláscsaládba tartozik, loglineáris modellnek nevezik. Ezen feltevés mellett az eloszlás becslése és a hipotézisről való döntés a becsült és megfigyelt eloszlás összehasonlítása útján a loglineáris elemzés.

Ha N független megfigyelésünk van, akkor a megfigyelt gyakoriságokat tartalmazó $(X(\omega), \omega\in\Omega)$ vektorváltozó nyilván polinomiális eloszlású lesz N és $(P(\omega), \omega\in\Omega)$ paraméterekkel. (Más eloszlást feltételezve az alábbi állítások lényegében igazak maradnak, ld. BISHOP, FIENBERG, HOLLAND (1975)).

A. TÉTEL. Az $\{R_\gamma, \gamma\in\Gamma\}$ interakciókkal meghatározott loglineáris eloszláscsaládra nézve

$$\{X(\omega_\gamma), \ \gamma\in\Gamma, \ \omega_\gamma\in\Omega_\gamma\}$$

elégséges statisztika.

A továbbiakban jelölje

$$\hat{\tilde{P}}(\omega) = N^{-1}X(\omega), \quad \omega\in\Omega$$

az ún. megfigyelt eloszlást, $(\hat{P}(\omega), \omega\in\Omega)$ pedig a maximum likelihood becslést az $\{R_\gamma, \gamma\in\Gamma\}$ interakciókkal meghatározott loglineáris eloszláscsaládból az adott minta mellett. Ekkor igaz a

B. TÉTEL. A maximum likelihood becslésnek azon marginálisai, amelyek kijelölt interakciók, megegyeznek a megfigyelt eloszlás megfelelő marginálisaival, azaz

$$\hat{P}(\omega_\gamma) = \hat{\tilde{P}}(\omega_\gamma), \quad \gamma\in\Gamma, \quad \omega_\gamma\in\Omega_\gamma.$$

A loglineáris eloszláscsalád vizsgálatát, visszatérve a kiinduló problémához, a következőkkel is indokolhatjuk:

Tekintsük az $\{R_\gamma, \gamma\in\Gamma\}$ marginálisokat, valamint azon eloszlások $\varepsilon_{\Gamma,S}$ halmazát, amelyeknek $\{R_\gamma, \gamma\in\Gamma\}$ marginális eloszlásai megegyeznek egy $(S(\omega), \omega\in\Omega)$ eloszlás $\{R_\gamma, \gamma\in\Gamma\}$ marginálisaival:

$$\varepsilon_{\Gamma,S} = \{Q: Q(\omega_\gamma) = S(\omega_\gamma), \ \gamma\in\Gamma, \ \omega_\gamma\in\Omega_\gamma\}.$$

A definícióból következik, hogy $\varepsilon_{\Gamma,S}$ nem üres, ti. legalábbis $(S(\omega), \omega\in\Omega)$ eleme.

Egy általános statisztikai elv, a minimális diszkrimináló információ elve l. KULLBACK (1959) azon $\tilde{P}$ eloszlás választását írja elő az $\varepsilon_{\Gamma,S}$ eloszláshalmazból, amelyre

nézve a tetszőleges T, U, Ω-n értelmezett eloszlások esetén a

$$D(T\|U) = \sum_{\omega\in\Omega} T(\omega)\log\frac{T(\omega)}{U(\omega)},$$

$$\left(\log\frac{0}{U(\omega)} = -\infty,\quad \log\frac{T(\omega)}{0} = \infty,\quad 0(\pm\infty) = 0\right)$$

képlettel értelmezett információs divergencia egy rögzített R referencia eloszláshoz képest minimális, azaz

$$D(\tilde{P}\|R) = \inf_{Q\in\varepsilon_{\Gamma,S}} D(Q\|R)$$

(az információs divergencia tulajdonságait illetően l. CSISZÁR, (1975)).

Számunkra az az eset a legfontosabb, amikor

$$\varepsilon_{\Gamma,S} = \varepsilon_{\Gamma,\hat{P}},$$

azaz a megfigyelt eloszlás eleme a kijelölt eloszláshalmaznak.

Ebben az esetben az MDI elv szokásos alkalmazása az, hogy referencia eloszlásnak az E egyenletes eloszlást választjuk, azaz azt a $(\tilde{P}(\omega), \omega\in\Omega)$ eloszlást tekintjük MDI becslésnek (l. GOOD, (1963)), amelyre

$$D(\tilde{P}\|E) \leqq D(Q\|E),\quad Q\in\varepsilon_{\Gamma,\hat{P}}.$$

C. TÉTEL. (KULLBACK (1959)). Ha $\tilde{P}$ létezik, E-re vonatkozó sűrűsége (2.1) alakban írható.

Látjuk tehát, hogy az MDI elv alkalmazása bizonyos értelemben természetesen vezet a loglineáris eloszláscsalád vizsgálatához (l. CSISZÁR (1975), GOKHALE, KULLBACK (1978)).

Ezzel kapcsolatban megjegyezzük még, hogy pl. GOODMAN (1970) a szórásanalízis modelljével való formális hasonlósággal indokolja a loglineáris modellek vizsgálatát.

Az ML és MDI becslés közötti kapcsolatra vonatkozik a

D. TÉTEL. Ha $\hat{P}$ és $\tilde{P}$ közül legalább az egyik létezik, akkor a másik is, és

$$\hat{P} = \tilde{P}.$$

Az ML és MDI becslés létezésére vonatkozóan az alábbi elégséges feltételt adhatjuk

E. TÉTEL. (BIRCH (1963)). Ha

$$X(\omega) > 0,\quad \omega\in\Omega,$$

akkor létezik egyértelműen meghatározott ML (és MDI) becslés.

A H. tétel mutatja, hogy ez a feltétel nem szükséges.

A becslési feladat más megfogalmazásait és a fenti tételek bizonyítását illetően l. CSISZÁR (1985), RUDAS (1985).

Láttuk, hogy a loglineáris eloszláscsalád (adott $\{R_\gamma, \gamma\in\Gamma\}$ interakciók mellett) olyan tulajdonságú diszkrét eloszlásokból áll, amelyeket bizonyos alacsonyabb dimenziós marginálisaik meghatároznak. Továbbra is nyitott a kérdés, hogy melyek azok a loglineáris eloszláscsaládok, amelyeknek kijelölt marginálisai csak a kompati-

bilitásra ügyelve, egyébként tetszőlegesen előírhatók. Az erre a kérdésre adható válasz azon az észrevételen alapszik, hogy a kijelölt marginálisok $\{R_\gamma, \gamma \in \Gamma\}$ rendszere egy hipergráf (l. alább). Ennek a „hipergráfos szemléletnek" egy másik alkalmazását diszkrét valószínűségeloszlások elemzésére l. TUSNÁDY (1982) dolgozatában. (A későbbiekben ismertetésre kerülő algoritmus is az ott említett vizsgálat, ti. a születési rendellenességek adatainak elemzése során született.)

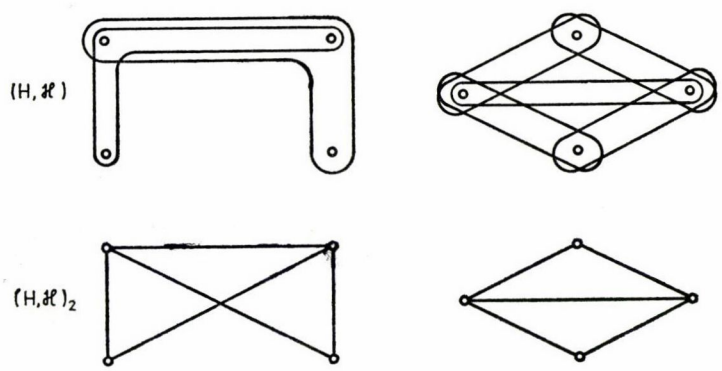
3. Maximum likelihood becslés

A ML becslések meghatározására egy régóta ismeretes iteratív algoritmust (DEMING, STEPHAN (1940)) szokásos használni. Az algoritmust loglineáris modellekre alkalmazta HABERMAN (1972), és ez az eljárás van beépítve a BMDP programcsomagba is (DIXON (1981)). A tapasztalat azt mutatja, hogy nagy kontingencia táblák esetén ennek az algoritmusnak az alkalmazása meglehetősen memória- és gépidőigényes.

A loglineáris eloszláscsaládok egy osztálya, az ún. direkt modellek esetén zárt képlet létezik a ML becslésre. A direkt modellek pontos definiálásához és a formula felírásához bizonyos gráfelméleti eszközök szükségesek. Ezeket ismertetjük az alábbiakban.

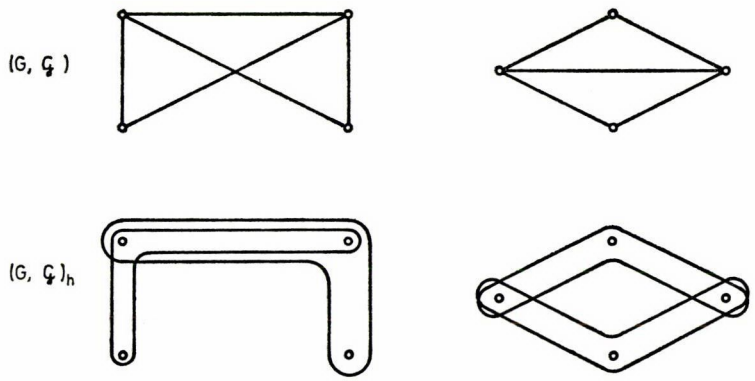
Minden $\{R_\gamma, \gamma \in \Gamma\}$ halmazrendszer meghatároz egy hipergráfot, amelynek csúcsai a $Z_1, \ldots, Z_p$ változók és az R_γ változókat pontosan akkor köti össze hiperél, ha $\gamma \in \Gamma$.

Legyen $(H, \mathcal{H})$ hipergráf, H a csúcsok, $\mathcal{H}$ a hiperélek halmaza. Ez meghatároz egy közönséges gráfot, amelyet a hipergráf 2-metszetének nevezünk, méghozzá a következőképpen. A 2-metszetben a csúcspontok ugyanazok, mint a hipergráfban, és két csúcsot pontosan akkor köt össze él, ha volt $\mathcal{H}$-ban olyan hiperél, amely tartalmazta a két csúcsot. A most definiált gráfot jelölje $(H, \mathcal{H})_2$. Az 1. ábrán két hipergráf 2-metszete látható.



1. ábra

Legyen $(G, \mathcal{G})$ gráf, G a csúcsok, $\mathcal{G}$ az élek halmaza. Ez a gráf definiál egy hipergráfot, melyben a csúcsok ugyanazok és $g \subset G$ pontosan akkor hiperél, ha a g-ben levő csúcsok egy maximális teljes részgráfot alkotnak $\mathcal{G}$-ben. Jelöljük ezt a hipergráfot $(G, \mathcal{G})_h$-val.



2. ábra

Ezt az eljárást a 2. ábrán láthatjuk, méghozzá az 1. ábra két alsó gráfjából kiindulva.

Egy $(H, \mathcal{H})$ hipergráf grafikus, ha

$$((H, \mathcal{H})_2)_h = (H, \mathcal{H}).$$

Tehát az 1. ábrán szereplő két hipergráf közül az első grafikus, a második nem.

Azt mondjuk, hogy a $(H, \mathcal{H})$ hipergráfot felbontottuk a $(H_1, \mathcal{H}_1)$ és a $(H_2, \mathcal{H}_2)$ hipergráfokra, ha $\mathcal{H}_1 \cap \mathcal{H}_2 = \emptyset$ és $\mathcal{H}_1 \cup \mathcal{H}_2 = \mathcal{H}$, H_i a $\mathcal{H}_i$-beli élek által tartalmazott csúcsokból áll, $i = 1, 2$; továbbá létezik $h_1^* \in \mathcal{H}_1$ és $h_2^* \in \mathcal{H}_2$ úgy, hogy

$$\left(\bigcup_{h_1 \in \mathcal{H}_1} h_1\right) \cap \left(\bigcup_{h_2 \in \mathcal{H}_2} h_2\right) = h_1^* \cap h_2^*.$$

Egy $(H, \mathcal{H})$ hipergráf dekomponálható, ha sorozatos felbontásokkal egyetlenélű hipergráfokra bontható.

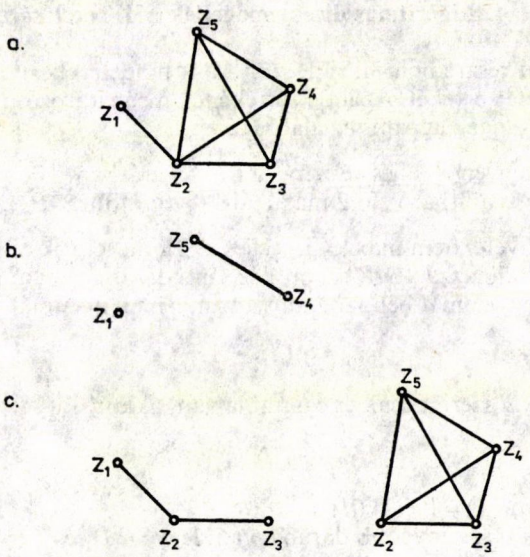
Egy loglineáris modellt direktnek nevezünk, ha a neki megfelelő hipergráf dekomponálható.

F. Tétel. (Lauritzen, Speed, Vijayan (1978)). Egy hipergráf akkor és csak akkor dekomponálható, ha grafikus, és 2-metszete trianguláris, azaz nem tartalmaz háromnál hosszabb kört húr nélkül.

Így direkt modellek vizsgálatakor elegendő hipergráfok helyett gráfokkal foglalkozni, hiszen egy grafikus hipergráf egyértelműen jellemezhető 2-metszetével. Az alábbiakban egy összefüggő gráf teljes részgráfjaihoz rendelünk egy értéket.

Legyen a vizsgált gráf $(G, \mathcal{G})$, $d \subset G$ teljes részgráf, azaz $\mathbf{P}(d) \subset \mathcal{G}$.

A $(G \setminus d)$ gráfot úgy kapjuk, hogy G-ből a d-beli csúcsokat, $\mathcal{G}$-ből az ilyeneket is tartalmazó éleket hagyjuk el. A $(G \setminus d)$ gráfnak több összefüggő komponense lehet. Minden ilyen komponenshez hozzávesszük a d-beli csúcsokat, és azokat az éleket, amelyek $(G, \mathcal{G})$-ben szerepeltek, és csak ebből a komponensből vagy d-ből tartalmaznak csúcsokat. A most definiált gráfokat nevezzük $(G, \mathcal{G})$ d-re vonatkozó darabjainak. Természetesen d minden ilyen darabban teljes részgráf. Jelölje $\mu(d)$ $(G, \mathcal{G})$ azon d-re vonatkozó darabjainak számát, amelyekben d nem maximális teljes részgráf. Végül legyen $\nu(d) = 1 - \mu(d)$.



3. ábra

A 3/a ábrán látható gráfban a Z_2 és Z_3 csúcsok teljes részgráfot alkotnak.

Ennek elhagyásával a gráf két komponensre esik (3/b ábra).

A $d=\{Z_2, Z_3\}$ teljes részgráfra vonatkozó darabok a 3/c ábrán láthatók. Közülük az elsőben d maximális teljes részgráf, a másodikban nem, ugyanis a Z_5 csúcsot hozzávehetjük d-hez, $\{Z_2, Z_3, Z_5\}$ teljes részgráf. Így $\mu(d)=1$ és $\nu(d)=0$.

G. TÉTEL. (LAURITZEN, SPEED, VIJAYAN (1978)) Legyen $(G, \mathcal{G})$ összefüggő gráf. $(G, \mathcal{G})$ akkor és csak akkor dekomponálható, ha

$$\sum_{(d, \mathbf{P}(d)) \subset (G, \mathcal{G})} \nu(d) = 1.$$

H. TÉTEL. (DARROCH, LAURITZEN, SPEED (1980), HABERMAN (1974)) Legyen a Γ loglineáris modell direkt. A neki megfelelő hipergráf 2-metszete, $(G, \mathcal{G})$ álljon a K_i, $i \in I$ összefüggő komponensekből. A 2-metszet egy d teljes részgráfjához tartozó, előbb definiált számértéket a K_i komponensben jelölje $\nu_i(d)$. Legyen N megfigyelésünk a kontigencia táblára. Tegyük fel végül, hogy $X(\omega_\gamma)>0$, $\gamma \in \Gamma$, $\omega_\gamma \in \Omega_\gamma$. Ekkor a modellnek megfelelő eloszlás ML becslése:

$$\hat{P}(\omega) = \prod_{i \in I} \prod_{(d, \mathbf{P}(d)) \subset (G, \mathcal{G})} \hat{\hat{P}}(\omega_d)^{\nu_i(d)}.$$

(Itt $\hat{\hat{P}}(\omega_d)$ ugyanazt jelenti, mint $\hat{\hat{P}}(\omega_\gamma)$, ha d éppen az R_γ-beli változókból áll, és hasonlóan értelmezhető, ha d nem maximális. ($d \not\subset K_i$ esetén $\nu_i(d)=0$.)

A fenti tétel azt mutatja, hogy egy direkt loglineáris eloszláscsalád elemei között az eloszláscsaládot definiáló marginálisoknak tetszőleges kompatibilis rendszeréhez van olyan eloszlás, amelynek ezek a kijelölt peremeloszlásai. A direkt loglineáris eloszláscsaládok tehát rendelkeznek a bevezetésben megfogalmazott tulajdonsággal.

4. Algoritmus direkt modellek ML becslésére

Az eljárást két részre bontjuk. Először adott modell teljes részgráfjaihoz meghatározzuk a $\nu(d)$ index értékét. Az index értékének meghatározása nem szükséges minden teljes részgráfra, igaz ugyanis az alábbi tétel.

4.1. TÉTEL. Legyen d teljes részgráf a $(G, \mathscr{G})$ gráfban. Ha d nem maximális teljes részgráf és nem is állítható elő 2 maximális részgráf metszeteként, akkor $\nu(d)=0$.

Bizonyítás. Mivel d nem maximális teljes részgráf, létezik teljes részgráf, amelyik valódi részként tartalmazza. $(G, \mathscr{G})$ azon d szerinti darabjában, amelyben a tartalmazó teljes részgráfnak egy nem d-beli szögpontja van, d nem maximális, ezért

$$\mu(d) \geqq 1. \tag{4.1}$$

Ha $(G, \mathscr{G})$-nek d szerint csak egy darabja van, akkor

$$\mu(d) \leqq 1, \tag{4.2}$$

(4.1) és (4.2) miatt $\mu(d)=1$ és $\nu(d)=0$.

Ha $(G, \mathscr{G})$-nek d szerint több darabja van, legyen, D_1 és D_2 két tetszőleges d szerinti darab. Belátjuk, hogy d legalább az egyikben maximális. Ekkor $\mu(d)\leqq 1$, ami (4.1)-gyel együtt bizonyítja az állítást.

Tegyük fel ugyanis, hogy d D_1-ben az $R_{\gamma_1}\subset\{Z_1, \ldots, Z_p\}$, D_2-ben pedig az $R_{\gamma_2}\subset\{Z_1, \ldots, Z_p\}$ szögpontok hozzávételével bővíthető, a teljesség megtartása mellett, de a két darabban más ilyen tulajdonságú szögpontok nincsenek (azaz R_{γ_1} és R_{γ_2} maximális). Ekkor $d\cup R_{\gamma_1}$ és $d\cup R_{\gamma_2}$ is maximális teljes részgráfok csúcsai. A két részgráf teljessége nyilvánvaló, $(G, \mathscr{G})$-ben pedig azért maximálisak, mert ha lenne egy olyan Z_u szögpont, amelyikkel például $d\cup R_{\gamma_1}$ bővíthető lenne a teljesség megtartásával, akkor ennek abban a darabban kellene lennie, amelyikből R_{γ_1}-et kaptuk, de ott R_{γ_1} maximális volt.

Akkor viszont $d=(d\cup R_{\gamma_1})\cap(d\cup R_{\gamma_2})$, ami ellentmond d-re vonatkozó feltevéseinknek. Ezért legfeljebb egy olyan darab lehet, amelyben d nem maximális.

A $\nu(d)$ meghatározására szolgáló eljárás az $\mathbf{M}$ $p\times|\Gamma|$ méretű mátrixból indul ki. $M(i,j)=1$, ha a Z_i változó szerepel a j-edik interakcióban, és $M(i,j)=0$, ha nem. Feltevéseink következtében $\mathbf{M}$-nek minden sorában van legalább egy 1, de nincs két olyan oszlop, hogy az egyikben néhány 1-et 0-val helyettesítve megkapjuk a másik oszlopot.

Az algoritmus a következő lépésekből áll:

1. Legyen K olyan természetes szám, hogy $2^{K-1}>p$.
2. Legyen $\mathbf{R}=\mathbf{M}$ és hajtsuk végre az $\mathbf{R}=\mathbf{R}\mathbf{R}^T$ utasítást K-szor.
3. $\mathbf{R}$ azon sorainak megfelelő változók, amelyekben a nullák ugyanott helyezkednek el, egy összefüggő komponensbe tartoznak.
4. Álljon a $\mathbf{T}$ mátrix $\mathbf{M}$ azon soraiból, amelyek egy (még nem vizsgált) komponensbe tartoznak.
5. Hajtsuk végre az alábbi lépéseket minden olyan (még nem vizsgált) d halmazra, amely vagy interakció a modellben, vagy két interakció metszete.
6. Ha $\mathbf{M}$-nek d által kijelölt sorait nem tartalmazza $\mathbf{T}$, akkor $\nu_T(d)=0$ és ugorjunk *14*-re. Ha tartalmazza:

7. T-ből elhagyjuk a d-nek megfelelő sorokat, így kapjuk az U mátrixot.
8. Meghatározzuk U összefüggő komponenseit (úgy, mint az *1*, *2*, *3* lépésben, M helyett U-t használva).
9. Legyen V U-nak egy (még nem vizsgált) összefüggő komponensébe tartozó soraiból álló mátrix.
10. Legyen W az a mátrix, amely M-nek d-beli soraiból és V-ből áll.
11. A W mátrixot vizsgáljuk. Ha van V-nek olyan r sora, amely rendelkezik azzal a tulajdonsággal, hogy minden d-beli s sorhoz létezik $j=j(s)$ oszlop, hogy $M(s,j)=1$ és $M(r,j)=1$, akkor $\mu_T(d)=\mu_T(d)+1$.
12. Ha még van nem vizsgált összefüggő komponens, vissza *9*-re.
13. Írjuk ki d-t és $v_T(d)=1-\mu_T(d)$ értékét.
14. Ha még van nem vizsgált d, vissza *5*-re.
15. Ha még van nem vizsgált komponens, vissza *4*-re.

A becslés alapjául szolgáló megfigyeléseket az A $N\times p$ méretű mátrixban helyezzük el. $A(i,j)$ az i-edik megfigyelés kategóriája a Z_j változó szerint; az egyes kategóriákat természetes számokkal jelöljük, 1-től C_j-ig.

Tegyük fel, hogy az $\Omega'\subset\Omega$ által tartalmazott cellák valószínűségeinek becslését szeretnénk meghatározni. A becsléseket az $e(\omega')$ vektorba írjuk be. Először legyen $e(\omega')=1$, $\omega'\in\Omega'$

A. Legyen d egy olyan, az előző eljárásból származó (még nem vizsgált) változóhalmaz, amelyre $v_T(d)\neq 0$.
B. Legyen $X(\omega_d)$ az a $\prod_{Z_i\in d} C_i$ koordinátájú vektor, amelyik az Ω_d-beli cellákat tartalmazza.
Először legyen $X(\omega_d)=0$, $\omega_d\in\Omega_d$.
C. A-nak egy (még nem vizsgált) soráról állapítsuk meg, hogy a d-beli változók szerint melyik ω_d cellába esik.
D. $X(\omega_d)$-nek ezen koordinátájú értékét növeljük meg 1-gyel.
E. Ha még van nem vizsgált sor, vissza C-re.
F. Ω'-nek egy (még nem vizsgált) ω' eleméről állapítsuk meg, hogy a d-beli változók szerint melyik ω_d cellába esik.
G. $e(\omega')=e(\omega')X(\omega_d)^{v_T(d)}$.
H. Ha még van nem vizsgált d, vissza A-ra.
I. $e(\omega')=e(\omega')/N^{|I|}$, ahol $|I|$ M összefüggő komponenseinek száma.

Az algoritmus helyességének belátásához a következő állításokra van szükségünk:

4.2. TÉTEL.
a) Azok a változók, amelyek R azon sorainak felelnek meg, amelyekben a nullák ugyanott helyezkednek el, egy összefüggő komponensbe tartoznak (*3*).
b) A W sorainak megfelelő változók egy d szerinti darabot alkotnak (*10*).
c) Ebben a darabban d nem maximális, ha teljesíti a *11*. lépés feltételét.

Bizonyítás.
a) Legyenek X és Y olyan mátrixok, amelyekben a sorok és az oszlopok egy gráf szögpontjainak felelnek meg, továbbá

$$X(i,j)>0$$

akkor és csak akkor teljesül, ha az i-edik csúcsból a j-edik csúcsba legfeljebb m lépésben el lehet jutni, és

$$Y(i, j) > 0$$

akkor és csak akkor, ha az i-edik csúcsból a j-edik csúcsba legfeljebb n lépésben el lehet jutni. Ekkor

$$XY^T(i, j) > 0 \tag{4.3}$$

akkor és csak akkor teljesül, ha az i-edik csúcsból a j-edik csúcsba legfeljebb $m+n$ lépéssel el lehet jutni. Ez azért igaz, mert (4.3) csak úgy teljesülhet, ha létezik olyan k, amelyre

$$X(i, k) > 0, \quad Y(j, k) > 0,$$

azaz amelybe az i-edik csúcsból legfeljebb m lépéssel el lehet jutni, és amelyből a j-edik csúcsba legfeljebb n lépéssel eljuthatunk.

Az algoritmusban szereplő $\mathbf{M}\,\mathbf{M}^T$ mátrix teljesíti az $\mathbf{X}$ -re és $\mathbf{Y}$-ra tett feltevéseinket, ha $m=n=1$, mert egyetlen lépéssel akkor juthatunk el egy csúcsból egy másikba, ha ugyanaz a maximális teljes részgráf tartalmazza őket. Legfeljebb azonban p lépésre lehet szükség.

Végül ha két sorban nem ugyanott helyezkednek el a nullák, akkor van olyan csúcs, amelybe az egyikből el lehet jutni, de a másikból nem, tehát a két sornak megfelelő csúcsok nem lehetnek egy összefüggő komponensben. Ha viszont a nullák ugyanott helyezkednek el a két sorban, akkor van olyan csúcs, amelybe mindkettőből el lehet jutni, és akkor ugyanabban a komponensben vannak.

b) Az $\mathbf{U}$ mátrix ugyanúgy írja le a d elhagyásával keletkező gráfot, mint az eredetit $\mathbf{M}$, azzal a különbséggel, hogy $\mathbf{U}$-ban azonos oszlopok is lehetnek. Könnyen látható, hogy ez az összefüggő komponensek meghatározására szolgáló eljárást nem befolyásolja.

