Alkalmazott matematikai lapok

1978/1-2

A MAGYAR TUDOMÁNYOS AKADÉMIA
MATEMATIKAI ÉS FIZIKAI TUDOMÁNYOK
OSZTÁLYÁNAK KÖZLEMÉNYEI

4.

KÖTET

AKADÉMIAI KIADÓ, BUDAPEST

A MAGYAR TUDOMÁNYOS AKADÉMIA
MATEMATIKAI ÉS FIZIKAI TUDOMÁNYOK OSZTÁLYÁNAK
ALKALMAZOTT MATEMATIKAI LAPJA

IV. 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.

Kéziratok a következő címre küldendők:

Prékopa András, felelős szerkesztő
1502 Budapest XI., Kende u. 13—17.

Ugyanerre a címre küldendő minden szerkesztőségi levelezés.

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 84 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 **4** (1978) 1—25

GYAKORI VELESZÜLETETT RENDELLENESSÉGEK ÖRÖKLŐDÉSMENETÉNEK VIZSGÁLATA

TUSNÁDY GÁBOR, TELEGDI LÁSZLÓ ÉS CZEIZEL ENDRE
Budapest

A veleszületett rendellenességek itt ismertetésre kerülő matematikai-genetikai vizsgálata az ún. GAMT- *(Gaussian Additive Multifactorial Thereshold-)* modellen alapul. A GAMT-modellben egy-egy rendellenességet egy normális eloszlású valószínűségi változó jellemez, melyet hajlamnak nevezünk. d-ed fokú rokonok hajlamainak $h^2/2^d$ a korrelációs együtthatója, ahol h^2 az örökölhetőségi együttható. A maximum likelihood-módszer segítségével elvégezzük a h^2 becslését és a modellnek a genetikai családvizsgálatok adataihoz való illesztését.

1. Bevezetés

A vizsgálat célja

Ismert tény, hogy jelenleg *Magyarországon* a születések mintegy 6—7%-ában veleszületett rendellenességgel kell számolni, továbbá egy sor veleszületett rendellenesség több családtagnál, halmozottan fordulhat elő. Ezért is fontos ezen betegségek öröklődésének vizsgálata. Az egyes rendellenességek öröklődésmenetének leírására az ún. GAMT-modellt választjuk. Ez feltételezi, hogy a vizsgált rendellenességnek mindenkire nézve van egy valós számmal kifejezhető mértéke, vagyis a betegséghez hozzá van rendelve egy L háttérváltozó. (Szokás ezt hajlamnak is nevezni.) A modell szerint L standard normális eloszlású valószínűségi változó, amelyre

$$L = G+E,$$

ahol G az additív genetikai, E a környezeti hatást jelenti, G és E független, 0 várható értékű, h^2 ill. $(1-h^2)$ szórásnégyzetű, normális eloszlású valószínűségi változók, továbbá az, hogy valaki beteg, azt jelenti, hogy rendellenességének mértéke nagyobb egy — a vizsgált egyedet tartalmazó populációra jellemző — küszöbnél. (h^2 az örökölhetőség mértéke, közismertebb néven örökölhetőségi együttható.)

A GAMT-modellben a rokonok rendellenességének mértékei a rokonsági foktól függően korreláltak: d-ed fokú rokonok rendellenessége mértékeinek $h^2/2^d$ a korrelációs együtthatója. Ezt figyelembe véve vizsgálatunk célja különböző rokoncsoportokra ill. ilyenek bizonyos összességeire a h^2 becslése és a becslés 95%-os szintű konfidenciaintervallumba foglalása volt, továbbá annak ellenőrzése, hogy a különböző h^2-becslések eltérése szignifikáns-e vagy sem. Ez utóbbi a modell tesztelését jelentette. Azon rendellenességekre, amelyekre a modell illeszkedett a genetikai családvizsgálat adataihoz, a kapott h^2-értékek jól használhatók voltak a genetikai tanácsadásban.

Alapfogalmak

A fogamzást követően a születésig (esetleg ezt követően bizonyos időn belül) strukturális, funkcionális és/vagy biokémiai fejlődési abnormitások alakulhatnak ki az embrióban (esetleg az újszülöttben). Az így létrejött fejlődési zavar következményét veleszületett rendellenességnek nevezzük.

A fejlődési zavarok megnyilvánulásai közül a morfológiai-strukturális jellegűeket ismerjük legrégebben és éppen ezért a legalaposabban. Az ún. *dysmorphogenesis* (SMITH, [5]) következménye a veleszületett fejlődési rendellenesség (*congenital malformation*, a továbbiakban CM).

A CM-eket súlyosságuk, gyakoriságuk, időbeni kialakulásuk, kóreredetük stb. alapján értékelhetjük:

(a) *Súlyosságuk* szerint major és minor CM-eket különíthetünk el. A *major* CM-ek az egyén egészsége, sőt, sokszor az élete védelmében orvosi beavatkozást igényelnek, esetleg halálosak. A *minor* CM-ek nincsenek befolyással az egyén életére és egészségére. Elkülönítésük a normál variációktól időnként nehézségbe ütközik.

(b) *Gyakoriságuk* alapján a CM-ek három csoportjáról beszélhetünk. A gyakori *(common)* CM-ek születéskori gyakorisága meghaladja az 1‰-et. A ritka CM-ek előfordulása nem éri el az 1/100 000 születésre eső gyakoriságot. A kettő között foglalnak helyet a közepes gyakoriságú CM-ek.

(c) Megjelenésük alapján elsősorban *izolált* és *multiplex* CM-eket különböztethetünk meg. Multiplex rendellenesség esetén az újszülöttben kettő vagy több független CM észlelhető.

Munkánk során a következő kilenc *major* CM-et vizsgáltuk: *anencephalia-spina bifida, ajakhasadék* szájpadhasadékkal vagy anélkül, veleszületett *hypertrophiás pylorus stenosis*, kamrai *septum defectus*, veleszületett *csípőficam*, veleszületett strukturális *dongaláb*, veleszületett *inguinalis sérv, hypospadiasis*, rejtett heréjűség. Valamennyi közülük *gyakori izolált* CM (ezentúl CICM).

A veleszületett rendellenességek vizsgálatának társadalmi jelentősége

A társadalmi haladással és az orvosi ellátás színvonalának javulásával párhuzamosan a legfőbb halálokok sorrendjében jelentős átrendeződés figyelhető meg. Ezen belül az egyik legjellemzőbb tendencia a CM-ek okozta halálozás előretörése. *Magyarországon* a hetvenes években a CM-ek már a tíz legfőbb halálok között találhatók. Még szembetűnőbb a halálokok átrendeződése a csecsemőhalálozáson belül: *Magyarországon* 1931 óta a CM-ek *részesedése* mintegy megtízszereződött.

A CM-ek abszolút gyakorisága ugyan nem mutat emelkedő tendenciát, a csökkenő korspecifikus halálozási arányszámokon belüli relatív gyakoriság-növekedésük azonban közegészségügyi és társadalmi jelentőségük fokozódását bizonyítja, fokozott aktivitásra sarkallva a betegellátásban és főleg a megelőzésben.

A mortalitási adatokhoz hasonló tendenciák figyelhetők meg a morbiditási statisztikákban is. Így pl. a gyermekosztályok beteganyagában egyre nagyobb hányadot képviselnek a CM-ek (WARKANY, [7]; MCCROY, [3]; ROBERTS et al., [4]; DAY és HOLMES, [1]). Hazánkban a vizsgált kilenc CICM a nyilvántartott össz-CM születéskori gyakoriság *51,5*%-át képezi. Valódi gyakoriságukat figyelembe véve részesedésük még magasabb is lehet, szemléletesen érzékeltetve a CICM-ek közegészségügyi jelentőségét.

2. Mérhető mennyiségek öröklődése

Az átörökítési érték

Munkánk elsődleges célja a CICM-ek kóreredetének vizsgálata volt a magyarországi genetikai családvizsgálatok alapján. Erre a célra a multifaktoriális modellt használtuk, amely lehetővé teszi a genetikai és környezeti hatások együttes figyelembevételét. A modell leírását a poligén rendszerek szabályszerűségeinek összefoglalásával kezdjük.

Akármilyen mérhető tulajdonságot választunk is, közvetlenül csak a fenotípusos *(phenotypical)* értéket tudjuk mérni. Jelöljük a kapott eredményt P-vel. Ez két nagy hatás eredője: a genotípusé és az erre rakódó környezeti hatásé. Jelöljük a vizsgált tulajdonságra ható *locus*-okon (génhelyeken) elhelyezkedő génpárok összességét Γ-val, a locusok számát S-sel. Γ-nak tehát $2S$ komponense van, jelöljük ezeket Γ_{ij}-vel, ahol i értéke 1-től S-ig fut és a különböző locusokat jelöli, j pedig 1 vagy 2 lehet. $j=1$ az egyed létrehozásában részt vevő anyai *(maternális)*, $j=2$ az apai *(faternális)* kromoszóma megfelelő locuson álló génjét jelöli. (Néha ezért a j értékeket M-mel ill. F-fel fogjuk jelölni.) Ezek szerint

$$\Gamma = \{\Gamma_{ij},\ 1 \leqq i \leqq S,\ 1 \leqq j \leqq 2\}.$$

Maga a Γ_{ij} gén az i-edik locuson előfordulható

$$A_{i1}, A_{i2}, \ldots, A_{im}$$

allélek közül kerülhet ki, ahol $m=m(i)$ a locus lehetséges alléljeinek a száma.

A P érték tehát a Γ genotípustól függő valószínűségi változó. A konkrét vizsgálati személyeken mérhető értékek akkor is eltérhetnek egymástól, ha azok genotípusa megegyezik. Legyen ezek rögzített Γ-hoz tartozó feltételes várható értéke

$$G = \mathcal{E}(P|\Gamma),$$

a $(P-G)$ különbségé E. Ekkor nyilván

$$P = G + E,$$

ahol $G=G(\Gamma)$ a genotípustól függő szám (ezt nevezzük genotípusos értéknek), E pedig a környezeti hatás, amelynek minden konkrét Γ mellett 0 a feltételes várható értéke.

Vizsgáljuk meg először a legegyszerűbb esetet, amikor $S=1$ és $m(1)=2$. Ekkor a locus-index elhagyható, tehát

$$\Gamma = \{\Gamma_F, \Gamma_M\},$$

és itt Γ_F, Γ_M az A_1, A_2 allélek közül kerülhet ki. Jelölje p az A_1 allél, $(1-p)$ az A_2 allél populációs gyakoriságát. Tegyük fel, hogy a vizsgált mennyiséget úgy skáláztuk, hogy a két homozigóta (A_1A_1 és A_2A_2) genotípusos értéke egymás ellentettje (a, ill. $-a$). A heterozigóta (A_1A_2) genotípushoz tartozó d érték (ami persze lehet negatív is) a-val egyenlő, ha A_1 domináns, $-a$-val, ha A_1 recesszív allél. Ha $d=0$, azt mondjuk, hogy nincs domináló hatás, $|d|>a$ pedig az ún. overdominanciát jelenti. A vizsgált mennyiség populációs átlaga a következő lesz:

$$M = p^2a + 2p(1-p)d - (1-p)^2a = (2p-1)a + 2p(1-p)d.$$

Nézzük meg, mit mondhatunk ebben az esetben a szülő—gyermek kapcsolatról. Tegyük fel, hogy a

$$\Gamma = \{\Gamma_F, \Gamma_M\}$$

genotípusú anyának gyermeke születik. Jelöljük ennek genotípusát $\tilde{\Gamma}$-mal, ebben $\tilde{\Gamma}_M$ az anyának vagy a Γ_M anyai, vagy a Γ_F apai génjével azonos, $\tilde{\Gamma}_F$ pedig (ha feltesszük, hogy nincs irányított párválasztás) $\tilde{\Gamma}_M$ értékétől függetlenül p valószínűséggel A_1, $(1-p)$ valószínűséggel A_2. Az utód genotípusos értékének feltételes várható értékére egyszerű számítással

$$\mathcal{E}[G(\tilde{\Gamma})|\tilde{\Gamma}_M = A_1] = M+(1-p)c, \quad \mathcal{E}[G(\tilde{\Gamma})|\tilde{\Gamma}_M = A_2] = M-pc$$

adódik, ahol

$$c = a+d(1-2p).$$

A kapott c mennyiség a génhelyettesítés átlagos hatása: ennyivel nő meg az utód feltételes várható értéke, ha az anyától A_2 helyett A_1 gént kap. (Számításunkban az anyának nem volt kitüntetett szerepe, csak a jelölés egyértelművé tétele érdekében választottuk a két szülő közül.)

FISHERtől ([2]) származik az az elképzelés, hogy válasszuk külön a genotípusos értékben azt a mennyiséget, amit abból az egyed át tud adni az utódnak, attól a tényezőtől, ami benne a génpár együttesétől függ és amit a génpár egyik tagja sem képes külön egyedül hordozni. Esetünkben ezek a tagok rendre a következők:

— az A_1A_1 genotípus értéke $a=M+2(1-p)c-2(1-p)^2d$;
— az A_1A_2 genotípus értéke $d=M+(1-2p)c+2p(1-p)d$;
— az A_2A_2 genotípus értéke $-a=M-2pc-2p^2d$.

Mindhárom genotípusos érték tehát három komponensre bontható:

i) az első tag a populációs átlag, ez mindhárom esetben M;

ii) a második tag az átörökítési érték, ez a genotípus két komponenséből additívan tevődik össze, A_1-hez $(1-p)c$, A_2-höz $-pc$ tartozik;

iii) a harmadik tag a génpár korrekciós értéke, ez rendre

$$-2(1-p)^2d, \quad 2p(1-p)d, \quad 2p^2d.$$

A kvantitatív jellegek vizsgálatakor a populációs átlagnak nincs különösebb jelentősége, a skála alkalmas megválasztásával ez mindig 0-val tehető egyenlővé. Nagyobb szerepet kap az öröklődés menetének leírásakor a variancia (V), ezért ezt fogjuk különböző komponensekre bontani.

A fenti példában az átörökítési értékből származó additív variancia-komponens

$$V_A = 2p(1-p)c^2$$

lesz, a génpár korrekciós értékéből származó domináló variancia pedig

$$V_D = 4p^2(1-p)^2d^2.$$

E kettő összege egyenlő a — szűkebb értelmezésű — genotípusos érték varianciájával:

$$V_G = V_A+V_D.$$

Most rátérünk az általános esetre. Legyen

$$\Gamma = \{\Gamma_{ij};\ i = 1, 2, \ldots, S;\ j = 1, 2\}$$

egy tetszőleges genotípus, amelyben Γ_{ij} az

$$\{A_{ik},\ k = 1, 2, \ldots, m(i)\}$$

allélek valamelyike és jelöljük $G(\Gamma)$-val a Γ-hoz tartozó genotípusos értéket. $G(\Gamma)$ tehát a populációs átlaga mindazon valaha élt vagy még meg sem született személyek fenotípusos értékének, akiknek a genotípusa Γ. (Feltesszük, hogy a vizsgált populáció az ún. *Hardy—Weinberg-egyensúlyban* van, azaz sem az egyes allélek gyakorisága, sem a fenotípusos értékeknek a genotípustól való függése nem változik az idő során.) Matematikailag $G(\Gamma)$ a P fenotípusos érték feltételes várható értéke azon feltétel mellett, hogy Γ értéke rögzített:

$$G(\Gamma) = \mathscr{E}(P|\Gamma).$$

(Itt és a továbbiakban a „populációs átlag", a „populációs gyakoriság" stb. megnevezések mindig a várható érték, valószínűség szinonímái.)

Jelöljük az i-edik locuson levő génpárt Γ_i-vel. [Γ_i tehát a $(\Gamma_{i1}, \Gamma_{i2})$ párt jelöli, amit időnként $(\Gamma_{iM}, \Gamma_{iF})$-fel is jelölünk.] Legyen $G_i(\Gamma_i)$ a $G(\Gamma)$ genotípusos érték populációs átlaga mindazokban a személyekben, akik genotípusában az i-edik locuson éppen a Γ_i génpár van, azaz legyen

$$G_i(\Gamma_i) = \mathscr{E}[G(\Gamma)|\Gamma_i], \quad i = 1, 2, \ldots, S.$$

A $G_i(\Gamma_i)$ mennyiséget a i-edik locus főhatásának nevezzük. A különböző locusok főhatásának az összegét $G_L(\Gamma)$-val jelöljük:

$$G_L(\Gamma) = \sum_{i=1}^{S} G_i(\Gamma_i).$$

A $[G(\Gamma) - G_L(\Gamma)]$ különbség azt mutatja meg, hogy milyen külön többletet jelent a genotípusos értékben a konkrét génstruktúra, az, hogy a génpárok együttesében az egyes tagoknak milyen a kölcsönhatásuk. Ezt a különbséget a locusok interakciójának nevezzük és $G_I(\Gamma)$-val jelöljük. A genotípusos érték tehát a következő összegre bontható:

$$G(\Gamma) = G_L(\Gamma) + G_I(\Gamma).$$

Fontos körülmény, hogy az itt szereplő két mennyiség szorzatának populációs átlaga mindig 0. Eszerint — normális eloszlást feltételezve — a két mennyiség független egymástól, ami nagymértékben megkönnyíti az öröklődésmenet vizsgálatát. De ha nem is normális a két mennyiség együttes eloszlása, az mindig igaz, hogy a két mennyiség összegének a varianciája egyenlő a komponensek varianciájának összegével. Jelöljük ezeket rendre V_G, V_L, V_I-vel, akkor tehát

$$V_G = V_L + V_I.$$

Hasonló felbontást végezhetünk el a V_L tagon belül is. Jelöljük az i-edik locus főhatásában az egyes génekre eső populációs átlagot rendre $G_{i1}(\Gamma_{i1})$, $G_{i2}(\Gamma_{i2})$-vel. (Ez különben azonos az $\mathscr{E}[G_i(\Gamma_i)|\Gamma_{ij}]$ feltételes várható értékkel.) Ezeknek a konk-

rét génekre számított átlagoknak az összege felel meg a FISHER által additív komponensnek nevezett mennyiségnek:

$$G_A(\Gamma) = \sum_{i=1}^{S} \sum_{j=1}^{2} G_{ij}(\Gamma_{ij}).$$

A domináló komponens a $(G_L - G_A)$ különbségnek felel meg (ez tulajdonképpen a génpárokon belüli interakcióknak az összege):

$$G_D(\Gamma) = G_L(\Gamma) - G_A(\Gamma).$$

Ha a megfelelő varianciákat V_A, V_D-vel jelöljük, a

$$P = G_A(\Gamma) + G_D(\Gamma) + G_I(\Gamma) + E$$

felbontásnak megfelelően kapjuk, hogy

$$V_P = V_A + V_D + V_I + V_E,$$

ahol V_P ill. V_E a fenotípusos érték ill. a környezeti hatás varianciája.

Poligén modellek

A poligén rendszerek konkrét leírására, a bennük szereplő locusok identifikálására még nemigen van mód. Pusztán a genetikai predikció szempontjából azonban nincs is erre feltétlenül szükség. Lehetséges ugyanis e rendszerek olyan átfogó modelljeit kialakítani, amelyek statisztikus vizsgálatához elég a variancia különböző komponenseinek az arányát megadni. A poligén rendszerek érvényesülésének magyarázatára ezért olyan modelleket dolgoztak ki, amelyek számos megszorítás ill. feltételezés mellett, absztrahált elméleti körülmények között érvényesülnek. E modellek legfőbb feltételezései a következők:

1) A különböző locusok alléljei a vizsgált mérhető mennyiségre kifejtett hatásuk alapján egy összefüggő *rendszert* alkotnak. Ez a poligén rendszer azonban csupán a funkcionális hatásban nyilvánul meg, mivel a rendszerhez tartozó locusok a különböző kromoszómákban véletlenszerűen helyezkednek el.

2) Az egyes locusokon levő *allélek száma tetszőleges,* hatásuk pedig eltérő *előjelű és nagyságú* is lehet.

3) A poligén rendszeren belül minden egyes locus hatása külön-külön elhanyagolható a teljes genotípusos hatáshoz képest. Éppen ezért a poligén rendszer tagjait *minor géneknek* is szokás nevezni. (Tehát kizárjuk az ún. *major* gének jelenlétét.)

4) A fenotípusos érték eloszlása, ezen túlmenően pedig a populáció adott tagjai fenotípusos értékének *együttes* eloszlása is *normális.*

5) Az adott poligén rendszer locusai egymástól *függetlenül* öröklődnek, vagyis a számcsökkentő osztódáskor az egyes locusok génpárjai egymástól *függetlenül szegregálódnak.* (Tehát nincs *linkage.*)

6) A génpárokon belüli interakció elhanyagolható, vagyis a *domináló variancia* 0-val egyenlő. A genotípusos érték varianciája csak az additív varianciából származik.

7) A *locusok interakciója elhanyagolható,* vagyis a genotípusos érték az egyes locusok főhatásának az összege. (Tehát nincs *episztázis.*)

8) A *környezeti hatás* független a genotípustól és a különböző személyekre ható környezetek is függetlenek egymástól.

9) Nincs *irányított párválasztás.*

10) A vizsgált jelleg nem befolyásolja a termékenységet és a halálozást, populációbeli gyakorisága állandó (pl. nincs *mutáció*).

Ezekből a feltevésekből a lehetséges modellek egymásra épülő szerkezete építhető fel. Egy-egy konkrét öröklődésmenet vizsgálatakor célszerű először a legegyszerűbb, tehát a legtöbb feltevést tartalmazó modell helyességét megvizsgálni és szükség esetén fokozatosan térni át a bonyolultabb, kevesebb feltevést tartalmazó modellekre. Azt a modellt, amelyben a fenti feltevések *mindegyike* teljesül és amelyben legdöntőbb a (normális) *G*auss-eloszlást adó, *a*dditív, *m*ultifaktoriális hatás, *GAM-modellnek* nevezzük. (A névadás a genetikai és környezeti hatások *együttes* figyelembevételére is utal.)

Feltevéseink — így a GAM-modell — helyességét általában háromféleképpen ellenőrizhetjük:

i) Illeszkedésvizsgálat: csak azt vizsgáljuk meg, hogy a minta adatai származhatnak-e az adott modellből.

ii) Hierarchikus vizsgálat: a vizsgált modellt beágyazzuk egy általánosabb modellbe és azt a hipotézist vizsgáljuk meg, hogy azon belül elfogadhatók-e a konkrét modell feltevései.

iii) Összehasonlító vizsgálat: két vagy több modell esetén azt vizsgáljuk, melyik áll közelebb a minta adataihoz.

A kovariancia és a korrelációs együttható

Az előző fejezetekben a poligén variáció jellegzetességeit a populáció vonatkozásában tárgyaltuk. A populáció azonban családokra bontható és az adott családon belüli személyek összehasonlító vizsgálata lehetőséget teremt a kvantitatív jellegek öröklődésének, a családtagok hasonlatosságának a vizsgálatára is.

Térjünk vissza arra a speciális esetre, amikor a poligén rendszer egyetlen locusból áll és a lehetséges allélek száma kettő. Az A_1, A_2 allélek gyakorisága p és $(1-p)$, az A_1A_1, A_1A_2, A_2A_2 genotípusok értéke rendre a, d, $-a$ volt. Ha tehát az egyik szülő genotípusos értéke a, tőle az utód biztosan A_1 gént kap, és ehhez a másik szülő p valószínűséggel A_1, $(1-p)$ valószínűséggel A_2 génnel társul. Ha tehát az egyik szülő genotípusa A_1A_1, akkor az utód genotípusos értéke p valószínűséggel a és $(1-p)$ valószínűséggel d. A további lehetőségek hasonlóan adódnak.

Láttuk, hogy a populációs átlag

$$M = (2p-1)a+2p(1-p)d,$$

ahonnan a genotípusos értékek kovarianciája

$$p^2(a-M)[pa+(1-p)d-M]+p(1-p)(d-M)[ap+d-a(1-p)-M]+$$
$$+(1-p)^2(-a-M)[pd-(1-p)a-M] = p(1-p)[a+(1-2p)d]^2.$$

A továbbiakban nagy hasznunkra lesz a következő észrevétel: ha az X_i, Y_j valószínűségi változók korrelációs együtthatója 0 minden $i \neq j$ mellett és r_i, ha $i=j$, továbbá a szórásukra

$$V_{X_i} = V_{Y_i} = V_i,$$

akkor az $X=(X_1+X_2+\ldots+X_n)$, $Y=(Y_1+Y_2+\ldots+Y_n)$ összegek kovarianciája

$$\mathrm{COV}_{XY} = r_1V_1+r_2V_2+\ldots+r_nV_n.$$

Ha még azt is feltesszük, hogy az összegek tagjai páronként korrelálatlanok, akkor

$$r_{XY} = \frac{r_1V_1+r_2V_2+\ldots+r_nV_n}{V_1+V_2+\ldots+V_n}.$$

Nevezetesen ha $r_1=r_2=\ldots=r_n=r$, akkor $r_{XY}=r$.

Ezek után már könnyen meghatározhatjuk a különböző rokonsági kapcsolatokra a fenotípusos értékek korrelációs együtthatóit. A felbontás az előző fejezetben talált

$$P = G_A(\Gamma)+G_D(\Gamma)+G_I(\Gamma)+E$$

összeg, amelynek tagjai — mint láttuk — valóban korrelálatlanok. Sőt, ha valamely rokon fenotípusos értékének hasonló felbontása

$$\tilde{P} = G_A(\tilde{\Gamma})+G_D(\tilde{\Gamma})+G_I(\tilde{\Gamma})+\tilde{E},$$

akkor a megfelelő varianciák rendre egyenlőek és a különböző jellegű mennyiségek korrelálatlanok. Elég tehát a különböző rokonsági kapcsolatok korrelációs együtthatóját megadni.

A családi közös gének

A számcsökkentő osztódáskor a gamétákba a szülő kromoszómáinak pontosan a fele kerül. Az, hogy egy-egy locus génpárjának melyik felét kapja az utód, a véletlentől függ. De az biztos, hogy a párból pontosan egy gén kerül át az utódba. Emiatt a szülő—gyermek párban pontosan a gének fele közös. Ilyen egyszerű összefüggés azonban csak e rokonsági kapcsolatban adható meg. Éppen azért, mert a gamétákba 1/2 valószínűséggel az anyai és 1/2 valószínűséggel az apai gén kerül, a nagyszülő—unoka párban a közös gének száma S locus esetén már 0 és S között tetszőleges szám lehet. Erről a számról éppen annyit mondhatunk, mintha feldobnánk S érmét és megszámolnánk az „írás"-ok előfordulását. A testvéreknél még bonyolultabb a helyzet: minden egyes locuson — egymástól függetlenül — rendre 1/4, 1/2, 1/4 annak a valószínűsége, hogy (i) a teljes génpár azonos, (ii) pontosan egy közös gén van, illetve (iii) egy közös gén sincs. (Itt és a továbbiakban a közös gén a család „eleve megegyező" identikus génjeit jelenti. Ezek az ún. „családi közös gének". A nem családi gének függetlenek egymástól, de bizonyos valószínűséggel szintén lehetnek identikusak. Ez utóbbiak a jellegükben azonos gének.) A jelenség azzal magyarázható, hogy az additív és domináló genotípusos értékek másként viselkednek a családfa „vertikális" és „horizontális" ágain. Az additív variancia híven követi a családi közös gének megoszlását, viszont a domináló variancia csak akkor mu-

tatható ki a kovarianciában, ha egyáltalán előfordulhat, tehát ha a teljes génpárok azonosak.

Ha figyelembe vesszük, hogy ezekre a hatásokra még a locusok interakciója és a környezeti hatások is rárakódhatnak, természetesnek látszik az az igény, hogy minden egyes rokonsági kapcsolatra közvetlenül a populációból határozzuk meg a korrelációs együttható értékét. Ezekben a vizsgálatokban a következő szempontokra lehetünk tekintettel:

1) Legfontosabb feladat a *környezeti* hatás mértékének a meghatározása.

2) A különböző kvantitatív jellegek populáción belüli korrelációjának tükröznie kell e mennyiségek *kölcsönhatását.*

3) Egy adott jelleg esetén megvizsgálhatjuk, hogyan változik a korrelációs együttható értéke a *rokonsági kapcsolat* függvényében.