Ezért a $\mathbf{W}$ mátrix sorainak megfelelő csúcsok éppen egy d szerinti darabba tartozók lesznek, a mátrix oszlopai pedig az ebben a darabban maximális teljes részgráfokat jelentik (azonos oszlopok itt is előfordulhatnak).

c) A *11.* lépés feltétele úgy fogalmazható, hogy létezik olyan nem d-beli csúcs a darabban, amelyik minden d-beli csúccsal össze van kötve. Ekkor viszont d valóban nem maximális ebben a darabban.

5. A loglineáris elemzés alkalmazásairól

Ebben a fejezetben azzal foglalkozunk, hogy milyen értelmezés adható az egyes loglineáris modelleknek.

A loglineáris eloszláscsalád (2.1) definíciójában nem marginális eloszlások, hanem tetszőleges, a marginálisokon értelmezett függvények szerepeltek. Mennyivel bővebb az így definiált eloszláscsalád, mint az, amelyikben csak marginális eloszlások szerepelnek? Az alábbi tétel azt mutatja, hogy legalábbis azon eloszláscsaládok esetében, amelyeket a változók halmazának egy diszjunkt partíciója határoz meg, semennyivel.

5.1. TÉTEL. Ha $\{R_\gamma, \gamma\in\Gamma\}$ olyan, hogy $\gamma_1, \gamma_2\in\Gamma$ esetén $R_{\gamma_1}\cap R_{\gamma_2}=\emptyset$, továbbá $\bigcup_{\gamma\in\Gamma} R_\gamma=\{Z_1, \ldots, Z_p\}$ és $(P(\omega), \omega\in\Omega)$ a megfelelő eloszláscsalád eleme, akkor

$$P(\omega) = \prod_{\gamma\in\Gamma} P(\omega_\gamma) \quad \gamma\in\Gamma, \quad \omega\in\Omega.$$

Bizonyítás. Legyen $P(\omega)$ olyan eloszlás, amely megfelel a (2.1) leglineáris modellnek úgy, hogy interakcióknak a feltételben megjelölt tulajdonságú R_γ részhalmazokat választjuk. Ekkor

$$P(\omega) = \prod_{\gamma \in \Gamma} f_\gamma(\omega_\gamma). \tag{5.1}$$

Belátjuk, hogy (5.1)-ből és az R_γ halmazok diszjunkt voltából következik, hogy

$$f_\gamma(\omega_\gamma) = P(\omega_\gamma). \tag{5.2}$$

Az (5.2) állítást az interakciók számára vonatkozó teljes indukcióval bizonyítjuk. Ha $|\Gamma|=1$, akkor $R_\gamma=\{Z_1, \ldots, Z_p\}$ és (5.2) nyilvánvalóan teljesül.

Tegyük fel, hogy minden, a $|\Gamma|$-nál kevesebb interakciót tartalmazó modell esetén teljesül (5.2). Legyen $\gamma^* \in \Gamma$ és $\omega^*_{\gamma^*} \in \Omega_{\gamma^*}$ rögzített. Jelölje $\sum^*$ azokra az ω-kra az összegzést, amelyek γ^* szerint $\omega^*_{\gamma^*}$-gal egyeznek meg.

$$P(\omega^*_{\gamma^*}) = \sum{}^* P(\omega) = \sum{}^* \prod_{\gamma \in \Gamma} f_\gamma(\omega_\gamma) = f_{\gamma^*}(\omega^*_{\gamma^*}) \sum{}^* \prod_{\substack{\gamma \in \Gamma \\ \gamma \neq \gamma^*}} f_\gamma(\omega_\gamma). \tag{5.3}$$

Legyen $R_{\gamma'} = \{Z_1, \ldots, Z_p\} \setminus R_{\gamma^*}$. Az $\{R_\gamma, \gamma \in \Gamma, \gamma \neq \gamma^*\}$ interakciók az $\Omega_{\gamma'}$ (marginális) táblázatra nézve ugyanolyan modellt alkotnak, mint a tétel feltételében szereplő modell az Ω táblázatra nézve. Az indukciós feltevés szerint:

$$f(\omega_\gamma) = P(\omega_\gamma)$$

minden $\gamma \in \Gamma$, $\gamma \neq \gamma^*$ esetén, és így

$$P(\omega_{\gamma'}) = \prod_{\substack{\gamma \in \Gamma \\ \gamma \neq \gamma^*}} f_\gamma(\omega_\gamma).$$

Ezért (5.3) így írható:

$$P(\omega^*_{\gamma^*}) = f_{\gamma^*}(\omega^*_{\gamma^*}) \sum{}^* P(\omega_{\gamma'}). \tag{5.4}$$

Mivel

$$R_{\gamma^*} \cap \Big(\bigcup_{\substack{\gamma \in \Gamma \\ \gamma \neq \gamma^*}} R_\gamma \Big) = \emptyset,$$

a $\sum^*$-gal jelölt összeadásban $\Omega_{\gamma'}$ minden cellája pontosan egyszer szerepel, azaz

$$\sum{}^* P(\omega_{\gamma'}) = 1.$$

Ez az állítás és (5.4) bizonyítja (5.2)-t.

A fenti állítás nyilvánvalóan megfordítható.

Ezután a loglineáris modelleknek egy másik osztályát szeretném bemutatni, amely szintén jól értelmezhető elemzési eredményeket ad.

A szokásos többváltozós regresszió analízis problémának a következő megfelelője fogalmazható meg diszkrét változókra: Legyen Z_1 a magyarázandó, $\{Z_2, \ldots, Z_p\}$ a magyarázó változók. Arra vagyunk kíváncsiak, hogy ismerve a populáció egy tagjának a $Z_2, \ldots, Z_p$ változók szerinti kategóriáját, mi Z_1 kategóriáinak feltételes valószínűsége.

Feltesszük, hogy ez a valószínűség a következő alakban áll elő

$$P(Z_1|Z_2, \ldots, Z_p) = g(Z_2, \ldots, Z_p) \prod_{i=2}^{p} h_i(Z_1, Z_i) \tag{5.5}$$

$$g\colon \Omega_2 \times \ldots \times \Omega_p \to \mathbf{R}$$

$$h_i\colon \Omega_1 \times \Omega_i \to \mathbf{R}, \quad i = 2, \ldots, p$$

Kissé általánosabban feltehetjük, hogy a $\{Z_2, \ldots, Z_p\}$ változók úgy vannak csoportokba osztva, hogy

$$P(Z_1|Z_2, \ldots, Z_p) = t(Z_2, \ldots, Z_p) \prod_{i=1}^{k} u_i(Z_1, R_{\delta_i}), \tag{5.6}$$

$$t\colon \Omega_2 \times \ldots \times \Omega_p \to \mathbf{R},$$

$$u_i\colon \Omega_1 \times \Big(\bigtimes_{Z_j \in R_{\delta_i}} \Omega_j\Big) \to \mathbf{R}, \quad i = 1, \ldots, k,$$

$$R_{\delta_i} \cap R_{\delta_j} = \emptyset, \quad i \neq j, \quad i, j = 1, \ldots, k.$$

Végül feltehetjük, hogy a $\{Z_2, \ldots, Z_p\}$ változók nem diszjunkt csoportokba vannak osztva

$$P(Z_1|Z_2, \ldots, Z_p) = v(Z_2, \ldots, Z_p) \prod_{i=1}^{m} w_i(Z_1, R_{\varepsilon_i}), \tag{5.7}$$

$$v\colon \Omega_2 \times \ldots \times \Omega_p \to \mathbf{R},$$

$$w_i\colon \Omega_1 \times \Big(\bigtimes_{Z_j \in R_{\varepsilon_i}} \Omega_j\Big) \to \mathbf{R} \quad i = 1, \ldots, m.$$

5.2. TÉTEL.

a) Az (5.5) modell ekvivalens azzal a loglineáris modellel, amelyben az interakciók:

$$R_{\gamma_1} = \{Z_2, \ldots, Z_p\}, \quad R_{\gamma_2} = \{Z_1, Z_2\}, \ldots, R_{\gamma_p} = \{Z_1, Z_p\}.$$

b) Az (5.6) modell ekvivalens azzal a loglineáris modellel, amelyben az interakciók:

$$R_{\gamma_1} = \{Z_2, \ldots, Z_p\}, \quad R_{\gamma_2} = \{Z_1\} \cup R_{\delta_1}, \ldots, R_{\gamma_{k+1}} = \{Z_1\} \cup R_{\delta_k}.$$

c) Az (5.7) modell ekvivalens azzal a loglineáris modellel, amelyben az interakciók:

$$R_{\gamma_1} = \{Z_2, \ldots, Z_p\}, \quad R_{\gamma_2} = \{Z_1\} \cup R_{\varepsilon_1}, \ldots, R_{\gamma_{m+1}} = \{Z_1\} \cup R_{\varepsilon_m}.$$

Bizonyítás. Az (5.5)—(5.7) modelleknek eleget tevő eloszlások leírhatók az állításban szereplő loglineáris modellekkel a feltételes valószínűség definícióját felhasználva.

Másrészt, ha egy eloszlás leírható az a) állításban szereplő loglineáris modellel, akkor legyen

$$g(Z_2, \ldots, Z_p) = \frac{f_{\gamma_1}(\omega_{\gamma_1})}{P(Z_2, \ldots, Z_p)}$$

és

$$h_i(z_1, z_i) = f_{\gamma_i}(\omega_{\gamma_i}), \quad i = 2, \ldots, p$$

(nyilvánvaló, hogy $P(Z_2, \ldots, Z_p)$ csak $\{Z_2, \ldots, Z_p\}$ függvénye).

Hasonlóan a b) esetben

$$t(Z_2, \ldots, Z_p) = \frac{f_{\gamma_1}(\omega_{\gamma_1})}{P(Z_2, \ldots, Z_p)}$$

és

$$u_i(Z_1, R_{z_i}) = f_{\gamma_{i+1}}(\omega_{\gamma_{i+1}}) \quad i = 1, \ldots, k.$$

A c) állítás ugyanúgy látható be, mint b).

A loglineáris elemzésnek gyakorlati feladatokra való alkalmazását illetően l. például: GOODMAN (1970), BISHOP, FIENBERG, HOLLAND (1975), GOKHALE, KULLBACK (1978), RUDAS (1982).

A loglineáris eloszláscsaládnak ebben a cikkben tárgyalt fogalma általánosítható arra az esetre is, amikor a peremeloszlások között egyaránt vannak diszkrét és folytonos eloszlások is (l. LAURITZEN, WERMUTH, 1984).

Végül szeretném megköszönni TUSNÁDY GÁBORnak a dolgozat megírásához nyújtott segítségét.

IRODALOM

[1] BIRCH, M. W. (1963) "Maximum likelihoood in three-way contingency tables", *J. Roy. Statist. Soc. Ser. B.* **25** 220—233.
[2] BISHOP, Y. M. M., FIENBERG, S. E., HOLLAND, P. W. (1975) *Discrete Multivariate Analysis* (MIT Press).
[3] CSISZÁR, I. (1975) "I-divergence geometry of probability distributions and minimization problems", *Ann. Probab.* **3** 146—158.
[4] CSISZÁR, I. „A minimális diszkrimináló információ elve; kontingencia táblák elemzése", in: *Móri, Székely* (szerk.) *Többváltozós statisztikai módszerek,* (Műszaki Könyvkiadó, megjelenés alatt).
[5] DARROCH, J. N., LAURITZEN, S. L., SPEED, T. P. (1980) "Markov fields and log-linear models for contingency tables", *Ann. Statist.* **8** 522—539.
[6] DEMING, W. E., STEPHAN, F. F. (1940) "On a least squares adjustment of a sampled frequency table when the expected marginal totals are known", *Ann. Math. Statist.* **11** 427—444.
[7] DIXON, W. J. (ed) (1981) *BMDP Statistical Software* (University of California Press).
[8] GOKHALE, D. V., KULLBACK, S. (1978) *The Information in Contingency Tables* (Marcel Dekker Inc.).
[9] GOOD, I. J. (1963) "Maximum entropy for hypothesis formulation especially for multidimensional contingency tables", *Ann. Math. Statist.* **34** 911—934.
[10] GOODMAN, L. A. (1970) "The multivariate analysis of qualitative data: interactions among multiple classifications", *J. Amer. Statist. Assoc.* **65** 226—256.
[11] HABERMAN, S. J. (1972) "Log-linear fit for contingency tables, Algorithm As 51", *Appl. Statist.* **21** 218—224.
[12] HABERMAN, S. J. (1974) *The Analysis of Frequency Data* (The University of Chicago Press).
[13] KULLBACK, S. (1959) *Information Theory and Statistics* (Wiley).
[14] LAURITZEN, S. L., SPEED, T. P. VIJAYAN, K. (1978) "Decomposable graphs and hypergraphs", Inst. Math. Statist. Univ. of Copenhagen, preprint No 9.
[15] LAURITZEN, S. L., WERMUTH, N. (1984) "Mixed interaction models", Aalborg Universitetscenter.
[16] RUDAS, T. (1982) *Kontingencia táblák elemzése,* ELTE BTK jegyzet.
[17] RUDAS, T. (1985) „Loglineáris elemzés", in: *Móri, Székely* (szerk.) *Többváltozós statisztikai módszerek,* (Műszaki Könyvkiadó, megjelenés alatt).
[18] TUSNÁDY, G. (1982) „Keverékek felbontása", *Matematikai Lapok,* **30** 1—3, 59—67.

(Beérkezett: 1983. szeptember 26.)
(Átdolgozva beérkezett: 1985. május 2.)

RUDAS TAMÁS
ELTE SZOCIOLÓGIAI INTÉZET
1052 BUDAPEST, PESTI BARNABÁS U. 1.

MAXIMUM LIKELIHOOD ESTIMATION OF DIRECT LOG-LINEAR MODELS

T. RUDAS

One of the most widely used methods of the statistical evaluation of discrete multivariate data is the log-linear analysis. This means fitting a member of a log-linear family of distributions to the observations. The log-linear families associated with direct log-linear models are such that the defining marginals of their members can be fixed, besides compatibility, arbitrarily. It is well known that direct log-linear models have closed form maximum likelihood estimates. In this paper a slight simplification of this formula, together with an easily programmeable algorithm, is given. As to the practical application of log-linear analysis independence and multiple regression interpretations of certain log-linear families are investigated.

Alkalmazott Matematikai Lapok **11** (1985) 363—369

LINEÁRIS PROGRAMOZÁSI FELADATOK MEGOLDÁSA VETÍTÉSES MÓDSZERREL

PAP GYULA RÓZSA GYÖRGY
Debrecen Hajdúböszörmény

Leírjuk, hogy milyen pontsorozat keletkezik, amikor lineáris egyenlőtlenségrendszer egy partikuláris megoldását ún. vetítéses módszerrel (az ellipszoid algoritmus speciális változatával) keresünk. Ezt felhasználva algoritmusokat adunk lineáris programozási feladatok megoldására.

1. Lineáris egyenlőtlenségrendszer egy partikuláris megoldásának keresése vetítéses módszerrel

Tekintsük az

$$\sum_{j=1}^{N} a_{ij}x_j \leqq b_i, \quad i = 1, 2, \ldots, M \tag{1.1}$$

lineáris egyenlőtlenségrendszert. Feltehetjük, hogy

$$\sum_{j=1}^{N} a_{ij}^2 > 0, \quad i = 1, 2, \ldots, M.$$

Definiáljunk $\mathbf{R}^N$-ben egy pontsorozatot a következő módon:

— induljunk ki egy tetszőleges $P_0=(x_1^{(0)}, \ldots, x_N^{(0)})\in\mathbf{R}^N$ pontból;

— ha $n=k\cdot M+l$, $1\leqq l\leqq M$ alakú, akkor vizsgáljuk meg, hogy a P_{n-1} koordinátái kielégítik-e az l-edik egyenlőtlenséget; ha igen, akkor legyen $P_n=P_{n-1}$, ha pedig nem, akkor P_n legyen a P_{n-1} vetülete a $\sum_{j=1}^{N} a_{lj}x_j=b_l$ hipersíkra. Vagyis a $P_n=(x_1^{(n)}, \ldots, x_N^{(n)})\in\mathbf{R}^N$ pont koordinátáit az alábbi egyszerű módon kapjuk:

$$x_j^{(n)} = \begin{cases} x_j^{(n-1)}, & \text{ha } q^{(n)} \geqq 0 \\ x_j^{(n-1)} + q^{(n)}a_{lj}, & \text{ha } q^{(n)} < 0 \end{cases} \quad j = 1, 2, \ldots, N, \tag{1.2}$$

ahol

$$q^{(n)} = \left(b_l - \sum_{j=1}^{N} a_{lj}x_j^{(n-1)}\right) \Big/ \left(\sum_{j=1}^{N} a_{lj}^2\right).$$

(Nyilván előnyös először „normálni" az egyenlőtlenségeket, azaz az i-ediket elosztani a $\sum_{j=1}^{N} a_{ij}^2>0$ mennyiséggel, hiszen ekkor a $q^{(n)}$ kiszámítása során már nem kell osztani vele.)

A vetítéses módszer a relaxációs módszernek (lásd pl. AGMON [1], MOTZKIN és SCHOENBERG [2]) egy speciális esete; ez utóbbi pedig tekinthető az ellipszoid algoritmus speciális változatának is (WALUKIEWICZ [3]).

Mint ismeretes, ha az (1.1) lineáris egyenlőtlenségrendszernek van megoldása, akkor a $\{P_n\}$ sorozat konvergens, és $\lim_{n\to\infty} P_n$ megoldása (1.1)-nek.

Most megvizsgáljuk mi történik akkor, ha (1.1)-nek nincs megoldása.

Vezessük be a következő jelöléseket. Jelölje $1\leqq l\leqq M$ esetén S_l a $\sum_{j=1}^{N} a_{lj}x_j=b_l$ hipersíkot, és $T_l\colon \mathbf{R}^N\to\mathbf{R}^N$ azt a transzformációt, amellyel $n=k\cdot M+l$, $1\leqq l\leqq M$ esetén $P_n=T_l(P_{n-1})$, vagyis ha P koordinátái kielégítik az l-edik egyenlőtlenséget, akkor $T_l(P)=P$, ha pedig nem elégítik ki, akkor $T_l(P)$ a P pontnak az S_l hipersíkra való vetülete. Jelölje $\mathcal{K}$ az (1.1) által meghatározott konvex poliédert.

1.1. TÉTEL. Ha (1.1)-nek nincsen megoldása, akkor

(i) a $\{P_n\}$ sorozat korlátos, de nem konvergens;
(ii) a $\{P_{kM+l},\ k\geqq 0\}$, $1\leqq l\leqq M$ részsorozatok konvergensek és a határértékek között legalább két különböző van;
(iii) jelölje $Q_l=\lim_{k\to\infty} P_{kM+l}$, $1\leqq l\leqq M$.

Ekkor érvényesek a

$$Q_2=T_2(Q_1),\quad Q_3=T_3(Q_2),\ \ldots,\ Q_M=T_M(Q_{M-1}),\quad Q_1=T_1(Q_M)$$

összefüggések, más szóval a $\{P_n\}$ sorozat egy ciklikus határhelyzethez tart.

Bizonyítás. Először is belátjuk, hogy ha $\{P_n\}$ konvergens, akkor van megoldása (1.1)-nek, például $\lim_{n\to\infty} P_n=Q$. Tegyük fel ugyanis, hogy $Q\notin\mathcal{K}$. Ekkor van olyan egyenlőtlenség, amelyet Q nem elégít ki. Vegyünk most a Q pont körül egy olyan kis G gömböt, mely nem metsz bele a Q által ki nem elégített egyenlőtlenségekhez tartozó hipersíkokba. Ha a $\{P_n\}$ sorozat bejut a G gömbbe, akkor a következő vetítések során nyilván előbb vagy utóbb kikerül a G gömbből, ami ellent mond annak, hogy $Q=\lim_{n\to\infty} P_n$. Tehát ha (1.1)-nek nincs megoldása, akkor a $\{P_n\}$ sorozat nem lehet konvergens.

A továbbiak bizonyításához szükség van a következő lemmára.

1.2. LEMMA. Legyen adott $\mathbf{R}^N$-ben hipersíkoknak egy véges $\mathcal{S}$ halmaza és egy O pont. Ekkor $\exists\Omega\subseteq\mathbf{R}^N$ kompakt, konvex halmaz úgy, hogy $O\in\Omega$, és bármely $S\in\mathcal{S}$ esetén Ω-ból nem vezet ki az S-re való vetítés.

A lemma bizonyítását a 2. szakasz tartalmazza. A lemma alapján a $\{P_n\}$ sorozat akkor is korlátos, ha (1.1)-nek nincs megoldása; így van torlódási pontja is.

Tekintsük most a $T_i\colon \mathbf{R}^N\to\mathbf{R}^N$, $i=1,\ldots,M$ folytonos leképezéseket és képezzük a $V_l=T_l\circ T_{l-1}\circ\ldots\circ T_1\circ T_M\circ\ldots\circ T_{l+1}$, $1\leqq l\leqq M$ szintén folytonos leképezéseket. A lemma szerint tetszőleges $P_0\in\mathbf{R}^N$ kezdőpont esetén $\exists\Omega\subseteq\mathbf{R}^N$ kompakt, konvex halmaz úgy, hogy $V_l(\Omega)\subseteq\Omega$, $1\leqq l\leqq M$. Így a *Schauder-féle fixpont-tétel* alapján minden $1\leqq l\leqq M$ esetén $\exists P_l^*\in\Omega$ fixpontja a V_l leképezésnek: $V_l(P_l^*)=P_l^*$.

Könnyen belátható, hogy a $\{P_{kM+l},\ k\geqq 0\}$ részsorozat nem távolodhat a P_l^* fixponttól:

$$d(P_l^*, P_{(k+1)M+l})\leqq d(P_l^*, P_{kM+l}),\quad k\geqq 0,\quad 1\leqq l\leqq M. \tag{1.3}$$

Ehhez elég utalni a V_l leképezések definiálására és a T_i leképezések „összehúzó" jellegére:

$$d(T_i(P'), T_i(P'')) \leqq d(P', P''), \quad 1 \leqq i \leqq M, \quad P', P'' \in \mathbf{R}^N. \tag{1.4}$$

Belátjuk most, hogy a $\{P_{kM+l}, k \geqq 0\}$ részsorozat bármely Q_l torlódási pontja szintén fixpontja a V_l leképezésnek. Indirekt módon tegyük fel ugyanis, hogy $V_l(Q_l) \neq Q_l$. Ekkor

$$d(P_l^*, V_l(Q_l)) < d(P_l^*, Q_l), \tag{1.5}$$

hiszen egyenlőség csak akkor lehetne, ha teljesülne

$$d(P_l^*, Q_l) = d(T_{l+1}(P_l^*), T_{l+1}(Q_l)) = \ldots = d(T_M \circ \ldots \circ T_{l+1}(P_l^*), T_M \circ \ldots \circ T_{l+1}(Q_l)) = \ldots$$
$$\ldots = d_l(V_l(P_l^*), V_l(Q_l)) = d(P_l^*, V_l(Q_l)),$$

viszont ez csak úgy lehetne, ha bármely $1 \leqq i \leqq M$ esetén vagy nem kell vetíteni az S_i hipersíkra a $T_{i-1} \circ \ldots \circ T_{l+1}(P_l^*)$ és $T_{i-1} \circ \ldots \circ T_{l+1}(Q_l)$ pontokat, vagy pedig az őket összekötő vektor párhuzamos az S_i hipersíkkal. Ebből viszont az következne, hogy $V_l(Q_l) = Q_l$. Tehát ha $V_l(Q_l) \neq Q_l$, akkor fennáll (1.5): $d(P_l^*, V_l(Q_l)) < d(P_l^*, Q_l)$. Viszont a V_l leképezés folytonossága miatt ha a P pont a Q_l elég kicsi környezetében van, akkor $V_l(P)$ a $V_l(Q_l)$-hez lesz közel, így ha Q_l körül elég kicsi környezetet veszünk, akkor abból egyrészt (1.5) miatt a részsorozat kilép, de (1.3) miatt oda vissza már nem juthat. Ez pedig ellentmond annak, hogy Q_l torlódási pont, így beláttuk, hogy a $\{P_{kM+l}, k \geqq 0\}$ részsorozat minden torlódási pontja fixpontja a V_l leképezésnek. Mivel pedig egy fixpont körül vett tetszőleges sugarú gömbből (1.3) miatt a részsorozat nem tud kijutni, így annak csak egy torlódási pontja lehet, vagyis konvergens.

2. Az 1.2. lemma bizonyítása

Vezessük be a következő jelöléseket. Jelölje a $P \in \mathbf{R}^N$ pontnak az $S \subseteq \mathbf{R}^N$ hipersíkra való vetületét P_s. Legyen $A, B \in \mathbf{R}^N$ esetén $\|\overrightarrow{AB}\|$ az $\overrightarrow{AB}$ vektor hossza, és ha $H \subseteq \mathbf{R}^N$ egy lineáris altér, akkor $(\overrightarrow{AB})_H$ illetve $(\overrightarrow{AB})_{H^\perp}$ az $\overrightarrow{AB}$ vektor H-val párhuzamos, illetve H-ra merőleges összetevője. Hipersíkok véges halmazát lineárisan függetlennek fogjuk nevezni, ha a normálvektoraik lineárisan független rendszert alkotnak.

Legyen adva tehát $\mathbf{R}^N$-ben hipersíkoknak egy véges $\mathscr{S}$ halmaza és egy O pont. Most $k = 1, 2, \ldots, N-1$-re minden $\{S_1, \ldots, S_k\} \subseteq \mathscr{S}$ lineárisan független hipersík-k-ashoz hozzárendelünk egy $D(S_1, \ldots, S_k)$ számot a következő rekurzív módon:

— legyen $k = 1$ esetén $D(S) = d(O, S)$, azaz $D(S)$ az S hipersíknak az O ponttól való távolsága.

— ha már definiálva vannak a $D(S_1, \ldots, S_l)$ mennyiségek $\{S_1, \ldots, S_l\} \subseteq \mathscr{S}$, $l = 1, 2, \ldots, k-1 \leqq N-2$ esetén, akkor $l = k$-ra a következő módon járunk el. Legyen $D(S_1, \ldots, S_k) = \infty$, ha valamely $\{i_1, i_l\} \subseteq \{1, \ldots, k\}, l \leqq k-1$ esetén $D(S_{i_1}, \ldots, S_{i_l}) = \infty$. Ha ezek mind végesek, akkor legyen $D(S_1, \ldots, S_k)$ az a minimális r, amelyre a

$$\{P \in \mathbf{R}^N: \|\overrightarrow{OP}\| \leqq r, \|(\overrightarrow{OP})_{S_{i_1}}\| \leqq \sqrt{r^2 - D^2(S_{i_1})}, \ldots, \|(\overrightarrow{OP})_{S_{i_1} \cap \ldots \cap S_{i_{k-1}}}\| \leqq$$
$$\leqq \sqrt{r^2 - D^2(S_{i_1, \ldots, S_{i_{k-1}}})}, \{i_1, \ldots, i_{k-1}\} \subseteq \{1, \ldots, k\}; \|(\overrightarrow{OP})_{S_1 \cap \ldots \cap S_k}\| = 0\} \tag{2.1}$$

halmazból nem lehet kijutni az $S_1, \ldots, S_k$ hipersíkokra való vetítésekkel. Ha nincs ilyen r, akkor legyen $D(S_1, \ldots, S_k) = \infty$.

Az 1.2 lemma következik az alábbi állításból.

2.1. ÁLLÍTÁS.

(i) A $D(S_1, \ldots, S_k)$, $1 \leqq k \leqq N-1$, $S_1, \ldots, S_k \in \mathcal{S}$ mennyiségek végesek.

(ii) Elég nagy R esetén az

$$\Omega^{(N)}(O, \mathcal{S}, R) = \{P \in \mathbf{R}^N : \|\overrightarrow{OP}\| \leqq R,\ \|(\overrightarrow{OP})_{S_1}\| \leqq \sqrt{R^2 - D^2(S_1)}, \ldots$$

$$\ldots,\ \|(\overrightarrow{OP})_{S_1 \cap \ldots \cap S_{N-1}}\| \leqq \sqrt{R^2 - D^2(S_1 \ldots, S_{N-1})};\ S_1, \ldots, S_{N-1} \in \mathcal{S} \text{ lineárisan függetlenek}\}$$

halmazból nem vezet ki az $\mathcal{S}$-beli hipersíkokra való vetítés.

(Az Ω halmaz konstrukciójának alapgondolata az, hogy egy elég nagy sugarú, O középpontú gömbből „lefaragjuk" azokat a részeket, ahonnan az $\mathcal{S}$-beli hipersíkokra való vetítésekkel ki lehetne jutni.)

Bizonyítás. N szerinti teljes induckióval történik.

I. $N=2$ *esetén* a $D(S) = d(O, S)$, $S \in \mathcal{S}$ mennyiségek végesek. Azt kell még belátni, hogy ha R elég nagy, akkor az

$$\Omega = \{P \in \mathbf{R}^2 : \|\overrightarrow{OP}\| \leqq R,\ \|(\overrightarrow{OP})_S\| \leqq \sqrt{R^2 - D^2(S)},\ S \in \mathcal{S}\}$$

halmazból nem vezet ki az $\mathcal{S}$-beli egyenesekre való vetítés, azaz hogyha $P \in \Omega$, $S \in \mathcal{S}$, akkor $P_S \in \Omega$. Először megmutatjuk, hogy $\|\overrightarrow{OP_S}\| \leqq R$. Vegyük ugyanis az $\overrightarrow{OP_S} = (\overrightarrow{OP_S})_S + (\overrightarrow{OP_S})_{S^\perp}$ merőleges felbontást. Mivel $(\overrightarrow{OP_S})_S = (\overrightarrow{OP})_S$ és $\|(\overrightarrow{OP_S})_{S^\perp}\| = d(O, S)$, így felhasználva azt, hogy $P \in \Omega$, kapjuk

$$\|\overrightarrow{OP_S}\|^2 = \|(\overrightarrow{OP_S})_S\|^2 + \|(\overrightarrow{OP_S})_{S^\perp}\|^2 = \|(\overrightarrow{OP})_S\|^2 + d^2(O, S) \leqq R^2.$$

Meg kell még vizsgálni azt, hogy ha $S_1 \in \mathcal{S}$, akkor teljesül-e

$$\|(\overrightarrow{OP_S})_{S_1}\| \leqq \sqrt{R^2 - D^2(S_1)}. \tag{2.2}$$

Ha most $S_1 \| S$, akkor $(\overrightarrow{OP_S})_{S_1} = (\overrightarrow{OP})_{S_1}$, így ekkor $P \in \Omega$ miatt teljesül (2.2). Ha pedig $S_1 \nparallel S$, akkor S_1 és S lineárisan függetlenek, és R növelésével elérhető, hogy teljesüljön

$$S \cap \{Q : \|(\overrightarrow{OQ})_{S_1^\perp}\| \leqq D(S_1)\} \subseteq S \cap \{Q : \|\overrightarrow{OQ}\| \leqq R\}, \tag{2.3}$$

hiszen a bal oldalon egy szakasz van. Most ha vesszük a (2.3) reláció komplementerét az S egyenesen belül, akkor azt kapjuk, hogy

$$S \cap \{Q : \|\overrightarrow{OQ}\| > R\} \subseteq S \cap \{Q : \|(\overrightarrow{OQ})_{S_1^\perp}\| > D(S_1)\},$$

amiből következik, hogy

$$S \cap \{Q : \|\overrightarrow{OQ}\| = R\} \subseteq S \cap \{Q : \|(\overrightarrow{OQ})_{S_1^\perp}\| \geqq D(S_1)\}.$$

Tehát $\|\overrightarrow{OQ_S}\| = R$ esetén $\|(\overrightarrow{OQ_S})_{S_1^\perp}\| \geqq D(S_1)$, amiből az $\overrightarrow{OQ_S} = (\overrightarrow{OQ_S})_{S_1} + (\overrightarrow{OQ_S})_{S_1^\perp}$

merőleges felbontás alapján kapjuk, hogy

$$\|(\overrightarrow{OQ_S})_{S_1}\|^2 = \|OQ_S\|^2 - \|(\overrightarrow{OQ_S})_{S_1^{\perp}}\|^2 \leqq R^2 - D^2(S_1).$$

Nyilván ebből következik, hogy $\|\overrightarrow{OP_S}\| \leqq R$ miatt teljesül (2.2), ha R elég nagy.

II. *Tegyük fel most,* hogy a 2.1 állítás igaz $k=2, 3, \ldots, N$-re. Bebizonyítjuk, hogy ekkor $k=N+1$-re is igaz. Először lássuk be a $D(S_1, \ldots, S_k)$, $1 \leqq k \leqq N$, $S_1, \ldots, S_k \in \mathcal{S}$ mennyiségek végességét. Ehhez azt kell belátni, hogy van olyan r, amelyre a (2.1) halmazból nem vezet ki az $S_1, S_2, \ldots, S_k$ hipersíkokra való vetítés. Viszont ezt a problémát tekinthetjük a $\{P \in \mathbf{R}^{N+1} : \|(\overrightarrow{OP})_{S_1 \cap \ldots \cap S_k}\| = 0\}$ k-dimenziós altérben is. Így a fenti tulajdonságú r létezése következik az indukciós feltevésből.

Azt kell még belátni, hogy elég nagy R esetén az $\mathcal{S}$-beli hipersíkokra való vetítés nem vezet ki az

$$\Omega^{(N+1)}(O, \mathcal{S}, R) = \{P \in \mathbf{R}^{N+1} : \|\overrightarrow{OP}\| \leqq R, \|(\overrightarrow{OP})_{S_1}\| \leqq \sqrt{R^2 - D^2(S_1)}, \ldots$$

$\ldots, \|(\overrightarrow{OP})_{S_1 \cap \ldots \cap S_N}\| \leqq \sqrt{R^2 - D^2(S_1, \ldots, S_N)}$, $S_1, \ldots, S_N \in \mathcal{S}$ lineárisan függetlenek$\}$

halmazból, azaz ha $P \in \Omega$, $S \in \mathcal{S}$, akkor $P_S \in \Omega$.

Először is $\|\overrightarrow{OP_S}\| \leqq R$ ugyanúgy következik, mint $N=2$ esetén. Most belátjuk, hogy ha $1 \leqq l \leqq N-1$, és $S_1, \ldots, S_l \in \mathcal{S}$ lineárisan függetlenek, akkor teljesül

$$\|(\overrightarrow{OP_S})_{S_1 \cap \ldots \cap S_l}\| \leqq \sqrt{R^2 - D^2(S_1, \ldots, S_l)}. \tag{2.4}$$

Ha $S \not\parallel S_1 \cap \ldots \cap S_l$, akkor jelölje P' a P pontnak a $\{Q : \|(\overrightarrow{OQ})_{S \cap S_1 \cap \ldots \cap S_l}\| = 0\}$ altérre való vetületét. Mivel ekkor $\|(\overrightarrow{OP'})_{S \cap S_1 \cap \ldots \cap S_l}\| = 0$, így $P \in \Omega$ figyelembevételével P' eleme a (2.1) halmaznak $k=l+1$, $S_{l+1}=S$, $N \to N+1$, $r=D(S, S_1, \ldots, S_l)$ szereposztással. Mivel ebből nem vezet ki az $S, S_1, \ldots, S_l$ hipersíkokra való vetítés, így P'_S is eleme lesz, amiből adódik, hogy

$$\|(OP'_S)_{S_1 \cap \ldots \cap S_l}\| \leqq \sqrt{D^2(S, S_1, \ldots, S_l) - D^2(S_1, \ldots, S_l)}. \tag{2.5}$$

Másrészt P'_S előáll úgy is, mint a P_S pont vetülete a $\{Q : \|(\overrightarrow{OQ})_{S \cap S_1 \cap \ldots \cap S_l}\| = 0\}$ altérre, így érvényes az

$$\overrightarrow{OP_S} = (\overrightarrow{OP_S})_{S \cap S_1 \cap \ldots \cap S_l} + \overrightarrow{OP'_S}$$

merőleges felbontás, amiből kapjuk az

$$(\overrightarrow{OP_S})_{S_1 \cap \ldots \cap S_l} = (\overrightarrow{OP})_{S \cap S_1 \cap \ldots \cap S_l} + (\overrightarrow{OP'_S})_{S_1 \cap \ldots \cap S_l}$$

merőleges felbontást, így végül

$$\|(\overrightarrow{OP_S})_{S_1 \cap \ldots \cap S_l}\|^2 = \|(\overrightarrow{OP})_{S \cap S_1 \cap \ldots \cap S_l}\|^2 + \|(\overrightarrow{OP'_S})_{S_1 \cap \ldots \cap S_l}\|^2 \leqq$$

$$\leqq (R^2 - D^2(S, S_1, \ldots, S_l)) - (D^2(S, S_1, \ldots, S_l) - D^2(S_1, \ldots, S_l)) = R^2 - D^2(S_1, \ldots, S_l),$$

vagyis ekkor fennáll (2.4).

Ha pedig $S \| S_1 \cap \ldots \cap S_l$, $1 \leqq l \leqq N$, akkor egyszerűen $(\overrightarrow{OP_S})_{S_1 \cap \ldots \cap S_l} = (\overrightarrow{OP})_{S_1 \cap \ldots \cap S_l}$, így $P \in \Omega$-ból közvetlenül adódik (2.4).

Már csak azt kell elérni R növelésével, hogy $P\in\Omega$ és $S, S_1, \ldots, S_N\in\mathscr{S}$ lineárisan független hipersíkok esetén teljesüljön a (2.4). Viszont a

$$(2.6)\qquad S\cap\{Q\colon \|(\overrightarrow{OQ})_{(S_1\cap\ldots\cap S_N)^\perp}\|\leqq D(S_1,\ldots,S_N)\}\subseteq S\cap\{Q\colon \|\overrightarrow{OQ}\|\leqq R\}$$

relációból — ugyanúgy, ahogy a kétdimenziós esetben is — következni fog (2.4). A (2.6) pedig elérhető R növelésével, ugyanis a baloldali halmazról kimutatható, hogy egy ellipszoid. Ehhez vegyünk fel egy koordinátarendszert $\mathbf{R}^{N+1}$-ben úgy, hogy az S hipersík egyenlete $x_{N+1}=0$ legyen. Az $S_1\cap\ldots\cap S_N$ egyenes pontjai $t\cdot\mathbf{v}$ alakúak, ahol $\|\mathbf{v}\|=1$, és tudjuk, hogy $S\not\parallel S_1\cap\ldots\cap S_N$ miatt $v_{N+1}\neq 0$. Mivel $\|(\overrightarrow{OQ})_{(S_1\cap\ldots\cap S_N)^\perp}\|$ éppen a $Q=(x_1,\ldots,x_{N+1})$ pontnak az $S_1\cap\ldots\cap S_N$ egyenestől való távolsága, így ez

$$\|(\overrightarrow{OQ})_{(S_1\cap\ldots\cap S_N)^\perp}\|=\sqrt{\|\overrightarrow{OQ}\|^2-\langle\overrightarrow{OQ},\mathbf{v}\rangle^2}=\sqrt{\sum_{k=1}^{N+1}x_k^2-\Big(\sum_{k=1}^{N+1}v_kx_k\Big)^2}.$$

Tehát a $\{Q\in\mathbf{R}^{N+1}\colon \|(\overrightarrow{OQ})_{(S_1\cap\ldots\cap S_N)^\perp}\|\leqq D(S_1,\ldots,S_N)\}$ „hiperhengernek" az S hipersíkkal való metszete

$$\sum_{k=1}^{N}x_k^2-\Big(\sum_{k=1}^{N}v_kx_k\Big)^2\leqq D^2(S_1,\ldots,S_N).$$

Ez pedig valóban ellipszoid, hiszen a baloldalon egy pozitív definit kvadratikus forma áll:

$$\sum_{k=1}^{N}x_k^2-\Big(\sum_{k=1}^{N}v_kx_k\Big)^2\geqq\sum_{k=1}^{N}x_k^2-\Big(\sum_{k=1}^{N}v_k^2\Big)\Big(\sum_{k=1}^{N}x_k^2\Big)=v_{N+1}^2\Big(\sum_{k=1}^{N}x_k^2\Big)>0,$$

hogyha $\sum_{k=1}^{N}x_k^2>0$, hiszen $v_{N+1}\neq 0$.

3. Alkalmazás lineáris programozási feladatok megoldására

Az 1.1. tétel alapján könnyen adódik algoritmus, mely eldönti az (1.1) lineáris egyenlőtlenségrendszerről, hogy van-e megoldása, és ha van, akkor előállít egyet.

Induljunk ki ugyanis egy tetszőleges $P_0\in\mathbf{R}^N$ pontból, és kezdjük el képezni a $\{P_n\}$ sorozatot. Jelöljünk ki ε_1 és ε_2 korlátokat, és ha $d(P_{kM}, P_{(k-1)M})<\varepsilon_1$, akkor vizsgáljuk meg, hogy a P_{kM} pont koordinátái kielégítik-e ε_2 hibát megengedve az (1.1) egyenlőtlenségrendszert. Ha igen, akkor van egy megoldásunk, ha nem, akkor úgy tekintjük, hogy (1.1)-nek nincs megoldása.

Ha most meg akarjuk oldani a

$$(3.1)\qquad \max\{\mathbf{cx}\colon \mathbf{Ax}\leqq\mathbf{b}\}$$

lineáris programozási feladatot, akkor alkalmazhatjuk akár a dualitási tételen alapuló módszert, akár a „célfüggvényfelezéses" módszert, akár a „jó irányok" módszerét (lásd pl. WALUKIEWICZ [3]) — mindegyik visszajátsza a problémát lineáris egyenlőtlenségrendszer (vagy -rendszerek) partikuláris megoldásának keresésére. Megjegyezhetjük, hogy a kereső algoritmus könnyen programozható, kevés adatot kell tárolni,

és ha az algoritmus nem talál megoldást, akkor a ciklikus határhelyzetben, amihez tart a pontsorozat, egy olyan részrendszeren ugrál, ami már maga is ellentmondásos; így arra is lehet információt nyerni, hogy mely egyenlőtlenségeket megváltoztatva lehet megoldhatóvá tenni a rendszert.

IRODALOM

[1] AGMON, S., "The relaxation method for linear inequalities", *Canadian J. of Math.* **6** (1954) 382—392.
[2] MOTZKIN, T. and SCHOENBERG, I. J., "The relaxation method for linear inequalities", *Canadian J. of Math.* **6** (1954) 393—404.
[3] WALUKIEWICZ, S., "The ellipsoid algorithm for linear programming", *MTA SZTAKI Tanulmányok* **152**/1983, 139—167.

(Beérkezett: 1984. október 19.)

PAP GYULA
KLTE MATEMATIKAI INTÉZETE
4010 DEBRECEN, EGYETEM TÉR 1.

RÓZSA GYÖRGY
TUNGSRAM RT.
4221 HAJDÚBÖSZÖRMÉNY

SOLVING OF LINEAR PROGRAMMING PROBLEMS WITH PROJECTION METHOD

GY. PAP AND GY. RÓZSA

Description of sequence of points generated by projection method for searching a particular solution of system of linear inequalities is given. Using this we get algorithms to solve linear programming problems.

Alkalmazott Matematikai Lapok **11** (1985) 371—383

A MATEMATIKA NÉHÁNY ALKALMAZÁSA A GEODÉZIÁBAN

GERGELY JÓZSEF és PERGEL JÓZSEFNÉ

Budapest

A dolgozatban először a matematika és a geodézia fejlődésének néhány közös vonására emlékeztetünk, majd a geodéziának a matematikai alkalmazások szempontjából érdekes feladatai közül sorolunk fel néhányat. Ismertetjük ezek megoldásához használatos fontosabb matematikai modelleket és ezek numerikus megoldási módszereit. Végül leírunk egy általunk megoldott szintezési feladatot.

1. Történeti áttekintés

A részletes történeti fejtegetések (lásd [23], [21]) helyett, itt csak arra szeretnénk rámutatni, hogy a két tudománynak milyen sok közös vonása van.

Már a történelem előtti időkben (a 3. évezredben) a babiloniak, akik jó matematikusok voltak, magas fokú csillagászati ismeretekkel is rendelkeztek. Számítani tudták a csillagok mozgását (felkelését, delelését, lenyugvását), ismerték a nap- és holdfogyatkozások periodikus visszatérését, ami pontos számolási készség nélkül nehezen elképzelhető. Az ókor természettudósai matematikusok és egyben geodéták is voltak. PYTHAGORASZ és ARISZTOTELESZ eredményei elsősorban a földmérés szükségleteiből fakadtak. Időszámításunk előtt 500 körül PYTHAGORASZ már tudta, hogy a Föld gömbölyű. ARISZTOTELESZ a Föld gömb alakjának bizonyításával is foglalkozott. Az alexandriai HERON (i.e. 3. évszázad) *„Metrika"* című könyvében különböző idomok területének és a térfogatának mérését és számítását írta le.

A hajózás a csillagoknak az égbolton való elhelyezkedésének meghatározását igényelte. Mindkettő a térgeometria magas szintű fejlesztését kívánta meg. A csillagászati méréseket és az ezekhez kapcsolódó számításokat nagy pontossággal kellett elvégezni. Már időszámításunk előtt primitív mérőeszközökkel végzett csillagászati mérések segítségével tűrhető közelítéssel kiszámították a Föld sugarát.

HENRY GELLIBRAND angol matematikus állapította meg elsőként (1635-ben), hogy a mágneses elhajlás értéke térben is és időben is változik.

NEWTON *„Philosophiae Naturalis Principia Mathematica"* című 1687-ben megjelent könyve nagymértékben hozzájárult a tudományos geodézia kialakulásához. Ebben a három kötetes műben fejti ki az általános gravitáció elméletét, a naprendszerben a bolygók mozgásának törvényeit, a *Newton-féle axiómákat* és a mozgásuk részletes matematikai tárgyalását. Ha ezek mellett emlékeztetünk arra is, hogy NEWTON fedezte fel a differenciál- és integrálszámítást is és számos ma is a nevével jelzett matematikai eljárást, képletet, összefüggést, akkor felmerülhet a kérdés, hogy vajon a matematikai ismeretek vezették a csillagászati (geodéziai) felismerésekhez, vagy fordítva ezek a felismerések késztették őt arra, hogy kidolgozza és tisztázza a jelenségek megmagyarázásához szükséges matematikai apparátust.

A matematika és a geodézia másik legnagyobb közös úttörője GAUSS volt. A két tudományág GAUSS munkásságában összefonódott. Például a modern számelmélet kezdetét jelentő 1801-ben megjelent *„Disquisitiones arithmetricae"* című művének megjelenésével egyidőben fedezte fel a *Ceres kisbolygót* is. GAUSS geodéziai munkássága követelte meg a legkisebb négyzetek módszerének kidolgozását. Az ugyancsak tőle származó *Gauss—Krüger koordináta rendszer* a geodéziában most is használatos.

2. A geodézia és a fotogrametria néhány feladata

Napjainkban a geodéziai kutatásokban és a geodézia gyakorlati feladataiban széles körben alkalmazzák a matematikát. Az alkalmazások egy jelentős részét azok a feladatok jelentik, ahol mérési eredményekre támaszkodva kiegyenlítő számítást hajtanak végre, azaz a meghatározandó mennyiségre közvetett vagy közvetlen ismételt mérésekből torzítatlan becslést próbálnak kapni.

Ebben a paragrafusban három, azonos matematikai feladatra visszavezethető geodéziai illetve fotogrametriai problémát ismertetünk.

a) *Geodéziai hálózatok számítása.* Az országban nagyon sok hitelesített hálózati pont van. Ezek a geodéziában rögzített, nyilvántartott és megjelölt pontok. A pontokhoz tartozó legfontosabb adatok: a pontszám, a pont két (vízszintes) vagy három (vízszintes és magassági) geodéziai koordinátája. A nyilvántartás első, harmad, negyed és ötödrendű alappontokat különböztet meg és első, harmad, és negyedrendű hálózatról beszélhetünk.

A hálózatszámítás célja a meglevő hálózat karbantartása, pontosítása és bővítése. Minden hálózatszámítás geodéziai méréseken alapszik. Méréseket végeznek az alappontok és az új pontok közt. Mérik a pontok távolságát és a két pont által meghatározott irányt. A mérések alapján az új pontoknak a koordinátáit kell kiszámítani. Az alábbiakban a vízszintes értelmű hálózat számításának modelljét adjuk meg. Legyenek a P_i pont kezdeti koordinátái (x_i^0, y_i^0). A számítás céljától függően fixálhatunk pontokat (ezek koordinátái ne változzanak) a többinek pedig „változásokat" adunk.

$$x_i = x_i^0 + dx_i, \quad y_j = y_i^0 + dy_i.$$

A P_i pontnak a P_j pontról vett távolsága

$$S_{ij} = ((x_j - x_i)^2 + (y_j - y_i)^2)^{1/2},$$

az irányszög pedig

$$T_{ij} = \operatorname{arc\,tg} \frac{y_j - y_i}{x_j - x_i}.$$

Legyen a két pont mért távolsága s_{ij}, iránya pedig t_{ij}. Írjuk fel a

$$G(\mathbf{u}) = \sum_{i,j} (S_{ij} - s_{ij})^2 + \sum_{i,j} (T_{ij} - t_{ij} - \sigma_i)^2$$

négyzetösszeget, ahol σ_i a tájékozási szög és az összegezés az összes mérésre vonatkozik. A $G(\mathbf{u})$ funkcionált a dx_i, dy_i és $d\sigma_i$ változók (amiket az $\mathbf{u}$ vektorban rendeztünk)

függvényének tekintve minimalizáljuk azt. A minimum szükséges feltételéből adódó

$$\frac{\partial G(\mathbf{u})}{\partial \mathbf{u}} = \mathbf{O} \tag{2.1}$$

nemlineáris egyenletrendszer megoldása adja a dx_i, dy_i és $d\sigma_i$ megváltozásokat.

Valójában a (2.1) egyenletrendszer linearizált közelítését oldjuk meg (az egyenletrendszerek levezetését lásd a [13] dolgozatban), ami a kiegyenlített hálózat jó közelítését szolgálja, ha elég jók az x_i^0, y_i^0 kiindulási értékek. A feladat megoldása ekkor a

$$\mathbf{Qu} = \mathbf{d} \tag{2.2}$$

normál egyenletrendszer megoldását igényli, ahol $\mathbf{Q}$ szimmetrikus, pozitív definit, többnyire ritka mátrix. Nagy hálózatok esetén a (2.2) egyenletrendszer nagyméretű lesz, így annak megoldása általában a nagyszámítógépekre kidolgozott ritkamátrixos megoldási módszert igényli. A most ismertetettektől eltérő feladatot jelent a magassági hálózat kiegyenlítése. Ezt egy konkrét feladat megoldása kapcsán dolgozatunk 4. fejezetében ismertetjük.

b) *Domborzatmodellezés.* Egy vizsgált terület geodéziai, fotogrammetriai, vagy kartográfiai úton meghatározott magassági értékei diszkrét pontokkan, a magassági részletpontokban adnak információt a terepről. Ez az alapmodell a felhasználók igényeit gyakran nem elégíti ki. Az alapmodellből új modell származtatására (levezetett modellre) van szükség. Ezek meghatározása álltalában az új pontok magasságainak meghatározására szolgáló interpolációs vagy approximációs eljárás. Az alapmodell magasságpontjai geodéziai koordinátákkal rögzített mérési pontok, amik az alapmodell szempontjából véletlen elhelyezkedésű pontoknak számítanak (lásd [11]).

A domborzatmodellezés célja lehet a magasságok olyan diszkrét pontokban való meghatározása, amiket valamilyen célnak megfelelően választunk ki. Például a terepen felvett négyzetháló csúcspontjaiban kívánjuk a magasságokat meghatározni, vagy egy vonal pontjaiban (például út, vasút építéséhez).

A domborzatmodellezés szolgálhat a terep domborzatának analitikus leírására is. Ez az előbbinél általánosabban használható fel, hiszen ebből tetszőleges pontban kiszámítható a magasság. Igény merülhet fel a terepi szintvonalak meghatározására vagy a terep lejtés inflexiós pontjainak meghatározására. Mindehhez a terep analitikus modellezése célszerű.

A digitális domborzatmodellezés egyik legtermészetesebb módszere a dinamikus felületek módszere. Ez azt jelenti, hogy az adott magassági értékekre kétváltozós n-edfokú polinomot fektettünk:

$$z = \sum_{i=0}^{n} \sum_{j=0}^{n-k} a_{ij} x^i y^j. \tag{2.3}$$

Adott pontok x, y, z koordinátáit (2.3)-ba helyettesítve az a_{ij} együtthatókat kiszámíthatjuk. A számítási eljárások különböznek aszerint, hogy milyen ismert pontokat vonunk be az együtthatók kiszámításába. Legelfogadhatóbb eljárás az, hogy az érvényességi tartomány (ahol a polinomot használni akarjuk) középpontjához legközelebbi a fokszámtól függő számú pontot választjuk. A bevont pontokat a középponttól vett távolságuk szerint súlyozzuk. A súlyok a távolságokkal vagy azok négyzeteivel fordítottan arányosak.

A (2.3) együtthatóinak számítására alkalmazható a legkisebb négyzetek módszere is. Ha (2.3)-ba az előbbinél több alappont koordinátái helyettesítjük be és legkisebb négyzetek módszere alkalmazásával kapott lineáris egyenletrendszer megoldásából kaphatjuk meg az a_{ij} együtthatókat.

A gyakorlatban több más közelítési mód mellett leginkább a

$$z = a+bx+cy+dxy \tag{2.4}$$

bilineáris és a

$$z = a+bx+cy+dxy+ex^2+fy^2$$

másodfokú polinommal való közelítés használatos.