A génkonfigurációk hatása

A különböző rokonsági kapcsolatban álló személyek mérhető jellegeiben kimutatható korrelációs értékeknek a rokonsági foktól való függése, mely szerint az I-, II-, III-adfokú rokonokban a korrelációs együttható értéke rendre 1/2, 1/4, 1/8 (ez az ún. „ős örökség törvénye"), a családi közös gének feleződésén alapul. Vizsgáljuk meg, milyen feltételek mellett érvényes ez a törvény és mivel helyettesíthető az általános esetben, amikor ezek a feltételek nem teljesülnek.

Legyen Γ egy specifikus poligén rendszer, az általa meghatározott mérhető P jelleg genotípusos értékét jelöljük $G(\Gamma)$-val. Legyen Γ_0 a Γ rendszer tetszőleges részrendszere; Γ_0-ban csak olyan locusok szerepelhetnek, amelyek Γ-ban is előfordulnak, de, szemben Γ-val, Γ_0-ban egy-egy locuson nem feltétlenül két gén foglalhat helyet. Előfordulhat, hogy Γ_0 valamelyik locusán csak az egyik gén van megadva. (Ez utóbbi esetben — az egyértelműség kedvéért — Γ-ban az anyai és apai kromoszómákat megkülönböztethetőnek tekintjük és Γ_0 definíciójához az is hozzátartozik, hogy az egyes locusokon az anyai vagy az apai gén szerepel-e.) Egy adott génrendszer említett részeit génkonfigurációknak *(configuration)* fogjuk nevezni. Tekinthetnénk a génkonfiguráció hatásának a

$$GE(\Gamma_0) = \mathscr{E}[G(\Gamma)|\Gamma_0]$$

feltételes várható értéket. Most azonban olyan $GC(\Gamma)$ értéket keresünk, amelyre

a) $\mathscr{E}[GC(\Gamma_0)|\Gamma_1]=0$ tetszőleges $\Gamma_1 \subset \Gamma_0$ mellett,

b) $G(\Gamma)$ előállítható a génfiguráció-hatások összegeként:

$$G(\Gamma) = \sum_{\Gamma_0 \subseteq \Gamma} GC(\Gamma_0).$$

Feltételezéseinkből következik, hogy ha Γ_0 az üres halmaz, akkor $GC(\Gamma_0)=\mathscr{E}[G(\Gamma)]$ a vizsgált mérhető jelleg populációs átlaga, amiről feltesszük, hogy 0. Ha Γ_0 egyetlen génből áll, akkor

$$GC(\Gamma_0) = GE(\Gamma_0),$$

ez a Γ_0 gén főhatása a Γ poligén rendszerben. Továbbmenve,

$$GC(\Gamma_0) = GE(\Gamma_0) - \sum_{\Gamma_1 \subset \Gamma_0} GC(\Gamma_1).$$

$GC(\Gamma_0)$ ennek alapján — lépésről lépésre — tetszőleges génkonfigurációra és végül magára Γ-ra is meghatározható. E felbontás előnye, hogy tagjai páronként korrelálatlanok, ami az egyes rokonsági kapcsolatokhoz tartozó korrelációs együttható meghatározását nagymértékben megkönnyíti.

Legyen $\Gamma, \tilde{\Gamma}$ két egymással tetszőleges rokonsági kapcsolatban álló személy genotípusa és jelöljük $P, \tilde{P}$-mal a megfelelő fenotípusos értékeket. Állítsuk elő a

$$G(\tilde{\Gamma}) = \sum_{\Gamma_0 \subseteq \tilde{\Gamma}} GC(\tilde{\Gamma}_0)$$

összegnek a Γ-ra vett feltételes várható értékét. Ezt egyrészt tagonként vehetjük, másrészt

$$\mathscr{E}[GC(\tilde{\Gamma}_0)|\Gamma] = 0,$$

kivéve, ha Γ-ban a $\tilde{\Gamma}_0$-nak megfelelő konfiguráció betöltése azonos $\tilde{\Gamma}_0$-mal. Ez utóbbi esetben viszont a feltételes várható érték természetesen maga $GC(\Gamma_0)$, tehát

$$\mathscr{E}[G(\tilde{\Gamma})|\Gamma] = \sum_{\Gamma_0 \subseteq \Gamma} p(\Gamma_0) GC(\Gamma_0),$$

ahol $p(\Gamma_0)$ annak a valószínűsége, hogy a Γ_0 konfiguráció Γ-ból sértetlenül került át $\tilde{\Gamma}$-ba. A kapott összefüggés lényege, hogy a $p(\Gamma_0)$ átmenetvalószínűségek a $G(\Gamma)$ függvénytől függetlenek és csupán a Γ, Γ_0 konfigurációk viszonyától, valamint a rokonsági kapcsolattól függenek.

Tekintve, hogy a testvéreket és a kettős unokatestvéreket kivéve minden rokonsági kapcsolatra $p(\Gamma_0)$ értéke 0 minden olyan konfigurációra, amelyben egy locuson két gén szerepel, az egy locuson levő génpárok kölcsönhatása csak a testvérek és a kettős unokatestvérek közötti korrelációban fordul elő. Minden más rokonsági kapcsolatban csak azok a konfiguráció-hatások mutathatók ki, amelyek locusonként egy gént érintenek. Ezeknek viszont $(1/2)^{ds}$ az átmenetvalószínűségük, ahol d a rokonsági fok*(degree)*, s pedig a konfiguráció mérete *(size)*, vagyis a konfigurációhoz tartozó gének száma. Emiatt

$$\mathscr{E}(\Gamma\tilde{\Gamma}) = \sum_{s=1}^{S} \frac{V_s}{2^{ds}},$$

ha nem testvérekről van szó, ahol V_s tetszőleges, *nem negatív* állandó, testvérek esetén pedig ehhez még egy tetszőleges, *nem negatív* δV_0 tag adódhat hozzá, ahol

$$0 \leqq \delta \leqq 1/4.$$

Tehát $V_G = \sum_{s=0}^{S} V_s$ miatt a fenotípusos értékek korrelációs együtthatója

$$r_{P\tilde{P}} = \frac{\varepsilon\delta V_0 + \sum_{s=1}^{S} V_s/2^{ds}}{\sum_{s=1}^{S} V_s + V_E},$$

ahol $\varepsilon = 1$, ha testvérekről van szó, különben 0.

Levezetésünk azt jelenti, hogy az „ős örökség törvénye" akkor és csak akkor érvényes, ha az egyes gének hatása additív és nincs környezeti hatás. Különben a korrelációs együttható fokozatosan elmarad az $(1/2)^d$ elvi értéktől.

Genetikai predikció

Legyenek $P_1, P_2, \ldots, P_n$ a P fenotípusos értékű személy rokonainak a fenotípusos értékei; kérdés, mit mondhatunk P-ről, ha csak a $P_1, P_2, \ldots, P_n$ értékeket ismerjük. Jelöljük P és P_i kovarianciáját c_i-vel, P_i és P_j kovarianciáját pedig C_{ij}-vel, a $(c_1, \ldots, c_n)'$ vektort $\mathbf{c}$-vel, a $\{C_{ij}\}_{i,j=1}^{n}$ mátrixot pedig $\mathbf{C}$-vel.

Amint könnyen kimutatható, a

$$\hat{P} = a_1 P_1 + \ldots + a_n P_n$$

lineáris becslés akkor adja a legkisebb szórású becslést P-re, ha az $\mathbf{a}=(a_1, \ldots, a_n)'$ vektor

$$\mathbf{a} = \mathbf{C}^{-1}\mathbf{c}$$

alakú. Ekkor a becslés hibája

$$V_{P-\hat{P}} = V_P - \mathbf{c}'\mathbf{C}^{-1}\mathbf{c}.$$

Ha most P a gyermek, P_1, P_2 a szülők fenotípusos értéke, akkor $a_1=a_2=1/2$, tehát a legjobb becslést a szülők átlaga adja és ennek a becslésnek a varianciája $(1/2)V_P$, feltéve, hogy a szülő-gyermek korreláció 1/2. Ha általában a korreláció $(1/2)^d$, ahol d a rokonság foka és P_1, P_2 a szülőknek, $P_3, \ldots, P_n$ pedig a gyermek egyéb rokonainak a fenotípusos értékei (de a rokonok közt nem fordulnak elő a gyermek leszármazottai), meglepő módon a legjobb becslés még mindig $(1/2)(P_1+P_2)$ lesz. Csak akkor fog a többi rokon értéke belekerülni ebbe a becslésbe, ha egyéb génkonfiguráció-hatások is fellépnek vagy a környezeti hatás nem elhanyagolható.

Genetikai determináció és örökölhetőségi együttható

A genetikai determináció vagy tágabb értelmezésű örökölhetőségi együttható értéke a genotípusos és fenotípusos varianciák hányadosa:

$$H = \frac{V_G}{V_P} = \frac{V_A+V_D}{V_A+V_D+V_E}.$$

Itt V_A, V_D, V_E rendre az additív, domináló és környezeti variancia.

A genetikai determináció meghatározására a genetikai családvizsgálatok és az ikeradatok nyújtanak lehetőséget. Tekintve azonban, hogy — testvérektől eltekintve — a domináló variancia elvileg nem mutatható ki a rokonok között, a locusok közti interakció mérése pedig gyakorlatilag lehetetlen, e jelen körülmények között olyan mérőszámra van szükségünk, mely egyrészt tükrözi a vizsgált jelleg átörökíthetőségét, másrészt egyértelműen meghatározható. Ez az additív variancia és a fenotípusos variancia hányadosa, amit a (szűkebb értelemben vett) örökölhetőség *(heritability)* mértékének vagy örökölhetőségi együtthatónak nevezünk és h^2-tel jelölünk:

$$h^2 = \frac{V_A}{V_P} = \frac{V_A}{V_A+V_D+V_E}.$$

Ennek értéke elvileg 0 és 1 közé esik. E határokon kívül eső érték elvileg lehetetlen, ezt azonban a becslésben nem szoktuk figyelembe venni. Egyrészt egy-egy becslés mindig tartalmaz bizonyos, véletlentől származó hibákat, amik akár e határok

túllépésére is vezethetnek (és ha ez a helyzet, a határok túllépése hasznos információt nyújthat épp e véletlen hatások nagyságára vonatkozóan), másrészt célszerű hagyni, hogy egy-egy modellben a konkrét minta kisodorjon minket az elvi határokon túlra, mert e jelenség ismételt fellépése esetén a modell helytelenségére következtethetünk és az eltérés iránya arról is tájékoztat minket, milyen irányba kell a modellt továbbfejleszteni.

Az örökölhetőség mértéke tehát annyiban különbözik a genetikai determinációtól, hogy csak az átörökíthető genetikai hatást foglalja magába, pontosabban szólva annak is csak azt a részét, amelyre még érvényes az „ős örökség törvénye". Abban az esetben, ha sem domináló variancia, sem locus-interakció, sem irányított párválasztás nincs, e két mennyiség egyenlő egymással. A h^2 értékét gyakorlatban az irányított párválasztás szokta elsősorban befolyásolni.

3. A küszöb-modell

Poligén öröklődés kvantitatív jellegek esetén

A *Mendel-szabályok* szerint öröklődő monogén ártalmak, az ún. *teratogén hatások,* majd a kromoszómarendellenességek megismerésével számos CM kóreredete vált ismertté. Tekintve azonban, hogy ezek a CM-ek meglehetősen ritkák voltak, az összes CM-nek csak mintegy 15%-ában tisztázódott ily módon az etiológia. Így mind sürgetőbbé vált a maradék, a döntő hányad kóreredetének tisztázása. A tisztázatlan eredetű CM-ek között főleg a CICM-ek fordultak elő, ezért fő feladatnak ezek epidemiológiai és genetikai családvizsgálata látszott.

A CICM-ekkel kapcsolatban az öröklődés szerepe már régen felmerült. Az ilyen betegek rokonságában ugyanis — mégpedig a rokonsági fokkal párhuzamosan egyre csökkenő mértékben — kifejezett családi halmozódást észleltek. Ugyanakkor öröklődésmenetük a *Mendel-szabályokkal* nem volt összeegyeztethető. Kromoszómarendellenességet nem, vagy legfeljebb szabályt erősítő kivételként lehetett a vizsgált betegekben kimutatni.

A CICM-ek kóreredetében a környezeti tényezők szerepe is nyilvánvaló volt. A területi és időbeni (pl. szezonális) eltéréseket, a személyi jellemzőktől (anyai életkor, szülési sorrend, méhen belüli pozíció stb.) és szociális-gazdasági tényezőktől függően eltérő gyakoriságokat csak külső hatásokkal lehetett megmagyarázni. Az ikervizsgálatok is a környezeti hatások szerepét igazolták.

Mindezek az ellentétes nézeteket valló kutatók időleges elkülönüléséhez vezettek: egyesek csak a genetikai, mások csak a környezeti tényezőkkel kívánták a CICM-ek kóreredetét megmagyarázni. Csak a hatvanas évek elején került sor e két szempont szintézisére: a küszöb-modellen alapuló multifaktoriális kóreredet kidolgozására. E modell általános esetére térünk most rá.

Tegyük fel, hogy a vizsgált populáció a *Hardy—Weinberg-egyensúly* állapotában van és jelöljük Γ-val a vizsgált CM-ért felelős poligén rendszert. Γ egy konkrét értéke mellett a populáció mindazon tagjaiban, akiknek a genotípusa éppen Γ, a CM fellépése legyen egy minden egyéb történéstől független véletlen hatás eredménye. Ennek megfelelően a CM kialakulásának a valószínűsége (kockázata) legyen $\pi(\Gamma)$. $\pi(\Gamma)$ tehát ugyanolyan függvénye Γ-nak, mint a $G(\Gamma)$ genotípusos érték, a köztük levő formális különbség csupán annyi, hogy $\pi(\Gamma)$ csak 0 és 1 közötti szám

lehet. A lényegi különbség azonban az, hogy $\pi(\Gamma)$ nem egy P kvalitatív jelleg feltételes várható értéke, hanem egy kvantitatív jelleg feltételes valószínűsége. (Formálisan ez a különbség persze megszüntethető a vizsgált kvalitatív jelleg indikátorfüggvényének bevezetésével, amely 1 vagy 0 aszerint, hogy van-e CM vagy sem. Ez azonban változatlanul hagyná azt a lényeges körülményt, hogy az így kapott P kvantitatív jellegnek továbbra is csak két értéke lehet, míg az eddig vizsgált mérhető mennyiségek tetszőleges sok lehetséges értékkel rendelkező, többnyire folytonos eloszlású változók voltak.)

A genetikai predikció ebben a modellben — adott családfák vizsgálatakor — a család egy tagjára nézve annak a feltételes valószínűségnek a meghatározását jelenti, hogy CM-je lesz, feltéve, hogy közeli rokonságának néhány tagjáról tudjuk, van-e hasonló CM-jük vagy sem. Most csak azt fogjuk vizsgálni, hogy egy betegnek vele meghatározott rokonsági kapcsolatban álló rokona rendellenes-e vagy sem. Amíg tehát a kvantitatív jellegek mellett a kovarianciát kellett meghatároznunk a különböző rokonsági kapcsolatban álló személyekre, most a feltételes valószínűség a kérdés. Mégpedig annak feltételes valószínűsége, hogy a vizsgált pár másik tagja is rendellenes lesz, feltéve, hogy az egyik CM-mel sújtott. Ezt a feladatot mi további feltevések mellett fogjuk megoldani.

A multifaktoriális küszöb-modell

A multifaktoriális küszöb-modell (GAMT) három lényeges összetevője a CM mértéke, a küszöb *(threshold)* és a korábbi GAM-modell. A GAMT-modellben feltételezzük, hogy a bináris CM-eknél minden konkrét személyhez hozzárendelhetünk egy számot, amely a rendellenességnek a vizsgált személyben jelenlevő mértékét, a hajlamot *(liability)* fejezi ki. Ehhez egy — az egész populációra jellemző — küszöb tartozik és a modell szerint akkor és csak akkor rendellenes a vizsgált személy, ha benne a rendellenesség mértéke nagyobb a küszöbnél. E mennyiségeket L-lel és T-vel jelöljük. Feltesszük azt is (ez pusztán technikai feltétel), hogy L populációs átlaga 0 és szórása 1. Ha tehát $L_1, L_2, \ldots, L_n$ egy család tagjaiban a vizsgált rendellenesség mértékei, akkor

$$\mathcal{E}(L_i) = 0, \quad \mathcal{E}(L_i^2) = 1, \quad \mathcal{E}(L_i L_j) = \frac{h^2}{2^d},$$

ahol h^2 a rendellenességre jellemző paraméter (az örökölhetőség mértéke), d az i, j személyek rokonsági foka. A h^2 paraméteren kívül a GAMT második paramétere a T küszöb. Ennek megfelelően a család i-edik tagjának akkor és csak akkor lesz rendellenessége, ha $L_i \geqq T$.

Matematikai szempontból magának a CM mértékének a feltételezése nem jelent semmi megszorítást. Legyen ugyanis $\pi(\Gamma)$ egy tetszőleges rendellenesség kialakulásának a valószínűsége, $F(t)$ tetszőleges eloszlásfüggvény és T tetszőleges valós szám. Ekkor a rendellenesség L hajlamát az

$$L = G(\Gamma) + E,$$

$$F[T - G(\Gamma)] + \pi(\Gamma) = 1$$

összefüggések jellemzik, ahol E $G(\Gamma)$-tól független, $F(t)$ eloszlásfüggvényű valószínűségi változó. F és T választása önkényes, de ha már egyszer ezeket választottuk, akkor $G(\Gamma)$ egyértelműen adódik. A GAMT-ban — hiszen a GAM-ra alapozzuk számításunkat — E-ről is, $G(\Gamma)$-ról is feltesszük, hogy normális eloszlásúak. (Mindkét feltevést a centrális határeloszlás-tétel támasztja alá.)

A CM mértékét és a küszöböt számos tényező befolyásolhatja, így pl. a nem és az ártalom súlyossága. Így a CICM-ek P értéke általában jellemző eltérést mutat a két nemben. Ez nemtől függő küszöb felvételével könnyen magyarázható. Kérdés, hogy maga az öröklődés mértéke függ-e a nemtől vagy sem. Általában felteszik, hogy nem függ. E feltételt mi is a modell részének tekintjük. Ez azonban nem szerves része a GAMT-nak és bármikor elvethető anélkül, hogy a modell egyéb részein változtatni kellene.

A GAMT fontos tulajdonsága, hogy a hajlamnak még a vizsgált betegek között is meghatározott eloszlása van. Ha a hajlam épp a küszöb felett van, a rokonok gyakorisága alacsonyabbra várható, mint ha a hajlam messze túlesne a küszöbön. Előfordulhat, hogy ez utóbbi körülmény a beteg állapotának súlyosságában is megfigyelhető, így a szabály közvetlenül igazolható. Ha azonban ez nincs is így, a rendellenesség családon belüli halmozódása alapján mindig képet alkothatunk a súlyosság mértékéről, ami a modell igen rugalmas alkalmazását teszi lehetővé.

A CM mértékét teljes genetikai meghatározottság esetén az érintett poligén rendszer összetétele határozza meg. Feltevésünk szerint a nagyszámú minor génből összetevődő poligén rendszernek csak a végső „tömeghatása" érvényesül a fenotípusban. Ez az eltérő jellegű és dózisú gének hatásának összege. Ezen belül lesznek olyanok, amelyek az adott rendellenesség megjelenésének az irányában hatnak. Az adott poligén rendszeren belül első közelítésben ezek száma határozza meg a CM mértékét.

Ha nem teljes a genetikai meghatározottság, a CM mértékének $G(L)$ genotípusos értéke azt fejezi ki, mennyi a vizsgált személy örökölt terheltsége, hajlama a vizsgált rendellenességre. Az $L=G(\Gamma)+E$ felbontás szerint a CM mértékében ehhez még környezeti hatás társul. Ezért ha félreértésre vezethet, nem használjuk a „hajlam" elnevezést erre a mennyiségre, mert az csak a $G(\Gamma)$ komponensre illik.

A genetikai családvizsgálatok

A *Mendel-szabályok* szerint öröklődő tulajdonságok konkrét családfák vizsgálatával általában még ellenőrizhetők voltak. Ennek továbbfejlesztését jelentette a *Weinberg-féle* [8] *proband módszer*. A modellek bonyolultságának és sztochasztikus jellegének fokozódásával az ellenőrzést egyre nehezebb volt elvégezni kisszámú családfák adatai alapján. Előtérbe kerültek a nagyméretű mintákkal dolgozó populációgenetikai módszerek. Ezek azonban (elsősorban számítástechnikai okokból) a családfák széttöredezését vonták maguk után. Nem térünk itt ki e módszerek számtalan buktatójára, hiszen ezek többsége a CM-ek vizsgálatakor jelentősen nem érezteti hatását. Meg kell azonban jegyeznünk, hogy az adatok általunk használt tömörítése, a feltételes gyakoriságok meghatározása nem jelenti a mintában rejlő lehetőségek maradéktalan kimerítését. Éppen a különböző modellek összehasonlító vizsgálata igényelné a vizsgálati adatok komplexebb kialakítását. A GAMT-modell illeszkedésvizsgálatához azonban adataink megfelelőek voltak.

A küszöb jellegű ártalmakban a kvantitatív jellegek szokásos mért értékei nem állnak rendelkezésre. Helyettük csak az ártalom gyakorisága határozható meg egyrészt a teljes népességben (ill. reprezentatív mintában), másrészt a vizsgált betegek adott rokonsági körében. (Ez utóbbi a genetikai családvizsgálat célkitűzése.) E gyakorisági értékek teremtik meg a szilárd alapot a küszöb-modell alkalmazásához. Maga a küszöb bizonyos betegségekben (pl. értelmi fogyatékosság, koraszülöttség, hypertensio) emberi elhatározás kérdése, másokban, köztük a CM-ekben természetes (teratológiai) produktum.

Magunk a következőképpen jártunk el. Vettük egy adott időszak és terület összes és meghatározott CICM-ben szenvedő szülötteit, ők voltak a vizsgált betegek. Első-, másod- és harmadfokú rokonságukban meghatároztuk az adott CICM előfordulását. Minden egyes rokonsági kapcsolaton belül meghatároztuk azoknak a személyeknek a számát, akikről információink voltak, továbbá ezen belül az ugyanazon CICM-ben szenvedők számát. Mindezek alapján számítható volt az adott CICM vizsgált populációbeli gyakorisága és a betegekkel meghatározott rokonsági kapcsolatban állók „ismételt" specifikus CICM-gyakorisága.

Előfordulhat, hogy a betegekkel adott rokonsági kapcsolatban állók a populációnak nem ugyanabból a rétegéből kerülnek ki, mint maguk a betegek. Ez a helyzet pl., ha a CICM gyakorisága függ a nemtől és a két réteg eltérő nemű. További gyakori problémát jelentenek a generációkülönbségből adódó diagnosztikai és terápiás eltérések. Éppen ezért a rokonoknak megfelelő réteg populációbeli gyakorisága nem feltétlen egyenlő a betegekével. A családvizsgálatok eredménye tehát különböző rétegekhez tartozó számokban foglalható össze. Ezek alapján határozhatjuk meg a modell h^2, T paramétereit és ellenőrizhetjük, illeszkedik-e a modell az adatokhoz vagy sem. Most erre térünk rá.

4. A számítógépes program leírása

A program jellemzői

A program a *Magyar Tudományos Akadémia* CDC 3300 típusú számítógépére íródott SIMULA 67 programozási nyelven. A futási idő betegségenként 3-4 perc, a memóriaigény 25 K. A program kb. 400 soros. Az adatokat lyukkártyán kell beolvasni, az eredmények protokoll-listán jelennek meg, egy betegség egy oldal. A program két könyvtári és két saját függvényeljárással dolgozik. Ezek a következők:

i) NORM (A, B). Jelölje $\Phi(z)$ ill. $\varphi(z)$ a standard normális eloszlás- ill. sűrűségfüggvényt. Tetszőleges A valós szám esetén NORM $(A, B)=\Phi(A)$ és $B=\varphi(A)$.

ii) NORMINV $(U)=\Phi^{-1}(U)$.

iii) JNORM $(\tilde{T}, T, r)$. Jelöljön $\tilde{T}$, T és r olyan valós számokat, melyekre

$$\tilde{T}, T \geqq 0, \quad |r| \leqq 0{,}5$$

teljesül, $\tilde{L}$ és L pedig olyan standard normális eloszlású valószínűségi változókat, amelyeknek r a korrelációs együtthatója, akkor

$$\text{JNORM}\,(\tilde{T}, T, r) = P(\tilde{L} \geqq \tilde{T}, \quad L \geqq T).$$

iv)

$$\text{CNORM}\,(\tilde{T}, T, r) = P(\tilde{L} \geqq \tilde{T} | L \geqq T),$$

ahol az r korrelációs együttható értéke tetszőleges.

A JNORM függvényeljárás leírása

A kétváltozós normális eloszlás számítására a következő eljárást dolgoztuk ki. Legyen

$$F(\tilde{T}, T, r) = P(\tilde{L} \geqq \tilde{T},\ L \geqq T),$$

$$Q(z) = 1-\Phi(z), \quad P_r(\tilde{T}, T) = \frac{F(\tilde{T}, T, r)}{Q(T)}.$$

$\tilde{L}$, L és r definíciójából következik, hogy $|r| \leqq 1/2$ és

$$\tilde{L} = rL + sW,$$

ahol W L-től független, standard normális eloszlású valószínűségi változó, $|s| \geqq \sqrt{3}/2$ és

$$r^2+s^2=1.$$

4.1. TÉTEL. Legyen

$$t = \frac{r}{s}, \quad Z = \frac{\tilde{T}-rT}{s},$$

akkor $|r|<1/\sqrt{2}$ esetén

$$P_r(\tilde{T}, T) = Q(Z)+\varphi(z)\sum_{n=1}^{\infty} a_n b_n \frac{t^n}{n!},$$

ahol a_n és b_n az

$$a_1 = 1, a_2 = Z, \ldots, a_n = Za_{n-1}-(n-2)a_{n-2};$$

$$b_1 = \frac{\varphi(t)}{Q(T)}-T, \quad b_2 = 1-Tb_1, \ldots, b_n = (n-1)b_{n-2}-Tb_{n-1}$$

rekurziókkal számolható.

Bizonyítás.

$$P_r(\tilde{T}, T) = \frac{1}{Q(T)}\int_T^{\infty} \varphi(v) \int_{\frac{1}{v}(\tilde{T}-rv)}^{\infty} \varphi(z)\, dz\, dv =$$

$$= Q(Z)+\frac{1}{Q(T)}\int_T^{\infty} \varphi(v) \int_0^{t(v-T)} \varphi(Z-z)\, dz\, dv.$$

Jelöljük $h(t)$-vel a második tagot, akkor

$$h(t) = \frac{1}{Q(T)}\int_0^{\infty} \varphi(T+u) \int_0^{tu} \varphi(Z-v)\, dv\, du.$$

Rögzített $\tilde{T}$ és T mellett fejtsük *Taylor-sorba* a $h(t)$ függvényt $t=0$ körül. Ekkor $h(0)=0$, továbbá

$$h'(t) = \frac{1}{Q(T)} \int_0^\infty u\varphi(T+u)\varphi(Z-tu)\,du,$$

$$\frac{h'(0)}{\varphi(Z)} = \frac{\varphi(T)}{Q(T)} - T = b_1 = 1b_1 = a_1 b_1.$$

A második deriváltra

$$h''(t) = \frac{1}{Q(T)} \int_0^\infty u^2\varphi(T+u)(Z-tu)\varphi(Z-tu)\,du$$

adódik, általában pedig

$$h^{(n)}(t) = \frac{1}{Q(T)} \int_0^\infty u^n\varphi(T+u)A_n(Z-tu)\,du,$$

ahol egyrészt $A_n(Z-tu)_{t=0}$ nem függ u-tól és $\varphi(Z)a_n$-nel egyenlő, másrészt

$$\int_0^\infty u^n\varphi(T+u)\,du = \int_0^\infty u^{n-1}(u+T)\varphi(T+u)\,du - T\int_0^\infty u^{n-1}\varphi(T+u)\,du = \ldots = b_n.$$

Ezek után bebizonyítjuk, hogy $|r|<1/\sqrt{2}$ esetén a *Taylor-sor* konvergens. Nyilvánvaló módon

$$|r| < \frac{1}{\sqrt{2}} \Leftrightarrow |t| < 1.$$

Legyen $c=\max(|Z|, T)$ és $1<d<\sqrt{1/|t|}$. Szükségünk lesz a következő nyilvánvaló állításra: létezik olyan n^* természetes szám, hogy

$$\begin{aligned} c^2d^2 + 2cd\sqrt{n-1} - (n-1) &\leqq d^4 n, \\ cd(d^2+1) + \sqrt{n-1} &\leqq d^4\sqrt{n+1} \end{aligned} \tag{1}$$

valamely n természetes számra akkor és csak akkor teljesül, ha $n>n^*$.

Legyen $K>0$ olyan, hogy $n\leqq\max(2, n^*)$ esetén fennálljon, hogy

$$\begin{aligned} |a_n b_n| &\leqq Kd^{2n} n!, \\ |a_n b_{n+1}|,\ |a_{n+1} b_n| &\leqq Kd^{2n+1} n!\sqrt{n+1}. \end{aligned} \tag{2}$$

Teljes indukcióval bizonyítható, hogy tetszőleges n-re teljesül (2). Emiatt tetszőleges n-re

$$\left|a_n b_n \frac{t^n}{n!}\right| \leqq K(d^2|t|)^n,$$

márpedig d definíciójából következik, hogy

$$d^2|t| < 1.$$

Az exponenciális túlcsordulás elkerülése végett a függvényeljárás a következő rekurzióval dolgozik:

$$e_1 = tb_1;$$

$$c_2 = Ze_1; \quad d_2 = \frac{t - Te_1}{2}; \quad e_2 = tZd_2;$$

$$\vdots \qquad\qquad \vdots$$

$$c_n = Ze_{n-1} - (n-2)td_{n-1}; \quad e_n = t\left[Zd_n + \frac{(n-2)t(Td_{n-1} - e_{n-2})}{n}\right]$$

$$d_n = \frac{tc_{n-1} - Te_{n-1}}{n}.$$

Teljes indukcióval belátható, hogy

$$e_n = a_n b_n \frac{t^n}{n!}.$$

Legyen $d = \sqrt[4]{1/|t|}$,

$$K_n = \max\{|a_n b_n|/d^{2n} n!,\ |a_n b_{n+1}|/d^{2n+1} n!\sqrt{n+1},\ |a_{n+1} b_n|/d^{2n+1} n!\sqrt{n+1}\},$$

$$K = \max_{n \leqq n^*} \{K_n\} \quad \text{és} \quad N = \min\{n:\ n > n^* \text{ és } K(\sqrt{|t|})^n < 0{,}5 \cdot 10^{-6}\}, \quad \text{akkor}$$

$$\text{JNORM}(\tilde{T}, T, r) = Q(T)\left[Q(Z) + \varphi(Z) \sum_{n=1}^{N} e_n\right].$$

A CNORM függvényeljárás leírása

$|r| < 0{,}5$ esetén

$$\text{CNORM}(\tilde{T}, T, r) = \frac{F(\tilde{T}, T, r)}{Q(T)}.$$

$|r| > 0{,}5$ esetén szükségünk lesz a következő, egyszerű számolással adódó állításokra:

$$F(\tilde{T}, T, r) + F(\tilde{T}, -T, -r) = Q(\tilde{T});$$

$$F(\tilde{T}, T, r) = F\left(\tilde{T}, \frac{T - \tilde{T}}{\sqrt{2(1-r)}}, -\sqrt{\frac{1-r}{2}}\right) + F\left(\frac{\tilde{T} - T}{\sqrt{2(1-r)}}, T, -\sqrt{\frac{1-r}{2}}\right);$$

$|r| > 0{,}5$ esetén $\sqrt{\dfrac{1-|r|}{2}} < 0{,}5.$

Ezek alapján $0{,}5 < r < 0{,}98$ esetén

$$\text{CNORM}(\tilde{T}, T, r) = \frac{1}{Q(T)}\left[F\left(\tilde{T}, \frac{T - \tilde{T}}{\sqrt{2(1-r)}}, -\sqrt{\frac{1-r}{2}}\right) + F\left(\frac{\tilde{T} - T}{\sqrt{2(1-r)}}, T, -\sqrt{\frac{1-r}{2}}\right)\right],$$

$0{,}5 < -r < 0{,}98$ esetén

$$\mathrm{CNORM}(\tilde{T}, T, r) = \frac{1}{Q(T)}\left[Q(\tilde{T}) - F\left(\tilde{T}, -\frac{\tilde{T}+T}{\sqrt{2(1+r)}}, -\sqrt{\frac{1+r}{2}}\right) - \right.$$

$$\left. - F\left(\frac{\tilde{T}+T}{\sqrt{2(1+r)}}, -T, -\sqrt{\frac{1+r}{2}}\right)\right];$$

$|r| \geqq 0{,}98$ esetén pedig

$$\mathrm{CNORM}(\tilde{T}, T, r) = \frac{Q(\max(\tilde{T}, T))}{Q(T)}.$$

Az adatok beolvasása

Mint már említettük, a program futtatásához az adatokat lyukkártyán kell beolvasni. Az adatkártya-csomag egy indító kártyával kezdődik. Ezen az 1. oszloptól a vizsgált betegségek számát kell megadni; jelöljük ezt ND-vel. Az indító kártya után következnek az egyes betegségekhez tartozó alcsomagok, összesen ND darab. Mindegyik alcsomag a betegség-kártyával kezdődik. Ennek beosztása a következő:

1. oszlop: 0;
11. oszloptól: betegség kódja (legfeljebb 8 karakter);
21. oszlop: 1, ha a betegség csak az egyik nemre jellemző, vagy ha a nemek ömlesztve szerepelnek, egyébként 2;
26. oszlop: 1, ha a vizsgált betegek figyelembe vett rokonai között szerepelnek az egypetéjű ikrek, egyébként 0;
31. oszlop: 1, ha a rokonok között szerepelnek a szülők, egyébként 0;
41. oszlop: 0, ha nem szerepelnek a testvérek, 2, ha a kétpetéjű ikrek és a nemiker testvérek külön szerepelnek, egyébként 1;
51. oszlop: 0, ha nem szerepelnek a szülők testvérei, 2, ha az apák és az anyák testvérei külön szerepelnek, egyébként 1;
61. oszlop: 0, ha nem szerepelnek a testvérek gyermekei, 2, ha a fiú- és leánytestvérek gyermekei külön szerepelnek, egyébként 1;
71. oszlop: 0, ha nem szerepelnek az unokatestvérek, 2, ha vagy az apai és az anyai, vagy a „nagybácsi"- és „nagynéni"-unokatestvérek külön szerepelnek, 4, ha a lehetséges négy eset mind külön-külön szerepel, egyébként 1.

Az egyes alcsomagokban a betegség-kártyát a rokon-kártyák követik. Ezek beosztása a következő:

1. oszlop: 1;
11. oszloptól: betegség kódja;
31. oszlop: B, ha a nemek ömlesztve vannak, egyébként M vagy F aszerint, hogy a vizsgált beteg fiú-e vagy lány;
35. oszloptól: százezrelékben a betegség relatív gyakorisága a beteg nemének és generációjának megfelelő kontrollcsoportban (egész szám);

41. oszloptól: rokon kódja (egypetéjű ikrek=MZ, szülő=P, kétpetéjű iker=SDZ, nem-iker testvér=SS, testvér kétpetéjű ikret is megengedve=SB, apa testvére=UM, anya testvére=UF, szülő testvére=UB, fiútestvér gyermeke=NM, leánytestvér gyermeke=NF, testvér gyermeke=NB, apai nagybácsi gyermeke=CMM, apai nagynéni gyermeke=CMF, apai unokatestvér=CMB, anyai nagybácsi gyermeke=CFM, anyai nagynéni gyermeke=CFF, anyai unokatestvér=CFB, „nagybácsi"-unokatestvér=CBM, „nagynéni"-unokatestvér=CBF, unokatestvér=CBB); (*)
46. oszlop: rokonsági fok;
51. oszlop: a rokon neme (B, M vagy F);
55. oszloptól: százezrelékben a betegség relatív gyakorisága a rokonságnak megfelelő kontrollcsoportban;
61. oszloptól: a megfelelő betegek megfelelő rokonainak száma;
71. oszloptól: ezen körben a betegség gyakorisága.

A rokon-kártyák sorrendje a következő: a 41. oszlopban álló karakter szerint (M, P, S, U, N, C sorrendben), ezen belül a beteg neme szerint (M, F), ezen belül a rokon neme szerint (M, F), ezen belül a 42—43. oszlopban álló karakterek szerint [ld. (*)].

A programtörzs leírása és a program outputja

Valamely konkrét veleszületett rendellenesség esetén jelölje p_1 és p_2 a betegség relatív gyakoriságát a vizsgált betegeknek megfelelő férfi, ill. női kontrollcsoportban, T_1 ill. T_2 az ezen adatok által a GAMT-modell szerint meghatározott küszöböket, akkor $v=1, 2$ esetén

$$T_v = \Phi^{-1}(1-p_v).$$

Mint láttuk, adataink lehetnek a betegek nulladfokú (egypetéjű ikrek), elsőfokú (szülők, testvérek), másodfokú (szülők testvérei, testvérek gyermekei) és harmadfokú (unokatestvérek) rokonairól. A zárójelben említett rokoncsoportokat rétegeknek nevezzük. $d=0, 1, 2, 3$ esetén jelöljük t_d-vel a vizsgált d-edfokú rétegek számát, akkor $t_0=0$ vagy 1, $t_1=0, 1$ vagy 2, $t_2=0, 1$ vagy 2, $t_3=0$ vagy 1. Legyen r^* az egyes rétegek sorszáma a megfelelő rokonsági fokú rétegek között, akkor testvérekre és testvérek gyerekeire $r^*=2$, egyébként $r^*=1$. $0 \leqq d \leqq 3$ és $1 \leqq r^* \leqq t_d$ esetén legyen

$$R = \sum_{v=0}^{d-1} t_v + r^*.$$

Nyilván $1 \leqq R \leqq 6$. Jelölje $\tilde{p}_{2R-1}$, $\tilde{T}_{2R-1}$, ill. $\tilde{p}_{2R}$, $\tilde{T}_{2R}$ a betegség relatív gyakoriságát és a küszöb értékét az R-edik rétegnek megfelelő férfi, ill. női kontrollcsoportban, akkor $v=0, 1$ esetén

$$\tilde{T}_{2R-v} = \Phi^{-1}(1-\tilde{p}_{2R-v}).$$

Legyen $n=1$ ill. $n=2$, ha fiú, $n=3$ ill. $n=4$, ha leánygyermekek férfi, ill. női rokonait vizsgáljuk, továbbá

$$i = 4(R-1)+n.$$

Nyilván $1 \leqq i \leqq 24$. Jelölje m_i, ill. M_i a vizsgált betegek összes, ill. rendellenes rokonainak számát az i-edik esetben. Mint láttuk, az egyes veleszületett rendellenességekre vonatkozóan a következő adatok állnak rendelkezésünkre: $p_1, p_2, \tilde{p}_{2R-1}, \tilde{p}_{2R}, m_i, M_i$. Megjegyezzük, hogy az $R=3, 4, 5, 6$ rétegekhez tartozó esetekben alesetek is lehetnek, ha a kétpetéjű ikrek, az apák és az anyák testvérei, a fiú- és leánytestvérek gyermekei, az apai és anyai és/vagy a „nagybácsi"- és „nagynéni"-unokatestvérek külön szerepelnek.

Az adatokból a h^2-et a *maximum likelihood módszerrel* becsüljük. Az egyszerűség kedvéért a *binomiális eloszlást* a *Poisson-eloszlással* közelítjük és feltesszük, hogy a különböző betegek rokonai között végzett megfigyelések függetlenek (tehát elhanyagoljuk a különböző gyerekek rokonai között esetleg fennálló rokoni kapcsolatokat).

Legyen

$$p_i(r) = P_r\left(\tilde{T}_{2R-i(\mathrm{mod}\,2)},\ T_{1+\mathrm{entier}\left(\frac{i-1}{2}\right)(\mathrm{mod}\,2)}\right),$$

$$\lambda_i(x) = m_i p_i(0{,}5^d x),\quad v_i(x) = \log \frac{\lambda_i(x)^{M_i} e^{-\lambda_i(x)}}{M_i!}.$$

Jelölje I az i-k egy olyan összességét, amely mellett a h^2-et becsülni kívánjuk, akkor a *likelihood függvény* I-re vonatkozólag

$$G_I(x) = \sum_{i \in I} v_i(x).$$

Jelöljük ML_I-vel a likelihood függvény maximumát és legyen $\hat{h}_I^2$ az a szám, ahol ez a maximum felvétetik (ez nyilván a megfelelő h_I^2 becslése). Az ehhez tartozó h_{IA}^2, h_{IF}^2 alsó és felső konfidenciahatárt a

$$G_I(x) = ML_I - \frac{1}{2}\chi_1^2(0{,}95) \tag{3}$$

egyenletből határozzuk meg.

Ha $I=\{i\}$ egyelemű, akkor

$$G_I(x) = v_i(x), \tag{4}$$

ez pedig

$$\lambda_i(x) = M_i$$

esetén maximális. Ebből következik, hogy

$$ML_I = ML_i = \log \frac{M_i^{M_i} e^{-M_i}}{M_i!},$$

$\hat{h}_I^2 = \hat{h}_i^2$ pedig a (4) egyenletből számolható.

A maximumkeresést és az egyetletmegoldást a program a következőképpen végzi: a (3) egyenlet megoldásánál $x=\hat{h}_I^2$-től mindkét irányban, a (4) egyenlet megoldásánál — mivel $\lambda_i(x)$ x-ben monoton növő $-x=0{,}6\cdot 2^d$-től lefelé, $I=\bigcup_{k=1}^{k^*} I_k$ melletti maximumkeresés esetén $\min_k \hat{h}_{I_k}^2$ és $\max_k \hat{h}_{I_k}^2$ között 0,1 lépésközzel lépegetve „durva" megoldást ill. maximumot keres, majd az ezen értékek köré vont 0,2 átmérőjű intervallumot 0,01 lépésközzel „végigjárva" finomítja őket.

A különböző h^2-becslések azonosságát a

$$h_I^2 \equiv h^2$$

hipotézis tesztelésével ellenőrizzük. $I = \bigcup_{k=1}^{k^*} I_k$ esetén

$$F_I = 2\left(\sum_{k=1}^{k^*} ML_{I_k} - ML_I\right)$$

a próbafüggvény, amely (k^*-1) szabadságfokú χ^2-eloszlást követ.

Ha $I=\{i\}$ olyan, hogy alesetek is vannak, akkor az egyes alesetekre a program a megfelelő $\hat{h}^2$, h_A^2, h_F^2 és ML értékeket pontosan úgy számolja ki, mint az eseteknél, az alesetek összevonását viszont másképpen végzi: az i-edik esethez tartozó alesetek $m_i^{(v)}$ és $M_i^{(v)}$ értékeinek összeadása révén előálló m_i, M_i értékekből $\hat{h}_i^2$, h_{iA}^2, h_{iF}^2 és ML_i értékét úgy számolja ki, mintha nem lennének alesetek. Az alesetek örökölhetőségi együtthatóinak összehasonlításához jelöljük S_1-gyel, S_2-vel, S_3-mal az alesetek ML, $m_i^{(v)}$, $M_i^{(v)}$ értékeinek összegét, akkor

$$F_i = 2\left\{S_1 - \sum_v \log \frac{(m_i^v)^{M_i^{(v)}}}{M_i^{(v)}!} - \left[ML_i - \log \frac{S_2^{S_3}}{S_3!}\right]\right\}$$

a próbafüggvény, amely az alesetek számánál 1-gyel kisebb szabadságfokú χ^2-eloszlást követ.

Rátérünk a program outputjának ismertetésére. Mint már említettük, egy betegség egy oldal (1. Táblázat). Ennek első része hierarchikus, amely aleset-, eset-, réteg-, rokonsági fok-sorokból és egy totál-sorból áll. Az aleset-sorok és az aleseteket nem tartalmazó esetek sorai a rokon-kártyák adataival kezdődnek (százezrelék helyett ezrelékkel), valamint a betegség ezrelékes relatív gyakoriságával a betegek rokonai között. Ezek a pozíciók az aleseteket tartalmazó esetek soraiban részben (a megfelelő m_i, M_i és $(1000 \cdot M_i/m_i)$ értékeket kiírja a program), a többi sorban teljesen üresen maradnak. Ezután valamennyi sorban a megfelelő $\hat{h}^2$, h_A^2, h_F^2, $-ML$ érték kerül kiírásra, majd az aleset-sorok és az aleseteket nem tartalmazó esetek sorainak kivételével a megfelelő F_i, ill. F_I a hozzátartozó szabadságfokkal. A sorok sorrendje a következő:

aleset-sorok
  eset-sor
    .
    .
    .
aleset-sorok
  eset-sor
    réteg-sor
aleset-sorok
    .
    .
    .
    réteg-sor
      rokonsági fok-sor

1. TÁBLÁZAT

S	*P*	*TR*	*DR*	$\bar{S}$	$\bar{P}$	*m*	*M*	*Q*	h^2	h^2_A	h^2_F	$-ML$	χ^2	*DF*
M	11,96	*P*	1	*M*	2,00	422	7	16,59	0,66	0,36	0,93	3,80		
M	11,96	*P*	1	*F*	9,00	422	20	47,39	0,58	0,40	0,76	4,84		
F	39,00	*P*	1	*M*	2,00	1345	9	6,69	0,42	0,15	0,68	4,06		
F	39,00	*P*	1	*F*	9,00	1345	46	34,20	0,56	0,42	0,70	5,68		
		P	*TOTAL*						0,56	0,46	0,65	19,94	1,57	3
M	11,96	*SS*	1	*M*	11,96	125	20	160,00	1,08	0,83	1,31	4,84		
M	11,96	*SS*	1	*F*	39,00	126	20	158,73	0,62	0,39	0,85	4,84		
F	39,00	*SS*	1	*M*	11,96	421	29	68,88	0,82	0,60	1,03	5,22		
F	39,00	*SS*	1	*F*	39,00	398	79	198,49	0,92	0,77	1,08	6,20		
		SS	*TOTAL*						0,87	0,77	0,96	28,76	7,65	3
			1	TOTAL					0,70	0,64	0,77	66,50	17,81	1
M	11,96	*UM*	2	*M*	2,00	683	2	2,93	0,20	−0,59	0,86	2,62		
M	11,96	*UF*	2	*M*	2,00	563	7	12,43	1,08	0,54	1,61	3,80		
						1246	9	7,22	0,72	0,29	1,12	4,06	3,99	1
M	11,96	*UM*	2	*F*	9,00	609	11	18,06	0,44	0,03	0,82	4,26		
M	11,96	*UF*	2	*F*	9,00	601	8	13,31	0,24	−0,21	0,65	3,94		
						1210	19	15,70	0,32	0,05	0,62	4,80	0,44	1
F	39,00	*UM*	2	*M*	2,00	1696	7	4,13	0,48	−0,06	0,98	3,80		
F	39,00	*UF*	2	*M*	2,00	1534	5	3,26	0,32	−0,29	0,87	3,48		
						3230	12	3,72	0,40	−0,00	0,77	4,34	0,16	1
F	39,00	*UM*	2	*F*	9,00	1473	25	16,97	0,48	0,16	0,79	5,06		
F	39,00	*UF*	2	*F*	9,00	1611	55	34,14	1,12	0,85	1,38	5,84		
						3084	80	25,94	0,84	0,65	1,05	6,22	9,00	1
		U	*TOTAL*						0,64	0,50	0,78	29,34	9,93	3
M	11,96	*CMB*	3	*M*	11,96	601	4	6,66	−0,64	−1,69	0,31	3,26		
M	11,96	*CFB*	3	*M*	11,96	544	7	12,87	0,08	−0,82	0,95	3,80		
						1145	11	9,61	−0,24	−0,92	0,39	4,26	1,15	1
M	11,96	*CMB*	3	*F*	39,00	667	58	86,96	1,28	0,83	1,74	5,90		
M	11,96	*CFB*	3	*F*	39,00	544	35	64,34	0,80	0,22	1,30	5,40		
						1211	93	76,80	1,04	0,72	1,41	6,38	2,02	1
F	39,00	*CMB*	3	*M*	11,96	1643	50	30,43	1,52	1,01	2,04	5,76		
F	39,00	*CFB*	3	*M*	11,96	1333	52	39,01	2,00	1,47	2,56	5,80		
						2976	102	34,27	1,76	1,38	2,13	6,26	1,57	1
F	39,00	*CMB*	3	*F*	39,00	1637	128	78,19	1,36	0,97	1,71	6,70		
F	39,00	*CFB*	3	*F*	39,00	1282	172	134,17	2,64	2,24	6,00	6,98		
						2919	300	102,77	1,92	1,70	2,22	7,54	21,70	1
		C	*TOTAL*						1,49	1,32	1,66	74,92	50,28	3
						1+2+3 *TOTAL*			0,78	0,73	0,84	241,64	70,88	2
						P+2+3 *TOTAL*			0,74	0,67	0,81	209,16	84,97	2
						S DIFFERENCE							3,72	1
							MM TOTAL		0,78	0,62	0,94	33,38	16,43	3
							MF TOTAL		0,61	0,49	0,72	15,35	9,86	3
							FM TOTAL		0,77	0,64	0,91	56,76	36,70	3
							FF TOTAL		0,89	0,80	0,97	107,22	81,57	3
							SEX DIFFERENCE						13,56	3

aleset-sorok
.
.
.
rokonsági fok-sor
.
.
.
rokonsági fok-sor
totál-sor.

Azoknál a veleszületett rendellenességeknél, ahol a vizsgált betegek figyelembe vett rokonai közt szerepelnek a testvérek, a totál-sor alatt egy, a testvéreket figyelmen kívül hagyó kvázi totál-sor áll. A két $\tilde{h}^2_{\text{totál}}$ azonossága hipotézisének teszteléséhez a program kiírja az 1 szabadságfokú χ^2 eloszlást követő

$$2(ML_{\text{kvázi totál}} + ML_{\text{testvér}} - ML_{\text{totál}})$$

próbafüggvény értékét és a szabadsági fokot.

A fenti, bizonyos értelemben horizontális hierarchián kívül a program vertikálisan is összevon, amennyiben a veleszületett rendellenesség mindkét nemre jellemző és a nemek nem ömlesztve szerepelnek: nevezzük az egyes rétegek rögzített n-hez tartozó eseteinek összességét az n-edik oszlopnak ($n=1, 2, 3, 4$), akkor a program a rétegekhez teljesen hasonlóan az oszlopokra is számolja és kinyomtatja a megfelelő $\hat{h}^2, h_A^2, h_F^2, -ML$ értékeket. Az oszlopokat a rokonsági fokhoz hasonlóan összevonva újra megkapnánk a totál-sorban szereplő $\hat{h}^2, h_A^2, h_F^2, -ML$ értékeket, de természetesen a program nem számolja ki ezeket mégegyszer. Az oszlopok örökölhetőségi együtthatóinak összehasonlításához viszont számolja és kiírja a 3 szabadságfokú χ^2-eloszlást követő

$$F = 2\left(\sum_{n=1}^{4} ML_{n\text{-edik oszlop}} - ML_{\text{totál}}\right)$$

próbafüggvény értékét és a szabadsági fokot.

A kilenc CICM outputját vizsgálva általános volt, hogy

$$h^2_{\text{testvér}} > h^2_{1.\text{fokú}},\quad h^2_{2.\text{fokú}} > h^2_{1.\text{fokú}},\quad h^2_{3.\text{fokú}} > h^2_{1.\text{fokú}}.$$

2. TÁBLÁZAT

S	*P*	*TR*	*DR*	*S̄*	*P̄*	*m*	*M*	*Q*	$\hat{h}^2$	h_A^2	h_F^2	$-ML$	χ^2	*DF*
M	11,96	*P*	1	*M*	2,00	422	7	16,59	0,66	0,36	0,93	3,80		
M	11,96	*P*	1	*F*	9,00	422	20	47,39	0,58	0,40	0,76	4,84		
F	39,00	*P*	1	*M*	2,00	1345	9	6,69	0,42	0,15	0,68	4,06		
F	39,00	*P*	1	*F*	9,00	1345	46	34,20	0,56	0,42	0,70	5,68		
		P	*TOTAL*						0,56	0,46	0,65	19,94	1,57	3
M	11,96	*SS*	1	*M*	11,96	125	20	160,00	0,82	0,67	0,95	4,84		
M	11,96	*SS*	1	F	39,00	126	20	158,73	0,52	0,34	0,70	4,84		
M	39,00	*SS*	1	*M*	11,96	421	29	68,88	0,68	0,52	0,84	5,22		
F	39,00	*SS*	1	*F*	39,00	398	79	198,49	0,76	0,66	0,87	6,20		
		SS	*TOTAL*						0,72	0,65	0,79	28,06	6,95	3
			1	*TOTAL*					0,66	0,60	0,71	54,40	6,41	1
M	11,96	*UM*	2	*M*	2,00	687	2	2,93	0,20	−1,95	0,72	2,62		
M	11,96	*UF*	2	*M*	2,00	563	7	12,43	0,88	0,47	1,26	3,80		
						1246	9	7,22	0,60	0,27	0,91	4,06	3,99	1
M	11,96	*UM*	2	*F*	9,00	609	11	18,06	0,40	0,03	0,70	4,26		
M	11,96	*UF*	2	*F*	9,00	601	8	13,31	0,20	−0,23	0,57	3,94		
						1210	19	15,70	0,32	0,05	0,54	0,48	0,44	1
F	39,00	*UM*	2	*M*	2,00	1696	7	4,13	0,44	−0,06	0,84	3,80		
F	39,00	*UF*	2	*M*	2,00	1534	5	3,26	0,28	−0,33	0,75	3,48		
						3230	12	3,72	0,36	0,00	0,67	4,34	0,16	1
F	39,00	*UM*	2	*F*	9,00	1473	25	16,97	0,44	0,15	0,70	5,06		
F	39,00	*UF*	2	*F*	9,00	1611	55	34,14	0,96	0,75	1,17	5,84		
						3084	80	25,94	0,76	0,58	0,91	6,22	9,00	1
		U	*TOTAL*						0,56	0,45	0,67	14,94	10,48	3
M	11,96	*CMB*	3	*M*	11,96	601	4	6,66	−0,72					
M	11,96	*CFB*	3	*M*	11,96	544	7	12,87	0,08	−1,03	0,85	3,80		
						1145	11	9,61	−0,24	−1,21	0,37	4,26	1,15	1
M	11,96	*CMB*	3	*F*	39,00	667	58	86,96	1,12	0,77	1,53	5,90		
M	11,96	*CFB*	3	*F*	39,00	544	35	64,34	0,72	0,22	1,17	5,40		
						1211	93	76,80	0,96	0,67	1,26	6,38	2,02	1
F	39,00	*CMB*	3	*M*	11,96	1643	50	30,43	1,36	0,93	1,78	5,76		
F	39,00	*CFB*	3	*M*	11,96	1333	52	39,01	1,76	1,31	2,18	5,80		
						2976	104	34,27	1,52	1,24	1,84	6,26	1,57	1
F	39,00	*CMB*	3	*F*	39,00	1637	128	78,19	1,20	0,91	1,54	6,70		
F	39,00	*CFB*	3	*F*	39,00	1282	172	134,17	2,24	1,97	2,56	6,98		
						2919	300	102,74	1,76	1,53	1,96	7,54	21,70	1
		C	*TOTAL*						1,32	1,18	1,46	78,20	53,56	3
						1+2+3 *TOTAL*			0,71	0,67	0,76	235,78	73,31	2
						P+2+3 *TOTAL*			0,71	0,64	0,77	207,66	79,65	2
						S DIFFERENCE							0,06	1
							MM TOTAL		0,70	0,57	0,82	29,66	12,71	3
							MF TOTAL		0,55	0,45	0,65	22,38	11,19	3
							FM TOTAL		0,69	0,58	0,81	55,84	35,78	3
							FF TOTAL		0,81	0,74	0,87	102,56	76,91	3
						SEX DIFFERENCE							15,68	3

Az első egyenlőtlenség a domináló variancia pozitív voltával megmagyarázható. A másik kettőt az együttes normalitás gyengítésével (csak a peremeloszlásokat feltételezve normálisnak, ld. TUSNÁDY et al., [6]) sikerült az egyenlőség felé közelítenünk (2. táblázat).

5. Genetikai tanácsadás

Az elmúlt 5 évben *Magyarországon* 3540 család esetében történt genetikai tanácsadás. A családok 62%-a veleszületett rendellenesség miatt jelentkezett; az utóbbiak mintegy 70%-a CICM volt. Az esetükben végzett genetikai tanácsadást — amely becslések szerint e családokban harmadára csökkentette a CICM-ek manifesztációját — nagymértékben a GAMT-modell ill. az általunk megadott h^2-értékek tették lehetővé.