Domborzatmodellezésre használhatjuk a véges elem módszert is, lásd [27]. Kiindulásul a terepet

$$R_{ij}(x, y) = (x_i \leqq x < x_{i+1},\ y_j \leqq y < y_{j+1})$$

téglalapokkal fedjük le aminek a csúcsaiba a h_{ij} magasságok ismertek (vagy valamilyen másodlagos modellel meg lehet határozni.) Az R_{ij} téglalapok lesznek a véges elemek.

Legyen

$$f_{ij}(x, y) = \begin{cases} a_1+a_2x+a_3y+a_4xy, & \text{ha} \quad (x, y)\in R_{ij} \\ 0, & \text{különben.} \end{cases} \tag{2.5}$$

A (2.5) bilineáris függvények a négyzetek oldallapjai mentén folytonosan csatlakoznak. Ezekkel mint próbafüggvényekkel felépíthető a végeselem módszer. Az abból nyert egyenletrendszer megoldása pedig adja a (2.5) együtthatóit.

Ha a (2.5) próbafüggvények helyett pontosabb, például kétváltozós harmadfokú spline-okkal dolgozunk, jobb közelítés érhető el:

$$f_{ij}(x, y) = \begin{cases} \sum_{i=0}^{3}\sum_{j=0}^{3} a_{ij}(x-x_j)^i(y-y_j)^j & \text{ha} \quad (x, y)\in R \\ 0, & \text{különben.} \end{cases}$$

Mint említettük a téglalapú végeselemek használata esetén ismerni kell a rácspontokban a h_{ij} magassági értékeket. Ezeket általában a mérési pontokból kell levezetni. Ez a közbülső számítás hibalehetőséget hoz be a számításba. Ezért célszerű a végeselemekként a mérési pontok által meghatározott háromszögeket választani és a végeselem módszert így felépíteni.

A felsorolt megoldási módszerek mellett még számos közelítési mód terjedt el.

c) *Fotogrammetria.* A tárgyak mértani adatainak meghatározásával foglalkozik a róluk készített fényképek segítségével ez a tudomány. Alkalmazási területe igen sokrétű: elsősorban a térképészet, de a régészet, építészet, képzőművészet, formatervezés, orvostudomány (röntgenfelvételek kiértékelése) is alkalmazza módszereit. Felhasználják árvízjárta és belvizes területek térképezésére, erdők faállományának, fafajták területi elosztásának meghatározására. Egyik matematikai vonatkozásai szempontjából legigényesebb felhasználása az alappontsűrítés. Itt néhány, a terepen mért pont adataiból, a fényképezőgép adataiból[1], két vagy több kép segítségével tetszőleges fényképi pont terepi koordinátáit kell meghatározni.

Ez a feladat egy sor műszaki és matematikai problémát vet fel. A matematikai feladat: adott műszerek és fényképanyag segítségével a lehető legpontosabban határozzuk meg az új pontok koordinátáit, minél kevesebb földi mérést igényelve. A használt matematikai modelltől függően (sorkiegyenlítés, blokk-kiegyenlítés stb.) nagyméretű lineáris vagy nemlineáris egyenletrendszert kell megoldani. A fényképezett területről fényképek sora, ill. több sorból álló fényképek csoportja ún. blokkja áll rendelkezésünkre. Ha csak a sort felépítő képeket kapcsoljuk össze, akkor sorkiegyenlítésről, ha az egész blokkhoz tartozó képeket használjuk, blokkiegyenlítésről beszélünk.

Napjainkban a blokk-kiegyenlítést használják inkább, mivel ez kevesebb terepen mért pontot ún. illesztőpontot igényel. A mérési adatok két típusát különböztetjük meg aszerint, hogy milyen fotogrammetriai műszerről kapjuk ezeket. A műszerek egy csoportjával csak a fényképen levő pont x, y koordinátáit tudják mérni. Azonos területről készült fényképpár segítségével ezekből a mérésekből és a fényképező kamera adataiból ki tudjuk számolni egy a terephez hasonló felület x, y, z koordinátáit. Ezt a felületet modellnek nevezzük. Részletesebben l. [16].

A műszerek másik csoportjánál főleg kézi beállítással, úgy igazítják a képpárt, hogy az eredetihez hasonló (pl. repülőgép-terep) helyzet álljon elő és így ezekről a műszerekről már az x, y, z modellkoordinátákat lehet leolvasni.

A modellekből a terepi koordinátákat úgy tudjuk kiszámolni, hogy a modelleket elforgatjuk, eltoljuk ill. méretarányukat változtatjuk. Az ismeretlenek meghatározásához az illesztőpontokat használjuk fel (l. [16]).

Nézzük részletesebben az egyik legelterjetebb kiegyenlítési módszert, a független modellekkel történő tömbkiegyenlítést (l. [3]).

Az egyes modellek transzformációs paramétereit úgy határozzuk meg, hogy a csatlakozópontokon és az illesztőpontokon levő maradékhibát minimalizáljuk.

A javítási egyenletek az i-edik illesztőpontra a j-edik modellen:

$$\begin{pmatrix} v_x \\ v_y \\ v_z \end{pmatrix}_{ij} = m_j A_j \begin{pmatrix} x \\ y \\ z \end{pmatrix}_{ij} + \begin{pmatrix} X_0 \\ Y_0 \\ Z_0 \end{pmatrix}_j - \begin{pmatrix} X \\ Y \\ Z \end{pmatrix}_i,$$

és a terepi mérésekkel nem rendelkező csatlakozópontokra

$$\begin{pmatrix} v_{x_{cs}} \\ v_{y_{cs}} \\ v_{z_{cs}} \end{pmatrix}_{ij} = m_j A_j \begin{pmatrix} x \\ y \\ z \end{pmatrix}_{ij} + \begin{pmatrix} X_0 \\ Y_0 \\ Z_0 \end{pmatrix}_j - \begin{pmatrix} X_{cs} \\ Y_{cs} \\ Z_{cs} \end{pmatrix}_i,$$

ahol m_j a méretarányténvező, A_j a forgatási mátrix, X_0, Y_0, Z_0 az eltolási értékek, X, Y, Z a terepi koordináták, x, y, z a modellkoordináták. A forgatási mátrix nemlineáris kapcsolatot létesít az ismeretlenek között.

Általában a nemlineáris egyenletrendszert linearizálva jutunk megoldáshoz, de nemlineáris egyenletrendszerrel is dolgoznak. Gyakran használják a konjugált gradiens módszert (l. [9]).

Másik módszer a sugárnyaláb eljárással végzett tömbkiegyenlítés. Alapgondolata H. SCHMIDTtől és HIRVONENtől származik. Itt magukból az x, y képkoordinátákból indulnak ki. A v_x, v_y javításokra a fényképi és a terepi pontok kollinearitási feltételei

alapján a következő összefüggésnek kell fennálni:

$$\begin{pmatrix} v_x \\ v_y \end{pmatrix} = -\frac{c}{c_1(X-X_0)+c_2(Y-Y_0)+c_3(Z-Z_0)} \begin{pmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{pmatrix} \begin{pmatrix} X-X_0 \\ Y-Y_0 \\ Z-Z_0 \end{pmatrix} - \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} \xi \\ \eta \end{pmatrix}.$$

Ismeretlenek képenként: a_i, b_i, c_i, $i=1,2,3$ forgatási elemek és az X_0, Y_0, Z_0 a vetítési középpont koordinátái, pontonként: X, Y, Z terepi koordináták (kivéve az illesztőpontoknál), tömbönként: ξ, η képfőpont koordináták és a c kamara állandó (l. [4], [25]).

Bonyolítják a feladatot a mérési hibák ill. a modellhibák. A mérési hibákat a következőképpen csoportosították (l. [14]):

1. Durva hibák: Olyan nagy értékű hibák, amelyek túllépik a mérés pontosságától megkövetelt határt.
2. Szabályos v. szisztematikus hibák, amik a méréseket azonos értelemben, bizonyos szabályossággal befolyásolják.
3. Véletlen hiba: a hiba 0 várható értékű része.

Az utóbbi évtizedben a szabályos v. szisztematikus hibákat vizsgálták nagyon részletesen. Nemzetközi együttműködés keretében négy tesztmező kiértékelése folyt különböző módszerekkel. A vizsgálat eredménye néhány ajánlott módszer, amely a tesztmezőn kielégítő eredményt adott. A kutatás tovább folyik ebben az irányban is a főkomponensanalízis, a faktoranalízis stb. módszereivel.

A *Nemzetközi Fotogrammetriai és Távérzékelési Társaság* (ISPRS) most a durva hibák kiszűrésére hirdetett nemzetközi vizsgálatot. A szisztematikus hibák kiküszöbölésénél az újabb paraméterek bevétele a matematikai modellbe, a durva hibák kiküszöbölésénél pedig a mérési hibákkal szemben stabil matematikai módszerek alkalmazása, a durva hibás pontok kidobása a legeredményesebben alkalmazott módszerek ([2]).

3. Matematikai modellek és numerikus módszerek

A geodézia és a fotogrammetria egyik alapfeladata a következő: n mérés alapján határozzuk meg az $\mathbf{x}$ ismeretlen értékét az

$$f(\mathbf{a}, \mathbf{x}) = 0$$

összefüggés alapján. Az f általában nemlineáris kapcsolatot jelent, $\mathbf{a}$ vektor, $\mathbf{x}$ pedig az ismeretlenek vektora. Az f függvény nemcsak a mérések és az ismeretlenek közötti konkrét kapcsolatot írhatja le, hanem bizonyos feltételeket is tartalmazhat mind az ismeretlenekre, mind a mérésekre vonatkozólag. Részletesen lásd [10]. A gyakorlatban általában linearizálják f-et, így

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

lineáris egyenletrendszert kapjuk ($\mathbf{A}$ $m \times n$-es mátrix).

Az ismeretlenek n száma és a mérések m száma között a következő relációk lehetnek:

I. $n > m$, azaz a megoldás határozatlan. Ugyancsak határozatlan lesz az egyenletrendszer, ha ugyan több mérés van, mint ismeretlen, de az ismeretlenek meghatározásához ez nem elég. Például ilyen feladatokhoz jutunk az ún. szabad háló-

zatok kiegyenlítésénél. Itt a megoldásnál többnyire a *Moore-Penrose inverzet* használják.

II. $n \leqq m$: Általában ezzel az esettel találkozunk a gyakorlati feladatoknál. A klasszikus esetre (mikor felteszik, hogy a mérések pontosak és függetlenek) már majdnem 200 éve ismert megoldási módszer, a legkisebb négyzetek módszere. Ha tekintetbe vesszük, hogy a mérések pontatlanok, látható, hogy ez a feladat tipikus nemkorrekt kitűzésű feladat, azaz az $\mathbf{x}$ megoldás erősen változhat a mérések kis változásakor is. Ilyen feladatok megoldása két részből áll:

a) A mérési adatok első feldolgozása, a mérési hibák kiszűrése: Ez a fotogrammetriában a szisztematikus képhibák figyelembevételével átlagtól való eltérés figyelésével, normalitásvizsgálatokkal valósul meg.

b) Olyan matematikai módszert kell keresni, amelyik stabil a mérési hibákra nézve. Felsorolunk néhányat a ma is használatos módszerek közül.

(i) A klasszikus legkisebb négyzetek módszere és közvetlen általánosításai (LEGENDRE 1806, GAUSS 1809). A

$$\sum (\mathbf{Ax}-\mathbf{u})^2 \to \min$$

legkisebb négyzetek módszere még ma is a legnépszerűbb kiegyenlítési módszer. Továbbfejlesztése két irányban történt: egyrészt a megoldás technikája változott (fázis módszerek), másrészt a mérések korreláltságának figyelembevételével finomították (súlyozott legkisebb négyzetek módszere). Ez a geodéziában a legelterjedtebb módszer ([17]).

(ii) A legkisebb összeg módszer a

$$\sum |\mathbf{Ax}-\mathbf{u}| \to \min$$

feltétel teljesítését követeli. Már a XVIII. században vizsgálták, de numerikus nehézségei miatt nem alkalmazták. Ma már hatékony lineáris programozási módszerek segítik ennek a modellnek a használóit ([18]).

(iii) Robusztus becslés, aminek a fogalmát 1948-ban KENDALL vezette be. Ezek olyan becslések, amelyek viszonylag érzéketlenek a mérések eloszlásfüggvényében korlátozott változásra.

Ezek közül az egyik:

$$\varphi(\mathbf{v}) \to \min$$

ahol

$$\varphi(\mathbf{v}) = \begin{cases} |\mathbf{v}|^2 & \text{ha} \quad |\mathbf{v}| \leqq 2\tau \\ 4\tau(|\mathbf{v}|-\tau) & \text{ha} \quad |\mathbf{v}| > 2\tau, \end{cases}$$

ahol $\mathbf{v}=\mathbf{Ax}-\mathbf{u}$, τ pedig a szórás.

Másik ilyen becslési elv a

$$\sum |\mathbf{v}|^p \to \min \quad 1 \leqq p \leqq 2.$$

Több robusztus becslési elvet vizsgáltak már geodéziai feladatokra. A tapasztalat szerint ezek jobbak, mint a klasszikus legkisebb négyzetek módszere, de közülük a legjobbat még nem sikerült kiválasztani ([22]).

(iv) Kombinált módszerek. Mivel az egyes módszerekkel nem mindig érnek el kielégítő eredményeket, gyakran megpróbálják felváltva használni azokat. Előfordul, hogy először legkisebb négyzetek módszerrel megoldják a feladatot, utána az outliereket (kiütő értékeket) levágják, majd újra elvégzik a kiegyenlítést. (*Dán módszer* [18]).

(v) *Kálmán—Bucy szűrő.* Bizonyos értelemben a legkisebb négyzetek módszere álltalánosításának tekinthető. Optimális szűrésre, ill. optimális előrejelzésre használható. Felteszik, hogy az állapotvektor időben változik és figyelembe vehető a mérések korreláltsága is. Rekurzív összefüggések segítségével határozhatjuk meg a t_{k+1}-beli állapotvektort a t_k-beli értékéből ([8]).

(vi) *Tyihonov-féle regularizációs módszer.* Azon az egyszerű észrevételen alapul, hogy ha az **u** mennyiségeket δ pontossággal mérjük és **A** pontosan adott, akkor a megoldásnak azon **x**-ek között kell lenni, amelyek eleget tesznek a

$$\|\mathbf{Ax}-\mathbf{u}\| \leqq \delta$$

egyenlőtlenségnek. Ezeket az **x**-eket **u**-val összemérhetőnek nevezzük. Az a feladat, hogy az összemérhető megoldások közül kiválasszuk azt, ami a feladat valódi megoldása. Hogy igazoljuk az eredményül kapott megoldás megbízhatóságát, „kvázireális" kísérletet kell végezni. Ez lehetővé teszi a zajtól függő eredmények pontosságának értékelését is. Maga TYIHONOV is eredményesen alkalmazta ezt a módszert geodéziai feladatok megoldására ([26]).

(vii) Legkisebb négyzetek módszerén alapuló kollokációs módszer. 1972-ben MORITZ a következő matematikai modellen alapuló kiegyenlítést használta, amit kollokációs módszernek neveztek el:

$$\mathbf{u} = \mathbf{Ax}+\mathbf{s}'+\mathbf{n},$$

ahol **u** jelöli a méréseket, **A** az úgynevezett alakmátrix, **x** az ismeretlen vektor, **s**′ a jel, **n** pedig a zaj vektora. Az **s**′ jel a matematikai modell hibájaként, az **n** zaj pedig mérési hibaként interpretálható. A módszert akkor lehet alkalmazni, ha jel második momentumai adottak. „Megoldóképletek" ismertek számos speciális esetre, sőt a megoldás megbízhatóságának kiszámítására is ([17]).

A fentiekből látható, hogy a geodézia és fotogrammetria problémái gyakran nem lineáris feladattal fogalmazhatók meg. Ha a nemlineáris feladatot linearizáljuk és a lineáris feladatot oldjuk meg, ez úgy tekinthető mint a nemlineáris feladat megoldásának első iterációs lépése.

Ezekután a fenti modellek többnyire nagyméretű, ritka mátrixú lineáris egyenletrendszer megoldását igénylik. Hogy a megoldásra hatékony módszert adhassunk, nagyon fontos a mátrix szerkezetének tanulmányozása. Már a 60-as években vizsgálták a fotogrammetriai blokk-kiegyenlítéskor használatos mátrix struktúráját és átjelölésekkel egyszerűbb alakra igyekeztek hozni. Általában nem a sáv szélességét próbálták csökkenteni, hanem blokkosítani, szétváló formára akarták hozni az egyenletrendszert, hogy az ismeretlenek egyes csoportjait külön kezelhessék.

A struktúravizsgálat után választhatjuk ki a megfelelő numerikus módszert. Felmerül a kérdés, iterációs vagy direkt módszert használjunk-e. Több összehasonlító vizsgálatot végeztek erre vonatkozóan ([9]).

Az iterációs módszerek nagy előnye, hogy a megoldást kívánt pontosságra lehet megadni, futás közben bizonyos hibaszűrési eljárások is használhatók és a kerekítések nem okoznak figyelemreméltó hibákat. A futást bármikor be lehet fejezni és újra lehet indítani jelentős költségnövekedés nélkül. Ezért idáig többnyire az iterációs módszereket használták ([7]).

Újabban néhány súlyos probléma merült fel a számításoknál. Észrevették, hogy iterációs módszereknél a konvergencia függ a kezdeti értékektől és a rendszer kondi-

cionáltságától. Gyengén kondicionált rendszerben nagyon sok iteráció kell és ráadásul nagyon nehéz megbízható kritériumot találni a számítás befejezésére. Ezért újra a direkt módszerek kerültek előtérbe. A számítógépek kapacitásának növekedése támogatja ezt az irányzatot. Másik jelentős lépés a direkt módszerek használhatósága terén a ritkamátrixos technika fejlődése. Ilyen programrendszerek már néhány hazai számítóközpontban is megtalálhatók, ill. fejlesztés alatt vannak. A [12]-ben ismertetett programrendszer továbbfejlesztéséről egy későbbi dolgozatban kívánunk beszámolni.

4. Függőleges kéregmozgási mérések kiegyenlítése

A szocialista országok geodéziai szolgálatainak többoldalú tudományos és műszaki együttműködése keretében a *Magyar Geodéziai Szolgálat* hét országot érintő *Kárpátok—Balkán terület* függőleges földkéregmozgási térképének elkészítését vállalta. Mivel csak az utolsó két szintezés idején történt meg a szomszédos országok szintezési hálózatainak összekapcsolása, így a kiegyenlítéshez az utolsó két szintezés adatait lehetett felhasználnia a közös vizsgálati hálózat összeállításánál. Hogy az országhatárokon keresztül meg legyen a közvetlen kapcsolat, feltételezik, hogy a földkéreg mozgásának sebessége egyenletes. Ugyancsak feltételezik, hogy a mérések normális eloszlásúak. Ebben az esetben a legkisebb négyzetek módszereivel torzítatlan becslést adhatunk a valószínűségi változók várható értékére, ezért ezt használtuk a kiegyenlítés elvégzéséhez,

Az előkészítő munkák során a következő elvet határozták el [15] alapján: A ki egyenlítést 3 lépésben kell elvégezni: először a magasságkülönbségekre, majd a magasságkülönbségekre és a sebességekre együttesen előbb 1 pontra (Nadap), azután a 13 tengerszintmérőre rögzített hálózat esetére. Ezeket a számításokat a szerzők végezték 1983-ban.

A vizsgált hálózat pontjait összekötő vonalain a mért magasságkülönbségekre támaszkodva felírtuk az úgynevezett javítási egyenleteket. Ezek megfogalmazásához vezettük be a következő jelöléseket: Legyen a k-adik magassági pont előzetesen becsült magassága H_k^0, a kiegyenlített magassága H_k, jelöljük a különbséget $dH_k = H_k - H_k^0$-val. A k és a j végpontú i-edik vonalon mért magasságkülönbség legyen h_{kj}. Képezzük az i-edik vonalra nézve a

$$z_i = H_k - H_j - h_{kj} = H_k^0 + dH_k - (H_j^0 + dH_j) - h_{kj}$$

különbségeket. Átrendezve a

$$z_i = dH_k - dH_j + h_i \tag{4.1}$$

javítási egyenletet kapjuk, ahol $h_i = H_k^0 - H_j^0 - h_{kj}$ számolható. A terület 317 magassági pontjára 445 vonalon írtuk fel a (4.1) javítási egyenleteket. Ezeket rendszerbe foglalva az $\mathbf{x}_1 = \{dH_k\}$ és a $\mathbf{h} = \{h_k\}$ jelölést használva a

$$\mathbf{z} = \mathbf{A}_1\mathbf{x}_1 - \mathbf{h}, \tag{4.2}$$

javítási egyenletrendszert kapjuk. Az egyenletrendszer alakmátrixa $\mathbf{A}_1$ $n = 317$ oszlopból és 445 sorból álló ritka mátrix. Az egyenletrendszerben szereplő dH_i ismeretleneket ($d\mathbf{H} = \{dH_k\}$) úgy határozzuk meg, hogy a

$$\mathbf{z}^T\mathbf{P}_1\mathbf{z} \to \min$$

feltétel teljesüljön, ahol $\mathbf{P}_1$ egy diagonális súlymátrix:

$$\mathbf{P}_1 = \left\{\frac{1}{L_i m_i^2}\right\},$$

ahol L_i az i-edik mérési vonal hossza, m_i pedig a súlyegység középhibája. A $\frac{d(\mathbf{z}^T P_1 \mathbf{z})}{d\mathbf{x}_1}$ differenciálhányadost egyenlővé téve 0-val, a

$$\mathbf{A}_1^T \mathbf{P}_1 \mathbf{A}_1 \mathbf{x}_1 + \mathbf{A}_1^T \mathbf{P}_1 \mathbf{h} = 0 \tag{4.3}$$

lineáris egyenletrendszer megoldását igényli. Az

$$\mathbf{S}_1 = \mathbf{A}_1^T \mathbf{P}_1 \mathbf{A}_1$$

normál mátrix szimmetrikus, pozitív definit ritka mátrix. A

$$\mathbf{b}_1 = \mathbf{A}_1^T \mathbf{P}_1 \mathbf{h}$$

jelölés bevezetésével a (4.3) egyenletrendszer

$$\mathbf{S}_1 \mathbf{x}_1 = \mathbf{b}_1 \tag{4.4}$$

alakú lesz.

A feladat második részeként magasságok és a pontok függőleges sebességének együttes kiegyenlítését hajtottuk végre. Ehhez 1950 és 1970-es méréseket használtunk fel. A javítási egyenletek felírásánál figyelembe kellett venni, hogy a vizsgálati idő alatt a pontok magassága is megváltozott, hiszen a sebességek éppen ezen változások alapján számolhatók. Ezért a pontok magasságát egy rögzített időpontra „*epochá*"-ra rögzítve adták meg — jelöljük ezeket $\bar{H}_i^0$-lal — a szintezési pontoknak a választott *epochára* vonatkozó kiegyenlített magasságát pedig H_i-vel. Legyen V_i^0 az i-edik pont függőleges irányú mozgássebességének az *epocha* idején felvett előzetes értéke, dV_i pedig ennek változása.

Két pont között egy kiválasztott *epochára* vonatkozó magasságkülönbséget megkapjuk a követő és az előző pontok kiegyenlített, a kiválasztott *epochára* vonatkozó abszolút magasságának különbségeként.

$$h_i = H_k - H_j = (\bar{H}_k^0 + dH_k) - (\bar{H}_j^0 + dH_j) = h_i'' + z_i' + t_i''((V_k^0 + dV_k) - (V_j^0 + dV_j)), \tag{4.5}$$

ahol h_i'' a k-adik és j-edik pontok közötti i-edik szakaszra vonatkozó mért magasságkülönbség, z_i' pedig ennek javítása, míg t_i'' a mérés és az *epocha* között eltelt idő.

Egy vonal kiegyenlített relatív mozgássebességét megkapjuk, ha a vonal két végpont abszolút sebességének különbségét képezzük.

$$v_i = V_k - V_j = (V_k^0 + dV_k) - (V_j^0 + dV_j) = \Delta V_i + z_i''. \tag{4.6}$$

A két szintezés közötti időszakban a mozgást lineárisnak tételezve fel, a relatív sebességet a bekövetkezett magasságváltozás és az eltelt idő hányadosaként kapjuk. Így a levezetett relatív mozgássebesség a következőképpen számítható

$$\Delta V_k = \frac{h_k'' - h_k'}{T_k},$$

ahol h_k' és h_k'' az első és a második szintezés mért magasságkülönbségei, T_k a két mérés közt eltelt idő.

A (4.5) és (4.6) átrendezésével a javítási egyenletek:

$$z_i' = dH_k - dH_j + t_i''(dV_j - dV_k) + w_i',$$
$$z_i'' = dV_k - dV_j + w_i'', \tag{4.7}$$

ahol a w_i' és w_i'' számolhatók:

$$w_i'' = V_k^0 - V_j^0 - \Delta V_i,$$
$$w_i' = \overline{H}_k^0 - \overline{H}_j^0 - h_i'' + t_i'' V_j^0 - t_i'' V_k^0. \tag{4.8}$$

A z_i' és z_i'' javításokat a $\mathbf{z}$ vektorba, a w_i' és w_i'' (4.7)-ből számolt mennyiségeket pedig a $\mathbf{w}$ vektorba rendezve javítási egyenletrendszerünk (4.7) alapján a következőképpen írható fel:

$$\mathbf{z} = \mathbf{A}_2\mathbf{x}_2 - \mathbf{w}, \tag{4.9}$$

ahol $\mathbf{A}_2$ az egyenletrendszer alakmátrixa, $\mathbf{x}_2$ pedig a dH_k és dV_k ismeretlen változások vektora.

Az egyenletrendszerben szereplő dH_k és dV_k ismeretleneket úgy határozzuk meg, hogy a (4.9)-ben meghatározott ellentmondások vektorára a $\mathbf{z}^T\mathbf{P}_2\mathbf{z}$ kvadratikus funkcionál minimális legyen, ahol $\mathbf{P}_2$ súlymátrix. $\mathbf{P}_2$ 2×2-es blokkokból álló blokkdiagonális mátrix. Az i-edik blokk

$$\mathbf{P}^{(i)} = \mathbf{Q}_i^{-1} = \begin{bmatrix} q_{i,i} & q_{i,i+1} \\ q_{i+1,i} & q_{i+1,i+1} \end{bmatrix}^{-1},$$

ahol

$$q_{i,i} = \frac{L_i m_i''^2}{c}, \quad q_{i+1,i+1} = \frac{L_i(m_i'^2 + m_i''^2)}{c\Delta T_i^2}, \quad q_{i,i+1} = q_{i+1,i} = \frac{L_i m_i''^2}{c\Delta T_i}$$

(L_i az i-edik mérési vonal hossza, m_i' és m_i'' az első és második szintezés középhibája, ΔT_i a két szintezés közt eltelt idő, c tetszőleges konstans). A minimum elérésének szükséges feltételéből

$$\frac{d(\mathbf{z}^T\mathbf{P}_2\mathbf{z})}{d\mathbf{x}_2} = (\mathbf{A}_2^T\mathbf{P}_2\mathbf{A}_2)\mathbf{x}_2 + \mathbf{A}_2^T\mathbf{P}_2\mathbf{w} = 0 \tag{4.10}$$

lineáris egyenletrendszer, normál egyenletrendszer adódik. Az egyenlet rendszerben szereplő

$$\mathbf{S}_2 = \mathbf{A}_2^T\mathbf{P}_2\mathbf{A}_2$$

normál mátrix szimmetrikus, pozitív definit rika mátrix.