IRODALOM

[1] DAY, N. AND HOLMES, L. B., "The incidence of genetic disease in a university hospital population", *Amer. J. Hum. Genet.* **25** (1973) 237.
[2] FISHER, R. A., "The correlation between relatives on the supposition of Mendelian inheritance", *Trans. Roy. Soc. Edinbourgh* **52** (1918) 399.
[3] MCCROY, W. W., "Child health in the United States", *Quart. Rev. Pediat.* **15** (1960) 94.
[4] ROBERTS, D. F., CHAVEZ, J. AND COURT, S. D. M., "The genetic component in child mortality", *Arch. Dis. Childh.* **45** (1970) 33.
[5] SMITH, D. W., "Dysmorphology (teratology)", *J. Pediat.* **69** (1966) 1150.
[6] TUSNÁDY, G., TELEGDI, L. AND CZEIZEL, E., "ML-fitting of multifactorial thereshold models", *Periodica Math. Hung.*, sajtó alatt.
[7] WARKANY, J., „Congenital malformations and pediatrics", *Pediatrics* **19** (1957) 725.
[8] WEINBERG, W., "Mathematische Grundlagen der Probandenmethode", *Z. indukt. Abstamm.-u. Vererb.-L.* **48** (1927) 179.

(Beérkezett: 1978. május 22.)

TUSNÁDY GÁBOR
MTA MATEMATIKAI KUTATÓ INTÉZET
1053 BUDAPEST, V., REÁLTANODA UTCA 13—15.

TELEGDI LÁSZLÓ
MTA SZÁMÍTÁSTECHNIKAI ÉS AUTOMATIZÁLÁSI KUTATÓ INTÉZET
1502 BUDAPEST, XI., KENDE UTCA 13—17.

CZEIZEL ENDRE
ORSZÁGOS KÖZEGÉSZSÉGÜGYI INTÉZET
1097 BUDAPEST, IX., GYÁLI ÚT 2—6.

ON THE GENETIC LAWS OF COMMON ISOLATED CONGENITAL MALFORMATIONS

G. TUSNÁDY, L. TELEGDI and E. CZEIZEL

The genetic investigation of congenital malformations is based on the so called *Gaussian additive multifactorial threshold* (GAMT) model. In the GAMT model the investigated malformation is characterized by a Gaussian variable, called the liability. The correlation coefficient of the liability of relatives is $h^2/2^d$, where h^2 is the heritability and d is the degre of relationship. For the sake of estimating h^2 and fitting the model to the data of genetic family studies, the maximum likelihood method is used.

Alkalmazott Matematikai Lapok **4** (1978) 27—34

MEGJEGYZÉSEK A NORMALIZÁLT BOLYONGÁS TORLÓDÁSI PONTJAINAK HALMAZÁRÓL

BENCZÚR ANDRÁS
Budapest

Legyen $S_n = \sum_{i=1}^{n} \xi_i$, ahol $\xi_i, i = 1, 2, \ldots$ független, azonos eloszlású valószínűségi változók sorozata. Az $n^{-\alpha} S_n$ sorozat torlódási pontjainak halmaza legyen $A(S_n, \alpha)$.

Két extrém eloszlás konstrukciója szerepel a dolgozatban, az egyikre $A(S_n, \alpha)$ 1 valószínűséggel csak a $-\infty, +\infty$ pontokból áll minden $\alpha \geqq 0$-ra, a másikra $A(S_n, \alpha)$ 1 valószínűséggel csak a $-\infty, 0, +\infty$ pontokból áll minden $\alpha > \frac{1}{2}$ esetén.

1. Bevezetés

Fontosnak tartom, hogy röviden ismertessem a dolgozat keletkezésének történetét.

1970-ben a *Matematikai Kutató Intézet* és a *Számítástechnikai Központ* közös valószínűségszámítási szemináriumán RÉVÉSZ PÁL az *oberwolfach*-i valószínűségszámítási kollokviumról tartott beszámolója hívta fel figyelmemet a témára. RÉVÉSZ PÁL szerint egyik legérdekesebb előadás, HARRY KESTEN előadása FELLER [2] VIII. §-ának 24. feladatát juttatta eszembe, melyet KRÁMLI ANDRÁSsal közösen oldottunk meg a *Számítástechnikai Központ Valószínűségszámítási Osztályának* szemináriumán fél évvel azelőtt. Így született meg a dolgozat, mely extrém eloszlásokra ad konstrukciót. A konstrukciót a már említett szemináriumon ismertettem. RÉVÉSZ PÁLtól megkaptam H. KESTEN [1] dolgozatát, melynek megjelenési helyét azóta sem ismerem. A konstrukcióm teljesen független a KESTEN által alkalmazott módszertől, ezért leírtam, s az *MTA III. Osztály Közleményeknek* leadtam. A *Közlemények* megszűnése miatt dolgozatom elkallódott. Minthogy abban az időben álltam át számítástechnikára, nem törődtem a cikk sorsával. Most, hogy kandidátusi vizsgáim miatt felfrissítettem valószínűségszámítási ismereteimet, észrevettem, hogy a dolgozat a valószínűségszámítás alkalmazásának korlátaira mutat rá. Mind a nagy számok törvényei, mind a határeloszlástételek az eloszlásra nézve véges megfigyelés alapján ellenőrizhetetlen feltételekre alapulnak. Az egyetlen feltétel, amit ellenőrizni, illetve biztosítani lehet: a korlátosság. Ekkor viszont a legszigorúbb tételek teljesülnek: a nagy számok erős törvénye és a centrális határeloszlástétel.

A dolgozat, melyet eredeti formájában közlök az alábbiakban, adalék ahhoz, hogy milyen meglepetések érhetnek bennünket, ha a nagy számok törvényét a korlátosság biztosítása nélkül használjuk.

2. A probléma kitűzése és az eredmények megfogalmazása

HARRY KESTEN [1] dolgozatában részletesen vizsgálja a normalizált véletlen bolyongás torlódási pontjainak halmazát. Bevezetésként röviden ismertetjük fontosabb eredményeit, melyek e cikk megírását inspirálták.

Legyen $S_n = \sum_{i=1}^{n} \xi_i$, ahol $\xi_i - k$, $i = 1, 2, \ldots$ független, azonos eloszlású valószínűségi változók. Az $n^{-\alpha} S_n$ sorozat torlódási pontjainak halmaza legyen

$$A(S_n, \alpha) = \bigcap_m \overline{\{n^{-\alpha} S_n : n \geqq m\}},$$

ahol a felülvonás a halmaz lezártját jelöli.

KESTEN megmutatja, hogy

2.1. TÉTEL. Ha $\gamma(n) \to \infty$, akkor létezik egy olyan nem véletlen zárt halmaz,

$$B = B(F, \{\gamma(n)\}),$$

hogy $\gamma(n)^{-1} S_n$ torlódási pontjainak halmaza 1 valószínűséggel B-vel egyenlő, ahol F a ξ_i valószínűségi változók eloszlásfüggvénye.

2.2. TÉTEL. Ha $F(0) - F(-0) < 1$, $0 < \alpha < \frac{1}{2}$ és $n^{-\alpha} S_n$-nek egy valószínűséggel van véges torlódási pontja, akkor $A(S_n, \alpha)$ minden valós számot tartalmaz.

2.3. TÉTEL. Tetszőleges zárt $C \subset \bar{R}$ halmazhoz, amely tartalmazza a $-\infty$ és $+\infty$ pontokat, található olyan F eloszlás, amelyre

$$B(F, \{n^{-1}\}) = C.$$

Ezen két utóbbi tétel kapcsán vetődik fel a kérdés, és veti is fel KESTEN dolgozatában, melyek a $B(F, \{n^{\alpha}\})$ halmaz lehetséges struktúrái $\alpha > \frac{1}{2}$, $\alpha \neq 1$ esetén. Ehhez a problémához járul hozzá jelen dolgozat két eloszlásfüggvény konstrukciójával, melyekre az $n^{-\alpha} S_n$ sorozat torlódási pontjainak halmaza

a) tetszőleges $\alpha \geqq 0$-ra csak a $-\infty$ és $+\infty$ pontokból áll,

b) tetszőleges $\alpha > \frac{1}{2}$ esetén csak a $-\infty$, 0 és $+\infty$ pontokból áll.

Az utóbbi példa mutatja, hogy a 2.2. tétel állítása $\alpha > \frac{1}{2}$ esetén nem teljesül.

3. Bizonyítás

A konstrukcióhoz az alapötletet FELLER [2] könyvének a VIII. §-ában található 24. feladata szolgáltatta. Jelölje $F^{n*}(x)$ az $F(x)$ eloszlásfüggvény önmagával való n-edik konvolúcióját. Az említett feladat a következőt állítja:

Lehetséges, hogy minden x-re

$$\limsup_{n \to \infty} F^{n*}(x) = 1 \quad \text{és} \quad \liminf_{n \to \infty} F^{n*}(x) = 0.$$

A feladat megoldásához útmutatás található:
Tekintsük a ξ valószínűségi változót, melynek eloszlása

$$P(\xi = (-1)^k a_k) = p_k.$$

Alkalmasan választva az a_k, p_k és n_k konstansokat, elérhető, hogy közel egy valószínűséggel a $\xi_1, \ldots, \xi_{n_k}$ sorozat legalább egy tagja egyenlő lesz $(-1)^k a_k$-val, és egy sem lesz abszolút értékben nagyobb a_k-nál.

Ekkor nyilván $|S_{n_k}| \geqq a_k - n_k a_{k-1}$, ahol

$$S_{n_k} = \sum_{i=1}^{n_k} \xi_i.$$

FELLER megoldásként az

$$n_k = (2k)!, \quad p_k \sim \frac{1}{(2k-1)!}, \quad a_k \sim (n_k)^k \tag{3.1}$$

konstrukciót adja.

Könnyen ellenőrizhető, hogy annak valószínűsége, hogy $\xi_1, \ldots, \xi_{n_k}$ közül legalább egy $(-1)^k a_k$-val egyenlő, az 1-hez konvergál $k \to \infty$ esetén, hiszen

$$1-(1-p_k)^{n_k} \sim 1-\left(1-\frac{1}{(2k-1)!}\right)^{2k!} \sim 1-e^{-2k}.$$

Annak valószínűsége pedig, hogy egyik ξ_i sem nagyobb abszolút értékben a_k-nál,

$$1-\left(1-\sum_{i=k+1}^{\infty} p_i\right)^{n_k} \sim 1-\left(1-\frac{1}{(2k+1)!}\, C_k\right)^{2k!} \sim 1-e^{-\frac{C_k}{2k+1}}$$

ami $C_k \to 1$ miatt valóban 0-hoz konvergál. Ebből a feladat állítása könnyen következik, csak annyit kell megjegyezni, hogy

$$a_k - n_k a_{k-1} \to \infty, \quad \text{ha} \quad k \to \infty.$$

A bizonyítás során azt mutattuk meg, hogy a „kritikus" n_k értékeknél az S_{n_k} összeg közel 1 valószínűséggel csak igen nagy abszolút értékű pozitív vagy negatív értékeket vehet fel. Ennél több is igaz az S_n sorozatra: 1 valószínűséggel nincs véges torlódási pontja, sőt, tetszőleges α esetén az $n^{-\alpha} S_n$ sorozat torlódási pontjainak halmaza a $-\infty$, $+\infty$ pontokból áll.

A bizonyítás során a torlódási pont definícióját a következő alakban fogjuk használni:
a b (nem feltétlenül véges) pont akkor és csak akkor torlódási pontja az $n^{-\alpha} S_n$ sorozatnak, ha a b pont tetszőleges U_b nyílt környezetére $\bigcap_m \{n^{-\alpha} S_n\}_{n=1}^{\infty} \cap U_b$ nem üres, amit a 2.1. tétel alapján $P\{\bigcap_m \{n^{-\alpha} S_n\}_{n=1}^{\infty} \cap U_b = \emptyset\} = 0$ alakban is kimondhatunk.

Térjünk most rá állításunk bizonyítására. Vezessük be az

$$A_k = \{|\xi_i| < a_k, \ i = 1, 2, \ldots, n_k\}$$

eseménysorozatot. Mint már mutattuk

$$P(A_k) = \left(1-\sum_{i=k}^{\infty} p_k\right)^{n_k} = \left(1-\frac{1}{(2k-1)!}\, C_{k-1}\right)^{(2k)!} \sim e^{-kC_{k-1}}.$$

Tehát a $\Sigma P(A_k)$ sor konvergens, így a *Borel—Cantelli lemma* alapján az A_k események közül csak véges sok következik be 1 valószínűséggel. Legyen most x és α adott. Megmutatjuk, hogy megadható olyan, 1 valószínűséggel véges $N(\omega)$ valószínűségi változó, hogy $n>N(\omega)$ esetén $|n^{-\alpha}S_n(\omega)|>x$. Legyen adott ω-ra K olyan nagy, hogy $k>K$ esetén $n_{k-1}^{-\alpha}(a_k-n_{k+1}a_{k-1})>x$ legyen és egyetlen A_k esemény se következzen be. Ilyen K megadható, hiszen

$$n_{k+1}^{-\alpha}(a_k-n_{k+1}a_{k-1}) \sim (2k+2)!^{-\alpha}\left((2k)!^k-(2k+2)!\,(2k-2)!^{k-1}\right) =$$

$$= (2k!)^{k-\alpha}[(2k+1)(2k+2)]^{-\alpha}\left(1-\frac{(2k+1)(2k+2)}{(2k(2k-1))^{k-1}}\right) \to \infty,$$

$k\to\infty$ esetén. Az így választott K-hoz tartozó n_K lesz a megfelelő $N(\omega)$. Valóban, hiszen $n>N(\omega)$ esetén, ha

$$n_k < n \leqq n_{k+1} \quad \text{és} \quad \max_{i=1,\ldots,n} |\xi_i| = a_L,$$

akkor

$$|n^{-\alpha}S_n(\omega)| \geqq (a_L-n_{k+1}a_{L-1})n_{k+1}^{-\alpha} \geqq (a_k-n_{k+1}a_{k-1})n_{k+1}^{-\alpha} \geqq$$
$$\geqq (a_K-n_{K+1}a_{K-1})n_{K+1}^{-\alpha} > x,$$

mert az A_k esemény nem következik be, s így $a_L \geqq a_k$.

Ezzel beláttuk, hogy véges torlódási pontja nem lehet az $n^{-\alpha}S_n$ sorozatnak. Az, hogy $-\infty$ és $+\infty$ is torlódási pontok, a *Feller feladat* bizonyításából következik.

Mielőtt rátérnénk a második példára, más formába írjuk az előbbi konstrukciót. Bontsuk fel a ξ valószínűségi változót független komponensekre, úgy, hogy az a_k értékeket az egyes komponensek vegyék fel a megfelelő p_k valószínűséggel. Legyen tehát

$$\xi = \sum_{i=1}^{\infty} \zeta^{(i)},$$

ahol $\zeta^{(i)}-k$ $(i=1, 2, \ldots)$ egymástól független valószínűségi változók, amelyek eloszlása

$$P\{\zeta^{(k)} = 0\} = 1-p_k, \quad P\{\zeta^{(k)} = (-1)^k a_k\} = P_k.$$

A (3.1) alatti p_k és a_k értékek figyelembevételével könnyen belátható, hogy a $\sum_{i=1}^{\infty} \zeta^{(i)}$ sor 1 valószínűséggel konvergens (csak véges sok $\zeta^{(i)}$ vesz fel zérustól különböző értéket) s a ξ változó eloszlásának megváltozása nem befolyásolja az S_n sorozat viselkedését. A konstrukció lényegét ezzel az átírással a következő módon lehet szemléltetni: a kritikus n_k értéktől kezdve az S_n összegben a $\zeta^{(k)}$ valószínűségi változó dominál, egészen addig, míg a következő, $\zeta^{(k+1)}$ valószínűségi változó csak zérus értéket vesz fel.

A következő példa is ezen alapszik, csak most a $\zeta^{(i)}$ változók eloszlása olyan lesz, hogy igen nagy abszolút értékű pozitív és negatív értékeket is felvesz $\zeta^{(i)}$, várható értéke 0 lesz, s amikor az S_n összegben valamilyen $n\in I$ intervallumon éppen $\zeta^{(k)}$ a domináns tag, előfordul $\sum_{j=1}^{n} \zeta_j^{(k)}=0$, de ez csak akkor következik már be, mikor a $j>k$ indexű $\zeta^{(j)}$-k járuléka az $n^{-\alpha}$-val való normálás miatt már nagy valószínűséggel igen kicsi.

Térjünk rá a konstrukcióra.

Legyen $\xi=\sum_{i=1}^{\infty}\zeta^{(i)}$, ahol $\zeta^{(k)}$-k független valószínűségi változók, a következő eloszlással

$$
\begin{aligned}
&P\{\zeta^{(k)}=(-1)^k a_k\}=p_k,\\
&P\{\zeta^{(k)}=(-1)^{k+1}a_k m_k\}=q_k=p_k m_k^{-1} \quad (m_k \text{ egész szám!}),\\
&P\{\zeta^{(k)}=0\}=1-p_k-q_k.
\end{aligned}
\tag{3.2}
$$

A továbbiakban tekintsük $\{\zeta_j^{(i)}\}$ valószínűségi változóknak egy teljesen független rendszerét, ahol minden rögzített i-re és j-re $\zeta_j^{(i)}$ a (3.2) eloszlást követi.

Legyen $\xi_j=\sum_{i=1}^{\infty}\zeta_j^{(i)}$.

Az a_k, p_k és q_k értékeket úgy kell megválasztani, hogy megadható legyen az

$$n_2<N_2<M_2<n_3<\ldots<n_k<N_k<M_k<\ldots$$

sorozat a következő tulajdonságokkal:

$$P\{A_k\}=P\left\{\text{minden } n\text{-re, ha } M_k\leqq n<n_{k+1},\ \sum_{i=1}^{n}\zeta_i^{(k)}\neq 0\right\}\leqq\frac{1}{2^{k!}},$$

$$P\{B_k\}=P\{\text{minden } n\text{-re, ha } n<N_k,\ \zeta_n^{(k)}=0\}\leqq\frac{1}{2^{k!}},$$

$$P\{C_k\}=P\left\{\sup_{n>M_{k+1}}\left|\frac{1}{\frac{1}{n^2}+\frac{1}{2^{(k+1)!}}}\sum_{i=1}^{n}\sum_{j=1}^{k}\zeta_i^{(j)}\right|>\frac{1}{2^{k!}}\right\}\leqq\frac{1}{2^{k!}},$$

$$P\{D_k\}=P\{\text{legalább egy } \zeta_n^{(j)}=0,\ n<n_{k+1},\ j\geqq k+1 \text{ esetén}\}\leqq\frac{1}{2^{k!}},$$

$$P\{E_k\}=P\{\text{legalább egy } \zeta_n^{(k)}=(-1)^{k+1}a_k m_k,\ n<M_k \text{ esetén}\}\leqq\frac{1}{2^{k!}}$$

s ezen kívül még $a_k=N_{k+1}^k a_{k-1}m_{k-1}$ teljesüljön.

Nézzük meg először, ha megadható ilyen $\{\xi_n\}$ valószínűségi változó sorozat, hogyan bizonyítható rá az állítás. Azt kell megmutatni, hogy az így konstruált $\{\xi_n\}$ sorozatra fennáll minden $\varepsilon>0$, $\alpha>1$ és $K>0$ esetén

(I) $$P\left\{\lim_{n\to\infty}\inf|n^{-\alpha}S_n|>\varepsilon\right\}=0,$$

(II) $$P\{\limsup n^{-\alpha}S_n<K\}=0,$$

(III) $$P\{\liminf n^{-\alpha}S_n>-K\}=0,$$

(IV) Tetszőleges $a\neq 0$-ra létezik δ, hogy

$$P\left\{\lim_{n\to\infty}\inf|n^{-\alpha}S_n-a|>\delta\right\}=1.$$

Bizonyítás: Az A_k, B_k, C_k, D_k és E_k események valószínűségeit úgy választottuk meg, hogy a $\sum P(A_k)$, $\sum P(B_k)$, $\sum P(C_k)$, $\sum P(D_k)$ $\sum P(E_k)$ sorok konvergáljanak. A *Borel—Cantelli lemma* szerint az A_k, B_k, C_k, D_k és E_k események közül 1 valószínűséggel csak véges sok következik be egyszerre. Az S_n sorozat $S_n(\omega)$ realizációjához megadható tehát olyan, 1 valószínűséggel véges $K(\omega)$, hogy $k>K(\omega)$ esetén az A_k, B_k, C_k, D_k és E_k események komplementerei következnek be. Adott α, ε és a-hoz $\left(\varepsilon<\frac{|a|}{4}\right)$ legyen K_1 olyan nagy, hogy

$$K_1>\alpha, \quad \frac{1}{2}+\frac{1}{2^{K_1!}}<\alpha, \quad \frac{1}{2^{K_1!}}<\varepsilon \quad \text{és}$$

$$N_{K_1+1}^{-\alpha}|a_{K_1}-N_{K_1}a_{K_1-1}m_{K_1-1}|>2|a|$$

teljesüljön.

(I) bizonyításához megmutatjuk, hogy tetszőleges $k>\max\left(K(\omega), K_1\right)$ esetén van olyan n, hogy $M_k<n<n_{k+1}$ és $|n^{-\alpha}S_n|<\varepsilon$. Az A_k komplementerének fennállása miatt ugyanis van olyan n, amelyre $\sum_{i=1}^{n}\zeta_i^{(k)}=0$, C_k komplementere miatt a k-nál kisebb indexű $\zeta^{(j)}$-k összegére

$$\left|\frac{1}{n^{\frac{1}{2}}+2^{-k!}}\sum_{i=1}^{n}\sum_{j=1}^{k-1}\zeta_i^{(j)}\right|<\left|\frac{1}{n^{\alpha}}\sum_{i=1}^{n}\sum_{j=1}^{k-1}\zeta_i^{(j)}\right|<\varepsilon,$$

míg a nagyobb indexű $\zeta^{(j)}$-k mindegyike 0-val egyenlő D_k komplementerének fennállása miatt, tehát $|n^{-\alpha}S_n|<\varepsilon$ valóban fennáll. (II) és (III) bizonyításához megmutatjuk $K=2|a|$ választással, hogy tetszőleges $k>\max\left(K(\omega), K_1\right)$ esetén minden n-re N_k és M_k között $(-1)^k n^{-\alpha}S_n>2|a|$. A B_k és E_k események komplementerei teljesülnek, tehát valamely $m<n$-re $\zeta_m^{(k)}=(-1)^k a_k$, de sem $\zeta_i^{(k)}=(-1)^{k+1}a_k m_k$, sem $\zeta_i^{(j)}\neq 0$, $j>k$ nem áll fenn semmilyen $i<n$-re, ezért

$$n^{-\alpha}(-1)^k S_n>(a_k-na_{k-1}m_{k-1})n^{-\alpha}>N_{K_1+1}^{-\alpha}(a_{K_1}-N_{K_1+1}a_{K_1-1}m_{K_1-1})>2|a|.$$

A (IV) tulajdonság megmutatásánál több esetet kell megkülönböztetni. Ha $k>$ $>\max\left(K_1, K(\omega)\right)$ és $N_k<n<M_k$ teljesül, a (II) és (III) bizonyítása során megmutattuk, hogy $|n^{-\alpha}S_n|>2|a|$, tehát itt a (IV)-beli egyenlőtlenség a $\delta<|a|$ választással teljesül.

Ha $k>\max\left(K(\omega), K_1\right)$ és $M_k\leqq n<N_{k+1}$, akkor vagy $\sum_{i=1}^{n}\zeta_i^{(k)}=0$, vagy $\sum_{i=1}^{n}\zeta_i^{(k)}\neq$ $\neq 0$. Az első esetben (I) bizonyításából kiderül, hogy $|n^{-\alpha}S_n|<\varepsilon$, hacsak $\zeta_i^{(k+1)}\neq 0$ nem teljesül valamely $i\leqq n$-re, ami az előző $N_{k+1}\leqq n\leqq M_{k+1}$ esetet jelenti lényegében.

A második esetben $\sum_{i=1}^{n}\zeta_i^{(k)}=l\cdot a_k$, ahol $l\neq 0$ egész szám. Ekkor — feltéve $\zeta_i^{(k+1)}=$ $=0$-t $i\leqq n$ esetén —

$$|n^{-\alpha}S_n|\geqq n^{-\alpha}|a_k-na_{k-1}m_{k-1}|\geqq 2|a|,$$

tehát $\delta<\frac{|a|}{2}$ választással minden n-re $(n>M_{K_1})$ teljesül az $|n^{-\alpha}S_n-a|>\delta$ egyenlőtlenség. Ezek után már csak a megfelelő p és q értékek megadása van hátra. A konstrukciót rekurzióval végezzük. Legyen $p_1=q_1=\frac{1}{2}$, továbbá n_1, N_1 és M_1-et válasz-

szuk 1-nek. Megadható a $P\{A_1\}$-re vonatkozó egyenlőtlenségnek eleget tevő n_2 érték. Ezek után válasszuk N_2-t olyan nagynak, hogy p_2-re nézve megadható legyen a

$$1-(1-p_2)^{n_2} \leqq \frac{1}{2^{3!}}, \quad (1-p_2)^{N_2} \leqq \frac{1}{2^{3!}}$$

egyenlőtlenségrendszer. Így megkaptuk N_2-t és p_2-t. N_2 ismeretében $a_1 = N_1$ legyen. Most már ismerjük $\zeta^{(1)}$ eloszlását, így a C_1 esemény definíciójában szereplő M_2 értéket megadhatjuk. (M_2 létezését a *Hajek—Rényi egyenlőtlenségből* lehet bizonyítani, lásd GNYEGYENKO [3]). Válasszuk most a q_2 értékét úgy, hogy eleget tegyen az $1-(1-q_2)^{M_2} < \frac{1}{2^{2!}}$ egyenlőtlenségnek, azaz

$$P\{E_2\} \leqq \frac{1}{2^{2!}}$$

legyen. Megadtuk már n_2, N_2, M_2, p_2 és q_2 értékét, ezek segítségével meghatározhatjuk n_3, N_3, a_2, M_3, p_3 és q_3 értékét. Általában, ha már ismert n_k, N_k, M_k, p_k és q_k, a következő módon juthatunk tovább:

p_k és q_k meghatározza a $P(A_k)$-ra vonatkozó egyenlőtlenségeknek eleget tevő n_{k+1}-et. n_{k+1} meghatározása után N_{k+1}-et és p_{k+1}-et az

$$1-(1-p_{k+1})^{n_{k+1}} \leqq 2^{-(k+2)!}, \quad (1-p_{k+1})^{N_{k+1}} \leqq 2^{-(k+2)!}$$

egyenlőtlenségrendszer megoldása adja. Legyen

$$a_k = N_{k+1}^k a_{k-1} \cdot m_{k-1}.$$

Ezzel megadjuk $\zeta^{(k)}$ eloszlását, s a *Hajek—Rényi egyenlőtlenség* alapján megadhatjuk a C_k esemény definíciójában szereplő M_{k+1} értéket. Most már csak q_{k+1} meghatározása van hátra, amit az

$$1-(1-q_{k+1})^{M_{k+1}} \leqq 2^{-(k+1)!}$$

egyenlőtlenség megoldásával nyerünk, ügyelve arra, hoyg $\frac{p_{k+1}}{q_{k+1}} = m_{k+1}$ egész szám legyen. Be kell még látnunk az A_k, B_k, C_k, D_k és E_k-ra vonatkozó egyenlőtlenségeket. Egyedül $P(D_k) \leqq 2^{-k!}$ szorul bizonyításra, mert a többi esetben n_k, N_k, M_{k+1} és q_k választása úgy történt, hogy $P(A_k)$, $P(B_k)$, $P(C_k)$ és $P(E_k)$ mind $2^{-k!}$-nál kisebb legyen. Mivel

$$P(D_k) \leqq \sum_{j=k+1}^{\infty} 1-(1-p_j)^{n_{k+1}} \leqq \sum_{j=k+1}^{\infty} 1-(1-p_j)^{n_j} \leqq \sum_{j=k+1}^{\infty} 2^{-(j+1)!} < 2^{-k!},$$

a konstrukció megfelelő voltát beláttuk.

Megjegyzések: Mindkét konstrukció végrehajtható n^α helyett tetszőleges $\gamma_n \to \infty$ sorozat esetén. Pontosabban: tetszőleges $\gamma_n \to \infty$ sorozathoz megadható olyan eloszlás, hogy a megfelelő $\gamma_n^{-1} S_n$ sorozat torlódási pontjainak halmaza csak a $-\infty$ és $+\infty$ pontokból áll; és ha még valamely $\delta > 0$-ra fennáll, $\gamma_n n^{-\frac{1}{2}-\delta} \to \infty$, akkor megadható olyan eloszlás is, amelyre a $\gamma_n^{-1} S_n$ sorozat torlódási pontjainak halmaza a $-\infty$, ∞ és 0 pontokból áll.

A konstrukciók során nem használtuk ki a 2.1. tételt; a torlódási pontok halmazát a $K(\omega)$ valószínűségi változó segítségével realizációnként adtuk meg.

Befejezésül egy problémát említünk: Igaz-e, hogy ha az $n^{-\alpha}S_n$ sorozatnak van legalább két torlódási pontja, akkor a $-\infty$ és $+\infty$ is torlódási pontok? $\alpha \leqq \frac{1}{2}$ esetén igaz az állítás, mi mondható $\alpha > \frac{1}{2}$-re? Igaz-e ugyanez az állítás tetszőleges $\gamma_n \to \infty$ sorozatra? (Az $\alpha < \frac{1}{2}$ esetén az állítás a 2.2. tételből következik, $\alpha = \frac{1}{2}$-re C. STONE [4] cikkéből következik.)

IRODALOM

[1] KESTEN, H., "The limit points of a normalized random walk", kézirat.
[2] FELLER, W., *An Introduction to Probability Theory and its Applications, Vol. II.* (John Wiley, New York, 1966).
[3] GNYEGYENKO, B. V., *Kursz tyeorii verojatnosztyej,* (FIZMATGIZ, Moszkva, 1961).
[4] STONE, C., "The growth of a random walk", *Ann. Math. Stat.* **40** (1969).

(Beérkezett: 1978. június 19.)

BENCZÚR ANDRÁS
MTA SZÁMÍTÁSTECHNIKAI ÉS AUTOMATIZÁLÁSI KUTATÓ INTÉZET
1132 BUDAPEST, VICTOR HUGO U. 18.

REMARKS ON THE LIMIT POINTS OF A NORMALIZED RANDOM WALK

A. BENCZÚR

Let $\{\xi_i\}_{i=1}^{\infty}$ be a sequence of independent, identically distributed random variables, and $S_n = \sum_{i=1}^{n} \xi_i$. The (random) set of accumulation points of $n^{-\alpha}S_n$ will be denoted by $A(S_n, \alpha)$.

In this paper we give two extrem distributions, for the first one $A(S_n, \alpha)$ consists only of points $-\infty$ and $+\infty$ with probability 1 for all $\alpha \geqq 0$, for the second one $A(S_n, \alpha)$ consists only of points $-\infty$, 0 and $+\infty$ with probability 1 for all $\alpha > \frac{1}{2}$.

Alkalmazott Matematikai Lapok 4 (1978) 35—94

MONTE CARLO MÓDSZEREK A TÖBBDIMENZIÓS TÉRBEN ELHELYEZKEDŐ HALMAZOK VALÓSZÍNŰSÉGÉNEK MEGHATÁROZÁSÁRA NORMÁLIS ELOSZLÁS ESETÉN[1]

DEÁK ISTVÁN
Budapest

Bevezetés

Sztochasztikus programozási modellek optimalizálási eljárása folyamán gyakran van szükség n-dimenziós halmazok valószínűségének kiszámítására. Hasonló valószínűségeket kell meghatározni egy sor többváltozós statisztikai problémában is. A gyakorlatban fellépő valószínűségeloszlások közül jelentőségben kiemelkedik a többdimenziós normális eloszlás.

Normális eloszlással kapcsolatos valószínűségek kiszámításának problémájával szerző a *Prékopa-féle* STABIL [102] sztochasztikus programozási modell számítógépes megoldása során találkozott. Mivel a gyakorlatban magasabb ($n \geqq 10$) dimenziós esetek kiszámítására is szükség van, így ezt a numerikus integrálási feladatot — dimenziószámra csak kevéssé érzékeny — *Monte Carlo módszerrel* akartuk megoldani. Több próbálkozás után [29], [30] közöltük az első olyan algoritmust [31], amely segítségével a valószínűségeket $n=50$ dimenzióig elfogadható pontossággal és a gyakorlat szempontjából megfelelő rövid idő alatt ki lehet számítani. Tudomásunk szerint ez volt az első olyan publikáció, amely $n>5$ dimenzióra működő algoritmust adott.

A dolgozatban több olyan algoritmust közlünk, amelyek alkalmasak a többdimenziós normális eloszlással kapcsolatos valószínűségek meghatározására. A többdimenziós tér halmazai valószínűségének kiszámítása gyakran az eloszlásfüggvény kiszámítására vezethető vissza; a közölt algoritmusokat erre az esetre dolgoztuk ki részletesen. Algoritmusaink nehézség nélkül átalakíthatók a többdimenziós t eloszlással kapcsolatos valószínűségek kiszámítására.

Algoritmusok használhatóságáról az algoritmus számítógépes programja és számos feladat megoldása után mondhatunk csak véleményt. Különösen igaz ez *Monte Carlo módszerek* esetén, mivel az algoritmus által megadott eredmény hibáját, valamint a kiszámításhoz szükséges időt általában nem tudjuk előre meghatározni. A jelen munka nagyon sok számítógépes kísérletezésen alapszik, a futási eredmények egy részét az egyes fejezetek végén közöltük.

Az értekezés felépítése a következő. Az 1. Fejezetben összefoglaljuk azokat az általános módszereket, amelyek felhasználásával a számítógépen véletlen mennyiségeket állíthatunk elő. A 2. Fejezetben az n-dimenziós gömbökben egyenletes eloszlású vektorokat generáló algoritmusokat írjuk le; az irodalomban hiányosan közölt algoritmusokat kiegészítettük. A 3. Fejezetben több új algoritmust közlünk stacionárius, n-dimenziós ellipszoidokban egyenletes eloszlású vektorok generálására. Ezek-

[1] A dolgozat, a már publikált részek lerövidítésétől eltekintve, a szerző azonos cím alatt benyújtott kandidátusi értekezésének első változata.

nek a módszereknek a felhasználásával a 4. Fejezetben stacionárius normális vektorok előállítására adunk gyors algoritmusokat.

Az 5. Fejezetben fő témánkkal, a többdimenziós normális eloszlás eloszlásfüggvényének *Monte Carlo kiszámításával* foglalkozunk. Az eddig publikált eredmények összefoglalása mellett új algoritmusokat is közlünk. Ezek segítségével az eddigieknél lényegesen gyorsabban lehet az eloszlásfüggvény konkrét értékeit kiszámítani. Az itt közölt legjobb algoritmusok felhasználásával két tizedesjegyre pontos eredményt kaphatunk $n=10$ dimenzióig 0,1 másodperc alatt, $n=20$ dimenzióig 1 másodperc alatt és $n=50$ dimenzióig 10 másodperc alatt. A sztochasztikus programozásban különösen fontos, egyhez közeli valószínűségek esetén $n=5$ és 15 dimenzió között mintegy tízszer gyorsabban lehet ilyen pontosságú eredményt kapni.

A Függelékben a számítógépes program néhány fontos részletére térünk ki; ezek felhasználásával a program reprodukálható. Megadjuk a véletlenszám generátorok gépi kódos listáját is; ezeknek a gyakran használt részeknek gépi kódban történt megírása nagy mértékben hozzájárult az algoritmusok gyorsaságához.

A használt számítástechnikai és algoritmikus tárgyalási módot az értekezés anyaga kívánta meg. Az algoritmusok leírásánál KNUTH [73] jelöléseit használtuk. A gépi kódban írt véletlenszám generátorok kivételével a programokat FORTRAN nyelven írtuk és a KSH ÁSzSz Honeywell 66/60 számítógépén futtattuk le (kivételt képeznek a 2. Fejezet futási eredményei, amelyeket az ÁSzSz Honeywell 66/20 gépén kaptunk).

Köszönettel tartozom PRÉKOPA ANDRÁSnak, aki felhívta figyelmemet a többdimenziós valószínűségek kiszámításának témakörére, a *Monte Carlo módszerek* fontosságára és munkámban mindvégig segített. Hálás vagyok SZÁNTAI TAMÁSnak és BENE BÉLÁnak, a velük folytatott beszélgetéseim gyümölcsözően hatottak munkámra.

1. FEJEZET

Véletlen számok generálása

1.1. Véletlen számok számítógépeken

Monte Carlo számításokban a különböző eloszlású valószínűségi változókat a realizációikkal helyettesítjük. Adott eloszlást követő számokat algoritmusok segítségével állítunk elő számítógépeken. A generált számokat, amelyeket egy kezdeti érték és az algoritmus teljesen meghatároz, véletleneknek tekintjük. Az előállított mennyiségeket pszeudo-véletleneknek is nevezik, mi azonban az irodalomban elterjedt szóhasználat szerint a véletlen szám *(random number)* kifejezést használjuk. Elektronikus zaj, vagy másmilyen fizikai jelenség felhasználásával is elő lehet valószínűségi változók realizációit állítani [86], [89], de ezeket a gyakorlatban — tudomásunk szerint — nem használják.

Igen fontos feladat olyan algoritmusokat, véletlenszám generátorokat készíteni, amelyek a számítógépeken gyorsan lefutnak, rövid idő alatt képesek egy új számot előállítani egy adott eloszlásból.

A témakör alig harminc éves, az első összefoglaló mű 1954-ben jelent meg (lásd [23]-nál). A számítógépek fejlődésével párhuzamosan újabb és újabb módszereket fejlesztettek ki véletlen számok generálására. Az első algoritmusokat NEU-

MANN [91] javasolta, a további módszerek kifejlesztése főleg MULLER, MARSAGLIA, AHRENS, DIETER és ATKINSON nevéhez fűződik. Az igen gazdag irodalomból — NANCE és OVERSTREET [90] 1972-ben ötszáz cikket és könyvet sorol fel irodalomjegyzékében — a következő műveket lehet kiemelni. HULL és DOBELL [64] 1962-ben megjelent összefoglaló cikke a véletlenszám generátorokat tárgyalja, HAMMERSLEY és HANDSCOMB [59] könyve a *Monte Carlo módszerekről,* alkalmazásokról és a véletlenszám generátorokról ad 1964-ig jó áttekintést. Az eredményeket elméleti oldalról foglalja össze HALTON 1970-ben megjelent cikkében [58]. A legfontosabb eloszlások véletlenszám generátorjait AHRENS és DIETER hasonlították össze számítástechnikai szempontból, eredményeik az [1], [2] és [3] cikkekben jelentek meg. A legújabb módszerek ATKINSON és PEARCE [8] 1976-os cikkében találhatók.

Az algoritmusok leírásában a KNUTH [73] által ajánlott jelölésrendszert követtük. Az $x \leftarrow y$ kifejezés jelenti az y változó értékének átadását az x változónak, ezt FORTRAN-ban az $x=y$ utasítással érhetjük el. A „Generáljuk u-t $g(x)$ sűrűségfüggvénnyel", (u^*-t, x-t, stb.) kifejezés azt jelöli, hogy a megfelelő szubrutin behívásával az u változónak egy $g(x)$ sűrűségfüggvényű véletlen szám értékét adjuk. Amennyiben a sűrűségfüggvény vagy az eloszlás nincsen specifikálva, akkor a [0, 1) intervallumban egyenletes eloszlásból generálunk egy számot. Az „Adjuk át x-t" kifejezés után egy RETURN kilépés következik be az algoritmusból.

A közölt algoritmusok egzaktak abban az értelemben, hogy egy $u[0, 1)$-ben egyenletes eloszlású valószínűségi változót az algoritmus pontosan a kívánt eloszlásúvá transzformál. A gyakorlatban az előállított számok eloszlása az u realizációinak minőségétől és a számítógép szóhosszúságától függ.

1.2. Egyenletes eloszlású számok generálása

Az ismert algoritmusok többsége egyenletes eloszlású számokat transzformál kívánt eloszlásúvá. Ezt a központi helyet az egyenletes eloszlás matematikai leírásának egyszerűsége és az a tény biztosítja, hogy ilyen eloszlású számok állíthatók elő a leggyorsabban.

Több algoritmus ismeretes egyenletes eloszlású számok generálására. Ezek közül a legismertebb a multiplikatív generátor, amely a következő. Legyen u_0 egy [0, 1) intervallumban adott szám, m és c két konstans, ekkor az

$$u_{i+1} = mu_i + c \pmod 1, \quad i = 0, 1, 2, \ldots \tag{1.1}$$

rekurzív összefüggés segítségével előállított u_i számokat [0, 1)-ben egyenletes eloszlásúnak tekintjük. Az m és c konstansok és az u_0 kezdeti érték megválasztására különböző javaslatok ismeretesek [37], [117]. Az (1.1) alakú kifejezést használó multiplikatív generátorok egy kevéssé ismert rossz tulajdonságára mutatott rá MARSAGLIA [81]. A generált számokból alkotott $(u_{j+1}, u_{j+2}, \ldots, u_{j+n})$, szám n-eseket az n-dimenziós tér pontjainak tekintve, ahol a j index n-nek egész számú többszöröse, az összes így előállított pont a tér viszonylag kis számú, egymással párhuzamos hipersíkján van rajta. (Egy 36 bites számítógépen $n=10$ dimenzióban az összes pontok legfeljebb 54 hipersíkon találhatók.) A tétel a kezdeti u_0 érték és az m és c konstansok választásától függetlenül igaz. A hipersík tulajdonságait DIETER [38], [40] vizsgálta.

Számítógépes programjainkban ezért egy eltolásokkal és összeadással véletlen számokat előállító generátort használtunk, melynek leírása és programlistája a Függelékben található.

Az egyenletes eloszlású véletlen számokat generáló algoritmusok közül a következők érdemelnek még említést. MACLAREN és MARSAGLIA [76] két multiplikatív generátor keverését ajánlja; az egyikkel egy 128 elemű tömböt töltünk fel, a másikkal ezek közül egyet véletlenszerűen kiválasztunk, míg az előzőt használjuk a tömb kiemelt elemének újragenerálására. TAUSWORTHE [130] közölt egy, a GF2 felett primitív polinomok elméletén alapuló módszert. Ennek általánosítását LEWIS és PAYNE [74] fogalmazták meg, a téma irodalmából TOOTILL [131], [132] és WHITLESEY [138] cikkét emeljük ki.

1.3. Tetszőleges eloszlású számok generálásának általános módszerei

Három olyan algoritmust írunk le, amelyek segítségével egyenletes eloszlású számokat tetszőleges eloszlásúvá lehet transzformálni.

Legyen az u valószínűségi változó egyenletes eloszlású a [0, 1) intervallumban és $F(x)$ egy eloszlásfüggvény. A

$$\xi = F^{-1}(u) \tag{1.2}$$

valószínűségi változó eloszlásfüggvénye $F(x)$, mivel $P\{\xi<x\}=P\{F^{-1}(u)<x\}=P\{u<F(x)\}=F(x)$. Az (1.2) egyenlőséget felhasználó módszert közvetlen vagy inverziós módszernek nevezzük:

1. Generáljuk u-t.
2. Adjuk át az $x \leftarrow F^{-1}(u)$ számot, mint $F(x)$ eloszlásfüggvényűt.

A módszert a gyakorlatban csak olyan eloszlásokra lehet alkalmazni, amelyek eloszlásfüggvénye invertálható. Példaként tekintsük az $F(x)=1-e^{-x}$ eloszlásfüggvényű exponenciális eloszlást. Ilyen eloszlású véletlen számokat állít elő a következő algoritmus:

1. Generáljuk u-t.
2. Adjuk át az $x \leftarrow -\ln u$ számot, mint exponenciális eloszlásút.

Az algoritmus helyessége az $x=F^{-1}(u)=-\ln(1-u)$ egyenlőségből látható; az $1-u$ változót a vele statisztikailag azonos u-val helyettesítettük.

Az elfogadás-elvetés *(acceptance-rejection)* módszerének leírásában nem törekszünk teljes általánosságra, csak olyan formában fogalmazzuk meg, ahogy használni fogjuk. Egy általánosabb megfogalmazás található például SZOBOL [150] könyvében.

Legyen $f(x)$ egy folytonos sűrűségfüggvény az $[a, b)$ intervallum felett. Legyen $h(x)$ egy másik sűrűségfüggvény, amelyre

$$f(x) \leqq ch(x) = g(x), \quad a \leqq x < b, \tag{1.3}$$

ahol $c \geqq 1$ konstans. Ha $h(x)$ sűrűségfüggvényű véletlen számokat tudunk generálni, akkor a következő algoritmus segítségével $f(x)$ sűrűségfüggvényű számokat generálhatunk.

1. Generáljuk a $h(x)$ sűrűségfüggvényű x számot az $[a, b)$ intervallumban.
2. [Elvetés.] Generáljuk u-t, ha $u \geqq f(x)/g(x)$, akkor menjünk vissza az 1. lépésre.
3. [Elfogadás.] Adjuk át x-t, mint $f(x)$ sűrűségfüggvényűt.

Az algoritmus helyessége a következőképpen látható be. Az elfogadott x értékek eloszlásfüggvénye

$$P\{x < r | u < f(x)/g(x)\} = \frac{P\{x < r,\, u < f(x)/g(x)\}}{P\{u < f(x)/g(x)\}} =$$

$$= \frac{\int_{-\infty}^{r} \int_{0}^{f(x)/g(x)} h(x)\, du\, dx}{\int_{-\infty}^{+\infty} \int_{0}^{f(x)/g(x)} h(x)\, du\, dx} = \frac{\int_{-\infty}^{r} f(x)\, dx}{\int_{-\infty}^{+\infty} f(x)\, dx},$$

mivel u egyenletes [0, 1)-ben és az u és x változók egymástól függetlenek.

A módszer akkor ajánlható, ha $h(x)$ sűrűségfüggvényű változókat könnyen tudunk generálni és ha a $P\{f(x)/g(x) \leqq u\}$ valószínűség — az elvetés valószínűsége kicsi. Az ilyen algoritmusok hatásfokának az elfogadás $P\{u \leqq f(x)/g(x)\}$ valószínűségét nevezik; ez megmutatja, hogy az 1. lépésben $h(x)$ sűrűségfüggvénnyel generált véletlen számot milyen valószínűséggel fogadjuk el.

A kompozíciós módszert BUTLER [23] közölte a következő formában. Legyen $f(x)$ egy sűrűségfüggvény, amely felírható az

$$f(x) = \int_{-\infty}^{+\infty} g_y(x)\, dH(y) \tag{1.4}$$

alakban, g_y egy $H(y)$ eloszlású y paramétertől függő sűrűségfüggvény. Az (1.4) egyenlőséget egy $f(x)$ sűrűségfüggvényű véletlen számot generáló algoritmus készítésére használhatjuk.

1. Generáljuk y-t a $H(y)$ eloszlásfüggvény szerint.
2. Generáljuk x-t a $g_y(x)$ sűrűségfüggvény szerint és adjuk át x-t mint $f(x)$ sűrűségfüggvényűt.

A kompozíciós módszert alkalmazta BUTCHER [21] normális eloszlású, BÁNKÖVI [10], [11] és BÉKÉSSY [13] gamma és béta eloszlású véletlen számok generálására. Diszkrét H eloszlásfüggvényre a felbontás és a generátor egyszerűbbé válik:

$$f(x) = p_1 g_1(x) + p_2 g_2(x) + \ldots + p_k g_k(x), \tag{1.5}$$

az algoritmus pedig a következő:

1. Generáljuk u-t. Ennek segítségével p_i valószínűséggel kiválasztjuk a g_i sűrűségfüggvényt.
2. Generáljuk x-t a $g_i(x)$ sűrűségfüggvény szerint és adjuk át x-t, mint $f(x)$ sűrűségfüggvényűt.

Ezt a módszert minden eloszlásra alkalmazhatjuk. BUTLER [22] adott erre egy algoritmust. Gyors algoritmusok készítéséhez két szempontot kell figyelembe venni.

1. Az 1. pontban a p_i, $i=1, \ldots, k$ diszkrét eloszlásból gyorsan tudjunk mintát venni — ezt például MARSAGLIA *Táblázat módszerével* [79] érhetjük el (lásd a következő pontot).

2. Legyen I az $i=1, \ldots, k$ indexeknek egy részhalmaza és legyen a $\sum_{i \in I} p_i$ valószínűség 1-hez közeli. Az algoritmus gyorsaságának egy szükséges feltétele, hogy a g_i, $i \in I$ sűrűségfüggvények szerint gyorsan tudjunk véletlen számokat generálni. Ezt

például úgy érhetjük el, hogy a g_i, $i \in I$ függvényeket konstansnak választjuk, vagyis a valamilyen $[a_i, b_i)$ intervallumban egyenletes eloszlás sűrűségfüggvényét vesszük. Más szavakkal ezt úgy fogalmazhatjuk, hogy az $f(x)$ sűrűségfüggvényt konstans függvények összegével (egy lépcsős függvénnyel) közelítjük meg, a g_i, $i \notin I$ függvények pedig az $f(x)$ függvénynek a lépcsősfüggvénytől való eltérését tartalmazzák.

A fenti szempontok alapján MARSAGLIA [82] adott egy algoritmust normális eloszlású számok generálására; ez jelenleg a leggyorsabb normális véletlenszám generátor. Ennek az ún. Téglalap—Szél—Vég *(Rectangle—Wedge—Tail)* algoritmusnak két hátránya van: nagyszámú konstans tárolását és bonyolult gépi kódos program írását igényli a számítógépes realizáció. A módszer leírása KNUTH [73] könyvében valamint AHRENS és DIETER [1] cikkében is megtalálható.

1.4. Marsaglia Táblázat módszere diszkrét eloszlású számok generálására

Legyen ξ egy diszkrét eloszlású valószínűségi változó, vagyis $P\{\xi = x_i\} = p_i$, $i = 1, \ldots, k$. Az inverziós módszer alkalmazásával a következő algoritmus adható.

1. Generáljuk u-t és legyen $i \leftarrow 1$.
2. [Vizsgálat.] Ha $p_1 + \ldots + p_i < u$, akkor legyen $i \leftarrow i+1$ és ismételjük meg ezt a lépést.
3. Adjuk át x_i-t.

Az i index meghatározására egyéb kereső eljárások is adhatók, de ezeknél nagy k esetén a számítási munka is nagy, $k \ln k$ nagyságrendű.

Diszkrét eloszlású véletlen számok előállítására MARSAGLIA [79] közölt egy gyors algoritmust. A formális leírás előtt egy számpéldával világítjuk meg az algoritmus működésének főbb lépéseit. Legyen a diszkrét eloszlás a következő valószínűségekkel és értékekkel adott:

valószínűség	érték
0,321	*a*
0,034	*b*
0,212	*c*
0,433	*d*

Ilyen eloszlású változók generálására a leggyorsabb módszer a következő. Tároljuk egy ezer elemű vektor 321 helyén az „*a*" értéket, 34 helyén a „*b*" értéket, 212 helyén a „*c*" értéket és 433 helyén a „*d*" értéket. Generáljunk egy véletlen indexet, azaz generáljuk az $u = 0, d_1 d_2 \ldots, d_z$ véletlen számot, ahol d_j jelöli a j-edik decimális jegyet és legyen az i index a $d_1 d_2 d_3 + 1$ szám. Adjuk át a vektor i-edik elemének tartalmát, mint kívánt eloszlásút. Ez az eljárás egy urnamodellel könnyen magyarázható. Tegyünk egy urnába ezer darab golyót, amelyek közül 321 darabon az „*a*" felírás van, 34 darabon a „*b*" felírás, 212-n „*c*" és 433-on „*d*" felírás. Ezek közül egyet véletlenszerűen kiválasztunk.

A memóriaigényt egy egyszerű módosítással lényegesen le lehet csökkenteni. Legyen három urnánk, amelyek a használt három számjegynek felelnek meg. Az első urnát az első számjegyek összegének megfelelő valószínűséggel választjuk ki, a másodikat a második számjegyek valószínűségével.

	Tartalom	Urna kiválasztásának valószínűsége
1. urna	3 db a, 2 db c, 4 db d,	0,9
2. urna	2 db a, 3 db b, 1 db c, 3 db d,	0,09
3. urna	1 db a, 4 db b, 2 db c, 3 db d,	0,01

A kiválasztást itt két lépésben végezzük el. Először egy urnát választunk ki a megfelelő valószínűséggel, majd ebből az urnából egy golyót emelünk ki véletlenszerűen (egyenletes eloszlás szerint).

Az algoritmus számítógépes realizációjában az előző változat ezer konstansával szemben csak 28 konstans tárolására van szükség. Az egyes urnák tartalmát különböző „szintekre" töltjük be, az egyes szinteket a kezdőcímek határolják el. A részletes leírás előtt megemlítjük, hogy tetszőleges alapú számrendszerben felírható az algoritmus. Mi a kettes számrendszerre adjuk meg a részleteket, mivel az eljárás így foglalja a legkevesebb memóriát, bár több összehasonlításra van szükség egy minta generálásához.

Legyenek adva a $p_i=0, b_{i1}b_{i2} \ldots b_{im}$, $i=1, \ldots, k$ valószínűségek, ahol b_{ij} az i-edik valószínűségben a j-edik bit értéke. A generálandó valószínűségi változó ξ, melyre $P\{\xi=x_i\}=p_i$, $i=1, \ldots, k$.

Előkészítés

1. Számítsuk ki a j-edik bitek $s_j \leftarrow \sum_{i=1}^{k} b_{ij}$, $j=1, \ldots, m$ oszlopösszegét.
2. [Határozzuk meg a t_j szintvalószínűségeket.] Legyen $t_j \leftarrow s_j/2^j$, $j=1, \ldots, m$, és az összegezett szintvalószínűség $t_j \leftarrow t_{j-1}+t_j$, $j=2, \ldots, m$.
3. [Az egyes szintek kezdőcíme c_i lesz.] Legyen $c_1 \leftarrow 0$ és $c_{i+1} \leftarrow c_i+s_i$, $i=1, \ldots, m-1$. Az index módosítók a $d_1 \leftarrow 0$, $d_j \leftarrow t_{j-1}2^j-c_j$, $j=2, \ldots, m$ számok.
4. [Betöltjük az A vektor elemeibe az x_i, $i=1, \ldots, k$ értékeket.] Legyen $q \leftarrow 0$ és $j \leftarrow 0$.
5. Növeljük meg az indexet, legyen $q \leftarrow q+1$.
6. Ezt a lépést az $i=1, \ldots, k$ indexek mindegyikére végezzük el: vizsgáljuk meg a b_{iq} bitet, ha $b_{iq}=1$, akkor $A(j) \leftarrow x_i$, $j \leftarrow j+1$ legyen.
7. Ha $q=m$ akkor vége az előkészítésnek, egyébként menjünk vissza az 5. lépésre.

A generáláshoz csak az A vektort, a t_j összegezett szintvalószínűségeket, valamint a d_j indexmódosítókat kell megőrizni. Megjegyezzük, hogy ha $s_j=0$ valamilyen $j=1, \ldots, r$ indexekre, akkor a megfelelő t_j és d_j értékekre nincs szükség.

Generálás

1. Generáljuk u-t egyenletesen [0, 1)-ben, ahol u bináris alakban $u=0, u_1u_2 \ldots u_m$ és legyen $j \leftarrow 1$.
2. [Szint megállapítása.] Ha $u<t_j$, akkor adjuk át az $A(u_1 \ldots u_j-d_j)$ számot.
3. Legyen $j \leftarrow j+1$ és menjünk vissza 2-re.

A leírt algoritmus igen gyors, ha gépi kódban írjuk meg a programot; csak összehasonlításokat igényel a szint megállapításához és egy adott indexű vektorkomponens memóriából való kiolvasását.

Természetesen a 4-es vagy 8-as alapú számrendszerben felírt valószínűségekből kiindulva egy kevesebb összehasonlítást igénylő algoritmus fogalmazható meg. MARSAGLIA eredetileg 8-as illetve 10-es alapú számrendszerben közölte az algoritmust, AHRENS és DIETER [1] cikkében pedig 4-es számrendszerre van megfogalmazva. Mivel a szükséges memória az A vektor tárolásához $s_1+\ldots+s_m$ szó, így a legkevesebb memóriát a fenti, bináris algoritmus igényli, ilyen formában is használtuk az *Ellipszoid módszerhez.*

1.5. Néhány véletlenszám generátor

Ebben a pontban néhány olyan véletlenszám generátort írunk le, amelyekre a továbbiakban szükségünk lesz. Az egyenletes, a normális és az exponenciális eloszlású véletlen számok generátorjait a Függelékben írjuk le.