A méréseket a résztvevő országok végezték saját területeiken. A vizsgálatban szereplő nemzetközi szintezési hálózat összhossza 35 000 km és három tenger *(Balti, Fekete és Adriai)* 13 tengerszint mérőjével (mareográfjával) van kapcsolatban.

A feladat harmadik részeként megoldottuk az együttes kiegyenlítést azzal a feltételezéssel is, hogy a tengerszint magasságokat 0-nak választottuk. A feladat megfogalmazását ebben az esetben is a (4.5), (4.6), (4.7), (4.8), (4.9) és (4.10) képletek szolgáltatják.

A fenti számítások elvégzése után a kiegyenlített mennyiségek kovarienciamátrixát ami arányos a normálmátrix inverzével is meghatároztuk, hogy a megbízhatóságra is következtethessünk, (lásd [10]).

A feladat megoldásának vázlata.

a) A mérési adatokból felépítettük az $\mathbf{A}_1$ ritka mátrixot (méretei 317×450) és $\mathbf{b}_1$ vektort.

b) Felépítettük az S_1 ritka mátrixot (317×317).

c) Az S_1 mátrixon sávredukciót hajtottunk végre (*E. H. Cuthill és J. McKee módszerrel*, lásd [6]). A továbbiakban az itt kialakult számozást használtuk.

Megjegyezzük, hogy a mérési adatokat úgy kaptuk, hogy a pontok beszámozását egy a megrendelők által kedvezőnek ítélt sorrend szerint alakították ki. Ekkor a sávszélesség 37 volt. A sávredukciós programunk a sávszélességet 25-re redukálta.

d) Megoldottuk a (4.4) normál egyenletet. A megoldáshoz itt és a későbbiekben is szimmetrikus ritkamátrixú lineáris egyenletrendszer megoldó programot használtunk, ami *Gauss eliminációval* dolgozik. Ennek ismertetését egy későbbi dolgozatban tervezzük. A program először faktorizálta a normál mátrixot, ami szimetrikus pozitív definit ritka mátrix. A program vizsgálata, hogy a faktorizálás közbeni főelem választás kell-e és szükség esetén a megfelelő cserék végrehajtására felkészült. Ellenőrzésként az egyenletrendszer megoldását elvégeztük más módszerrel is. A normál mátrixot sávmátrixnak tekintve elvégeztük a sávmátrix faktorizálását a [12] dolgozatban ismertetett programmal is. A kétféleképpen kapott eredmények teljesen (7 jegy pontossággal) megegyeztek.

A sávredukciót valójában csak ezen összehasonlítás miatt hajtottuk végre. A ritkamátrixos megoldó program működéséhez a sávredukció nem szükséges. A számolási idők összehasonlításából a ritkamátrixos megoldás kedvezőbb volt mint az egyenletrendszer megoldására alkalmazott sávmátrixos megoldás.

A feladat második és harmadik részének számítási vázlata:

a) felépítettük az $\mathbf{A}_2$ mátrixot (634×900) és a $\mathbf{w}$ vektort,

b) elkészítettük az S_2 mátrixot, (634×634-es) szimmetrikus ritka mátrix,

c) megoldottuk a (4.10) egyenletrendszert.

Mindhárom esetben az egyenletrendszer megoldása mellett számolnunk kellett az S_1, ill. az S_2 mátrix inverzének főátlóbeli és a mellette levő elemeit is, amikből a megoldás középhibái kaphatók meg.

A számításokat az ASzSz Honeywell 66/20-as számítógépén végeztük. A programok FORTRAN nyelven készültek.

A második esetben az egyenletrendszer mérete 634 volt. S_2 mátrix felső fele (a szimmetricitás miatt csak ezt tároltuk) kezdetben 2800 nem 0 elemet tartalmazott. Az elimináció folytán ez 18 000-re növekedett. Az S_2 mátrix faktorizálása (ritkamátrixos technikával) és az egyenletrendszer megoldása 1,3 percet vett igénybe, az inverz elemeinek számítása kb. 20 percet.

A számítási eredmények alapján megkezdődhetett a földkéreg mozgás tudományos elemzése. Az értékelés, a következtetések kidolgozása a megfelelő mozgástérképek elkészítése még folyamatban van.

IRODALOM

[1] ACKERMANN, F., "Über Matrizen Strukturen bei Blockausgleichungen," *Photogrammetria* XIX (1962).

[2] ACKERMANN, F., Report of working Group III/1 Identification and Elimination of Gross and Systematic Errors.

[3] ACKERMANN, F., *Aerotriangulation with Independent Models* (Verlag Des Institut für angenandte Geodasie, Frankfurt A. M., 1972).

[4] ALPÁR, GY., Légiháromszögelési hálózatok matematikai modelljének vizsgálata, Doktori értekezés, 1974.
[5] ALPÁR, GY., „Légiháromszögelési tömbök kiegyenlítési eljárásainak összehasonlítása", *Geod. és Kart.* 1974/1.
[6] ARANY, I., SZÓDA, L., „Ritka szimmetrikus mátrixok sávszélesség redukciója", *Információ-Elektronika* 4 1973, 273—282.
[7] BÁN, I., „Iterációs módszerek lineáris rendszerekre", *SZTAKI Tanulmányok* 1976/56.
[8] BRAMMER, K. and SIFFLING, G., *Kalman-Bucy Filter* (1982, Nauka).
[9] CARLSON, E. and HALJALA, S., "Iterative methods for solving large photogrammetric normal equations", *The Photogrammetric Journal of Finland* 1974.
[10] DETREKÖI, Á., *Kiegyenlítőszámítások* (Budapest, 1973).
[11] DIVÉNYI, P. és TARASZOVA, G., A digitális domborzatmodellezés néhány módszere, Budapest, Földmérési Intézet, 1983.
[12] GERGELY, J., „Módszerek és programok ritka mátrixokra", *AML* 6 (1980) 407—442.
[13] GERGELY, J., „Geodéziai hálózatok kiegyenlítése", *Geodéziai és Kartográfiai* 35 (1983) 18—21.
[14] HAZAY, I., *Kiegyenlítőszámítások* (Tankönyvkiadó, 1966).
[15] HAZAY, I., „A vertikális kéregmozgási hálózatok kiegyenlítése", *Geodézia és Kartográfia* 1967/5.
[16] HOMORÓDI, L., *Fotogrammetria*, BME Jegyzet.
[17] KRAKIWSKY, E. J., A synthesis of recent advances in the method of least squares, 1975.
[18] KRARUP, T. and JUHL, J., "Götterdämmerung over least squares adjustment", Presented paper, ISP-Congress, 1980.
[19] MIHAILOVIC, K., "Adjustment of geodetic networks", XV. International congres of surveyors, June 6—14, 1977.
[20] MOLNÁR, L., „Eredményeink a fotogrammetriai tömbkiegyenlítésben", *FÖMI Tudományos közlemények*, 1972.
[21] RÉDEY, I., *A geodézia története* (Tankönyvkiadó, Bp., 1966).
[22] REY, W. J. J., *Robust Statistical Methods* (Springer Verlag, 1978).
[23] RIBNYIKOV, K. A., *A matematika története* (Tankönyvkiadó, Bp., 1974).
[24] SOMOGYI, J. „Gondolatok a tömbkiegyenlítéshez", *Geod. és Kart.* 1974/4.
[25] SOMOGYI, J., Tömbkiegyenlítések hazai alkalmazása gazdaságosság és pontosság figyelembevételével, Doktori értekezés, 1976.
[26] TYIHONOV, A. N., BOLSAKOV, V. A., NEJMAN, Ju. N., „Nyekonektnüje Zadacsi geodézii", *Geodezija i Aerofotoszjomka*, 1980.
[27] ZÁVOTI, J., "Digital map construction", *Acta Geodaet., Geoghys et Montanist. Acad. Sci. Hung.* 16 (1981) 237—244.

(Beérkezett: 1984. május 7.)

(Átdolgozva beérkezett: 1984. november 1.)

GERGELY JÓZSEF ÉS PERGEL JÓZSEFNÉ
FÖLDMÉRÉSI INTÉZET
1051 BUDAPEST, GUSZEV U. 19.

SOME APPLICATIONS OF MATHEMATICS IN GEODESY

J. GERGELY and I. PERGEL

First some common characteristics of the history of mathematics and geodesy are reminded. Several tipical geodetic problems and tasks are discussed. Mathematical models and methods are outlined for their solutions.

A concret solved leveling task is written, too.

Alkalmazott Matematikai Lapok 11 (1985) 385—398

A VÉGES CRISS-CROSS MÓDSZER IRÁNYÍTOTT MATROIDOKON

TERLAKY TAMÁS

Budapest

Cikkünkben a BLAND és LAS-VERGNAS [2, 4] által definiált irányított matroid és tulajdonságainak ismertetése után megmutatjuk a lineáris programozási feladatok és az irányított matroidok kapcsolatát.

A criss-cross módszer egy speciális változata segítségével konstruktív bizonyítást adunk *Minty színezési lemmájára* és a *Farkas lemma* általánosítására.

Végül bemutatjuk a criss-cross módszer általános alakját, melynek segítségével konstruktíven bizonyítjuk az irányított matroidokon adott általános dualitás tételt.

1. Bevezetés

Cikkünkben az irányított matroidokra alkalmazzuk a véges criss-cross módszert. A lineáris programozási feladatokra megfogalmazott criss-cross módszer és végességének bizonyítása TERLAKY [10, 11] dolgozataiban található. Mivel a criss-cross módszer, mint cikkünkben is bemutatjuk, irányított matroidokon is véges, így jól látható, hogy a criss-cross módszer a lineáris programozási feladatok kombinatorikus tulajdonságain alapszik.

A bevezető fejezetben röviden összefoglaljuk a matroidok azon alaptulajdonságait, melyeket a továbbiakban felhasználunk. Bizonyítás nélkül közöljük a különböző axiómarendszerek ekvivalenciáját.

A második fejezetben definiáljuk az irányított matroidot és ismertetjük néhány fontos tulajdonságát. Az irányított matroidok elméletének kidolgozása BLAND és LAS-VERGNAS [2, 4] nevéhez fűződik.

A lineáris programozási feladatok és az általuk generált irányított matroidok kapcsolatát mutatjuk be a harmadik fejezetben. Az ebben a részben közölt eredmények BLAND [2] cikkében találhatók. Így bemutatjuk, hogy az irányított matroid a lineáris programozás kombinatorikus absztrakciójaként is felfogható. Ennek az absztrakciónak a lehetőségét ROCKAFELLAR [9] vetette fel. Számos példa található irányított és nem irányítható matroidokra BLAND és LAS-VERGNAS [4] cikkében.

A negyedik fejezetben a criss-cross módszer egy speciális alakja segítségével konstruktív bizonyítást adunk *Minty színezési lemmájára* és a *Farkas lemma* általános alakjára. Konstruktív bizonyításunk új. *Minty színezési lemmáját* BLAND [2] bizonyította először, de bizonyítása induktív volt.

A criss-cross módszer általános alakját az ötödik fejezetben közöljük, melynek segítségével véges lépésben tudunk generálni optimális irányított ciklusokat, illetve kociklusokat. Konstruktívan bizonyítjuk az általános dualitási tételt, melyet elő-

ször BLAND [2] bizonyított. BLAND a tétel bizonyításához konstruált pivotálási szabályban csak a pivot elem létezését tudta bizonyítani a bázistábla egy adott oszlopában, nem tudta a pivot elemet közvetlenül megadni. Ezzel szemben a criss-cross módszer közvetlenül megadja a pivot elemet.

Mielőtt az irányított matroidot definiálnánk, néhány szót kell ejtenünk a matroidok alaptulajdonságairól. A matroidok a mátrixok és vektorterek lineáris függetlenségi tulajdonságainak absztrakciói (WHITNEY [14]). A matroidelméleti alapismeretek elsajátíthatók LAWLER [5] könyvéből, LOVÁSZ [6] vagy TUTTE [13] cikkéből. A számunkra is fontos matroidelméleti dualitás szép tárgyalása található MINTY [7] cikkében.

Az alábbiakban röviden közöljük a matroidok definícióját független halmazokon, bázisokon, illetve ciklusokon keresztül. Ismert, hogy ezek a definíciók ekvivalensek (TUTTE [13]).

Legyen $E=\{e_1, \ldots, e_n\}$ egy véges halmaz, és jelölje $P(E)$ az E halmaz részhalmazainak a halmazát. Legyen $\mathcal{F} \subset P(E)$ az E halmaz bizonyos részhalmazainak az összessége.

1.1. Definíció. Az $F \in \mathcal{F}$ halmazokat *függetleneknek* nevezzük és az $M=(E, \mathcal{F})$ párt *matroidnak* nevezzük, ha

1. $\emptyset \in \mathcal{F}$,
2. $F_1 \in \mathcal{F}$ és $F_2 \subset F_1$, akkor $F_2 \in \mathcal{F}$,
3. $F_1, F_2 \in \mathcal{F}$ és $\|F_1\| > \|F_2\|$, akkor van olyan $e \in F_1 \setminus F_2$, hogy $F_2 \cup \{e\} \in \mathcal{F}$.

1.2. Definíció. Egy $C \subset E$ halmazt *ciklusnak* nevezünk, ha $C \notin \mathcal{F}$, de minden $F \subsetneq C$ esetén $F \in \mathcal{F}$.

Jelöljük $\mathcal{C}$-vel az E-n értelmezett ciklusok halmazát.

1.3. Definíció. Az $M=(E, \mathcal{C})$ pár matroid és a $C \in \mathcal{C}$ halmazok a matroid ciklusai, ha

(a) 1. $C_1, C_2 \in \mathcal{C}$ és $C_1 \subset C_2$, akkor $C_1 = C_2$,
2. $C_1, C_2 \in \mathcal{C}$ és $e_i \in C_1 \setminus C_2$, $e_j \in C_1 \cap C_2$, akkor van olyan $C_3 \subset \mathcal{C}$, hogy $e_i \in C_3 \subset (C_1 \cup C_2) \setminus \{e_j\}$.

(b) Ekkor a matroid független halmazai azok az $F \subset E$ halmazok, melyek nem tartalmaznak ciklust.

1.4. Definíció. A $B \subset E$ halmazt *bázisnak* nevezzük, ha $B \in \mathcal{F}$ és nincs olyan $F \in \mathcal{F}$, hogy $B \subsetneq F$.

Jelöljük $\mathcal{B}$-vel az E-n értelmezett bázisok halmazát.

1.5. Definíció. Az $M=(E, \mathcal{B})$ pár matroid és a $B \in \mathcal{B}$ halmazok a matroid bázisai, ha

(a) 1. $B_1, B_2 \in \mathcal{B}$ esetén $\|B_1\| = \|B_2\|$,
2. $B_1, B_2 \in \mathcal{B}$ és $e_i \in B_1$, akkor létezik olyan $e_j \in B_2$, hogy $(B_1 \setminus \{e_i\}) \cup \{e_j\} \in \mathcal{B}$.

(b) Ekkor a matroid független halmazai a B bázisok és ezek részhalmazai.

Könnyen belátható, hogy ha $\mathcal{B}^* = \{B^* | B^* = E \setminus B$ valamely $B \in \mathcal{B}$ esetén$\}$, akkor $M^* = (E, \mathcal{B}^*)$ szintén matroid.

1.6. Definíció. Az $M^* = (E, \mathcal{B}^*)$ matroidot az $M=(E, \mathcal{B})$ matroid *duálisának* nevezzük.

Megjegyezzük, hogy a duális matroid ciklusait *kociklusoknak* nevezzük, valamint a matroid bázisainak elemszámát *a matroid rangjának* nevezzük.

A törlés és összehúzás művelete, egy adott matroidra alkalmazva, egy újabb matroidot eredményez.

1.7. Definíció. Ha az $M=(E, \mathscr{C})$ matroidot az $\tilde{M}=(\tilde{E}, \tilde{\mathscr{C}})$ matroiddal helyettesítjük, ahol $\tilde{E}=E\setminus e$, $\tilde{\mathscr{C}}=\{C|C\in\mathscr{C}$ és $e\notin C\}$, akkor azt mondjuk, hogy az e elemet *töröltük* M-ből.

1.8. Definíció. Ha az $M=(E, \mathscr{C})$ matroidot az $\tilde{M}=(\tilde{E}, \tilde{\mathscr{C}})$ matroiddal helyettesítjük, ahol $\tilde{E}=E\setminus e$, $\tilde{\mathscr{C}}=\{C\setminus\{e\}|C\setminus\{e\}\neq\emptyset$, $C\in\mathscr{C}$ és nincs olyan $C_0\in\mathscr{C}$, hogy $C_0\setminus\{e\}\subsetneq C\setminus\{e\}\}$, akkor azt mondjuk, hogy az e elemet *összehúztuk* M-ben.

Ismert, hogy egy elem összehúzásával illetve törlésével ismét matroidot kapunk, valamint a törlésnek, illetve összehúzásnak a duális matroidban összehúzás illetve törlés felel meg.

2. Az irányított matroid definíciója és alaptulajdonságai

Ebben a fejezetben az irányított matroid BLAND [2] által adott definícióját, és az irányított matroidok azon alaptulajdonságait foglaljuk össze, melyek a criss-cross módszer végrehajtásához, illetve végességének bizonyításához szükségesek.

Legyen $E=\{e_1, \ldots, e_n\}$ egy véges halmaz. *Előjeles halmaznak* nevezünk egy $X=(X^+, X^-)$ halmazpárt, ha X^+, $X^-\subset E$ és $X^+\cap X^-=\emptyset$. Az $X=(X^+, X^-)$ előjeles halmazt úgy is tekinthetjük, mintha az $\overline{X}=X^+\cup X^-$ halmazt osztottuk volna pozitív és negatív elemekre. Ha $X=(X^+, X^-)$ egy előjeles halmaz, akkor X ellentettjének a $(-X)$ előjeles halmazt nevezzük, ahol $(-X)^+=X^-$ és $(-X)^-=X^+$. Az $Y=\pm X$ jelölést használjuk, ha vagy $Y=X$ vagy $Y=-X$. Ha $O=\{X_1, \ldots, X_p\}$ előjeles halmazok egy rendszere E-n, akkor $\overline{O}=\{\overline{X}_1, \ldots, \overline{X}_p\}$-vel jelöljük a megfelelő (nem előjeles) halmazrendszert.

2.1. Definíció. Legyenek O és O^* előjeles halmazok rendszerei E-n. Az $M=(E, O)$ és $M^*=(E, O^*)$ *párokat duális irányított matroidoknak* nevezzük, ha az alábbi négy feltétel fennáll.

(a) $\overline{O}$, illetve $\overline{O}^*$ ciklusai illetve kociklusai az $\overline{M}=(E, \overline{O})$ illetve $\overline{M}^*=(E, \overline{O}^*)$ duális matroidoknak.

(b) $X\in O\Rightarrow -X\in O$ és $Y\in O^*\Rightarrow -Y\in O^*$.

(c) $X_1, X_2\in O$ és $\overline{X}_1=\overline{X}_2\Rightarrow X_1=\pm X_2$,

$Y_1, Y_2\in O^*$ és $\overline{Y}_1=\overline{Y}_2\Rightarrow Y_1=\pm Y_2$,

(d) $X\in O$, $Y\in O^*$ és $\overline{X}\cap\overline{Y}\neq\emptyset$ esetén

$$(X^+\cap Y^+)\cup(X^-\cap Y^-)\neq\emptyset \quad \text{és} \quad (X^+\cap Y^-)\cup(X^-\cap Y^+)\neq\emptyset.$$

Ha $M=(E, O)$ és $M^*=(E, O^*)$ duális irányított matroidok, akkor az M és M^* matroidot *irányított matroidnak* nevezzük, és O, illetve O^* az M, illetve M^* matroid egy irányítása. Az M^* irányított matroidot az M irányított matroid duálisának nevezzük. A matroid dualitás tulajdonságaiból azonnal következik, hogy minden irányított matroidnak létezik és egyértelmű a duálisa, és $M^{**}=M$.

A (d) feltételt *ortogonalitási feltételnek* nevezzük. A továbbiakban az X és Y előjeles halmazokat *ortogonálisaknak* nevezzük, ha $\overline{X} \cap \overline{Y} = \emptyset$ vagy (d) fennáll.

BLAND és LAS VERGNAS ([4] 2.1—2.2. tétel) bizonyította, hogy a (d) ortogonalitási feltétel ekvivalens az alábbi (d') feltétellel.

(d') Minden $X_1, X_2 \in O$, $e' \in (X_1^+ \cap X_2^-) \cup (X_1^- \cap X_2^+)$ és $e'' \in (X_1^+ \setminus X_2^-) \cup (X_1^- \setminus X_2^+)$ esetén van olyan $X_3 \in O$, hogy $X_3^+ \subset (X_1^+ \cup X_2^+) \setminus \{e'\}$, $X_3^- \subset (X_1^- \cup X_2^-) \setminus \{e'\}$ és $e'' \in \overline{X}_3$.

Hasonlóan O^* esetén is.

A fejezet hátralevő részében a lineáris algebrából jól ismert bázistáblának megfelelő bázistábla konstrukciót ismertetjük irányított matroidokra, majd bizonyítás nélkül közöljük a bázistábla alaptulajdonságait. Az alábbi bázistábla konstrukció szintén BLAND [2] cikkében található.

Legyen $\mathcal{B}$ az $\overline{M}$ matroid bázisainak halmaza és legyen m az $\overline{M}$ matroid rangja. Jól ismert, hogy tetszőleges $B = \{e_{b_1}, \ldots, e_{b_m}\} \in \mathcal{B}$ bázis esetén minden e_{b_i}-hez ($i = 1, \ldots, m$) egyértelműen létezik $\overline{Y}_{b_i} \in \overline{O}^*$ kociklus $\overline{M}^*$-ban úgy, hogy $\overline{Y}_{b_i} \cap B = \{e_{b_i}\}$. Az $\{\overline{Y}_{b_1}, \ldots, \overline{Y}_{b_m}\}$ halmazt az $E \setminus B$ duál bázishoz tartozó *kociklusok alaprendszerének* nevezzük. Legyen Y_{b_i} $i = 1, \ldots, m$ az az egyértelműen létező irányított kociklus, melyre $e_{b_i} \in Y_{b_i}^+$ és $\overline{Y}_{b_i} \cap B = \{e_{b_i}\}$.

Legyen $T(B)$ az $\{Y_{b_1}, \ldots, Y_{b_m}\}$ irányított kociklusok alaprendszerének előjeles incidencia mátrixa, azaz a $T(B)$ mátrix sorait az Y_{b_i} irányított kociklusok előjeles incidencia vektorai alkotják. Hasonlóan a [10, 11] dolgozatokban bevezetett jelölésekhez, az Y_{b_i} irányított kociklushoz tartozó sort $T(B)$-ben a b_i-edik sornak nevezzük (a szokásos i-edik sor elnevezés helyett). Így a $T(B)$ mátrix τ_{ij} eleme azt jelenti, hogy az $Y_i \in \{Y_{b_1}, \ldots, Y_{b_m}\}$ irányított kociklusban milyen az e_j elem előjele.

3. A lineáris programozási feladat és az irányított matroid kapcsolata

Ebben a fejezetben először megmutatjuk, hogy egy altér miként definiál egy irányított matroidot. Az ilyen matroidokat *reprezentálható matroidoknak* nevezzük. Számos példa található reprezentálható, nem reprezentálható, irányítható és nem irányítható matroidokra BLAND és LAS-VERGNAS [4] cikkében.

Az ebben a fejezetben közölt eredmények BLAND [2] cikkében találhatók, itt csak azért ismételjük meg őket, mivel jól szemléltetik az irányított matroid fogalmát. Így könnyebben érthetőek lesznek későbbi állításaink és betekintést nyerünk a lineáris programozási feladatok kombinatorikus tulajdonságaiba is.

Bizonyítani fogjuk a lineáris programozási feladatok extremális megoldásainak és a megfelelő irányított matroid irányított ciklusainak kapcsolatát. Először néhány definíciót közlünk.

3.1. Definíció. Az $x = (\xi_1, \ldots, \xi_n) \in R^n$ vektor *tartóján* az $S(x) = \{i | \xi_i \neq 0, i = 1, \ldots, n\}$ halmazt értjük.

3.2. Definíció. Az $x \in R^n$ vektor *előjeles tartóján* az $(S^+(x), S^-(x))$ párt értjük, ahol $S^+(x) = \{i | \xi_i > 0, i = 1, \ldots, n\}$, $S^-(x) = \{i | \xi_i < 0, i = 1, \ldots, n\}$.

3.3. Definíció. Az $x \in \mathcal{L} \subset R^n$ ($\mathcal{L}$ altér) vektort *elemi* vagy *minimális tartójú* vektornak nevezzük, ha minden $y \in \mathcal{L}$, $S(y) \subseteq S(x)$ esetén $S(y) = S(x)$.

1. Példa. Legyen A egy tetszőleges m-szer n-es mátrix. Legyen $\mathscr{L}=\lin(A)$ az A mátrix sorai által kifeszített altér, és jelölje $\mathscr{L}^\perp$ az $\mathscr{L}$ altér ortogonális kiegészítőjét R^n-ben. Az $M=(E, O)$ és $M^*=(E, O^*)$ duális irányított matroidok, ahol $E=\{1, \dots, n\}$,

$O=\{(S^+(x), S^-(x)) | x$ elemi vektor az $\mathscr{L}^\perp$ altérben$\}$,

$O^*=\{(S^+(x), S^-(x)) | x$ elemi vektor az $\mathscr{L}$ altérben$\}$.

Könnyen ellenőrizhető, hogy az így definiált M és M^* matroidok kielégítik az irányított matroid axiómáit.

2. Példa. Legyen $\bar{A}$ egy tetszőleges m-szer $(n-1)$-es, teljes sorrangú mátrix (rang $(A)=m$). Tekintsük az alábbi formában adott lineáris programozási feladatot.

$$\max \xi_2$$

feltéve, hogy

$$b = \bar{A}\bar{x} \tag{3.1}$$

$$\xi_3, \dots, \xi_n \geqq 0,$$

ahol $\bar{x}=(\xi_2, \xi_3, \dots, \xi_n)$.

Legyen $E=\{1, \dots, n\}$, $A=[-b, \bar{A}]$ m-szer n-es mátrix és $\mathscr{L}=\lin(A)$. Legyen továbbá $P=\{\bar{x} | b=\bar{A}\bar{x},\ \xi_3, \dots, \xi_n \geqq 0\}$ és $\hat{P}=\{x | Ax=0,\ \xi_1=1,\ \xi_3, \dots, \xi_n \geqq 0\}$. Jól ismert, hogy egy-egy értelmű megfeleltetés van P és $\hat{P}$ poliéderek pontjai közt.