Legyen x egy normális eloszlású valószínűségi változó, ekkor az $x^2/2$ valószínűségi változó gamma (1/2) eloszlású lesz, ugyanis

$$P\{x^2/2 < z\} = P\{|x| < \sqrt{2z}\} = \frac{2}{\sqrt{2\pi}} \int_0^{\sqrt{2z}} e^{-y^2/2}\, dy = \frac{1}{\sqrt{\pi}} \int_0^z e^{-v}\, v^{1/2-1}\, dv.$$

A megfelelő algoritmus gamma (1/2) eloszlású véletlen számok generálására

N algoritmus

1. Generáljunk egy normális eloszlású x számot.
2. Adjuk át a $w_1 \leftarrow x^2/2$ értéket, mint gamma (1/2) eloszlásút.

A multiplikatív módszerrel [13] egész paraméterű gamma eloszlásból lehet mintát sorsolni: az $\ln(u_1 u_2 \ldots u_n)$ n paraméterű gamma eloszlású, ahol $u_1, \ldots, u_n$ [0, 1)-ben egyenletes. Így gamma $((m-1)/2)$ eloszlású számok generálása a multiplikatív módszer és a fentebbi *N algoritmus* egyesítésével oldható meg.

G algoritmus

1. [Számítsuk ki a paraméter egész részét.] Legyen $j \leftarrow [(m-1)/2]$, $i \leftarrow 0$ és $q \leftarrow 1$.
2. Ha $j=(m-1)/2$, akkor legyen $w \leftarrow 0$ és menjünk 4-re.
3. Generáljunk egy gamma (1/2) eloszlású w számot az *N algoritmussal.*
4. Ha $i=j$, akkor menjünk a 6. lépésre.
5. Generáljuk u-t, számítsuk ki az $i \leftarrow i+1$, $q \leftarrow qu$ értékeket és menjünk vissza 4-re.
6. Adjuk át a $w_2 \leftarrow w - \ln q$ számot, mint gamma $((m-1)/2)$ eloszlásút.

A következő két módszert bizonyítás nélkül írjuk le. Mindkettő AHRENS és DIETER [3] algoritmusa, az első gamma eloszlású számokat generál normális és exponenciális eloszlású számok segítségével, a második pedig szimmetrikus béta eloszlású véletlen számokat állít elő normális eloszlásúakból.

GO algoritmus (a paraméterű gamma eloszlás, $a>2{,}53$)

1. Legyen $\mu \leftarrow a-1$, $V \leftarrow a^{1/2}$, $\sigma^2 \leftarrow a+1{,}63299316\,V$, $\sigma \leftarrow (\sigma^2)^{1/2}$, $W \leftarrow \sigma^2/\mu$, $d \leftarrow$ $\leftarrow 2{,}44948974$ és $b \leftarrow \mu+d$.
2. Generáljuk u-t, ha $u \leqq 0{,}0095722652$, akkor menjünk a 8. lépésre.
3. Generáljuk s-t a standard normális eloszlásból és legyen $x \leftarrow \mu+\sigma s$. Ha $x<0$ vagy $x>1$, akkor menjünk vissza a 2. lépésre.
4. Generáljuk u-t és legyen $S \leftarrow s^2/2$. Ha $s \leqq 0$, akkor menjünk a 6. lépésre.
5. Ha $u \leqq 1-S((1-2s/V)W-1)$, akkor adjuk át x-t, egyébként menjünk 7-re.
6. Ha $u \leqq 1-S(W-1)$, akkor adjuk át x-t.
7. Ha $\ln u > \mu(1+\ln(x/\mu))-x+S$, akkor menjünk vissza a 2. lépésre, egyébként adjuk át x-t.
8. Generáljuk s-t a standard exponenciális eloszlásból és legyen $x \leftarrow b(1+s/d)$.
9. Generáljuk u-t. Ha $\ln u > \mu(2+\ln(x/\mu)-x/b)+3{,}72032849-b-\ln(\sigma d/b)$, akkor menjünk 2-re egyébként adjuk át x-t.

BS algoritmus (a paraméterű szimmetrikus béta eloszlás, $a>1{,}5$)

1. Legyen $A \leftarrow a-1$ és $t \leftarrow (A+A)^{1/2}$.
2. Generáljuk s-t a standard normális eloszlásból és számítsuk ki az $x \leftarrow \frac{1}{2}(1+s/t)$ értéket.
3. Ha $x<0$ vagy $x>1$, akkor menjünk vissza 2-re.
4. Generáljuk u-t. Ha $u \leqq 1-s^4/(8a-12)$, akkor adjuk át x-t.
5. Ha $u \geqq 1-s^4/(8a-8)+\frac{1}{2}(s^4/(8a-8))^2$, akkor menjünk 2-re.
6. Ha $\ln u > A \ln(4x(1-x))+s^2/2$, akkor menjünk 2-re.
7. Adjuk át x-t.

Természetesen béta (1/2, 1/2) eloszlású véletlen számokat az *N algoritmus* segítségével állítunk elő, nem a *BS algoritmussal*. Legyen w_1 és w_2 az *N algoritmus* által generált gamma (1/2) eloszlású, akkor $w_1/(w_1+w_2)$ éppen béta (1/2, 1/2) eloszlást követ. Megjegyezzük még, hogy gamma (n/2) eloszlású valószínűségi változó kétszerese egy n szabadságfokú χ^2 eloszlású valószínűségi változó.

Az ebben a pontban leírt algoritmusokat mind FORTRAN nyelven írtuk meg, mivel itt a gyorsaság nem játszott lényeges szerepet.

A *GO* és *BS algoritmusok* előkészítő műveleteit adott paraméter esetén külön hajtottuk végre, ez gyorsította a generálást. Kis paraméterértékekre a multiplikatív módszerrel kombináltuk ezeket az algoritmusokat. A futási időket (μsec) a következő táblázatban adtuk meg, ezek az AHRENS és DIETER által publikált adatoktól nem térnek el lényegesen. Az a paraméter értéke adott n-re a *GO algoritmusnál* $a=n/2$, a *BS algoritmusnál* $a=(n-1)/2$ volt.

1.1 TÁBLÁZAT

Gamma és béta eloszlású számok generálásának ideje (μsec.)

Alg. \ n	4	10	20	50
GO	268	427	371	337
BS	491	206	180	183

2. FEJEZET

Egyenletes eloszlású véletlen vektorok generálása n-dimenziós gömbben és gömbfelületen

2.1. Jelölések és az algoritmusok felosztása

Minden algoritmust változó n dimenziószámra, az origó körüli

$$S = \left\{ \mathbf{x} \middle| \sum_{i=1}^{n} x_i^2 \leqq 1 \right\}$$

egységgömbre és annak

$$F = \left\{ \mathbf{x} \middle| \sum_{i=1}^{n} x_i^2 = 1 \right\}$$

felületére írtunk le, másmilyen gömbök esetére az algoritmusok egyszerű lineáris transzformációval módosíthatók. Attól függően, hogy az S tartományban vagy annak F felületén generál egyenletes eloszlású vektorokat az algoritmus, belső vagy felületi algoritmusról beszélünk.

Az S gömbben egyenletes eloszlású vektorok hossza x^n eloszlásfüggvényű — ezt például a sűrűségfüggvény polár koordinátákra való átírásával lehet belátni. Így a felületi algoritmusok belső algoritmusokká alakíthatók a következő véletlen sugárhossz generálási lépéssel:

SI. [Legyen $\mathbf{y}$ egy felületi algoritmus eredménye.]
Generáljuk u-t egyenletesen [0, 1)-ben, legyen $r \leftarrow u^{1/n}$ és adjuk át az $\mathbf{x} \leftarrow r\mathbf{y}$ vektort, mint egyenletes eloszlású S-ben.

Belső algoritmusok felületi algoritmussá alakíthatók, ha a véletlen pontokat a gömbfelületre vetítjük ki.

IS. [Legyen $\mathbf{x}$ egy belső algoritmus eredménye.]
Számítsuk ki az $s \leftarrow \left(\sum_{i=1}^{n} x_i^2 \right)^{-1/2}$ értéket és adjuk át az $\mathbf{y} \leftarrow s\mathbf{x}$ vektort, mint az F felületen egyenletes eloszlású.

Megjegyezzük, hogy az $u^{1/n}$ mennyiséget $u^{1/n} = e^{\ln u/n}$ alakba írva és az $\ln u$ valószínűségi változót egy vele statisztikailag egyenértékű változóval — egy exponencális eloszlású valószínűségi változó -1-szeresével — helyettesítve a véletlen sugárhossz generálása gyorsabbá tehető. Ekkor ugyanis az $1/n$ kitevőjű hatvány kiszámítása helyett egy exponenciális függvényértéket kell csak kiszámítani (ami az előzőnél általában háromszor gyorsabb a számítógépeken).

2.2. Felületi algoritmusok

Első módszerként a HICKS és WHEELING [61] által közölt induktív vetítéses algoritmust írjuk le. A módszer induktív, az i dimenziós egységgömb felületén generált egyenletes eloszlású pontot egy megfelelő faktorral beszorozva a gömb belsejébe transzformálja, majd az $i+1$ dimenziós gömb felületére vetíti oly módon, hogy az egyenletességet megőrzi. Az algoritmust a következő formában közölték.

IP algoritmus (Induktív vetítés)

IP1. [Az egy dimenziós gömb felületén egyenletes pont generálása.] Generáljuk u-t, ha $u<0{,}5$ akkor legyen $y_1 \leftarrow -1$, egyébként $y_1 \leftarrow +1$.

IP2. Legyen $m \leftarrow 2$.

IP3. [Véletlen előjel generálása.] Generáljuk u-t, legyen $v \leftarrow 2u-1$ és $u \leftarrow |v|$. (Ekkor u ismét egyenletes eloszlású [0, 1)-ben.) Ha $v>0$, akkor legyen $s \leftarrow +1$, egyébként $s \leftarrow -1$.

IP4. Oldjuk meg a következő egyenletet r_m-re

$$\int_0^{r_m} z^{m-2}/(1-z^2)^{1/2}\,dz \Big/ \int_0^1 z^{m-2}/(1-z^2)^{1/2}\,dz = u.$$

IP5. [Transzformáljuk a pontot a gömb belsejébe.] Számítsuk ki az $y_i \leftarrow r_m y_i$, $i=1, \ldots, m-1$ értékeket és legyen $y_m \leftarrow s(1-r_m^2)^{1/2}$.

IP6. Ha $m<n$, akkor növeljük meg a paramétert: $m \leftarrow m+1$ és menjünk vissza IP3.-ra.

IP7. Adjuk át az $\mathbf{y}=(y_1, \ldots, y_n)$ vektort, mint F-en egyenletes eloszlásút.

Az IP4 lépés $m=2$ és $m=3$ esetében r_m-re közvetlenül kiszámítható

IP4$_2$. Legyen $r_2 \leftarrow \sin(\pi \cdot u/2)$,

IP4$_3$. Legyen $r_3 \leftarrow (1-u^2)^{1/2}$.

Az $m>3$ esetben az eloszlásfüggvény invertálása helyett egy másik módszert javasolunk a

$$g_m(z) = d_m z^{m-2}/(1-z^2)^{1/2}, \quad m = 4, \ldots, n$$

sűrűségfüggvényű r_m szám generálására (a d_m konstans az $1/d_m = \int_0^1 z^{m-2}/(1-z^2)^{1/2}dz$ összefüggésből határozható meg.) Az $x=z^2$ helyettesítést elvégezve a g_m sűrűségfüggvényekből a

$$h_m(x) = (d_m/2)x^{(m-1)/2-1}(1-x)^{1/2-1}$$

függvényeket kapjuk az r_m^2 valószínűségi változó sűrűségfüggvényére, amelyek béta $((m-1)/2, 1/2)$ sűrűségfüggvények. Tehát ha w_1 egy gamma (1/2) és w_2 egy gamma $((m-1)/2)$ eloszlású véletlen szám, akkor az

$$r_m = (w_2/(w_1+w_2))^{1/2}$$

a kívánt g_m sűrűségfüggvényű lesz.

Összefoglalva a fentieket, az IP4. lépést helyettesítsük az alábbival

IP4′. Generáljunk egy gamma (1/2) eloszlású w_1 számot és egy gamma $((m-1)/2)$ eloszlású w_2 számot és legyen $r_m \leftarrow (w_2/(w_1+w_2))^{1/2}$.

A két gamma eloszlású véletlen szám generálására az előző fejezetben közölt *N* és *G algoritmusok* használhatók. Természetesen más módszerek is használhatók béta$((m-1)/2, 1/2)$ eloszlású számok generálására, de az itt kiválasztott algoritmusok FORTRAN nyelven megírhatók és csak az $n>40$ dimenziós esetekben lesznek számottevően lassabbak a többi béta algoritmusnál [3], [8].

A teljes *IP algoritmus* végrehajtásához szükséges munka így n darab béta eloszlású véletlen szám generálása, $2n$ négyzetgyökvonás és $n(n-1)/2$ szorzás.

MULLER [88] egy egyszerű módszert adott annak alapján, hogy az n-dimenziós, független normális eloszlású komponensekből álló vektor által meghatározott irány egyenletes eloszlású az egységgömb felületén.

NO algoritmus (Normális megközelítés)

NO1. Generáljunk n darab független, normális eloszlású $x_1, \ldots, x_n$ számot.

NO2. Számítsuk ki az $s \leftarrow \left(\sum_{i=1}^{n} x_i^2\right)^{-1/2}$ értéket.

NO3. Adjuk át az $\mathbf{y} \leftarrow s\mathbf{x}$ vektort mint egyenletes eloszlásút F-en.

Az algoritmus helyessége azon a tényen alapszik, hogy a független, normális komponensekből álló vektor sűrűségfüggvénye az n-dimenziós egységgömb felületén konstans. A módszer egyszerű, n darab normális szám generálását, $2n$ szorzást és egy négyzetgyökvonást igényel csak.

2.3. Belső algoritmusok

A jól ismert elfogadás-elvetés módszer a következőképpen fogalmazható meg.

AR algoritmus (Elfogadás-elvetés)

AR1. Generáljuk az u_i, $i=1, \ldots, n$ számokat egyenletesen $[0, 1)$-ben és legyen $x_i \leftarrow 2u_i - 1$, $i=1, \ldots, n$.

AR2. [Ellenőrzés.] Ha $\sum_{i=1}^{n} x_i^2 > 1$, akkor menjünk vissza az AR1. lépésre (elvetés).

AR3. Adjuk át az $\mathbf{x}$ vektort, mint S-ben egyenletes eloszlásút.

Nagy n esetében az elfogadás valószínűsége igen kicsi, mivel ez az S térfogatának és az n-dimenziós 2 oldalhosszúságú kocka térfogatának a hányadosa, vagyis

$$\frac{2\pi^{n/2}}{n\Gamma(n/2)} \Big/ 2^n = \frac{\pi^{n/2}}{2^{n-1} n\Gamma(n/2)}$$

Így $n=2$ dimenzióban annak a valószínűsége, hogy az AR1 lépésben generált vektort elfogadjuk 0,785, $n=10$ dimenzióban pedig csak 0,0025. Ezért az AR algoritmust — bár nagyon egyszerű — csak $n \leqq 5$ dimenzióban érdemes használni.

A következő algoritmus előkészítéseként tekintsük az

$$\begin{aligned} x_1 &= r\cos\varphi_1 \cos\varphi_2 \ldots \cos\varphi_{n-2} \cos\varphi_{n-1} \\ x_2 &= r\cos\varphi_1 \cos\varphi_2 \ldots \cos\varphi_{n-2} \sin\varphi_{n-1} \\ x_3 &= r\cos\varphi_1 \cos\varphi_2 \ldots \sin\varphi_{n-2} \\ &\vdots \\ x_{n-1} &= r\cos\varphi_1 \sin\varphi_2 \\ x_n &= r\sin\varphi_1 \end{aligned} \tag{2.1}$$

alakú polár transzformációt, ahol $-\pi/2 \leqq \varphi_i \leqq \pi/2$, $i=1, \ldots, n-2$, $0 \leqq \varphi_{n-1} \leqq 2\pi$ és $0 \leqq r \leqq 1$. A transzformáció *Jacobi determinánsa* ([70])

$$J = r^{n-1} \cos^{n-2} \varphi_1 \cos^{n-3} \varphi_2 \ldots \cos \varphi_{n-2}. \tag{2.2}$$

Mivel a generálni kívánt vektor sűrűségfüggvénye az $(x_1, \ldots, x_n)$ koordinátarendszerben konstans, ezért a sűrűségfüggvény a $(\varphi_1, \ldots, \varphi_{n-1}, r)$ koordinátarendszerben éppen a (2.2) kifejezés.

PT algoritmus (Polár transzformáció)

PT1. [Sugárhossz generálása.] Generáljuk u-t és legyen $r \leftarrow u^{1/n}$.
PT2. Generáljuk u-t és legyen $\varphi_{n-1} \leftarrow 2\pi u$.
PT3. Generáljuk a φ_i, $i=1, \ldots, n-2$ szögeket a $[-\pi/2, \pi/2)$ intervallumban $f_i(x) = c_i \cos^{n-i-1} x$ sűrűségfüggvény szerint, ahol

$$1/c_i = \int_{-\pi/2}^{\pi/2} \cos^{n-i-1} x \, dx.$$

PT4. [Transzformáljuk vissza az eredményt derékszögű koordinátarendszerbe.] A kapott $r, \varphi_1, \ldots, \varphi_{n-1}$ számokból a (2.1) polár transzformáció segítségével számítsuk ki az $\mathbf{x}=(x_1, \ldots, x_n)$ vektort és adjuk át $\mathbf{x}$-t, mint S-ben egyenletes eloszlásút.

Tekintsük a φ_i valószínűségi változót, melynek sűrűségfüggvénye $\cos^{n-i-1} x$. A $\cos \varphi_i$ valószínűségi változó sűrűségfüggvénye $d_i y^{n-i-1}/(1-y^2)^{1/2}$ valamilyen d_i konstanssal. Az *IP algoritmust* a *PT algoritmussal* összehasonlítva látjuk, hogy az IP4-ben generált r_m éppen $\cos \varphi_m$ a *PT algoritmusban,* az IP5 lépés megfelel a PT4 lépésben a (2.1) kifejezésnek. Tehát a *PT algoritmus* a véletlen sugárhossz generálási lépéssel kibővített *IP algoritmus.* Megjegyezzük, hogy az *IP algoritmust* szerzői egészen más gondolatmenettel bizonyították.

A *PT algoritmust* számítógépen nem futtattuk; eredeti formájában a sinus és cosinus függvények használata miatt lényegesen lassabb lenne az IP algoritmusnál, a φ_i szögek generálása helyett a $\cos \varphi_i$ mennyiségeket generálva pedig az *IP algoritmust* kapnánk.

2.4. Számítógépes tapasztalatok

A felületi algoritmusokat belső algoritmusokká alakítottuk és az egyetlen belső algoritmust felületivé, amint ezt a fejezet elején említettük.

Az egyes algoritmusok futási idejét a következő módon kaptuk. A

```
    DØ 1 I = 1,1000
    CALL XXXX
1 CØNTINUE
```

ciklus idejéből kivontuk az üres ciklus idejét és az eredményt 1000-rel osztottuk. A programokat a Honeywell 66/20 gépen futtattuk le, az időket μsec-ban adtuk meg a következő táblázatokba foglalva.

2.1 TÁBLÁZAT

Felületi algoritmusok ideje (μsec.)

Alg. \ $n=$	2	3	4	5	10	15	20
IP	1081	1582	3130	4973	14 890	27 220	41 575
NO	851	1089	1330	1572	2 780	4 040	5 277
AR	798	1263	2312	4635	569 700	—	—

2.2 TÁBLÁZAT

Belső algoritmusok ideje (μsec.)

Alg. \ $n=$	2	3	4	5	10	15	20
IP	1963	2509	4112	6002	16 115	28 673	43 294
NO	1637	1872	2118	2356	3 582	4 826	6 060
AR	398	840	1863	4155	563 900	—	—

A táblázatokban közölt időeredmények ugyanazt az eredményt mutatják, mint amit az egyes algoritmusok használata esetén szükséges műveletek számának egymáshoz hasonlításából kapunk. Az $n>5$ dimenziós esetben csak az *NO algoritmust* érdemes használni. Az $n\leqq 5$ dimenziós esetben egyértelmű ajánlás nem tehető. Az $n=2$ dimenzióban NEUMANN [91] racionális formuláin alapuló algoritmus a leggyorsabb [35], $n=3$ dimenzióban pedig KNOP [72], [120] algoritmusa.

3. FEJEZET

Egyenletes eloszlású vektorok generálása n-dimenziós ellipszoidban

3.1. Az egységgömbben egyenletes vektorok transzformálása

Megvizsgáljuk, hogy az n-dimenziós

$$E = \{\mathbf{x} | \mathbf{x}' \mathbf{R}^{-1} \mathbf{x} \leqq 1\}$$

ellipszoidban hogyan lehet egyenletes eloszlású vektorokat előállítani, ahol $\mathbf{R}$ egy korreláció mátrix, azaz pozitív definit és szimmetrikus. Ekkor E valóban hiperellipszoid. Ismert tétel szerint [48] az $\mathbf{R}$ mátrix

$$\mathbf{R} = \mathbf{T}\mathbf{T}'$$

alakban írható, ahol T egy alsó háromszög mátrix, a diagonális feletti elemek zérusok. Ekkor igaz az

$$E = \{\mathbf{x} | \mathbf{x} = \mathbf{Ty},\ \mathbf{y} \in S\}$$

összefüggés, ahol S az egységgömb. Az E ellipszoidban egyenletes eloszlású vektorok generálása a következő módon végezhető

EI. [Legyen x egyenletes eloszlású S-ben.] A $\mathbf{z} \leftarrow \mathbf{Tx}$ vektor egyenletes eloszlású E-ben.

Az algoritmus helyessége a következő gondolatmenetből látható be. Legyen D_1 és D_2 két térfogatelem S-ben, melyek térfogata T_1 és T_2, $T_1 = T_2$, ekkor

$$P\{\xi \in D_1\} = P\{\xi \in D_2\} = \frac{T_1}{V} = \frac{T_2}{V},$$

ahol V az S egységgömb térfogata, ξ pedig egyenletes eloszlású valószínűségi vektorváltozó S-ben. Legyen

$$D_1^* = \{\mathbf{x} | \mathbf{x} = \mathbf{Ty},\ \mathbf{y} \in D_1\},$$

$$D_2^* = \{\mathbf{x} | \mathbf{x} = \mathbf{Ty},\ \mathbf{y} \in D_2\},$$

ekkor az $\eta = T\xi$ jelölést használva

$$P\{\eta \in D_1^*\} = P\{T^{-1}\eta \in D_1\} = P\{\xi \in D_1\} = \frac{T_1}{V} = P\{T^{-1}\eta \in D_2\} = P\{\eta \in D_2^*\}$$

egyenlőség igaz. T affin leképezés, így aránytartó, tehát a térfogatok arányát is megtartja, így a generált vektorok valóban egyenletesek E-ben.

DEFINÍCIÓ. Legyen y egyenletes eloszlású az S gömb felszínén. A Ty vektor eloszlását az E hiperellipszoid térfogatára nézve egyenletesnek nevezzük.

Vegyük észre, hogy egy E ellipszoidban egyenletes eloszlású x vektor által meghatározott irány éppen a fenti eloszlású, az ellipszoid térfogatára nézve egyenletes. Ha pedig a q vektor az E térfogatára nézve egyenletes eloszlású, akkor $u^{1/n}\mathbf{q}$ az E ellipszoidban egyenletes eloszlású, hiszen

$$u^{1/n}\mathbf{q} = u^{1/n}\mathbf{Ty} = \mathbf{T}(u^{1/n}\mathbf{y}),$$

ahol y F-en egyenletes, u pedig [0, 1)-ben egyenletes.

3.2. Bázispont csere algoritmusa

Ez a pont és a fejezet további pontjai tulajdonképpen a 4. Fejezet előkészítését szolgálják. Az ebben a fejezetben leírt módszereket használva a 4. Fejezetben gyors algoritmusokat adunk stacionárius normális vektorok generálására. Ezek pedig az értekezés fő témájában, a normális eloszlásfüggvény kiszámításában kerülnek felhasználásra, vagyis ezeknek a módszereknek az igazi haszna csak az 5. Fejezetben mutatkozik meg: az algoritmusok gyorsasága miatt az eloszlásfüggvény rövid idő alatt számítható.

A módszereket először n-dimenziós gömbök esetére fogalmazzuk meg felületi algoritmusokként, majd az n-dimenziós ellipszoidok esetére módosítjuk az algoritmusokat.

Tegyük fel, hogy van egy módszerünk, például az *NO algoritmus,* amelynek segítségével az S gömb F felületén egyenletes eloszlású pontokat tudunk generálni. Ekkor a következő *BC algoritmus* segítségével stacionárius, egyenletes eloszlású vektorokat tudunk előállítani.

A *BC algoritmus* lényegében a következő módon működik. Először is az *NO algoritmus* segítségével generáljuk az F felületen az $\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(M)}$ vektorokat, ahol M tetszőleges, de rögzített pozitív egész. Ezek közül egy vektor kitüntetett szerepet fog játszani, ezt pivotvektornak nevezzük és $\mathbf{b}$-vel is jelöljük, míg a többi vektort bázisvektornak nevezzük. Kezdetben legyen $\mathbf{y}^{(1)}=\mathbf{b}$ a pivotvektor. A pivotvektor és az első, második, ..., $(M-1)$-edik bázisvektor felhasználásával — egy, a továbbiakban leírandó alapmódszert használva — összesen $M-1$ darab n-dimenziós vektort állítunk elő (ezeket fogjuk majd feladatok megoldására használni). Ezután az $\mathbf{y}^{(1)}, \ldots$ $\ldots, \mathbf{y}^{(M)}$ vektorrendszert felfrissítjük; bizonyos rendszer szerint egyiküket elhagyjuk, helyébe az *NO algoritmussal* egy új, az előzőektől független, F-en egyenletes eloszlású vektort generálunk. Ezt pivotvektornak használva az eljárást megismételjük előlről, stb.

Az algoritmus formális leírása a következő. Előkészületként állítsuk be a pivotvektor indexét $q \leftarrow 1$, a bázisvektor indexét $I \leftarrow 1$ és generáljuk az $\mathbf{y}^{(1)}, \ldots, y^{(M)}$ vektorokat egyenletesen az F felületen.

BC Bázispontcsere algoritmusa

1. Indexellenőrzés. Legyen $I \leftarrow I+1$. Ha $I=M+1$, akkor visszaállítjuk az indexet $I \leftarrow 1$.
2. [Felhasználtuk az összes bázispontot?] Ha $I=q$, akkor menjünk a 4. lépésre.
3. Az alábbi A alapmódszerek egyikével generáljunk egy $\mathbf{y}$ vektort az $\mathbf{y}^{(I)}$ bázisvektor és az $\mathbf{y}^{(q)}$ pivotvektor felhasználásával, vagyis legyen $\mathbf{y} \to c_1 \mathbf{y}^{(I)} + c_2 \mathbf{y}^{(q)}$, ahol c_1 és c_2 legfeljebb az $\mathbf{y}^{(I)}$ és $\mathbf{y}^{(q)}$ vektoroktól, valamint a felhasznált A alapmódszertől függhet. Menjünk az 5. lépésre.
4. [Bázispontcsere, új pivotvektor generálása.] Legyen $q \leftarrow q+1$, ha ekkor $q=M+1$, akkor legyen $q \leftarrow 1$. Állítsuk be az $I \leftarrow q$ kezdeti értéket. Generáljunk egy egyenletes eloszlású pontot az $\mathbf{y}^{(q)}$ vektor helyébe új pivotvektornak. Menjünk vissza az 1. lépésre.
5. Adjuk át az $\mathbf{y}$ vektort, mint F-en egyenletes eloszlásút.

Vizsgáljuk meg, hogy a *BC bázispontcsere algoritmus* használata esetén milyen eloszlásuk van a generált vektoroknak. Ha minden bázisvektort csak egyszer használnánk fel, tetszőleges A alapmódszert használva, a generált vektorok F-en egyenletes eloszlású független vektorok lennének. A *BC algoritmus* 1. lépésében az A alapmódszer tetszőlegesen választható, de rögzített.

Néhány jelölést vezetünk be. Az $\mathbf{y}^{(1)}$ mint pivotvektor és az $\mathbf{y}^{(2)}$ mint bázisvektor felhasználásával előállított vektort $\mathbf{x}^{(1)}$-el jelöljük, az $\mathbf{y}^{(1)}$ és $\mathbf{y}^{(3)}$ felhasználásával kapott vektort $\mathbf{x}^{(2)}$-vel, stb. Az $\mathbf{x}^{(1)}=\mathbf{x}^{(1)}(\mathbf{y}^{(1)}, \mathbf{y}^{(2)})$ felírásban a két paraméter közül az elsőt használjuk pivotvektornak; egyszerűség kedvéért az $\mathbf{x}^{(1)}=\mathbf{x}^{(1)}(1, 2)$

jelöléssel is fogunk élni. Az $\mathbf{y}^{(1)}$ vekorral, mint pivot vektorral generált vektorokat egy vektorba foglaljuk össze: $\mathbf{X}_1=(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(M-1)})$. A következő $\mathbf{y}^{(2)}$ pivotvektorral generált vektorokat $\mathbf{X}_2=(\mathbf{x}^{(M)}, \ldots, \mathbf{x}^{(2M-2)})$-vel jelöljük stb. Az $\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(M)}$ vektorok mindegyikét pivotvektornak használva a $\mathbf{Z}_1=(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_M)$ vektort kapjuk, ezt nevezzük a generált vektorok első blokkjának, a további blokkokat a $\mathbf{Z}_k=(\mathbf{X}_{M(k-1)+1}, \ldots, \mathbf{X}_{Mk})$, $k=2,3, \ldots$ vektorok adják.

A generálás folyamatát, valamint azt, hogy melyik vektor melyiktől függ, a következő táblázatból lehet látni. Az $\mathbf{x}^{(i)}$ vektorok indexelését blokkonként újrakezdjük.

Első blokk

$$\mathbf{X}_1 = (\mathbf{x}^{(1)}(1, 2), \mathbf{x}^{(2)}(1, 3), \mathbf{x}^{(3)}(1, 4), \ldots, \mathbf{x}^{(M-2)}(1, M-1), \mathbf{x}^{(M-1)}(1, M))$$

$$\mathbf{X}_2 = (\mathbf{x}^{(M)}(2, 3), \mathbf{x}^{(M+1)}(2, 4), \mathbf{x}^{(M+2)}(2, 5), \ldots, \mathbf{x}^{(2M-3)}(2, M), \mathbf{x}^{(2M-2)}(2, 1))$$

$$\mathbf{X}_3 = (\mathbf{x}^{(2M-1)}(3, 4), \mathbf{x}^{(2M)}(3, 5), \mathbf{x}^{(2M+1)}(3, 6), \ldots, \mathbf{x}^{(3M-4)}(3, 1), \mathbf{x}^{(3M-3)}(3, 2))$$

$$\vdots$$

$$\mathbf{X}_{M-1} = (\mathbf{x}^{(M^2-3M+3)}(M-1, M), \mathbf{x}^{(M^2-3M+4)}(M-1, 1), \ldots$$
$$\ldots, \mathbf{x}^{(M^2-2M)}(M-1, M-3), \mathbf{x}^{(M^2-2M+1)}(M-1, M-2))$$

$$\mathbf{X}_M = (\mathbf{x}^{(M^2-2M+2)}(M, 1), \mathbf{x}^{(M^2-2M+3)}(M, 2), \ldots$$
$$\ldots, \mathbf{x}^{(M^2-M-1)}(M, M-2), \mathbf{x}^{(M^2-M)}(M, M-1))$$

Második blokk

$$\mathbf{X}_{M+1} = (\mathbf{x}^{(1)}(1, 2), \mathbf{x}^{(2)}(1, 3), \ldots, \mathbf{x}^{(M-2)}(1, M-1), \mathbf{x}^{(M-1)}(1, M))$$

$$\mathbf{X}_{M+2} = (\mathbf{x}^{(M)}(2, 3), \mathbf{x}^{(M+1)}(2, 4), \ldots, \mathbf{x}^{(2M-3)}(2, M), \mathbf{x}^{(2M-2)}(2, 1))$$

$$\vdots$$

Mivel a pivotvektorokat állandóan újra generáljuk (a második sorban levő $\mathbf{y}^{(2)}$ vektor nem azonos az első sorban levő $\mathbf{y}^{(2)}$ vektorral) ezért a második blokk kiindulásánál használt vektorrendszer — az $\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(M)}$ vektorok — teljesen különbözik az első blokk kiindulási vektorrendszerétől.

Tekintsük most a következő meghatározást.

DEFINÍCIÓ. Az $\mathbf{X}_k$ sztochasztikus folyamatot szorosabb értelemben stacionáriusnak nevezzük, ha az

$$\mathbf{X}_{k_1+j}, \mathbf{X}_{k_2+j}, \ldots, \mathbf{X}_{k_n+j}$$

vektorsorozat eloszlása j-től független.

A táblázat után tett megjegyzésből következik, hogy a $\mathbf{Z}_k$ és $\mathbf{Z}_{k+2}$ vektorok függetlenek, valamint az, hogy a $\mathbf{Z}_k$, $\mathbf{Z}_{k+1}$ vektorok együttes eloszlása tetszőleges k-ra azonos, ezért igaz az

1. TÉTEL. A $\mathbf{Z}_1, \ldots, \mathbf{Z}_k, \ldots$ vektorok szorosabb értelemben vett stacionárius folyamatot alkotnak.

Belátható az ennél erősebb

2. TÉTEL. Az $\mathbf{X}_1, \ldots, \mathbf{X}_k, \ldots$ vektorok szorosabb értelemben vett stacionárius folyamatot alkotnak.

Bizonyítás. Legyen $\mathbf{X}_k$ és $\mathbf{X}_{k+j}$ két vektor a sorozatból. Ha $j \geqq M$, akkor ezek a vektorok egymástól függetlenek. Ha $j < M$, akkor az $\mathbf{X}_k$ előállításához használt vektorokat átindexelve azonnal látható, hogy az $\mathbf{X}_k$ és $\mathbf{X}_{k+j}$ vektorok együttes eloszlása az $\mathbf{X}_1$, $\mathbf{X}_{1+j}$ vektorok együttes eloszlásával azonos. —

3.3. Alapmódszerek

Néhány eljárást adunk, amelyek egy pivotvektor és egy vagy több bázisvektor felhasználásával új vektorokat generálnak.

A determinisztikus tükrözéses módszerben a bázisvektort tükrözzük a pivotvektor által megadott irányra. Az algoritmus leírása előtt két összefüggést fogalmazunk meg.

1. LEMMA. Legyen $\mathbf{x}$ és $\mathbf{y}$ két pont az S gömb F felszínén. A két pont közti húr hosszát h-val jelölve

$$\mathbf{x}'\mathbf{y} = 1 - h^2/2$$

Bizonyítás. Az $\mathbf{x}$ és $\mathbf{y}$ vektorok által bezárt szöget α-val jelölve ismert trigonometriai azonosság segítségével

$$\mathbf{x}'\mathbf{y} = \cos\alpha = 1 - 2\sin^2(\alpha/2) = 1 - h^2/2,$$

mivel a h hosszúságú húrhoz tartozó $\alpha/2$ nagyságú kerületi szögre $\sin(\alpha/2) = h/2$. —

3. TÉTEL. Az $\mathbf{y}^{(I)}$ vektornak a $\mathbf{b}$ vektor által meghatározott irányra való tükörképe

$$\mathbf{y} = (2 - s^2)\mathbf{b} - \mathbf{y}^{(I)},$$

ahol $\mathbf{y}^{(I)}$, $\mathbf{b}$, $\mathbf{y} \in F$, az s változó pedig az $\mathbf{y}^{(I)}$ és $\mathbf{b}$ közti húr hossza.

Bizonyítás. Nyilvánvaló az

$$\mathbf{y} + \mathbf{y}^{(I)} = \lambda\mathbf{b} \tag{3.1}$$

összefüggés valamilyen λ konstansra. Mivel $\mathbf{y}^{(I)}$ és tükörképe $\mathbf{b}$-vel azonos szöget zárt be, ezért (3.1)-t $\mathbf{b}$-vel skalárisan szorozva, az 1. Lemmát felhasználva

$$\lambda = \mathbf{b}'\mathbf{y} + \mathbf{b}'\mathbf{y}^{(I)} = 2\mathbf{b}'\mathbf{y}^{(I)} = 2(1 - s^2/2),$$

amivel a tételt bebizonyítottuk. —

A 1. Determinisztikus tükrözés alapmódszere

1. [Számítsuk ki az $\mathbf{y}^{(I)}$ és $\mathbf{b}$ pontok közötti húr hosszának négyzetét.] Legyen $s^2 \leftarrow \sum_{j=1}^{n} (y_j^{(I)} - b_j)^2$ és $c \leftarrow 2 - s^2$.
2. Legyen $\mathbf{y} \leftarrow c\mathbf{b} - \mathbf{y}^{(I)}$.

Az alapmódszer kifejezés most és a továbbiakban is azt jelenti, hogy a *BC bázispontcsere algoritmus* generálás részének 3. lépése helyett az alapmódszerben leírt lépéseket hajtjuk végre.

A következő alapmódszerben, a perturbált tükrözés módszerében egy **b**-től véletlen húrhossznyi távolságra levő pontot választunk ki az $\mathbf{y}^{(I)}$ és **b** vektorok által meghatározott főkörből (ugyanis az $\alpha_1\mathbf{y}^{(I)}+\alpha_2\mathbf{b}$ kétdimenziós sík egy főkört metsz ki F-ből). A részletes leírás előtt két tételt bizonyítunk.

4. TÉTEL. Jelöljük s-sel egy olyan húr hosszát, melynek végpontjai egyenletes eloszlásúak az F felszínen. Az $s^2/4$ valószínűségi változó sűrűségfüggvénye

$$f(x) = \frac{\Gamma(n-1)}{\{\Gamma((n-1)/2)\}^2} \cdot x^{(n-1)/2-1}(1-x)^{(n-1)/2-1}$$

vagyis béta $((n-1)/2, (n-1)/2)$.

Bizonyítás. A véletlen húr hosszának eloszlását többen meghatározták (lásd KENDALL és MORAN [71]). Mi ALAGAR [5] egyszerű gondolatmenetét használjuk fel a bizonyításban.

Az általánosság megszorítása nélkül feltehetjük, hogy a húr egyik végpontja rögzített, legyen ez az $(1, 0, \ldots, 0)$ pont. A másik végpontnak a $(\xi_1/\sigma, \xi_2/\sigma, \ldots, \xi_n/\sigma)$ pontot vesszük, ahol $\xi_1, \ldots, \xi_n$ független standard normális eloszlású valószínűségi változók, $\sigma^2=\xi_1^2+\ldots+\xi_n^2$. A két pont közti s távolságra azt kapjuk, hogy

$$s^2 = (1-\xi_1/\sigma)^2+\xi_2^2/\sigma^2+\ldots+\xi_n^2/\sigma^2 = 2-2\xi_1/\sigma.$$

A ξ_1^2/σ^2 valószínűségi változó béta $(1/2, (n-1)/2)$ eloszlású. Felírható, hogy

$$P\{s^2/4 < q\} = P\{1/2-\xi_1/(2\sigma) < q\} = P\{\xi_1^2/\sigma^2 > (1-2q)^2\}.$$

Az utolsó kifejezést deriválva a kívánt eredményt kapjuk, ugyanis valamilyen C konstanssal:

$$\left[C\int_{(1-2q)^2}^{1} x^{-1/2}(1-x)^{(n-3)/2}\,dx\right]' = C(1-2q)^{-1}(4q-4q^2)^{(n-3)/2}\,2(1-2q) =$$

$$= 2C[4q(1-q)]^{(n-3)/2}.$$

A következő tétel segítségével az $\mathbf{y}^{(I)}$ pontnak a **b** vektor által meghatározott irányra való perturbált tükörképét határozhatjuk meg.

5. TÉTEL. Tekintsük az egymástól s távolságra levő, F-en elhelyezkedő $\mathbf{y}^{(I)}$ és **b** pontokon áthaladó F-beli főkört. A főkörön **b**-től η távolságra levő pontok közül az $\mathbf{y}^{(I)}$ ponttól távolabb levő **y** pont a következő:

$$\mathbf{y} = c_1\mathbf{y}^{(I)}+c_2\mathbf{b}, \tag{3.2}$$

ahol $$c_1 = -\sqrt{\frac{1-(1-\eta^2/2)^2}{1-(1-s^2/2)^2}}, \quad c_2 = 1-\eta^2/2-c_1(1-s^2/2).$$

Bizonyítás. Az $\mathbf{y}=c_1\mathbf{y}^{(I)}+c_2\mathbf{b}$ vektorok az $\mathbf{y}^{(I)}$ és **b** vektorok által meghatározott síkban vannak. Az $\mathbf{y}'\mathbf{y}=1$ egyenlőség biztosítja, hogy F-ben van az **y** vektor, ezért

$$1 = \mathbf{y}'\mathbf{y} = c_1^2+2c_1c_2\mathbf{b}'\mathbf{y}^{(I)}+c_2^2. \tag{3.3}$$

Másrészt a (3.2) egyenlőséget **b**-vel skalárisan szorozva

$$\mathbf{b}'\mathbf{y} = c_1\mathbf{b}'\mathbf{y}^{(I)}+c_2. \tag{3.4}$$

Az 1. lemma szerint $\mathbf{b}'\mathbf{y}^{(I)} = 1 - s^2/2$, valamint feltevésünk miatt $\mathbf{b}'\mathbf{y} = 1 - \eta^2/2$ igaz. Ezeket az összefüggéseket a (3.3), illetőleg (3.4) egyenlőségbe helyettesítve a

$$c_1^2 + 2c_1c_2(1 - s^2/2) + c_2^2 = 1$$
$$c_1(1 - s^2/2) + c_2 = 1 - \eta^2/2$$

egyenletrendszert kapjuk a c_1 és c_2 konstansokra. Az egyenletrendszer megoldása a tételben megadott értékeket szolgáltatja. —

A2. Perturbált tükrözés alapmódszere

1. [Számítsuk ki az $\mathbf{y}^{(I)}$ és $\mathbf{b}$ pontok közti húr hosszának négyzetét.] Legyen $s^2 \leftarrow \sum_{j=1}^{n} (y_j^{(I)} - b_j)^2$ és $d_1 \leftarrow 1 - s^2/2$.
2. [Generáljuk egy véletlen húrhosszúság felének négyzetét.] Generáljunk egy béta $((n-1)/2, (n-1)/2)$ eloszlású β véletlen számot és legyen $d_2 \leftarrow 1 - 2\beta$.
3. Legyen $c_1 \leftarrow -\sqrt{(1-d_2^2)/(1-d_1^2)}$, $c_2 \leftarrow d_2 - d_1c_1$ és számítsuk ki az $\mathbf{y} \leftarrow c_1\mathbf{y}^{(I)} + c_2\mathbf{b}$ vektort.

Az *A2. alapmódszernél* kevesebb számítást igénylő, de kevésbé „véletlenszerű" vektorokat kaphatunk, ha a bázisvektor és a pivotvektor közti húr felezőpontjához tartozó sugarat választjuk ki.

6. TÉTEL. Legyen $\mathbf{x}$ és $\mathbf{y}$ az S gömb F felszínén két pont, távolságukat jelöljük s-sel. Az $\mathbf{x}+\mathbf{y}$ vektor irányába mutató sugár

$$\mathbf{z} = (\mathbf{x}+\mathbf{y})/\sqrt{4-s^2}.$$

Bizonyítás. A keresett sugárra $\lambda\mathbf{z} = \mathbf{x}+\mathbf{y}$ fennáll valamilyen λ konstanssal. Az $\mathbf{x}$ és $\mathbf{y}$ közti húr felezőpontja azonos az $\mathbf{x}+\mathbf{y}$ vektor hosszának felezőpontjával, így a *Pitagorasz tételt* alkalmazva $(\lambda/2)^2 + (s/2)^2 = 1$, ahonnan λ kifejezhető. —

A3. Húrfelezés alapmódszere

1. Legyen $s^2 \leftarrow \sum_{j=1}^{n} (y_j^{(I)} - b_j)^2$ és $c \leftarrow 1/\sqrt{4-s^2}$.
2. Legyen $\mathbf{y} \leftarrow c(\mathbf{y}^{(I)} + \mathbf{b})$.

Még két alapmódszert közlünk. Az egyik a perturbált tükrözéshez hasonlít, csak most a bázisvektor és a pivotvektor által meghatározott főkörből egyenletes valószínűséggel választunk ki egy pontot, vagyis az új pont távolsága a pivotvektortól béta (1/2, 1/2) eloszlású (ugyanis ez felel meg egy kétdimenziós körben generált véletlen húrhossznak). A másik alapmódszerben a pivotvektor mellett nem egy, hanem két bázisvektort használunk fel egy új vektor előállításához. Ennek az előnye az előző módszerekhez képest az, hogy az n-dimenziós térben elhelyezkedő F $(n-1)$-dimenziós felületnek nem egy kétdimenziós síkban elhelyezkedő körén választjuk meg az új pontot, hanem a két bázispont és a pivotvektor által meghatározott háromdimenziós alterében (a degenerációtól eltekintünk).

Pontosabban fogalmazva az utolsó alapmódszer a következő. Legyen a két bázispont $\mathbf{y}^{(I)}$ és $\mathbf{y}^{(I+1)}$, a pivotvektor pedig $\mathbf{b}$. Az $\mathbf{y}^{(I)}$ és $\mathbf{y}^{(I+1)}$ bázisvektorok által meghatározott főkörön egy $\mathbf{z}$ pontot választunk egyenletes valószínűséggel (béta (1/2, 1/2) eloszlású véletlen húrhosszúsággal, mint *A4*-ben). A kapott $\mathbf{z}$ pont, mint bázisvektor és a $\mathbf{b}$ pivotvektor felhasználásával végezzük el a perturbált tükrözést.

A4. Egyenletes pont alapmódszere

1. [Számítsuk ki az $\mathbf{y}^{(I)}$ és $\mathbf{b}$ távolságának négyzetét.]
   Legyen $s^2 \leftarrow \sum_{j=1}^{n} (y_j^{(I)} - b_j)^2$ és $d_1 \leftarrow 1 - s^2/2$.
2. [Generáljunk egy kétdimenziós körben véletlen húrhosszt.] Generáljunk egy béta (1/2, 1/2) eloszlású β változót és legyen $d_2 \leftarrow 1 - 2\beta$.
3. [Meghatározzuk a szorzókonstansokat.] Legyen $c_1 \leftarrow -\sqrt{(1-d_2^2)/(1-d_1^2)}$, $c_2 \leftarrow d_2 - d_1 c_1$ és számítsuk ki $\mathbf{y} \leftarrow c_1\mathbf{y}^{(I)} + c_2\mathbf{b}$.

A5. Egyenletes pont perturbált tükrözése

1. [Számítsuk ki az $\mathbf{y}^{(I)}$ és $\mathbf{y}^{(I+1)}$ távolságának négyzetét.]
   Legyen $s^2 \leftarrow \sum_{j=1}^{n} (y_j^{(I)} - y_j^{(I+1)})^2$ és $d_1 \leftarrow 1 - s^2/2$.
2. [Generáljunk egy kétdimenziós körben véletlen húrhosszt.] Generáljunk egy béta (1/2, 1/2) eloszlású β változót és legyen $d_2 \leftarrow 1 - 2\beta$.
3. [Meghatározzuk az egyenletes pontot.] Legyen $c_1 \leftarrow -\sqrt{(1-d_2^2)/(1-d_1^2)}$, $c_2 \leftarrow d_2 - d_1 c_1$ és számítsuk ki a $\mathbf{z} \leftarrow c_1\mathbf{y}^{(I)} + c_2\mathbf{y}^{(I+1)}$ vektort.
4. [Számítsuk ki a $\mathbf{z}$ bázispont és a $\mathbf{b}$ pivotvektor távolságát.] Legyen $s^2 \leftarrow \sum_{j=1}^{n} (z_j - b_j)^2$ és $d_1 \leftarrow 1 - s^2/2$.
5. [Generáljunk egy n-dimenziós véletlen húrhosszt.] Generáljunk egy béta $((n-1)/2, (n-1)/2)$ eloszlású β véletlen számot és legyen $d_2 \leftarrow 1 - 2\beta$.
6. [Számítsuk ki a $\mathbf{z}$ bázispont perturbált tükrözését.] Legyen $c_1 \leftarrow -\sqrt{(1-d_2^2)/(1-d_1^2)}$, $c_2 \leftarrow d_2 - d_1 c_1$ és határozzuk meg az $\mathbf{y} \leftarrow c_1\mathbf{z} + c_2\mathbf{b}$ vektort.

A leírt alapmódszerek mintájára további bázispontcsere és alapmódszer algoritmusok konstruálhatók. A *BC algoritmus* 1. lépésében a felhasználandó $\mathbf{y}^{(I)}$ bázisvektor és $\mathbf{b}$ pivotvektor választható véletlenszerűen. Használhatjuk k darab ($k < M$) bázispontnak a lineáris kombinációját is. Ilyen és hasonló módosításokkal az algoritmusok munkaigényesebbekké válnak, de a generált vektor kiválasztása kevésbé determinisztikus.

Végül egy tételt bizonyítunk be az *A2 alapmódszer* által generált $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(M-1)}$ vektorokra.

7. TÉTEL. Használjuk a *BC algoritmus* első lépésében az *A2 alapmódszert*. Az így előállított $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(M-1)}$ vektorok egymástól függetlenek és egyenletes eloszlásúak az F gömbfelületen.

Bizonyítás. Tekintsük az egységnyi hosszú **t** tengelyből és az erre merőleges $(n-1)$-dimenziós A altérből álló rendszert. (Az A hipersíknak **t** egy normálvektora.) Ebben a rendszerben egy **x** vektort — amely az S gömb F felületén van — felírhatunk két adat függvényeként. A **t** tengelytől való s távolság és egy $(n-1)$-dimenziós A altérbeli, egységhosszúságú **a** vektor teljesen meghatározza **x**-t, ha az **a** vektor az **x** vektornak A-ra való vetülete irányába mutató egységvektor, tehát $\mathbf{x}=\mathbf{x}(s, \mathbf{a})$. Ha az **x** vektor egyenletes eloszlású F-ben, akkor az $s^2/4$ valószínűségi változó béta $((n-1)/2, (n-1)/2)$ eloszlású, az **a** vektor pedig egyenletes eloszlású az $A \cap F$ $(n-1)$-dimenzióban elhelyezkedő $(n-2)$-dimenziós gömbfelületen.

Tekintsük most az *A2 alapmódszer* és egy rögzített **b** pivotvektor felhasználásával előállított $\mathbf{x}^{(i)}$, $i=1, \ldots, M-1$ vektorokat. Írjuk fel ezeket a **b** tengelyű fentebbi koordinátarendszerben: $\mathbf{x}^{(i)}=\mathbf{x}^{(i)}(s_i, \mathbf{a}_i)$. Az s_i mennyiség csak az *A2 algoritmusban* generált β_i véletlen számtól függ, ez minden i-re különböző. Az $\mathbf{a}_i$ vektor csak **b**-től és a felhasznált $\mathbf{y}^{(i+1)}$ bázisvektortól függ, de eloszlása csak $\mathbf{y}^{(i+1)}$-től függ, ez pedig egyenletes F-ben, így vetülete egyenletes eloszlású $A \cap F$-ben. Így ebben a koordinátarendszerben az $\mathbf{x}^{(i)}$ vektorok koordinátái függetlenek egymástól, tehát az $\mathbf{x}^{(i)}$, $i=1, \ldots, M-1$ vektorok is függetlenek. —

3.4. Az alapmódszerek ellipszoid esetére

Az alapmódszerek és a *BC bázispontcsere algoritmus* felhasználásával F-ben egyenletes eloszlású stacionárius vektorokat tudunk generálni.

A generált **y** vektor mindegyik módszerben az $\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(M)}$ vektorok lineáris kombinációja volt, vagyis $\mathbf{y}=c_1\mathbf{y}^{(i)}+c_2\mathbf{y}^{(j)}$ valamilyen c_1, c_2 konstansokkal egy i, j indexpárra. Az E ellipszoid felületén levő **Ty** vektor előállításához elegendő ismernünk az $\mathbf{y}^{(i)}$ vektorok $\mathbf{y}^{(i)*}=\mathbf{T}\mathbf{y}^{(i)}$, $i=1, \ldots, M$ transzformáltjait, mert $\mathbf{T}\mathbf{y}=c_1\mathbf{y}^{(i)*}+c_2\mathbf{y}^{(j)*}$. Tehát ha az új pivotvektor generálásakor annak transzformáltját is számítjuk és tároljuk, akkor a soron következő $M-1$ darab **y** vektor transzformálását nem kell ténylegesen végrehajtani. Így $M-1$ vektor generálásához csak egyszer kell egy $n(n+1)/2$ szorzást igénylő mátrix-szorzást elvégezni.

Az algoritmusok ellipszoid módszerekké való átalakítása a következő módosításokkal hajtható végre.

1. A bázispontcsere algoritmus előtt az $\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(M)}$ vektorok generálása után számítsuk ki és tároljuk az $\mathbf{y}^{(i)*}=\mathbf{T}\mathbf{y}^{(i)}$, $i=1, \ldots, M$ vektorokat is.

2. A *BC algoritmus* 4. lépésében az $\mathbf{y}^{(q)}$ vektor generálása után az $\mathbf{y}^{(q)*}=\mathbf{T}\mathbf{y}^{(q)}$ vektort is kiszámítjuk.

3. Az alapmódszerekben az $\mathbf{y} \leftarrow c_1\mathbf{y}^{(I)}+c_2\mathbf{b}$ vektor helyett az $\mathbf{y} \leftarrow c_1\mathbf{y}^{(I)*}+c_2\mathbf{b}^*$ vektor számítandó.

Az így generált vektorok egyenletes eloszlású stacionárius vektorok az E ellipszoid térfogatára nézve; E-ben egyenletes eloszlású vektorok generálása pedig egy véletlen sugárhossz generálási lépés hozzáadásával végezhető.

Az $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(M)}, \ldots$ vektorok sorozata korrelált. A korreláció csökkenthető néhány, kevés műveletet igénylő módosítással. Gömbi algoritmusok esetében véletlen előjelekkel lehet ellátni a generált vektor komponenseit.

M1. [Legyen **y** egy gömbi algoritmus eredménye.] Generáljuk u-t és jelöljük u_i-vel az i-edik bitet. Ha $u_i=0$, akkor változtassuk meg az **y** vektor i-edik komponensének előjelét, $y_i \leftarrow -y_i$, $i=1, \ldots, n$.

Ellipszoid algoritmusok esetén egy ennél több számítást igénylő módosítást lehet alkalmazni. Az E ellipszoid szimmetrikus az R mátrix $\mathbf{v}^{(i)}$ sajátvektorjaira; a generált vektort az egyik véletlenül választott sajátvektorra tükrözhetjük.

2. LEMMA. Legyenek a $\mathbf{v}^{(i)}$, $i=1, \ldots, n$ egységvektorok az $\mathbf{R}$ mátrix sajátvektorjai. Egy $\mathbf{y}$ vektornak a $\mathbf{v}^{(i)}$ sajátvektor által meghatározott irányra való tükörképe

$$\mathbf{z} = 2(\mathbf{v}^{(i)'}\mathbf{y})\mathbf{v}^{(i)} - \mathbf{y}.$$

Bizonyítás. Nyilvánvalóan igaz az $\mathbf{y}+\mathbf{z}=c\mathbf{v}^{(i)}$ egyenlőség valamilyen c konstanssal. Mivel $\mathbf{y}-\mathbf{z}$ merőleges $\mathbf{v}^{(i)}$-re, ezért

$$0 = \mathbf{v}^{(i)'}(\mathbf{y}-\mathbf{z}) = \mathbf{v}^{(i)'}(\mathbf{y}-c\mathbf{v}^{(i)}+\mathbf{y}) = 2(\mathbf{v}^{(i)'}\mathbf{y})-c,$$

amely a keresett összefüggést adja.

M2. [Legyen $\mathbf{y}$ egy ellipszoid algoritmus eredménye.] Generáljuk u-t és legyen $i \leftarrow [(n+1)u]+1$. Válasszunk le még egy bitet u-ról előjelnek, vagyis legyen $u \leftarrow (n+1)u-[(n+1)u]$, ha most $u<0{,}5$, akkor legyen $s \leftarrow +1$, egyébként $s \leftarrow -1$. Ha $i=0$, akkor adjuk át az $\mathbf{y} \leftarrow s\mathbf{y}$ vektort, egyébként pedig az $\mathbf{y} \leftarrow s(2(\mathbf{v}^{(i)'}\mathbf{y})\mathbf{v}^{(i)}-\mathbf{y})$ vektort.