Legyen az $M=(E, O)$ irányított matroid az $\mathscr{L}^\perp=\{x | Ax=0\}$ altér által (az 1. példában adottak szerint) definiált irányított matroid.

Az alábbi tétel szerint az M matroid „megengedett" ciklusai és a P poliéder extremális pontjai között egy-egy értelmű megfeleltetés létezik.

3.1. TÉTEL. Egy $\bar{x} \in R^{n-1}$ vektor akkor és csak akkor extremális megoldása a (3.1) feladatnak, ha $x=(1, \bar{x})$ esetén az $X=(S^+(x), S^-(x))$ előjeles halmazra fennáll, hogy

$$X \in O, \quad 1 \in X^+, \quad X^- \subset \{2\} \tag{3.2}$$

és

$$(\bar{X} \setminus \{1\}) \cup \{2\} \text{ független } \bar{M}\text{-ben.}$$

Bizonyítás. Tegyük fel, hogy az X előjeles halmaz kielégíti a (3.2) feltételeket. Így X előjeles tartója egy $x \in \mathscr{L}^\perp$, $\xi_1 > 0$ elemi vektornak. Könnyen belátható, hogy egy elemi vektor skalár-szorosa is elemi vektor, valamint, hogy az azonos tartójú elemi vektorok egymás skalár-szorosai. Így feltehetjük, hogy $\xi_1=1$ és így x egyértelmű. Mivel $X^- \subset \{2\}$, így $\xi_3, \dots, \xi_n \geqq 0$, azaz $x \in \hat{P}$. Belátjuk, hogy x extremális pontja $\hat{P}$-nek.

Tegyük fel, hogy van olyan $x^1, x^2 \in \hat{P}$ pont, hogy $x=\frac{1}{2}x^1+\frac{1}{2}x^2$. Mivel $x^i \in \hat{P}$, így $\xi_1^i=1$ és $\xi_j^i \geqq 0$, $j=3, \dots, n$, $i=1, 2$ és így $S(x^i) \subseteq S(x) \cup \{2\}$. Mivel $(\bar{X} \setminus \{1\}) \cup \{2\}$ független M-ben, így az A mátrix $(S(x) \setminus \{1\}) \cup \{2\}$ indexekhez tartozó oszlopai függetlenek R^m-ben. Mivel R^m-ben a fenti oszlopok segítségével a b vektor előállítása egyértelmű, így $x=x^1=x^2$ azaz x extremális pontja $\hat{P}$-nek, s így az $\bar{x}=(\xi_2, \dots, \xi_n)$ pont extremális pontja P-nek.

Fordítva, ha $\bar{x}=(\xi_2, \dots, \xi_n)$ extremális pontja P-nek, akkor az $x=(1, \xi_2, \dots, \xi_n)$ pont extremális pontja $\hat{P}$-nek. Így nyilván $X=(S^+(x), S^-(x))$ esetén $1 \in S^+(x)$, $S^-(x) \subset \{2\}$.

Először megmutatjuk, hogy x elemi vektor, azaz $X \in O$. Tegyük fel indirekt, hogy van olyan $x' \in \mathscr{L}^{\perp}$, hogy $S(x') \subsetneq S(x)$. Ha $\xi_1' = 0$, akkor $x \pm \varepsilon x' \in \hat{P}$ ($\varepsilon > 0$ kicsi), ellentmondva annak, hogy x extremális $\hat{P}$-ben. Így feltehetjük, hogy $\xi_1' = -1$. Ekkor viszont $x'' = x + x' \in \mathscr{L}^{\perp}$ esetén $\xi_1'' = 0$, ami a fentiek miatt lehetetlen, így x elemi vektor, azaz $X \in O$.

Végül megmutatjuk, hogy $(\bar{X} \setminus \{1\}) \cup \{2\}$ független $\bar{M}$-ben. Tegyük fel az ellenkezőjét, ekkor van olyan $0 \neq z \in \mathscr{L}^{\perp}$, melyre $S(z) \subset (\bar{X} \setminus \{1\}) \cup \{2\} = (S(x) \setminus \{1\}) \cup \{2\}$. Ekkor alkalmas, kicsi $\varepsilon > 0$ esetén $x \pm \varepsilon z \in \hat{P}$, ami ellentmond x extremalitásának. Tételünket így beláttuk.

Megjegyezzük, hogy tételünk tulajdonképpen annak a jól ismert állításnak felel meg, hogy egy megengedett megoldás akkor és csak akkor extremális, ha bázismegoldás.

A (3.1) lineáris programozási feladat duálisa az alábbi.

$$\min yb$$

feltéve, hogy

$$ya_2 = 1$$

(3.3)

$$ya_j \geqq 0, \quad j = 3, \ldots, n$$

ahol $y \in R^m$.

Tételünkkel ekvivalens állításként kapjuk, hogy egy-egy értelmű megfeleltetés van a (3.3) feladat extremális megoldásai és az $Y \in O^*$ irányított kociklusok közt, melyekre

$$Y \in O^*, \quad 2 \in Y^+, \quad Y^- \subset \{1\}$$

és

(3.4) $$(\bar{Y} \setminus \{2\}) \cup \{1\} \quad \text{független} \quad M^*\text{-ban}.$$

A fenti állítás igazolását az olvasóra hagyjuk. Csupán azt jegyezzük meg, hogy ha y extremális megoldása a (3.3) feladatnak, akkor a neki megfeleltetett $Y \in O^*$ irányított kociklus az alábbi:

$$Y = (S^+(z), S^-(z)), \quad \text{ahol} \quad z = (yb, ya_2, ya_3, \ldots, ya_n).$$

A fentiekkel azt is megmutattuk, hogy a duális lineáris programozási feladatok segítségével definiált irányított matroidok egymás duálisai.

4. Alternatíva tételek bizonyítása a criss-cross módszerrel

Ebben a fejezetben a criss-cross módszer egy speciális változata segítségével irányított matroidokon adott alternatíva tételeket bizonyítunk. Így bizonyítjuk a lineáris algebrából jól ismert *Farkas lemma* és a gráfelméletből jól ismert *Minty-féle színezési tétel* általánosítását.

Legyen $E = \{e_1, \ldots, e_n\}$ valamint $M = (E, O)$ és $M^* = (E, O^*)$ duális irányított matroidok.

4.1. Definíció. Az $X \in O$ irányított ciklust *megengedettnek* nevezzük, ha $e_1 \in X^+$ és $X^- = \emptyset$.

Az alábbi feladat megoldására adunk algoritmust ebben a fejezetben.

M.1. Feladat. Keressünk megengedett irányított ciklust, ha létezik, illetve mutassuk ki, hogy nem létezik irányított ciklus M-ben, amely megengedett.

A továbbiakban a 2. részben adott bázistáblát használjuk fel. Feltesszük, hogy $e_1 \notin B$, és így a $T(B)$ tábla e_1-oszlopa ad egy X_1 irányított ciklust. Így az M.1. feladat az alábbi módon is megfogalmazható:

Keressünk olyan bázistáblát, melyben $(-X_1) \in O$ megengedett irányított ciklus, illetve olyan bázistáblát, mely bizonyítja, hogy nincs megengedett irányított ciklus. Először egy egyszerű lemmát bizonyítunk.

4.1. LEMMA. Ha valamely bázistábla és $e_k \in B$ esetén $\tau_{k1} = +1$ és $\tau_{ki} \in \{0, +1\}$ ha $e_i \notin B$, akkor nem létezik $X \in O$ megengedett irányított ciklus.

Bizonyítás. Indirekt tegyük fel, hogy van $X \in O$ megengedett irányított ciklus, azaz $X = (X^+, X^-)$, $e_1 \in X^+$, $X^- = \emptyset$. Tekintsük az $e_k \in B$ elemhez tartozó $Y_k \in O^*$ irányított kociklust. Ekkor $e_1 \in \overline{X} \cap \overline{Y}_k \neq \emptyset$ és $e_1 \in (X^+ \cap Y_k^+) \cup (X^- \cap Y_k^-) \neq \emptyset$, de $(X^+ \cap Y^-) \cup (X^- \cap Y^+) = \emptyset$ mivel $X^- = Y^- = \emptyset$. Így ellentmondásba kerültünk az X és Y_k ciklusok ortogonalitásával, lemmánkat beláttuk.

Legyen adott egy $T(B)$ bázistábla ($e_1 \notin B$). Az algoritmusunkat definiáló pivotálási szabály az alábbi.

M.1. Pivotálási szabály

(a) (i) Ha $\tau_{i1} \in \{-1, 0\}$ minden $e_i \in B$ esetén, akkor $(-X_1)$ megengedett irányított ciklus. Az M.1. feladatot megoldottuk, eljárásunk véget ért.

(ii) Ha (i) nem áll fenn, legyen $r = \min \{i \mid \tau_{i1} = +1,\ e_i \in B\}$.

(b) (i) Ha $\tau_{rj} \in \{0, +1\}$ minden $e_j \notin B$ esetén, akkor 4.1. lemma szerint nincs megengedett irányított ciklus. Az M.1. feladatot megoldottuk, eljárásunk véget ért.

(ii) Ha (i) nem áll fenn, legyen $s = \min \{j \mid \tau_{rj} = -1,\ e_j \notin B\}$. Az e_r elem távozik, az e_s elem bekerül a bázisba. Pivotáljuk az (r, s) helyen. $\bar{B} = (B \cup \{e_s\}) \setminus \{e_r\}$.

Folytassuk az eljárást az új $T(\bar{B})$ bázistáblával. A bázistábla 4. tulajdonsága (BLAND [2]) szerint $\bar{B}$ is bázis. Az e_1 elem nem báziselem az eljárás során nyilvánvalóan.

Eljárásunk az (a) rész (i) vagy a (b) rész (i) eseténél ér véget, mindkét esetben megoldottuk M.1. feladatot. Az M.1. feladat megoldásához (mivel véges sok bázis van) mindössze azt kell bebizonyítanunk, hogy eljárásunk nem ciklizálhat, azaz tetszőleges B bázis legfeljebb egyszer fordulhat elő eljárásunk során.

4.1. TÉTEL. Az M.1. pivotálási szabály alkalmazásával ciklizálás nem fordulhat elő.

Bizonyítás. Tegyük fel indirekt, hogy eljárásunk ciklizál, azaz egy B bázisból indulva ismét a B bázishoz jutunk. Legyen $E^c = \{e_i \mid e_i$ kikerül a bázisból a ciklus során$\}$. Megjegyezzük, hogy $e_i \notin E^c$ esetén e_i vagy végig bázis, vagy végig nem bázis elem volt a ciklizálás során. Legyen $q = \max \{i \mid e_i \in E^c\}$.

Vizsgáljuk azt a két helyzetet, amikor e_q bekerül a bázisba és amikor távozik a bázisból. Legyen B' és B'' az előbb említett két bázis, és különböztessük meg $'$ illetve $''$-vel a $T(B')$ és $T(B'')$ tábla elemeit. Legyen e_r a bázist elhagyó elem, amikor e_q bekerül a bázisba és legyen e_s a bázisba bejövő elem, amikor e_q távozik a bázisból. Nyilvánvaló, hogy $q > 1$, $r, s < q$ és $e_r, e_s \in E^c$.

Tekintsük az Y_r' irányított kociklust és az X_1'' irányított ciklust, melyek az M.1. pivotálási szabály alapján az alábbi tulajdonságokkal rendelkeznek.

(1') $e_q \in Y_r'^-$ (1'') $e_q \in X_1''^+$

(2') $e_1 \in Y_r'^+$ (2'') $e_1 \in X_1''^-$

(3') $\overline{Y}_r' \subset (E \setminus B') \cup \{e_r\}$ (3'') $\overline{X}_1'' \subset B'' \cup \{e_1\}$

(4') $Y_r'^- \cap E^c = \{e_q\}$ (4'') $X_1''^+ \cap E^c = \{e_q\}$

Így $e_q \in \overline{X}_1'' \cap \overline{Y}_r' \neq \emptyset$ és $e_q \in (X_1''^+ \cap Y_r'^-) \cup (X_1''^- \cap Y_r'^+) \neq \emptyset$, viszont $(X_1''^+ \cap Y_r'^+) \subset E^c \cup \{e_1\}$ és $(X_1''^- \cap Y_r'^-) \subset E^c \cup \{e_1\}$ és így (1', 1'', 2', 2'', 4', 4'') tulajdonságok szerint mindkét halmaz üres. Ez ellentmond az irányított ciklusok és kociklusok ortogonalitásának, tételünket beláttuk.

Az M.1. pivotálási szabály által definiált algoritmus a criss-cross módszer egy speciális alakja. A criss-cross módszer általános alakját az 5. fejezetben tárgyaljuk.

A criss-cross módszer végességének közvetlen következménye a *Farkas lemma* egy általánosítása.

1. KÖVETKEZMÉNY Legyenek $M=(E, O)$ és $M^*=(E, O^*)$ duális irányított matroidok, $e \in E$ tetszőleges. Az alábbi alternatívák közül egy és csak egy áll fenn.

(a) van olyan $X \in O$ irányított ciklus, melyre $e \in X^+$, $X^- = \emptyset$ vagy

(b) van olyan $Y \in O^*$ irányított kociklus, hogy $e \in Y^+$, $Y^- = \emptyset$.

Bizonyítás. A ciklusok és kociklusok ortogonalitása miatt (a) és (b) egyidejűleg nem állhat fenn.

Számozzuk E elemeit úgy, hogy $e_1 = e$ legyen, ekkor a criss-cross módszer két kimenetele adja tételünk bizonyítását.

Hasonlóképpen nyerjük a *Minty-féle színezési tétel* általánosításának egy bizonyítását. Bizonyításunk konstruktív, míg BLAND [2] bizonyítása induktív.

2. KÖVETKEZMÉNY. Osszuk az E halmazt három R, G, W diszjunkt részre és legyen $e \in R$. Az alábbi alternatívák közül egy és csak egy áll fenn.

(a) van olyan $X \in O$, hogy $e \in \overline{X} \subset R \cup G$ és $X^- \cap R \neq \emptyset$, vagy

(b) van olyan $Y \in O^*$, hogy $e \in \overline{Y} \subset R \cup W$ és $Y^- \cap R = \emptyset$.

Bizonyítás. Használva a matroidelméletből jól ismert törlés és összehúzás műveletét (töröljük W-t és húzzuk össze G-t) tételünket visszavezethetjük az 1. következményre.

Természetesen az M.1. pivotálási szabály alkalmas módosításával is bizonyítható lett volna a 2. következmény. A bizonyításnak ezt a változatát most nem részletezzük.

5. A criss-cross módszer álltalános alakja és a dualitástétel bizonyítása

Legyen $E=\{e_1, \ldots, e_n\}$ és $M=(E, O)$ és $M^*=(E, O^*)$ duális irányított matroidok.

5.1. Definíció. Az $X \in O$ irányított ciklust *primál megengedettnek* nevezzük, ha $e_1 \in X^+$ és $X^- \subset \{e_2\}$.

5.2. Definíció. Az $Y \in O^*$ irányított kociklust *duál megengedettnek* nevezzük, ha $e_2 \in Y^+$ és $Y^- \subset \{e_1\}$.

Megjegyezzük, hogy a fenti megengedett halmazokat extremálisaknak is nevezzük. Használatos még a bázis megengedett elnevezés is. Az elnevezés indoklása új fogalmak bevezetését igényelné (BLAND [2]) így ettől itt, most eltekintünk.

5.3. Definíció. Az X és Y előjeles halmazokat *komplementárisaknak* nevezzük, ha $\overline{X} \cap \overline{Y} \subset \{e_1, e_2\}$.

5.4. Definíció. Az $X \in O$ primál megengedett irányított ciklust *optimálisnak* nevezzük, ha van olyan $Y \in O^*$ duál megengedett irányított kociklus, hogy X és Y komplementáris halmazok.

5.5. Definíció. Az $Y \in O^*$ duál megengedett irányított ciklust *optimálisnak* nevezzük, ha van olyan $X \in O$ primál megengedett irányított ciklus, hogy X és Y komplementáris halmazok. Ekkor nyilván X is optimális.

Az alábbi feladatot oldjuk meg ebben a fejezetben.

M.2. Feladat. Keressünk $X \in O$ és $Y \in O^*$ primál, illetve duál optimális irányított ciklust illetve kociklust, vagy bizonyítsuk be, hogy nem létezik optimális irányított ciklus, illetve kociklus.

Ebben a fejezetben is a 2. fejezetben bemutatott bázistáblát fogjuk felhasználni.

Feltesszük a továbbiakban, hogy $e_1 \notin B$ és $e_2 \in B$. Pivotálási szabályunk meg fogja őrizni ezt a tulajdonságot. A $T(B)$ tábla e_1-oszlopa egy $X_1 \in O$ irányított ciklust, az e_2-sora pedig egy $Y_2 \in O^*$ irányított kociklust ad. A lineáris programozási terminológiának megfelelően a tábla e_1-oszlopát *megoldás oszlopnak*, a tábla e_2-sorát *célfüggvény sornak* is nevezhetjük.

A $T(B)$ tábla optimális, ha $(-X_1) \in O$ és $Y_2 \in O^*$ primál illetve duál megengedettek, mivel ekkor $(-X_1)$ és Y_2 komplementárisak is $(\overline{X}_1 \subset B \cup \{e_1\}, \overline{Y}_2 \subset \subset (E \setminus B) \cup \{e_2\})$ azaz optimálisak.

Ha $\{e_2\} \in O$, akkor a fenti kívánalmaknak megfelelő tábla nem létezik, így feltesszük azt is, hogy $\{e_2\} \notin O$.

Az alábbi két lemma bizonyítja, hogyan látható egy $T(B)$ táblából, hogy nem létezik duál, illetve primál megengedett kociklus, illetve ciklus.

5.1. LEMMA. Ha valamely bázistábla és $e_k \notin B$, $k \neq 1$ esetén $\tau_{2k} = -1$ és $\tau_{ik} \in \{-1, 0\}$ $e_i \in B$ esetén, akkor nem létezik $Y \in O^*$ duál megengedett irányított kociklus.

Bizonyítás. Tegyük fel indirekt, hogy létezik $Y \in O^*$ duál megengedett irányított kociklus, azaz $Y = (Y^+, Y^-)$, $e_2 \in Y^+$, $Y^- \subset \{e_1\}$. Tekintsük a $T(B)$ tábla e_k-oszlopának megfelelő X_k irányított ciklust. Ekkor $e_2 \in \overline{X}_k \cap \overline{Y} \neq \emptyset$ és $e_2 \in (X_k^+ \cap Y^-) \cup \cup (X_k^- \cap Y^+) \neq \emptyset$, de $(X_k^+ \cap Y^+) \cup (X_k^- \cap Y^-) = \emptyset$ mivel $X_k^+ = \emptyset$, $e_1 \notin \overline{X}_k$ és $Y^- \subset \subset \{e_1\}$. Ez ellentmond X_k és Y irányított ciklusok ortogonalitásának, lemmánkat beláttuk.

5.2. LEMMA. Ha valamely bázistábla és $e_k \in B$, $k \neq 2$ esetén $\tau_{k1} = +1$ és $\tau_{ki} \in \{0, +1\}$ $e_i \notin B$ esetén, akkor nem létezik $X \in O$ primál megengedett irányított ciklus.

Bizonyítás. Tegyük fel indirekt, hogy létezik $X \in O$ primál megengedett irányí-

tott ciklus, azaz $X=(X^+, X^-)$, $e_1 \in X^+$, $X^- \subset \{e_2\}$. Tekintsük a $T(B)$ tábla e_k-sorának megfelelő Y_k irányított kociklust. Ekkor $e_1 \in \overline{X} \cap \overline{Y}_k \neq \emptyset$ és $e_1 \in (X^+ \cap Y_k^+) \cup (X^- \cap Y_k^-) \neq \emptyset$, de $(X^+ \cap Y_k^-) \cup (X^- \cap Y_k^+) = \emptyset$ mivel $Y_k^- = \emptyset$, $e_2 \notin \overline{Y}_k$ és $X^- \subset \{e_2\}$. Ez ellentmond az X és Y_k irányított ciklusok ortogonalitásának, lemmánkat beláttuk.

Amennyiben nem létezik primál megengedett ciklus vagy duál megengedett kociklus, akkor 5.4. és 5.5. definíciók értelmében nem létezik sem primál sem duál optimális ciklus, illetve kociklus.

Ha adott egy B bázis és $T(B)$ bázistábla, melyre $e_2 \in B$ és $e_1 \notin B$, akkor a criss-cross módszert definiáló pivotálási szabály az alábbi.

M.2. Pivotálási szabály

(a) (i) Ha $\tau_{2j} \in \{0, +1\}$ $j=2, \ldots, n$ és $\tau_{i1} \in \{-1, 0\}$ $e_i \in B$, $i \neq 2$, akkor a $(-X_1)$ irányított ciklus primál, az Y_2 irányított kociklus duál megengedett, azaz optimálisak. Az M.2. feladatot megoldottuk, eljárásunk véget ért.

(ii) Ha (i) nem áll fenn, legyen
$k = \min \{i \mid \tau_{2i} = -1 \text{ vagy } \tau_{i1} = +1,\ i = 3, \ldots, n\}$.

(b) (i) Ha $\tau_{2k} = -1$ és $\tau_{ik} \in \{-1, 0\}$ minden $e_i \in B$ esetén, akkor az 5.1. lemma szerint nincs $Y \in O^*$ duál megengedett irányított kociklus. Az M.2. feladatot megoldottuk, eljárásunk véget ért.

(ii) *Primál transzformáció*
Ha (i) nem áll fenn, legyen $r = \min \{i \mid \tau_{ik} = +1,\ e_i \in B\}$. Pivotáljunk az (r, k) helyen, az e_r elem távozik, az e_k elem kerül be a bázisba. $\bar{B} = (B \cup \{e_k\}) \setminus \{e_r\}$.

(c) (i) Ha $\tau_{k1} = +1$ és $\tau_{ki} \in \{0, +1\}$ $e_i \notin B$ esetén, akkor az 5.2. lemma szerint nincs $X \in O$ primál megengedett irányított ciklus. Az M.2. feladatot megoldottuk, eljárásunk véget ért.

(ii) *Duál transzformáció*
Ha (i) nem áll fenn, legyen $s = \min \{j \mid \tau_{kj} = -1,\ e_j \notin B\}$. Pivotáljunk a (k, s) helyen, az e_k elem távozik, az e_s elem kerül be a bázisba, $\bar{B} = (B \cup \{e_s\}) \setminus \{e_k\}$.

Folytassuk az eljárást az új $T(\bar{B})$ bázistáblával. A bázistábla 4. tulajdonsága értelmében $\bar{B}$ is bázis. Az eljárás során nyilván $e_2 \in B$ és $e_1 \notin B$.

Eljárásunk az (a) rész (i), (b) rész (i) vagy a (c) rész (i) eseténél ér véget. Az (a) rész (i) esetében optimális irányított ciklusokat kaptunk, a (b) rész (i) és a (c) rész (i) esetében pedig nem léteznek optimális irányított ciklusok. Az M.2. feladat megoldásához csak azt kell belátnunk, hogy az M.2. pivotálási szabállyal definiált criss-cross módszer nem ciklizálhat, mivel véges sok különböző $B \in \mathcal{B}$ bázis létezik.

5.1. TÉTEL. Az M.2. pivolálási szabály által definiált criss-cross módszer nem ciklizál, azaz véges lépésben véget ér.

Bizonyítás. Tegyük fel indirekt, hogy az eljárás ciklizál, azaz egy B bázisból indulva ismét a B bázishoz jutunk. Legyen $E^c = \{e_i \mid e_i$ elhagyja a bázist a ciklizálás során$\}$. Megjegyezzük, hogy $e_i \notin E^c$ esetén e_i vagy végig báziselem vagy végig nem báziselem volt. Jelölje $q = \max \{i \mid e_i \in E^c\}$.

Tekintsük azt a két helyzetet, amikor e_q bekerül a bázisba és amikor e_q távozik a bázisból. Legyen ebben a két helyzetben B', illetve B'' a két bázis. Különböz-

tessük meg ′, illetve ″-vel a $T(B')$, illetve $T(B'')$ táblák elemeit. Legyen e_r a bázisból kilépő elem, amikor e_q bekerül a bázisba, és e_s a bázisba belépő elem, amikor e_q kikerül a bázisból. Nyilvánvaló, hogy $q>2$, $r,s<q$ és $e_r, e_s \in E^c$.

Az alábbi négy esetet kell megkülönböztetnünk.

(α) e_q primál transzformációnál kerül be és primál transzformációnál kerül ki a bázisból.

(β) e_q primál transzformációnál kerül be és duál transzformációnál kerül ki a bázisból.

(γ) e_q duál transzformációnál kerül be és primál transzformációnál kerül ki a bázisból.

(δ) e_q duál transzformációnál kerül be és duál transzformációnál kerül ki a bázisból.

Vizsgáljuk a fenti négy esetet, be fogjuk bizonyítani, hogy egyik eset sem lehetséges, azaz ciklizálás nem fordulhat elő.

(α) Az e_q elem primál transzformációnál B' bázis esetén jön be a bázisba és e_r távozik, valamint az e_q elem primál transzformációnál B'' bázis esetén távozik a bázisból és e_s kerül be a bázisba.

Az M.2. pivotálási szabály alapján $Y_2' \in O^*$ és $X_s'' \in O$ az alábbi tulajdonságokkal rendelkezik.

(1′) $e_q \in Y_2'^-$ (1″) $e_q \in X_s''^+$

(2′) $e_2 \in Y_2'^+$ (2″) $e_2 \in X_s''^-$

(3′) $Y_2'^- \cap E^c = \{e_q\}$ (3″) $X_s''^+ \cap E^c = \{e_q\}$

(4′) $\overline{Y}_2' \subset (E \setminus B') \cup \{e_2\}$ (4″) $\overline{X}_s'' \subset B'' \cup \{e_s\}$

Az (1′, 1″) tulajdonságok szerint $e_q \in \overline{X}_s'' \cap \overline{Y}_2' \neq \emptyset$. A (4′, 4″) tulajdonságokat felhasználva $\overline{X}_s'' \cap \overline{Y}_2' \subset [B'' \cup \{e_s\}] \cap [(E \setminus B') \cup \{e_2\}] \subset E^c \cup \{e_2\}$, és így (3′, 3″) szerint $(X_s''^+ \cap Y_2'^+)$, $(X_s''^- \cap Y_2'^-) \subset \{e_2, e_q\}$ melyből (1′, 1″, 2′, 2″) felhasználásával kapjuk, hogy $X_s''^+ \cap Y_2'^+ = \emptyset$ és $X_s''^- \cap Y_2'^- = \emptyset$, ami ellentmond X_s'' és Y_2' ortogonalitásának. Így ez az eset nem lehetséges.

(β) Az e_q elem primál transzformációnál B' bázis esetén jön be és e_r távozik a bázisból, valamint az e_q elem duál transzformációnál B'' bázis esetén távozik a bázisból, amikor e_s jön be a bázisba. Tekintsük az X_1', X_1'' irányított ciklusokat és az Y_2', Y_2'' irányított kociklusokat.

Az X_1', X_1'' irányított ciklusok és az Y_2', Y_2'' irányított kociklusok az alábbi tulajdonságokkal rendelkeznek.

(X1) $X_1'^+ \cap E^c = \emptyset$ (Y1) $Y_2'^- \cap E^c = \{e_q\}$

(X2) $\overline{X}_1' \subset B' \cup \{e_1\}$ (Y2) $\overline{Y}_2' \subset (E \setminus B') \cup \{e_2\}$

(X3) $X_1''^+ \cap E^c = \{e_q\}$ (Y3) $Y_2''^- \cap E^c = \emptyset$

(X4) $\overline{X}_1'' \subset B'' \cup \{e_1\}$ (Y4) $\overline{Y}_2'' \subset (E \setminus B'') \cup \{e_2\}$

(X5) $e_1 \in X_1'^- \cap X_1''^-$ (Y5) $e_2 \in Y_2'^+ \cap Y_2''^+$.

A 2. fejezetben az ortogonalitási feltétellel ekvivalens feltételként adott (d') feltétel szerint ($X_1 = X_1''$, $X_2 = -X_1'$, $e' = e_1$, $e'' = e_q$, illetve $-Y_2'$, Y_2'', e_2, e_q szerep-

osztással) az alábbi tulajdonságokkal rendelkező $X \in O$ irányított ciklust és $Y \in O^*$ irányított kociklust nyerjük a fent felsorolt tulajdonságok alapján.

(1′) $e_1 \notin \overline{X}$ (1″) $e_2 \notin \overline{Y}$

(2′) $e_q \in X^+$ (2″) $e_q \in Y^+$

(3′) $X^+ \subset \{e_q\} \cup B' \cup (B'' \setminus E^c)$ (3″) $Y^+ \subset \{e_q\} \cup [(E \setminus B') \setminus E^c] \cup (E \setminus B'')$

(4′) $X^- \subset (B'' \setminus \{e_q\}) \cup (B' \setminus E^c)$ (4″) $Y^- \subset [(E \setminus B') \setminus \{e_q\}] \cup [(E \setminus B'') \setminus E^c]$.

A (2′, 2″) tulajdonságok szerint $e_q \in \overline{X} \cap \overline{Y} \neq \emptyset$, de (2′, 2″, 3′, 4″) szerint

$$X^+ \cap Y^- \subset \{\{e_q\} \cup B' \cup (B'' \setminus E^c)\} \cap \{[(E \setminus B') \setminus \{e_q\}] \cup [(E \setminus B'') \setminus E^c]\} = \emptyset,$$

és (2′, 4′, 2″, 3″) szerint

$$X^- \cap Y^+ \subset \{(B'' \setminus \{e_q\}) \cup (B' \setminus E^c)\} \cap \{\{e_q\} \cup [(E \setminus B') \setminus E^c] \cup (E \setminus B'')\} = \emptyset.$$

Ez ellentmond X és Y ortogonalitásának, így ez az eset sem lehetséges.

(γ) Az e_q elem duál transzformációnál B' bázis esetén jön be és e_r távozik a bázisból, valamint az e_q elem primál transzformációnál B'' bázis esetén távozik a bázisból amikor e_s jön be a bázisba.

Az M.2. pivotálási szabály alapján $Y'_r \in O^*$ és $X''_s \in O$ az alábbi tulajdonságokkal rendelkezik.

(1′) $e_1 \in Y_r'^+$ (1″) $e_2 \in X_s''^-$

(2′) $e_q \in Y_r'^-$ (2″) $e_q \in X_s''^+$

(3′) $Y_r'^- \cap E^c = \{e_q\}$ (3″) $X_s''^+ \cap E^c = \{e_q\}$

(4′) $\overline{Y}'_r \subset (E \setminus B') \cup \{e_r\}$ (4″) $\overline{X}''_s \subset B'' \cup \{e_s\}$

A (2′, 2″) tulajdonságok alapján $e_q \in \overline{X}''_s \cap \overline{Y}'_r \neq \emptyset$, valamint (4′, 4″) és E^c definíciójának felhasználásával kapjuk, hogy $\overline{X}''_s \cap \overline{Y}'_r \subset [B'' \cup \{e_s\}] \cap [(E \setminus B') \cup \{e_r\}] \subset E^c$. Így (3′, 3″) miatt $Y_r'^+ \cap X_s''^+ \subset \{e_q\}$ és $Y_r'^- \cap X_s''^- \subset \{e_q\}$, melyekből (2′, 2″) alapján kapjuk, hogy $Y_r'^+ \cap X_s''^+ = \emptyset$ és $Y_r'^- \cap X_s''^- = \emptyset$, ami ellentmond Y'_r és X''_s ortogonalitásának. Tehát ez az eset sem lehetséges.

(δ) Az e_q elem duál transzformációnál B' bázis esetén jön be és e_r távozik a bázisból, valamint az e_q elem duál transzformációnál B'' bázis esetén távozik a bázisból, amikor e_s jön be a bázisba.

Az M.2. pivotálási szabály alapján $Y'_r \in O^*$ és $X''_1 \in O$ az alábbi tulajdonságokkal rendelkezik.

(1′) $e_q \in Y_r'^-$ (1″) $e_q \in X_1''^+$

(2′) $e_1 \in Y_r'^+$ (2″) $e_1 \in X_1''^-$

(3′) $Y_r'^- \cap E^c = \{e_q\}$ (3″) $X_1''^+ \cap E^c = \{e_q\}$

(4′) $\overline{Y}'_r \subset (E \setminus B') \cup \{e_r\}$ (4″) $\overline{X}''_1 \subset B'' \cup \{e_1\}$.

Az (1′, 1″) tulajdonságok szerint $e_q \in \overline{X}''_1 \cap \overline{Y}'_r \neq \emptyset$, valamint (4′, 4″) és E^c definíciójának felhasználásával kapjuk, hogy $\overline{X}''_1 \cap \overline{Y}'_r \subset [B'' \cup \{e_1\}] \cap [(E \setminus B') \cup \{e_r\}] \subset$

$\subset E^c \cup \{e_1\}$. Így (3′, 3″) miatt $Y_r'^+ \cap X_1''^+ \subset \{e_1, e_q\}$ és $Y_r'^- \cap X_1''^- \subset \{e_1, e_q\}$, melyekből (1′, 1″, 2′, 2″) felhasználásával kapjuk, hogy $Y_r'^+ \cap X_1''^+ = \emptyset$ és $Y_r'^- \cap X_1''^- = \emptyset$, ami ellentmond Y_r' és X_1'' ortogonalitásának. Ez az eset sem lehetséges.

Mivel mind a négy lehetséges eset ellentmondásra vezetett, így beláttuk, hogy ciklizálás nem fordulhat elő, azaz tételünket bebizonyítottuk.

Tételünk közvetlen következménye az irányított matroidokon adott általános dualitás tétel (BLAND [2] 3.5. tétel), melyre így egy új algoritmikus bizonyítást nyertünk.

KÖVETKEZMÉNY. Legyenek $M=(E, O)$ és $M^*=(E, O^*)$ duális irányított matroidok. Legyen $e', e'' \in E$. Az alábbi alternatívák közül pontosan egy áll fenn.

(a) Létezik olyan $X \in O$, melyre $e' \notin \overline{X}$, $e'' \in X^+$ és $X^- = \emptyset$, vagy létezik olyan $Y \in O^*$, melyre $e' \in Y^+$, $e'' \notin \overline{Y}$ és $Y^- = \emptyset$.

(b) Létezik $X \in O$ és $Y \in O^*$, melyekre $e' \in \overline{X}^+$, $X^- \subset \{e''\}$, $e'' \in Y^+$, $Y^- \subset \{e'\}$ és $\overline{X} \cap \overline{Y} \subset \{e', e''\}$.

Bizonyítás. Az irányított ciklusok és kociklusok ortogonalitásából következik, hogy (a) és (b) nem állhat fenn egyidejűleg.

Ha $\{e''\} \in O$, akkor nyilvánvalóan (a) áll fenn.

Ha $\{e''\} \notin O$, akkor $e_1 = e'$ és $e_2 = e''$ választással alkalmazhatjuk a criss-cross módszert, melynek három kimenete adja állításunk bizonyítását, mivel a criss-cross módszer véges.

Megjegyezzük, hogy a fenti következmény szavakban kimondva azt jelenti, hogy ha primál megengedett irányított ciklus és duál megengedett irányított kociklus is létezik, akkor optimális irányított ciklus, illetve kociklus is létezik.

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 combinatorial abstraction of linear programming", *Journal of Combinatorial Theory (B)* **23** (1977) 33—57.

[3] BLAND, R. G., "New finite pivoting rules for the simplex method", *Mathematics of Operation Research* **2** (1977) 103—107.

[4] BLAND R. G. and M. LAS VERGNAS, "Orientability of matroids", *Journal of Combinatorial Theory (B)* **24** (1978) 94—123.

[5] LAWLER, E. L., *Kombinatorikus optimalizálás, hálózatok és matroidok* (Műszaki Könyvkiadó, Budapest, 1982).

[6] LOVÁSZ, L., „A matroidelmélet rövid áttekintése", *Matematikai Lapok* **22** (1971) 249—267.

[7] MINTY, G. J., "On the axiomatic foundations of the theories of directed linear graphs. Electrical networks and network programming", *Journal of Mathematics and Mechanics* **15** (1966) 385—520.

[8] PRÉKOPA, A., *Lineáris programozás I.* (Bolyai János Matematikai Társulat, Budapest, 1968).

[9] ROCKAFELLAR, R. T., "The elementary vectors of a subspace of R^N", in *Combinatorial Mathematics and its Applications*, Proceedings of the Chapel Hill Conference, 1967, eds. *R. G. Bore and T. A. Dowling*, University of North Carolina Press, 1969, 104—127.

[10] TERLAKY, T., „A criss-cross módszer lineáris programozási feladatok megoldására és végességének bizonyítása", *Alkalmazott Matematikai Lapok* **10** (1984) 289—296.

[11] TERLAKY, T., "A convergent criss-cross method", *Mathematics of Operationsforschung und Statistics, Ser. Optimization* **16** (1985) 683—690.

[12] TERLAKY, T., "A finite criss-cross method for oriented matroids", *Journal of Combinatorial Theory (B)*.

[13] TUTTE, W. T., "Lectures on matroids", *Journal of Research National Bureau Std. B.* **69** (1965) 1—47.

[14] WHITNEY, H., "On the abstract properties of linear dependence", *American Journal of Mathematics* 57 (1935), 509—533.
[15] ZIONTS, S., "The criss-cross method for solving linear programming problems", *Management Science* **15** (1969) 426—445.
[16] ZIONTS, S., "Some empirical tests of the criss-cross method", *Management Science* **19** (1972) 406—410.

(Beérkezett: 1985. március 1.)

TERLAKY TAMÁS
ELTE TTK OPERÁCIÓKUTATÁSI TANSZÉK
1088 BUDAPEST, MÚZEUM KRT. 6—8.

A FINITE CRISS-CROSS METHOD FOR ORIENTED MATROIDS

T. TERLAKY

Our paper presents a new finite criss-cross method for oriented matroids. Starting from a neither primal nor dual feasible tableau, we reach primal and dual optimal oriented circuits in finite number of steps, if they exist. If there is no optimal tableau then we show that there is no primal feasible oriented circuit or there is no dual feasible oriented cocircuit.

We also give a constructive proof for the generalization of *Minty's painting lemma.*

Alkalmazott Matematikai Lapok **11** (1985) 399—408

AZ EXTREMÁLIS SAJÁTVEKTOROK MEGHATÁROZÁSA ELIMINÁCIÓS MÓDSZERREL — AZ ÁLTALÁNOSÍTOTT WARSHALL ALGORITMUS EGY JAVÍTÁSA

HEGEDŰS GÁBOR

Budapest

A matematika gazdasági alkalmazása gyakran igényli az extremális sajátérték feladat megoldását (legrövidebb utak meghatározása, ütemezési feladatok stb.). Az ismert megoldó algoritmusok kétlépcsősek [5]: először meghatározzák a sajátértéket, majd az egység (additív esetben nulla) sajátértékűvé transzformált mátrix sajátvektorait. (Számos feladat csak az utóbbi lépést igényli.) Ha elő kell állítani az összes extremális sajátvektort, akkor egy műveletigény szempontjából igen jó megoldás az *általánosított Warshall algoritmus* [9], vagy valamilyen közvetett alkalmazása [7].

Az általános alakú extremális egyenletrendszerek [3]-ban közölt megoldó algoritmusa a lineáris egyenletrendszerek esetére közismert eliminációs módszer [2] megfelelője és rendelkezik az utóbbi hátrányos tulajdonságaival is: a generált vektorok száma rohamosan nőhet. Ha ezt az algoritmust a sajátvektor egyenletre alkalmazzuk, akkor a vektorok száma nem nő, sőt csökkenhet; a kapott algoritmus műveletigénye a legkedvezőtlenebb esetben eléri ugyan az *általánosított Warshall algoritmusét*, de tárolóigénye mindig kisebb.

1. Bevezetés

Az extremális algebrát e lap hasábjain ismertető [7] alkalmazásra vonatkozó példáinak zöme az extremális sajátérték feladathoz kapcsolódik (optimális állapot-sorozatok vagy legrövidebb utak meghatározása, beruházási háló kritikus útjainak megkeresése, munkafolyamatok ütemezése). Befejező része e feladat megoldásával foglalkozik; új algoritmusokat ismertet mind a sajátértékek, mind a sajátvektorok meghatározására. Az utóbbi esetében kulcsszerepet játszik a főátlójukban csak egységelemet (additív esetben nulla) tartalmazó stabilis mátrixok $(n-1)$-edik hatványának meghatározása (ami [7]-nek nem tárgya); amire az irodalom (pl. [5]) az *általánosított Warshall algoritmust* ajánlja.

[5] nyomán egy n-rendű négyzetes $\mathbf{A}=(\alpha_{ij})$ mátrixot — multiplikatív maximális tér ([7]) esetében — akkor nevezünk stabilisnak, ha

$$\lambda(\mathbf{A})=\bigvee_{k=1}^{n}\bigvee_{\substack{i_1,\ldots,i_k\\ i_{k+1}=i_1}}\sqrt[k]{\prod_{j=1}^{k}\alpha_{i_j i_{j+1}}}=1, \tag{1.1}$$

ahol $\bigvee$ a maximum operátort jelöli. Ha egy stabilis mátrix főátlójában csupa egységelem áll és $\mathbf{x}=(\xi_i)$, $\mathbf{y}=(\eta_i)$ esetén

$$\langle \mathbf{x}, \mathbf{y}\rangle = \bigvee_{i=1}^{n}\xi_i\eta_i, \tag{1.2}$$

akkor $r=n-1$-re $\mathbf{A}^{r+1}=\mathbf{A}^r$, ahol n a mátrix rendje. A stabilis mátrixok egy extremális sajátértéke az egységelem (additív esetben nulla), a hozzá tartozó sajátvektorok

az

$$\mathbf{Ax} = \mathbf{x} \tag{1.3}$$

egyenlet megoldásai.

Az (1.3) alakú egyenletek (skaláris felírásban egyenletrendszerek) az általános extremális egyenletrendszerek speciális esetei; az utóbbiak általános megoldásával foglalkozik [3]. A [3]-ban közölt algoritmus a lineáris egyenletrendszerek vagy egyenlőtlenségrendszerek (általában nemnegatív) megoldásainak meghatározására szolgáló eliminációs módszer (l. [8] vagy [2]) extremális algebrai megfelelője és létrejöttében kulcsszerepet játszott a [7]-ben bevezetett extremális előjel fogalma.

A cikkben a példákig az extremális algebra következő modelljét használjuk: skalárjaink nemnegatív valós számok (jelük görög betű), az extremális összeadás a maximum operáció ($\vee$), az extremális szorzás a közönséges szorzás, a mátrixok szorzása ezzel összhangban az (1.2) skalárszorzaton alapul.

2. A homogén extremális egyenletrendszerek megoldása

A [3]-ban közölt leírás elkerüli az extremális előjel használatát, megkönnyítve ezáltal a fogalmat nem ismerő olvasó dolgát, de vállalva bizonyos bonyodalmakat is. (A pozitív tagok az egyenlet egyik oldalán, a negatívak a másik oldalán, a neutrálisak pedig mindkét oldalán szerepelnek). A következő — csak a homogén esetre vonatkozó — ismertetés feltételezi az extremális előjeles számokkal végzett műveletek ismeretét, ezért először erről szólunk néhány szót.

Minden nem nulla szám három féle extremális előjel valamelyikével rendelkezhet: lehet pozitív (+), negatív (−) és neutrális (#); a nulla mindig neutrális ($0 = \#0$), a többi számot előjel nélkül pozitívnak tekintjük. Két, előjelétől ektekintve különböző, szám extremális összege a nagyobb szám eredeti előjelével. Előjelüktől eltekintve egyenlő számok extremális összegének numerikus része a közös numerikus rész, előjele pedig a közös előjel, ha a számok egyenlők és #, ha előjeleik különbözőek. Két szám extremális szorzatának numerikus része a tényezők numerikus részének szorzata. Előjele #, ha legalább egy tényezőjéé #, +, ha a két tényező előjele megegyezik és nem #, és végül −, ha különböznek és egyik sem #.

Az extremális előjellel rendelkező számoknak az extremális összeadással összhangban levő részbenrendezése szerint egy szám nagyobb egy másiknál, ha előjelüktől eltekintve is nagyobb, vagy ha azoktól eltekintve ugyanakkora ugyan, de előjele #, míg a másiké + vagy −. Az azonos numerikus értékhez tartozó pozitív és negatív számok nem összehasonlíthatók (összegük neutrális). Az így kiterjesztett számfogalom keretei között a signum függvény ugyanolyan kapcsolatban van az alapműveletekkel, mint a klasszikus esetben (értékét a nulla kiterjesztéseként felfogható neutrális számokon 0-nak vesszük): $\operatorname{signum}(\xi\eta) = \operatorname{signum}\xi \cdot \operatorname{signum}\eta$; $\xi > \eta$ esetén pedig $\operatorname{signum}(\xi \vee \eta) = \operatorname{signum}\xi$.

Néhány példa az elmondottakra minden további megjegyzés nélkül:

$$+1 \vee -2 = -2, \qquad \#1 \vee -3 = -3,$$

$$+1 \vee +1 = +1, \qquad \#2 \vee \#2 = \#2, \qquad +1 \vee -1 = +1 \vee \#1 = \#1,$$

$$+2 \cdot +2 = +4, \qquad -2 \cdot -2 = +4, \qquad -2 \cdot +3 = -6,$$

$$-2 \cdot \#3 = \#6; \qquad +1 < -2, \qquad -2 < \#2, \qquad +1 \| -1.$$

Az extremális előjel alkalmazása lehetővé teszi számos az euklideszi terekben használatos fogalom extremális algebrai megfelelőjének definiálását, analóg tételek bizonyítását. [7]-ben bőven találunk erre példát; itt csak néhányra lesz szükségünk.

2.1. *Definíció.* A C halmaz kúp, ha $c_i \in C$ és pozitív vagy 0 α_i, $i=1,2$ esetén $\alpha_1 c_1 \vee \alpha_2 c_2 \in C$.

2.2. *Definíció.* Az α_i, $i=0, \ldots, n$ paraméterekhez tartozó sík (pozitív, illetve negatív féltér) egyenlete:

$$\text{signum}\left(\bigvee_{i=1}^{n} \alpha_i \xi_i \vee \alpha_0\right) = 0 \quad (=+1, \text{ ill. } =-1). \tag{2.1}$$

Az $\mathbf{x}=(\xi_i)$ vektor akkor és csak akkor fekszik a síkban (a megfelelő féltérben), ha kielégíti a (2.1) egyenletet.

Számunkra csak a valós (extremális értelemben pozitív és nulla komponensekből álló) vektorok érdekesek, paramétereink azonban előjelesek lesznek.

Most röviden ismertetjük a homogén lineáris egyenletrendszerek nemnegatív megoldásainak explicit előállítására szolgáló eliminációs módszert. Legyen $\mathbf{A}=(\alpha_{ij})$ $i=1, \ldots, m$, $j=1, \ldots, n$ valós mátrix; keressük azt a $\mathbf{H}=(\eta_{jk})$ $j=1, \ldots, n$, $k=1, \ldots, K$ nemnegatív mátrixot, melynek minden oszlopa megoldása az

$$\mathbf{Ax} = \mathbf{0} \tag{2.2}$$

egyenletrendszernek és rendelkezik azzal a tulajdonsággal, hogy (2.2) minden megoldása előállítható $\mathbf{x}=\sum_{k=1}^{K} \beta_k \mathbf{h}_k$ alakban, ahol $\beta_k \geqq 0$ és $\mathbf{h}_k$ a $\mathbf{H}$ mátrix k-adik oszlopa.

A megoldás a következő (l. pl. [8] és [2]). A nemnegatív vektorok halmaza egy origó csúcspontú kúp. A (2.2) egyenletrendszer minden sora egy-egy origón átmenő sík egyenlete; egy ilyen sík és egy origó csúcspontú kúp metszete ismét egy origó csúcspontú kúp. Az összes nemnegatív vektort tartalmazó kúpból indulva, sorra képezve a metszeteket a (2.2) egyenlet soraival, végül a keresett kúphoz jutunk.

Az induló kúp tehát az n-dimenziós egységvektorok nemnegatív kombinációiból áll. Ha vizsgáljuk egy $\mathbf{p}_1, \ldots, \mathbf{p}_k$ vektorokkal kifeszített kúp és egy $\mathbf{ax}=0$ egyenletű sík metszetét, bizonyítható, hogy azt kifeszítik a következő vektorok:

$$\{\mathbf{p}_i;\ i \in I\} \cup \{\beta_i \mathbf{p}_j + \beta_j \mathbf{p}_i;\ i \in L,\ j \in J\},$$

ahol $\mathbf{ap}_i=0$ $i \in I$ (a síkba eső vektorok), $\mathbf{ap}_j>0$ $j \in J$ (a pozitív féltérbe eső vektorok) és $\mathbf{ap}_i<0$ $i \in L$ (a negatív féltérbe eső vektorok); továbbá $\beta_j=|\mathbf{ap}_j|$ $j \in J \cup L$. Ha ilyen módon határozzuk meg az előbb leírt kúp-sorozat elemeit, akkor a (2.2) egyenlet minden sorának számításba vétele után a kifeszítő vektorok éppen egy kívánt tulajdonságú $\mathbf{H}$ mátrix oszlopai. Ez a mátrix általában sok felesleges oszlopot is tartalmaz, de ha minden lépésben elhagyjuk a felesleges vektorokat [2], akkor is a gyakorlati alkalmazást lényegében kizáró mennyiségű oszlop marad. (Közvetett alkalmazásai megtalálhatók különféle dekompozíciós módszerekben.)

Térjünk át az extremális eset vizsgálatára! Keressük a

$$\text{signum}\ (\mathbf{a}_i \mathbf{x}) = 0, \quad i = 1, \ldots, m \tag{2.3}$$

egyenletrendszer csak pozitív és nulla komponenseket tartalmazó $\mathbf{x}=(\xi_j)$ $j=1,\ldots,n$ megoldásait.

A kúp és sík metszetéről lineáris esetben állítottak igazak extremális esetben is [7], bonyolultabb azonban a metszetet kifeszítő vektorok rendszere [3]. Ha a $\mathbf{p}_1,\ldots,\mathbf{p}_k$ vektorokkal kifeszített kúpot metszi a signum $(\mathbf{ax})=0$ egyenletű sík, akkor a metszet kúpot a következő vektorok feszítik ki:

$$(2.4)\quad \{\mathbf{p}_i;\ i\in I\cup N\}\cup\{\beta_i\mathbf{p}_j\vee\beta_j\mathbf{p}_i;\ i\in L,\ j\in J\}\cup\{\beta_i\mathbf{p}_j\vee\beta_j\mathbf{p}_i;\ i\in J\cup L,\ j\in N\}$$

ahol $I=\{i;\ \mathbf{ap}_i=0\}$, $N=\{i;\ \mathbf{ap}_i\neq 0,\ \text{signum}\,(\mathbf{ap}_i)=0\}$, $J=\{i;\ \text{signum}\,(\mathbf{ap}_i)=1\}$, $L=\{i;\ \text{signum}\,(\mathbf{ap}_i)=-1\}$ és $\beta_i=|\mathbf{ap}_i|$ ($|\alpha|$ jelöli az előjelétől megfosztott α-t, ami pozitív, ha α nem nulla, de $|0|=0=\#0$).

A fentiek bizonyítása nem túl bonyolult, [3]-ban megtalálható. Az újabb vektorcsoport megjelenése a metszetet kifeszítő vektorok között arra utal, hogy extremális esetben a helyzet rosszabb, mint a szinte kilátástalannak ítélt lineáris esetben. A következő fejezet — egy életképes alkalmazást bemutatva — ezt az utalást cáfolja.

3. Az extremális sajátvektorok meghatározása

Foglalkozzunk az (1.3) egyenlet megoldásával először csak abban az esetben, ha az A mátrix stabilitásán (1.1) kívül az $\alpha_{ii}=1$, $i=1,\ldots,n$ feltétel is teljesül! Ennek az esetnek a kezelésére alkalmas az $\mathbf{A}^{n-1}$-t szolgáltató *általánosított Warshall algoritmus* [9]:

Határozzuk meg az $\mathbf{A}^{(k)}=(\alpha_{ij}^{(k)})$, $k=0,\ldots,n$ mátrixsorozatot

$$\mathbf{A}^{(0)}=\mathbf{A}$$

$$(3.1)\quad \alpha_{ij}^{(k+1)}=\alpha_{ij}^{(k)}\vee\alpha_{ik+1}^{(k)}\alpha_{k+1j}^{(k)}\quad(\alpha_{ii}^{(k+1)}=1\ \text{marad}).$$

Az (1.3) egyenlet összes megoldását szolgáltató $\mathbf{H}$ mátrix a sorozat záró eleme ($\mathbf{H}=\mathbf{A}^{n-1}=\mathbf{A}^{(n)}$).

Alkalmazzuk ugyanerre a feladatra a homogén extremális egyenletrendszerek eliminációs megoldó algoritmusát! A (2.3) egyenletben szereplő $\mathbf{a}_i$ vektor az A mátrix i-edik sorától annyiban tér el, hogy a főátlóbeli elemnek megfelelő helyen 1 helyett $\#1$-t tartalmaz ($\bigvee_{j=1}^{n}\alpha_{ij}\xi_j=\xi_i$ helyett a vele egyenértékű

$$\text{signum}\Big(\bigvee_{\substack{j=1\\ j\neq i}}^{n}\alpha_{ij}\xi_j\vee(+1\vee-1)\xi_i\Big)=0$$

a követelmény).

Jelölje $\mathbf{H}^{(k)}=(\eta_{ij}^{(k)})$ az első k feltételt kielégítő vektorok kúpját generáló mátrixot. (Ezzel összhangban $\mathbf{H}^{(0)}$ az n-rendű egységmátrix). Nem nehéz belátni, hogy a két

algoritmus között — megfelelő indexeléssel — fennáll a következő kapcsolat:

$$(\mathbf{a}_k\mathbf{H}^{(k-1)})_j = \begin{cases} \alpha_{kj}^{(k-1)}, & \text{ha } k \neq j \\ \#1, & \text{ha } k \neq j, \end{cases} \tag{3.2}$$

$$\eta_{ij}^{(k)} = \begin{cases} \alpha_{ij}^{(k)}, & \text{ha } i \leqq k \\ \delta_{ij}, & \text{ha } i > k \ (\textit{Kronecker } \delta) \end{cases} \tag{3.3}$$

így $\eta_{kj}^{(k)} = (\mathbf{a}_k\mathbf{H}^{(k-1)})_j$, ha $k \neq j$, mert (3.1) szerint $\alpha_{kj}^{(k)} = \alpha_{kj}^{(k-1)} \vee \alpha_{kk}^{(k-1)}\alpha_{kj}^{(k-1)} = \alpha_{kj}^{(k-1)}$.

Valóban, $\mathbf{H}^{(0)}$ egységmátrix ((3.3) $k=0$ esetén), ami $\mathbf{a}_1$-gyel balról szorozva ismét $\mathbf{a}_1$ ((3.2) $k=1$ esetén). A bizonyítás induktív folytatása érdekében vizsgáljuk meg, hogyan keletkezik $\mathbf{H}^{(k-1)}$-ből $\mathbf{H}^{(k)}$. (Az $\mathbf{a}$ vektor szerepében $\mathbf{a}_k$, a $\mathbf{p}_i$ vektorok szerepében $\mathbf{H}^{(k-1)}$ oszlopai állnak!) Induktív feltevésünk (3.3) $k-1$-re és (3.2) k-ra. (3.2) miatt — (2.4)-gyel összhangban — $L=\emptyset$ és $N=\{k\}$, így az új vektorok az $I \cup N$-beli indexekhez és az $N \times J$-beli párokhoz tartoznak. N egyelemű, ezért az utóbbiak átvehetik a J-beli indexet; az új mátrix is n oszlopot fog tartalmazni.

$$\mathbf{H}^{(k-1)} = \begin{pmatrix} 1 & \eta_{12}^{k-1} & \dots & \eta_{1k}^{k-1} & \dots & \eta_{1n}^{k-1} \\ \eta_{21}^{k-1} & 1 & \dots & \eta_{2k}^{k-1} & \dots & \eta_{2n}^{k-1} \\ \vdots & \vdots & & \vdots & & \vdots \\ \eta_{k-1,1}^{k-1} & \eta_{k-1,2}^{k-1} & \dots & \eta_{k-1,k}^{k-1} & \dots & \eta_{k-1,n}^{k-1} \\ 0 & 0 & \dots & 1 & \dots & 0 \\ \vdots & \vdots & & \vdots & & \vdots \\ 0 & 0 & \dots & 0 & \dots & 1 \end{pmatrix}$$

$$\mathbf{a}_k\mathbf{H}^{(k-1)} = (\eta_{k1}^{(k)} \quad \eta_{k2}^{(k)} \quad \dots \quad \#1 \quad \dots \quad \eta_{kn}^{(k)}).$$

Számítsuk ki $\mathbf{H}^{(k)}$ j-edik oszlopát! Ha $j=k$, akkor ez az oszlop megegyezik $\mathbf{H}^{(k-1)}$ j-edik oszlopával:

— $i=k$ esetén $\eta_{kk}^{(k)}=1$, de (3.1) szerint is $\alpha_{kk}^{(k)}=1$;

— $i<k$ esetén $\eta_{ik}^{(k)}=\eta_{ik}^{(k-1)}$, de (3.1) szerint is $\alpha_{ik}^{(k)}=\alpha_{ik}^{(k-1)} \vee \alpha_{ik}^{(k-1)}\alpha_{kk}^{(k-1)}=\alpha_{ik}^{(k-1)}$.

Ha $j \in I$, azaz $\alpha_{kj}^{(k-1)}=0$, akkor szintén marad az eredeti vektor, $\eta_{ij}^{(k)}=\eta_{kj}^{(k-1)}$, $i=1, \dots, n$, de $i=k$ esetén (3.1) szerint is $\alpha_{ij}^{(k)}=\alpha_{ij}^{(k-1)} \vee \alpha_{ik}^{(k-1)}\alpha_{kj}^{(k-1)}=\alpha_{ij}^{(k-1)}$.

Ha $j \in J$, akkor (2.4) és az indukciós feltevés szerint $\eta_{ij}^{(k)}=\eta_{ij}^{(k-1)} \vee \eta_{ik}^{(k-1)}\alpha_{kj}^{(k-1)}$, ami $i=k$ esetén megegyezik a (3.1) formulával, $i>k$ esetén pedig $\eta_{ik}^{(k-1)}=0$, ezért $\eta_{ij}^{(k)}=\eta_{ij}^{(k-1)}$.

A két algoritmus közötti (3.2) és (3.3)-ban megfogalmazott összefüggés bizonyításából hátra van még (3.2) igazolása, feltéve hogy (3.3) $k-1$-re teljesül.

$$(\mathbf{a}_k\mathbf{H}^{(k-1)})_j = \bigvee_{i=1}^{k-1} \alpha_{ki}\eta_{ij}^{(k-1)} \vee \delta_{kj}\alpha_{kj} \vee \#\delta_{kj},$$

ahol δ_{kj} akkor 1, ha δ_{kj} nulla és viszont. Azt kell megmutatnunk, hogy $\alpha_{ki}\alpha_{ik}^{(k-1)} \leqq 1$ és $\alpha_{kj}^{(k)}=\alpha_{kj}^{(k-1)}=\bigvee_{i=1}^{k-1} \alpha_{ki}\alpha_{ij}^{(k-1)} \vee \alpha_{kj}$. (Utóbbi elég volna $j \neq k$-ra; mindkét állítás a (3.1) algoritmus egy-egy tulajdonságát fogalmazza meg).

(3.1) alapján nyilvánvaló, hogy $\alpha_{ij}^{(k)} = \prod_{h=1}^{s_{ijk}} \alpha_{i_h i_{h+1}}$ egy megfelelően választott $s_{ijk}+1$ elemű $\{i_h\}$ indexsorozattal, melyben $i_1 = i$ és $i_{s_{ijk}+1} = j$. Ugyanakkor $\mathbf{A}$ stabilitása (1.1) miatt $j=k$ esetén teljesül $\alpha_{ki} \prod_{h=1}^{s_{ikk}} \alpha_{i_h i_{h+1}} \leqq 1$. (Ha $s_{ikk} \geqq n$, akkor a tényezők indexei között ismétlődésnek kell lennie és a szomszédos ismétlődések közötti részletszorzatok felső korlátja 1.)

A másik összefüggést — némileg általánosítva — teljes indukcióval bizonyítjuk. Állításunk: $\alpha_{ij}^{(k)} = \bigvee_{h=1}^{k} \alpha_{ih}\alpha_{hj}^{(k)} \vee \alpha_{ij}$.

$k=0$ esetén az egyenlőség mindkét oldalán α_{ij} áll. Tegyük fel, hogy $k-1$-re fennáll az egyenlőség. (3.1) és az indukciós feltevés alapján

$$\alpha_{ij}^{(k)} = \bigvee_{h=1}^{k-1} \alpha_{ih}\alpha_{hj}^{(k-1)} \vee \alpha_{ij} \vee \Big(\bigvee_{h=1}^{k-1} \alpha_{ih}\alpha_{hk}^{(k-1)} \vee \alpha_{ik}\Big)\alpha_{kj}^{(k-1)} =$$

$$= \bigvee_{h=1}^{k-1} \alpha_{ih}(\alpha_{hj}^{(k-1)} \vee \alpha_{hk}^{(k-1)}\alpha_{kj}^{(k-1)}) \vee \alpha_{ik}\alpha_{kj}^{(k-1)} \vee \alpha_{ij}.$$

Ismét (3.1)-et és $\alpha_{kj}^{(k)} = \alpha_{kj}^{(k-1)}$-t felhasználva azt kapjuk, hogy

$$\alpha_{ij}^{(k)} = \bigvee_{h=1}^{k-1} \alpha_{ih}\alpha_{hj}^{(k)} \vee \alpha_{ik}\alpha_{kj}^{(k)} \vee \alpha_{ij} = \bigvee_{h=1}^{k} \alpha_{ih}\alpha_{hj}^{(k)} \vee \alpha_{ij}.$$

Megmutattuk, hogy stabilis és főátlójukban csak egységelemet tartalmazó mátrixok esetén az (1.3)-mal egyenértékű (2.3) egyenletrendszer eliminációs módszerrel történő megoldásának k-adik iterációjában tulajdonképpen az *általánosított Warshall algoritmus* ugyancsak k-adik iterációjának eredményeként előálló mátrix első k sorát határozzuk meg. Következésképpen végeredményben ezzel a módszerrel is az $\mathbf{A}^{n-1}$ extremális hatvány mátrixhoz jutunk. (A leírt algoritmus formális „transzponáltja" annak, amit [4] a *Jordan elimináció* analogonjaként közöl.)

A sajátvektorok eliminációs módszerrel történő meghatározása n lépésből áll; a k-adik lépésben egy $n \times n$-s mátrix első k sorának elemeit kell kiszámolni ($\mathbf{a}_k\mathbf{H}^{(k-1)}$ lényegében megegyezik $\mathbf{H}^{(k)}$ k-adik sorával). Ha az I indexhalmaz (2.4) minden lépésben üres, akkor $\mathbf{a}_k\mathbf{H}^{(k-1)}$ egy elemének meghatározása $k-1$ műveletet (szorzás — összehasonlítás párt) igényel és $n-1$ ilyen elemet kell meghatározni; $\mathbf{H}^{(k)}$ első $k-1$ sorának elemeit további egy-egy művelettel kapjuk meg és $n-1$ oszlop elemeit kell így meghatározni. Ezt a $2(n-1)(k-1)$ műveletet k szerint 1-től n-ig összegezve $n(n-1)^2$-et kapunk, ami az algoritmus teljes műveletigénye. Éppen ennyi műveletet igényel az *általánosított Warshall algoritmus* (3.1) is, ha ott se számoljuk ki feleslegesen a k-adik iterációban változatlanul maradó k-adik sort. (A $\mathbf{H}^{(k)}$ mátrixok teljes főátlója csak egységelemeket tartalmaz, ez kihasználható a számításigény további csökkentésére, de ugyanez áll az $\mathbf{A}^{(k)}$ mátrixokra is.)

A $\mathbf{H}^{(k)}$ mátrix első k oszlopa közül egyesek kifejezhetők lehetnek a többiek pozitív kombinációjaként. Ha csak az extremális sajátvektorokra van szükségünk és nem a teljes hatványmátrixra, akkor ezek elhagyhatók, így elvileg csökken a számítás-

igény. Valójában az ilyen lehetőségek felderítésének nehézségei [5] kizárják ezt a csökkentést. Kézi számolásnál ugyanakkor könnyű észrevenni, hogy két oszlop azonos.

Térjünk át az általános eset vizsgálatára, azaz vessük el az $\alpha_{ii}=1$, $i=1, \ldots, n$ feltételt! Az (1.3) egyenlet megoldásait szolgáltató $\mathbf{H}$ mátrix ekkor a $\bigvee_{k=1}^{n} \mathbf{A}^k$ hatványmátrix összegből meghatározható [5]. $\bigvee_{k=1}^{n} \mathbf{A}^k = \mathbf{A}\left(\bigvee_{k=0}^{n-1} \mathbf{A}^k\right)$, ahol $\mathbf{A}^0$ az n-rendű egységmátrix. (Jele $\mathbf{E}$, additív esetben főátlójának elemei nullák, többi eleme $-\infty$). Ugyanakkor $(\mathbf{E} \vee \mathbf{A})^{n-1} = \bigvee_{k=0}^{n-1} \mathbf{A}^k$. [5] javaslata az, hogy határozzuk meg a főátlójában csak egységelemeket tartalmazó $\mathbf{E} \vee \mathbf{A}$ mátrix $(n-1)$-edik hatványát pl. az *általánosított Warshall algoritmussal,* majd a kapott mátrixot szorozzuk $\mathbf{A}$-val. [7] javaslata feltételezi, hogy ismert az (1.1)-ben szereplő maximális tagok indexhalmazainak egyesítése. (Egyébként algoritmust is ad meghatározására). Ennek az ismeretnek a birtokában mutat lehetőséget arra, hogy a feladatot kisebb mátrixok hatványozásával oldjuk meg. (Akkor igazán gazdaságos [7] módszere, ha az extremális sajátvektorok meghatározását meg kell előznie a sajátérték meghatározásának is, mert akkor az említett indexhalmaz természetes módon adódik.)

Az eliminációs módszer alkalmazása esetén annyi a változás az előző esethez képest, hogy a (2.3) egyenletben szereplő $\mathbf{a}_k$ vektoroknak az $\mathbf{A}$ mátrix k-adik sorától való eltérése csak akkor lesz egy $\neq 1$ megjelenése a főátlóbeli elemnek megfelelő helyen, ha ott 1 szerepel, egyébként -1 kerül ugyanoda. Ha egy ilyen k-adik komponensében -1 $\mathbf{a}_k$ vektorral szorozzuk a $\mathbf{H}^{(k-1)}$ mátrixot és az első $k-1$ sor alapján kapott érték, amit -1-gyel extremálisan összeadunk, kisebb 1-nél, akkor $N \neq \emptyset$ és $L=\{k\}$ adódhat. Ez azt eredményezi, hogy $\mathbf{H}^{(k-1)}$ k-adik oszlopa nem kerül bele a $\mathbf{H}^k$ mátrixba, így tovább se számolunk vele, egyébként minden marad a régiben. A csökkenés lehetőségének az az ára, hogy $\mathbf{a}_k\mathbf{H}^{(k-1)}$ k-adik komponensét is meg kell határozni minden olyan k-ra, melyre $\alpha_{kk} \neq 1$.

Ha $\alpha_{ii} \neq 1$ valamely i-re, akkor ezt az indexet az első helyre permutálva legalább egy oszlop végig hiányozni fog a $\mathbf{H}^{(k)}$ mátrixokból ($\mathbf{a}_1\mathbf{H}^{(0)}=\mathbf{a}_1$; $\alpha_{11}=-1$). Ilyenkor a helyzet kedvezőbb, mint [5] algoritmusának alkalmazása esetén, de az eliminációs módszer akkor igazán előnyös, amikor több oszlop esik ki, mert az a műveletszám radikális csökkenésével járhat. (Az elképzelhető legkedvezőbb esetben minden iterációban eggyel csökken az oszlopok száma és az utolsó iteráció előtt megkapjuk a sajátvektorok kúpját egyedül kifeszítő vektort, aminek létezését már [1] bizonyította. $\mathbf{a}_k\mathbf{H}^{(k-1)}$ meghatározása ekkor $(k-1)(n-k+1)$ műveletet igényel, $\mathbf{H}^{(k)}$ első $k-1$ soráé $(k-1)(n-k)$-t, azaz összesen $\frac{1}{6}(n-1)(2n^2-n-6)$ műveletre van csak szükség). Erre a lehetőségre következtethetünk esetleg a feladat természetéből, vagy a stabilitást definiáló (1.1) egyenlet maximális tagjait meghatározó indexsorozatokból ([1] ennek alapján vezeti vissza az általános esetet az $\alpha_{ii}=1$, $i=1, \ldots, n$ esetre).

A példák előtt foglalkozzunk még azzal a kérdéssel, hogy mikor célszerű az eliminációs módszert alkalmazni. Az *általánosított Warshall algoritmus* helyett mindig használható, mert műveletigénye akkor se nagyobb, ha a teljes $\mathbf{A}^{n-1}$ mátrixot elő kell állítani. (Például azért, mert nem érdemes meghatározni a kihagyható oszlopokat.) Ha egy stabilis, de főátlójában nem csak egységelemeket tartalmazó mátrix sajátvektorait kell meghatározni, akkor a feladat visszavezetése a módosított mátrix $(n-1)$-

edik hatványának meghatározására biztosan kevésbé gazdaságos, sőt — ha nem volt előzőleg szükség a sajátérték meghatározására a stabilis mátrix létrehozása érdekében — az elimináció előnyösebb a [7]-ben ajánlott sémánál is.

Nem mellékes körülmény az sem, hogy helyigénye egy mátrixnyi, eredménye létrehozható az induló mátrix helyén (a *Warshall algoritmus* hátrányaként szokták emlegetni nagy helyigényét [6]).

4. Példák

Az első példa az additív minimális térben egy 3×3-as, főátlójában egységelemeket tartalmazó stabilis A mátrix négyzetének meghatározása, azaz három pont páronkénti irányított távolságának kiszámítása az A mátrixba foglalt közvetlen távolságok alapján. (Extremális összeadás a minimum operáció, extremális szorzás a hagyományos összeadás; egységelem a valós nulla, nulla elem a $+\infty$.) Az adódó mátrixok és vektorok rendre

$$\mathbf{A}=\begin{pmatrix}0&2&4\\3&0&1\\1&1&0\end{pmatrix}\quad \mathbf{H}^{(0)}=\begin{pmatrix}0&\infty&\infty\\\infty&0&\infty\\\infty&\infty&0\end{pmatrix}\quad \mathbf{H}^{(2)}=\begin{pmatrix}0&2&3\\3&0&1\\\infty&\infty&0\end{pmatrix}$$

$$\mathbf{a}_1\mathbf{H}^{(0)}=(\#0\ \ 2\ \ 4)\qquad \mathbf{a}_3\mathbf{H}^{(2)}=(1\ \ 1\ \ \#0)$$

$$\mathbf{H}^{(1)}=\begin{pmatrix}0&2&4\\\infty&0&\infty\\\infty&\infty&0\end{pmatrix}\quad \mathbf{H}^{(3)}=\begin{pmatrix}0&2&3\\2&0&1\\1&1&0\end{pmatrix}$$

$$\mathbf{a}_2\mathbf{H}^{(1)}=(3\ \ \#0\ \ 1)\qquad \mathbf{A}^2=\mathbf{H}^{(3)}$$

A második példa az additív maximális térben egy 6×6-os stabilis A mátrix sajátvektorainak meghatározása. (Extremális összeadás a maximum operáció, extremális szorzás a hagyományos összeadás; egységelem a valós nulla, nulla elem a $-\infty$. Az extremális „előjelet" — hogy a hagyományos előjeltől elkülönöljön — a szám után írjuk.) Ez a példa a [7]-ben szereplő ütemezési feladat megoldásának utolsó lépése (az ott szereplő adatokkal, de eliminációs módszerrel). A soronlevő iterációt igyekszünk mindig az A mátrix olyan sorával végrehajtani, amellyel esélyünk van az oszlopok számának csökkenésére. A következő mátrixok és vektorok adódnak:

$$\mathbf{A}=\begin{pmatrix}0&0&-\infty&-\infty&-\infty&-\infty\\0&-3&-\infty&-\infty&-\infty&-\infty\\-\infty&4&0&1&1&3\\-\infty&3&-\infty&-2&-4&1\\-\infty&-1&-\infty&-\infty&-2&2\\1&4&-\infty&-\infty&-3&-1\end{pmatrix}$$

$$\mathbf{H}^{(0)}=\begin{pmatrix}0&-\infty&-\infty&-\infty&-\infty&-\infty\\-\infty&0&-\infty&-\infty&-\infty&-\infty\\-\infty&-\infty&0&-\infty&-\infty&-\infty\\-\infty&-\infty&-\infty&0&-\infty&-\infty\\-\infty&-\infty&-\infty&-\infty&0&-\infty\\-\infty&-\infty&-\infty&-\infty&-\infty&0\end{pmatrix}$$

$$\mathbf{a}_6\mathbf{H}^{(0)} = (\ 1 \quad 4 \quad -\infty \quad -\infty \quad -3 \quad 0-)$$

$$\mathbf{H}^{(1)} = \begin{pmatrix} 0 & -\infty & -\infty & -\infty & -\infty \\ -\infty & 0 & -\infty & -\infty & -\infty \\ -\infty & -\infty & 0 & -\infty & -\infty \\ -\infty & -\infty & -\infty & 0 & -\infty \\ -\infty & -\infty & -\infty & -\infty & 0 \\ 1 & 4 & -\infty & -\infty & -3 \end{pmatrix} \qquad \mathbf{H}^{(4)} = \begin{pmatrix} 0 & -\infty \\ 0 & -\infty \\ -\infty & 0 \\ 5 & -\infty \\ 6 & -\infty \\ 4 & -\infty \end{pmatrix}$$

$$\mathbf{a}_5\mathbf{H}^{(1)} = (3 \quad 6 \quad -\infty \quad -\infty \quad 0-) \qquad \mathbf{a}_3\mathbf{H}^{(4)} = (7 \quad 0\#)$$

$$\mathbf{H}^{(2)} = \begin{pmatrix} 0 & -\infty & -\infty & -\infty \\ -\infty & 0 & -\infty & -\infty \\ -\infty & -\infty & 0 & -\infty \\ -\infty & -\infty & -\infty & 0 \\ 3 & 6 & -\infty & -\infty \\ 1 & 4 & -\infty & -\infty \end{pmatrix} \qquad \mathbf{H}^{(5)} = \begin{pmatrix} 0 & -\infty \\ 0 & -\infty \\ 7 & 0 \\ 5 & -\infty \\ 6 & -\infty \\ 4 & -\infty \end{pmatrix}$$

$$\mathbf{a}_4\mathbf{H}^{(2)} = (\ 1 \quad 5 \quad -\infty \quad 0-) \qquad \mathbf{a}_1\mathbf{H}^{(5)} = (0\# \quad -\infty)$$

$$\mathbf{H}^{3} = \begin{pmatrix} 0 & -\infty & -\infty \\ -\infty & 0 & -\infty \\ -\infty & -\infty & 0 \\ 1 & 5 & -\infty \\ 3 & 6 & -\infty \\ 1 & 4 & -\infty \end{pmatrix} \qquad \mathbf{H}^{(6)} = \begin{pmatrix} 0 & -\infty \\ 0 & -\infty \\ 7 & 0 \\ 5 & -\infty \\ 6 & -\infty \\ 4 & -\infty \end{pmatrix}.$$

$$\mathbf{a}_2\mathbf{H}^{(3)} = (\ 0 \quad 0- \quad -\infty)$$

Befejezésül példát mutatunk arra, hogy $\alpha_{ii} \neq 1$, $i = 1, \ldots, n$ esetén se feltétlenül csökken k növekedtével a $\mathbf{H}^{(k)}$ mátrixok oszlopainak száma a második iterációtól kezdve. Ebben a példában visszatérünk a multiplikatív maximális térhez

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix} \qquad \mathbf{H}^{(0)} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

$$\mathbf{a}_1\mathbf{H}^{(0)} = (-1 \quad 1 \quad 1 \quad 1)$$

$$\mathbf{H}^{(1)} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \qquad \mathbf{H}^{(3)} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}$$

$$\mathbf{a}_2\mathbf{H}^{(1)} = (\#1 \;\; 1 \;\; 1) \qquad \mathbf{a}_4\mathbf{H}^{(3)} = (1 \;\; 1 \;\; \#1)$$

$$\mathbf{H}^{(2)} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \qquad \mathbf{H}^{(4)} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}$$

$$\mathbf{a}_3\mathbf{H}^{(2)} = (1 \;\; \#1 \;\; 1)$$

Az első iterációban — szükségszerűen — $k \in L$ volt, a többiben $k \in N$ adódott, ezért az oszlopok száma nem csökkent automatikusan tovább. Ha a megoldás során felismerjük, hogy valamelyik oszlop benne van a többiek által kifeszített kúpban, akkor az a megoldás halmaz csökkenése nélkül elhagyható. (Példánkban $\mathbf{H}^{(3)}$ és $\mathbf{H}^{(4)}$ azonos oszlopai közül elég egyet megőrizni a következő iteráció illetve az eredmény számára.)

IRODALOM

[1] Воробьев, Н. Н., «Экстремальная алгебра положительных матриц», *Elektronische Informationsverarbeitung und Kybernetik* **3** (1967) 39—71.

[2] Черников, С. Н., *Линейные Неравенства* (Наука, 1968).

[3] BUTKOVIC, P., HEGEDŰS, G., "The elimination method for finding all salutions of the system of linear equations over an extremal algebra", *Ekonomicko-matematicky Obzor* **20** (1984).

[4] CARRÉ, B. A., "An algebra for network routing problems", *J. Inst. Math. Appl.* **7** (1971) 273—294.

[5] CUNINGHAME—GREEN, R., *Minimax Algebra* (Springer-Verlag, 1979).

[6] GALLO, G., PALLOTTINO, S., "A new algorithm to find the shortest paths between all pairs of nodes", *Discrete Applied Mathematics* **4** (1982) 23—35.

[7] HEGEDŰS, G., „Extremális Algebra", *Alkalmazott Matematikai Lapok* **8** (1982) 341—380.

[8] PRÉKOPA, A., *Lineáris Programozás I.* (Bolyai János Matematikai Társulat, 1968).

[9] ROBERT, P., FERLAND, J., "Généralisation de l'algorithme de Warshall", *Rev. Francaise Informat. Opérationelle* **2** (1968) 71—85.

(Beérkezett: 1984. május 9.)

HEGEDŰS GÁBOR
ÉPÍTÉSGAZDASÁGI ÉS SZERVEZÉSI INTÉZET
1027 BUDAPEST, CSALOGÁNY U. 11.

DETERMINATION OF THE EXTREMAL EIGENVECTORS BY AN ELIMINATION ALGORITHM — AN IMPROVEMENT OF THE GENERALIZED WARSHALL ALGORITHM

G. HEGEDŰS

A new elimination algorithm for finding all extremal (minimax) eigenvectors of matrices is presented. First it is used for finding the distances between all pairs of nodes of a directed network. From the point of the number of operations it has the same efficiency, as the generalized Warshall algorithm [9]. The matrix of the distances may be stored in the same place as the matrix of the given direct distances. In more general cases (i. e. the assumption $\alpha_{ii}=0$ is dropped) most of algorithms need more operations. Our new algorithm is an application of the elimination method for solving systems of extremal equations [3] and needs less operations and no more storage.

A kiadásért felelős az Akadémiai Kiadó és Nyomda főigazgatója
Műszaki szerkesztő: Sándor István
A kézirat a nyomdába érkezett: 1985. október 31. — Terjedelem: 15,05 (A/5 ív)
86-4242 — Szegedi Nyomda — F. v.: Surányi Tibor 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: 64 Ft

ISSN 0133—3399

TARTALOMJEGYZÉK

INDEX

ALKALMAZOTT MATEMATIKAI LAPOK

A MAGYAR TUDOMÁNYOS AKADÉMIA MATEMATIKAI ÉS FIZIKAI TUDOMÁNYOK OSZTÁLYÁNAK KÖZLEMÉNYEI

XI. KÖTET

AKADÉMIAI KIADÓ, BUDAPEST
1985

TARTALOMJEGYZÉK

INDEX

86-4242 — Szegedi Nyomda — F. v.: Surányi Tibor igazgató