További memóriát igényel MACLAREN és MARSAGLIA [76] egydimenziós véletlenszám generátorok esetén alkalmazott ötletének felhasználása. Az elgondolás lényege az, hogy a generált vektorokat nem előállításuk sorrendjében, hanem véletlen sorrendben használjuk fel.

Legyen $\mathbf{A}=(\mathbf{a}_1, \ldots, \mathbf{a}_N)$ egy $N \times n$-es mátrix. Az N számot célszerű egy $4M$-nél nagyobb 2 hatványnak választani például $N=128$. Az $\mathbf{A}$ mátrix $\mathbf{a}_1, \ldots, \mathbf{a}_N$ vektorjait feltöltjük a véletlen vektort generáló algoritmus N darab egymásutáni értékével. Míg az előző két módosítást a véletlen vektor generálásának elvégzése után kellett beilleszteni, most a véletlen vektort generáló algoritmus hívása a módosításon keresztül történik.

M3. [Az $\mathbf{A}$ mátrixban tároltuk a gömb vagy ellipszoid algoritmus által generált első N vektort.] Generáljuk u-t és legyen $i \leftarrow [Nu]+1$. Generáljunk az *alapmódszer+BC algoritmussal* egy $\mathbf{y}$ vektort. Az $\mathbf{a}_i$ vektor helyébe töltsük be az $\mathbf{y}$ vektort, míg $\mathbf{a}_i$ előző tartalmát adjuk át.

A leírt módosítások közül az *M1* és *M3* gömb esetén, az *M2* és *M3* pedig ellipszoid esetén egyidejűleg is alkalmazható. A módosítások lényeges tulajdonsága, hogy — az alapmódszerekhez hasonlóan — az n-dimenziószámtól csak lineárisan függő számú művelet végrehajtását igénylik.

Néhány megjegyzést fűzünk a közölt algoritmusokhoz. Az egydimenziós véletlenszám generátorok esetében több módszer ismeretes arra, hogyan lehet „memórián időt venni", vagyis több adat tárolásával az algoritmus futását gyorsabbá tenni. Példa erre a *Marsaglia Táblázat módszer*, valamint MARSAGLIA *RWT*, AHRENS és DIETER FL_5 algoritmusa normális eloszlás esetén. Többdimenziós vektorok generálása esetén erre gyakorlati megvalósítást nem találtunk. A gömbben, illetőleg ellipszoidban egyenletes eloszlás esetén ezt az átváltást csak részben sikerült elvégezni; a generált vektorok nem függetlenek, hanem csak több egymásutáni vektor együttese lesz stacionárius, de lényegesen gyorsabban tudjuk ezeket előállítani.

Az elvégzendő műveletek száma egy vektor generálása esetén durván egy négyzetgyökvonás, $2n$ szorzás és $2n$ összeadás, valamint egy normális eloszlású vektor generálási munkáinak $(M-1)$-ed része $\left(\text{ez kb. } \frac{n(n+1)}{2(M-1)} \text{ szorzás és összeadás}\right)$.

Véleményünk szerint az előállított vektorokat független vektorként is használhatjuk. Érvelésünk a következő. Az egydimenziós esetben a [0, 1) intervallumban előállított, egyenletesnek mondott számok nem függetlenek egymástól, eloszlásuk eltér az egyenletestől [37], [76], mégis úgy használjuk ezeket mint független, egyenletes eloszlású véletlen számokat. A közölt algoritmusok által generált vektorokról ugyanezek a tulajdonságok elmondhatók; de módosítások alkalmazásával a tulajdonságok javíthatók. Egy n-dimenziós ellipszoidban egyenletes eloszlású vektorok előállítása egy módosításokkal ellátott alapmódszer segítségével még mindig gyorsabb, mint az az egzakt módszer használata. A futási idők közötti különbség a dimenziószám növekedésével együtt nő.

3.5. Egy részletes algoritmus

Egyetlen algoritmust írunk le részletesen ellipszoid felszíni algoritmusként a *BC algoritmust+az A3 alapmódszert.* Azért esett választásunk erre a módszerre, mert az ennek a módszernek a felhasználásával előállított vektorokat nemcsak gyorsan lehetett megkapni, hanem a végső célunk, a többdimenziós normális eloszlásfüggvény kiszámításában is a legjobb eredményt adták (lásd az 5.5. pontban).

Az algoritmust abban a formában írjuk le, ahogy a továbbiakban használtuk; a *BC bázispontcsere algoritmust* az *A3 alapmódszerrel* egybeépítve és egyszerre $(M-1)$ darab vektort generálva, hiszen a 2. Tétel szerint ezek lesznek stacionáriusak.

Előkészítés

1. Generáljuk az S gömb F felszínén egyenletes eloszlású $\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(M)}$ vektorokat és számítsuk ki az $\mathbf{y}^{(i)*} \leftarrow \mathbf{T}\mathbf{y}^{(i)}$ vektorokat is. Legyen $q \leftarrow 1$.

Generálás

1. [Beállítjuk a bázisvektor indexét.] Legyen $I \leftarrow 0$.
2. Legyen $I \leftarrow I+1$. Ha $I=q$, akkor ismételjük meg ezt a lépést előlről. Ha $I=M$, akkor menjünk 6-ra.
3. Legyen $J \leftarrow I$. Ha $I>q$, akkor $J \leftarrow J-1$.
4. [Számítsuk ki a húr hosszúságát.] Határozzuk meg az $s^2 \leftarrow \sum_{j=1}^{n} (y_j^{(I)} - y_j^{(q)})^2$ mennyiséget és legyen $c \leftarrow 1/\sqrt{4-s^2}$.
5. [Töltsük be a generált vektort.] Határozzuk meg az $\mathbf{x}^{(J)} \leftarrow c(\mathbf{y}^{(I)*} + \mathbf{y}^{(q)*})$ vektort és menjünk vissza a 2. lépésre.
6. [Új bázisvektort generálunk.] Legyen $q \leftarrow q+1$, ha $q=M+1$, akkor legyen $q \leftarrow 1$. Generáljunk egy F-ben egyenletes eloszlású $\mathbf{y}$ vektort és legyen $\mathbf{y}^{(q)} \leftarrow \mathbf{y}$, $\mathbf{y}^{(q)*} \leftarrow \mathbf{T}\mathbf{y}$. Adjuk át az $\mathbf{x}^{(i)}$, $i=1, \ldots, M-1$ vektorokat.

3.6. Számítógépes tapasztalatok

A következő ellipszoid felületi algoritmusok számítógépes futási idejét adjuk meg msec-ban kifejezve (Honeywell 66/60).

1. Az *A2 alapmódszer* és a *BC bázispontcsere algoritmus* segítségével generált $(M-1)$ darab vektor előállításának ideje (az előző pontban az *A3 módszer* esetére leírt módhoz hasonlóan).
2. Az *A3+BC algoritmussal* $(M-1)$ darab vektor előállítása (lásd az előző pontban).
3. Az *A3+BC algoritmussal* 1 darab vektor előállítása.
4. Az egzakt módszer segítségével 1 darab vektor előállítása ellipszoidban: az *NO módszer+a mátrixtranszformálás* (lásd a 3.1. pontban).
5. Az egzakt gömbi algoritmus ideje: az *NO algoritmussal* egy darab vektor előállítása.

Az M számot néhány számítógépes próba után $M=20$-nak választottuk. Tapasztalataink szerint $M>20$ esetén a vektorok előállítási ideje nem csökkenthető lényegesen.

3.1 TÁBLÁZAT

Ellipszoid felületi algoritmusok ideje (msec)

Algoritmus \ $n=$	5	10	15	20	30	40	50
A2, $M-1$ darab	15,6	23,8	33,1	39,3	63,7	87,9	114,3
A2, $M-1$ darab	9,9	17,7	26,2	33,1	52,9	77,4	101,6
A3, 1 darab	0,5	0,9	1,4	1,7	2,8	4,1	5,3
NO, ellipszoid	1,0	2,6	4,7	7,4	14,8	25,2	36,7
NO, gömb	0,6	1,1	1,6	2,1	3,1	4,1	5,1

A táblázatból látható, hogy módszerünk segítségével az egzakt gömbi algoritmus futási idejével megegyező idő alatt tudunk ellipszoidban vektorokat előállítani. Az *A3 módszer* segítségével 2—7-szer gyorsabban tudunk vektorokat generálni, mint a leggyorsabb egzakt módszerrel és az idők aránya n növekedésével nő ($n=5$-re 2, $n=20$-ra 4,3 és $n=50$-re 7).

4. FEJEZET

Normális eloszlású vektorok generálása

4.1. Mátrixtranszformációs módszer

Három módszer ismeretes n-dimenziós, normális eloszlású véletlen vektorok generálására [12], [65], [119]. A mátrixtranszformációs módszerben a korreláció mátrixot trianguláris mátrixok szorzatára bontjuk fel. A második módszer az elsőtől csak abban tér el, hogy a korreláció mátrix felbontásánál nem törekszünk három-

szög alakú mátrixokat előállítani. A harmadik módszer a többdimenziós normális sűrűségfüggvénynek egydimenziós feltételes sűrűségfüggvények szorzatára való felbontásán alapszik. BARR és SLEZAK [12], valamint HURST és KNOP [65] számítógépes tapasztalatai szerint az első módszer a leggyorsabb — bár lényeges különbség a futási idők között nincsen — ezért csak ezt ismertetjük.

Legyen η egy n-dimenziós valószínűségi vektorváltozó, melynek komponensei függetlenek és normális eloszlásúak 0 várható értékkel és 1 szórással. Legyen feladatunk a

$$\varphi(\mathbf{x}) = (2\pi)^{-n/2}|\mathbf{R}|^{-1/2}\exp\left\{-\frac{1}{2}\mathbf{x}'\mathbf{R}^{-1}\mathbf{x}\right\} \tag{4.1}$$

sűrűségfüggvényű ξ vektorok előállítása, ahol $\mathbf{R}$ egy pozitív definit, szimmetrikus mátrix, a korreláció mátrix. Ismert összefüggés szerint az $\mathbf{R}$ mátrix felírható [48]

$$\mathbf{R} = \mathbf{T}\mathbf{T}' \tag{4.2}$$

alakban, ahol $\mathbf{T}$ egy alsó háromszög mátrix. A $\mathbf{T}\eta$ vektor (4.1) sűrűségfüggvényű [144]. A $\mathbf{T}$ mátrix t_{ij} elemeit a (4.2) egyenletrendszerből kaphatjuk meg, $t_{ij}=0$, ha $i<j$, egyébként

$$t_{i1} = r_{i1}/\sqrt{r_{11}}, \quad 1 \leqq i \leqq n,$$

$$t_{ii} = \sqrt{r_{ii} - \sum_{k=1}^{i-1} t_{ik}^2}, \quad 1 < i \leqq n,$$

$$t_{ij} = \left(r_{ij} - \sum_{k=1}^{j-1} t_{ik}t_{jk}\right)/t_{jj}, \quad 1 < j < i \leqq n.$$

Egy $\mathbf{m}$ várható értékű, $\mathbf{S}$ szórású véletlen vektor a következő módon kapható

$$\boldsymbol{\varphi} = (\mathbf{S}^{1/2}\mathbf{T})\boldsymbol{\eta} + \mathbf{m},$$

ahol az $\mathbf{S}$ mátrix elemeire $s_{ij}=0$, ha $i \neq j$.

4.2. Az Ellipszoid módszer

Legyen $\varphi(\mathbf{x})$ a (4.1) alatti sűrűségfüggvény, amelyet

$$\varphi(\mathbf{x}) = p_1 e_1(\mathbf{x}) + p_2 e_2(\mathbf{x}) + \ldots + p_k e_k(\mathbf{x}) + p_{k+1} r_1(\mathbf{x}) + p_{k+2} r_2(\mathbf{x}) \tag{4.3}$$

alakban írunk fel, ahol $p_1 + \ldots + p_{k+2} = 1$ és az e_i, $i=1, \ldots, k$, r_1, r_2 függvények n-dimenziós sűrűségfüggvények. A (4.3) egyenlőség alapján az egydimenziós kompozíciós módszerhez hasonlóan eljárás adható $\varphi(\mathbf{x})$ sűrűségfüggvényű vektorok generálására. Válasszuk ki a jobb oldali sűrűségfüggvények valamelyikét p_i valószínűséggel, a kiválasztott sűrűségfüggvény szerint generáljunk egy vektort és ezt adjuk át, mint $\varphi(\mathbf{x})$ sűrűségfüggvényűt. Az eljárás elvileg egyszerű, nehézséggel csak ott találkozunk, hogyan válasszuk meg a (4.3) jobb oldalán levő függvényeket, hogy szerintük véletlen vektorokat könnyen tudjunk generálni.

Az algoritmus részletes leírása előtt néhány jelölést vezetünk be. Legyen

$$E_i = \{\mathbf{x} | \varphi(\mathbf{x}) \geqq m_i\}, \quad i = 1, \ldots, k, \tag{4.4}$$

ahol $0 < m_1 < \ldots < m_k < m_{k+1} = \varphi(\mathbf{0})$ a $[0, \varphi(\mathbf{0})]$ intervallum egy felosztása.

Az E_i halmazok egymáshoz hasonló hiperellipszoidok, közös középpontjuk az origó, mivel

$$E_i = \{\mathbf{x} | \varphi(\mathbf{x}) \geqq m_i\} = \left\{\mathbf{x} | \mathbf{x}' \mathbf{R}^{-1} \mathbf{x} \leqq -2 \ln \left(m_i (2\pi)^{n/2} |\mathbf{R}|^{1/2}\right)\right\}.$$

Definiáljuk ezek után az e_i függvényt konstansnak az E_i ellipszoidban, azaz

$$e_i(\mathbf{x}) = \begin{cases} \dfrac{1}{V_i}, & \mathbf{x} \in E_i, \\ 0, & \mathbf{x} \notin E_i, \end{cases}$$

ahol $V_i = \int\limits_{E_i} d\mathbf{x}$ az E_i hiperellipszoid térfogata [70]. A p_i valószínűségeket a

$$p_i = V_i (m_i - m_{i-1}), \quad i = 1, \ldots, k$$

egyenletekből határozzuk meg. Megfelelő nagy k számot és a $[0, \varphi(\mathbf{0})]$ intervallumnak egy jó $m_1, \ldots, m_k$ felosztását véve a $p_1 + \ldots + p_k$ összeg 1-hez közeli. Definiáljuk a μ_i konstans szorzókat a

$$\mu_i = \frac{c_i}{c_1}, \quad i = 1, \ldots, k \tag{4.5}$$

egyenletek segítségével, ahol c_i az E_i hiperellipszoid konstansa, azaz $c_i^2 = -2 \ln \left(m_i (2\pi)^{n/2} |\mathbf{R}|^{1/2}\right)$. Ezeknek a szorzóknak a felhasználásával

$$E_i = \{\mathbf{x} | \mathbf{x} = \mu_i \mathbf{y}, \ \mathbf{y} \in E_1\}, \quad i = 1, \ldots, k$$

írható. Az E_i ellipszoidokat tehát az E_1 ellipszoid és a μ_i szorzók meghatározzák. Így az e_i sűrűségfüggvény kiválasztása helyett beszélhetünk a μ_i szorzók véletlenszerű választásáról. Tároljuk a diszkrét eloszlású μ_i szorzókat egy *Marsaglia Táblázatban*; $P\{\mu = \mu_i\} = p_i$.

A $p_{k+1} r_1(\mathbf{x})$ és $p_{k+2} r_2(\mathbf{x})$ függvényeket úgy határozzuk meg, hogy a (4.3) egyenlőség teljesüljön; az első lesz a $\varphi(\mathbf{x})$ sűrűségfüggvény széle, a második a vége. Definiáljuk a

$$p_{k+1} r_1(\mathbf{x}) = \begin{cases} \varphi(\mathbf{x}) - \sum\limits_{i=1}^{k} p_i e_i(\mathbf{x}), & \mathbf{x} \in E_1 \\ 0, & \mathbf{x} \notin E_1 \end{cases} \tag{4.6}$$

$$p_{k+2} r_2(\mathbf{x}) = \begin{cases} 0, & \mathbf{x} \in E_1 \\ \varphi(\mathbf{x}) & \mathbf{x} \notin E_1 \end{cases} \tag{4.7}$$

függvényeket. A p_{k+2} konstans meghatározható abból a feltételből, hogy $r_2(\mathbf{x})$ sűrűségfüggvény:

$$p_{k+2} = \int\limits_{R^n - E_1} \varphi(\mathbf{x}) \, d\mathbf{x} = \int\limits_{c_1^2}^{+\infty} k_n(z) \, dz,$$

ahol $k_n(z)$ az n szabadságfokú χ^2 eloszlás sűrűségfüggvénye. Így $p_{k+1} = 1 - p_{k+2} - \sum\limits_{i=1}^{k} p_i$ is meghatározható.

Egy megfelelő $m_1, \ldots, m_k$ felosztás meghatározása önmagában is nehéz számítástechnikai feladat. Leírását itt nem közöljük, csak vázlatosan foglaljuk össze az adaptív algoritmust [36]. Felosztjuk a $[0, \varphi(\mathbf{0})]$ intervallumot s egyenlő részre. Megtartjuk azokat az ellipszoidokat, amelyekre a megfelelő p_i valószínűség egy előírt kis valószínűségnél ($\sim$0,01 vagy 0,002) kisebb. Jelöljük k_1-gyel a megmaradt ellipszoidok számát. A $\left[0, \varphi(\mathbf{0})\left(1-\frac{k_1}{s}\right)\right]$ intervallumot újra s egyenlő részre osztva folytatjuk az eljárást. Az intervallum felosztását akkor fejezzük be, ha a $\varphi(\mathbf{x})$ sűrűségfüggvény végének (az $r_2(\mathbf{x})$ függvénynek) egy előírt valószínűségnél kisebb lesz a valószínűsége.

Ezek után leírhatjuk a részletes algoritmust normális eloszlású vektorok generálására.

1. [Kiválasztjuk valamelyik e_i függvényt.] Generáljuk u-t. Ha $u > \sum_{i=1}^{k} p_i$, akkor menjünk a 4. lépésre, egyébként a *Marsaglia Táblázatból* az u szám segítségével kiválasztunk egy μ_i szorzót.
2. Generáljunk egy E_1 hiperellipszoidban egyenletes eloszlású $\mathbf{z}$ vektort.
3. Számítsuk ki az $\mathbf{x} \leftarrow \mu_i \mathbf{z}$ vektort és menjünk a 6. lépésre.
4. [Generáljunk egy vektort az eloszlás széléből $r_1(\mathbf{x})$ sűrűségfüggvény szerint.] Ha $u > \sum_{i=1}^{k+1} p_i$, akkor menjünk az 5. lépésre, egyébként generáljunk egy $\mathbf{x}$ vektort az $r_1(\mathbf{x})$ sűrűségfüggvénnyel az elfogadás-elvetés módszerének felhasználásával és menjünk a 6. lépésre.
5. [Generáljunk egy vektort az eloszlás végéből.] Generáljunk egy z vektort, amely egyenletes eloszlású az E_1 ellipszoid térfogatára nézve. Az eloszlásfüggvény invertálásával generáljunk egy k véletlen számot, amely n szabadságfokú χ eloszlású és c_1-nél nagyobb [53]. Legyen $\mathbf{x} \leftarrow k\mathbf{z}$.
6. Adjuk át az $\mathbf{x}$ vektort, mint $\varphi(\mathbf{x})$ sűrűségfüggvényűt.

Az *Ellipszoid módszert* egy ún. *Húr módszerrel* együtt közöltük [36]. A *Húr módszer* egy közelítő eljárás n-dimenziós hiperellipszoidokban egyenletes eloszlású számok generálására. Az *Ellipszoid+Húr módszer* $n=5$—15 dimenziókban az előző pontban leírt mátrixtranszformálási módszernél kétszer-háromszor gyorsabban generált egy vektort.

4.3. A többdimenziós polár módszer és általánosítások

Két részre bontjuk fel a generálási tevékenységet: a vektor irányát és hosszát külön állítjuk elő.

Egy $\mathbf{R}$ korrelációmátrixszú normális eloszlású ξ vektor által meghatározott irány az

$$E = \{\mathbf{x} | \mathbf{x}' \mathbf{R}^{-1} \mathbf{x} \leqq 1\}$$

ellipszoid térfogatára nézve egyenletes eloszlású. Így ilyen eloszlású irányt egy ellipszoid felületi algoritmussal állíthatunk elő.

A $\varphi(\mathbf{x})$ sűrűségfüggvényű ξ vektornak az $s^2 = \xi' \mathbf{R}^{-1} \xi$ függvénye által meghatározott s mennyiséget nevezzük a vektor hosszának. Ez az $\mathbf{R}=\mathbf{I}$ független esetben a

vektor közönséges euklideszi hossza. Könnyen látható, hogy az így definiált s^2 valószínűségi változó n szabadságfokú χ^2 eloszlást követ, mivel

$$s^2 = \xi' \mathbf{R}^{-1} \xi = \eta' \mathbf{T}' \mathbf{R}^{-1} \mathbf{T} \eta = \eta' \eta = \sum_{i=1}^{n} \eta_i^2,$$

ahol η a független, normális eloszlású komponensekkel bíró valószínűségi vektorváltozó.

A polár módszer algoritmusa így a következő lesz:

1. [Legyen **y** egyenletes eloszlású az S gömb F felszínén.] Legyen $\mathbf{x} \leftarrow \mathbf{T}\mathbf{y}$, generáljuk az r véletlen számot az n szabadságfokú χ^2 eloszlásból.
2. Adjuk át az $\mathbf{x} \leftarrow \sqrt{r}\mathbf{x}$ vektort, mint $\varphi(\mathbf{x})$ sűrűségfüggvényűt.

A 3. fejezetben adott módszerek által generált egyenletes eloszlású stacionárius vektorokat használva a polár módszerben, eredményül stacionárius normális vektorokat kapunk.

A továbbiakban néhány megjegyzést teszünk arra vonatkozólag, hogyan lehet a közölt vektorgenerálási módszereket más eloszlásokra alkalmazni. Az *Ellipszoid módszer* természetes módon felhasználható néhány más eloszlásból származó vektor generálására is. Ezek az elliptikusan szimmetrikus eloszlások, melyek sűrűségfüggvénye csak a változók kvadratikus alakjától függ (lásd [67], IV. vol. 296. és [84]). Az ilyen eloszlások sűrűségfüggvénye a normális sűrűségfüggvényhez hasonlóan dekomponálható.

Elliptikusan szimmetrikus eloszlások közül a legismertebb a

$$h(\mathbf{x}) = \frac{\Gamma\left((\nu+n)/2\right)}{(\pi\nu)^{n/2}\Gamma(\nu/2)\,|\mathbf{R}|^{1/2}} \left(1+\nu^{-1}\mathbf{x}'\mathbf{R}^{-1}\mathbf{x}\right)^{-(\nu+n)/2} \tag{4.8}$$

sűrűségfüggvényű többdimenziós t eloszlás, ahol ν a megfelelő $\varkappa$ eloszlású változó szabadságfoka. Természetesen erre az eloszlásra a polár módszer is alkalmazható, csak a „hossz" eloszlása lesz más: egy F eloszlású valószínűségi változó négyzetgyöke.

Az ismert egydimenziós kompozíciós módszerek (MARSAGLIA *RWT* és AHRENS—DIETER FL_5 algoritmusa) esetében a sűrűségfüggvény értelmezési tartományát osztják fel k részre és mindegyik kis részintervallumban felvesznek egy sűrűségfüggvényt, amely szerint könnyű mintát generálni. Ez a „függőleges" felosztás a k darab intervallum és a hozzájuk tartozó sűrűségfüggvény adatainak tárolását igényli. Ilyen dekomponálást n dimenzióban alkalmazva k^n darab függvényt, illetőleg ugyanennyi n-dimenziós kockát kellene nyilvántartani, ami gyakorlatilag már $n \geqq 3$ esetén is kivitelezhetetlen.

Az általunk az *Ellipszoid módszerben* alkalmazott elgondolás a sűrűségfüggvény értékkészletét osztja fel k részre (vízszintes felbontás). Ez több dimenzióban is alkalmazható, az eljárás a dimenziószámtól nem függ, csak a sűrűségfüggvény alakjától. Ennek a vízszintes dekomponálásnak a felhasználásával lehetőség nyílik többdimenziós vektorokat gyorsan generáló algoritmusok készítésére. Az elliptikusan szimmetrikus sűrűségfüggvények mellett még egy lehetséges alkalmazási területet emelünk ki: az $f(\mathbf{x}) = \psi(Q(\mathbf{x}))$ sűrűségfüggvényű logaritmikusan konkáv eloszlásokat, ahol a $Q(\mathbf{x})$ függvény konvex. Ezeknek a sztochasztikus programozásban igen fontos szerepet betöltő eloszlásoknak a sűrűségfüggvényéről tudjuk, hogy a nívóhalmazaik

korlátos konvex halmazok (lásd [99], p. 55). Ezeken a halmazokon egyenletes eloszlások deflniálhatók és a (4.3) dekompozícióban ezeknek az eloszlásoknak a sűrűségfüggvényét vehetjük. A számítástechnikai megvalósítás feltétele az, hogy az $\{\mathbf{x}|f(\mathbf{x})\geqq c\}$ halmazban gyorsan tudjunk egyenletes eloszlású pontokat előállítani. Erre — a hívóhalmazokat téglatestbe vagy ellipszoidba foglalva — az elfogadás-elvetés módszerének felhasználásával a konvexitás miatt lehetőség van.

4.4. Számítógépes tapasztalatok

A polár módszer egyszerű felépítésű és kevés memóriát igényel. Az *Ellipszoid módszer* nem egyöntetű, bonyolult előkészítést igényel, sok memóriát foglal, viszont χ^2 eloszlású számok generálása helyett egy x^n eloszlású véletlen szám előállítását igényli. A számítógépes tapasztalatok a fentieket igazolják, az *Ellipszoid módszer* kissé gyorsabb a polár módszernél. A 3.1. táblázat és a következő 4.1. táblázat összehasonlításából látható, hogy egy irány előállítása több ideig tart, mint egy normális vektor előállítása, tehát mind az *Ellipszoid módszer,* mind a polár módszer lassabb a mátrixszorzásnál.

Két módszerre adjuk meg a futási időket a 4.1. táblázatban: a mátrixszorzásos módszerre és az *A3 módszer* által generált irányokat használó polár módszerre. Ez utóbbi módszer stacionárius normális vektorokat állít elő; gyorsabban, mint a FRANKLIN [49] által stacionárius vektorok generálására javasolt módszer, amely egy mátrixszorzást igényel, így a mátrixszorzásos módszernél nem gyorsabb.

4.1 TÁBLÁZAT

A független és a stacionárius normális vektorok előállításának ideje (msec)

Algoritmus \ $n=$	5	10	15	20	30	40	50
Mátrixszorzás	0,8	2,2	4,2	6,9	13,9	23,4	34,8
A3+polár	0,8	1,3	1,8	2,1	3,2	4,4	5,7

5. FEJEZET

A többdimenziós normális eloszlás eloszlásfüggvényének kiszámítása

A sztochasztikus programozásban és a többváltozós statisztikai analízisben gyakran szükség van többdimenziós halmazok valószínűségének meghatározására normális eloszlás esetén. PRÉKOPA sztochasztikus programozási modelljei [101], különösen a STABIL modell [102], [103] és víztározási modellek [105], [106], valamint humángenetikai vizsgálatok [15] számítógépes feldolgozása során kell ilyen valószínűségeket meghatározni. Ebben a fejezetben több algoritmust közlünk az

$$I = \Phi(\mathbf{h}) = \int_{-\infty}^{h_1} \dots \int_{-\infty}^{h_n} \varphi(\mathbf{x})\, d\mathbf{x} \tag{5.1}$$

integrál kiszámítására, ahol Φ az n-dimenziós $\mathbf{0}$ várható érték vektorú, $\mathbf{R}$ korreláció mátrixszú normális eloszlás eloszlásfüggvénye, $\varphi(\mathbf{x})$ pedig a sűrűségfüggvénye, azaz

$$\varphi(\mathbf{x}) = (2\pi)^{-n/2} |\mathbf{R}|^{-1/2} \exp\left\{-\frac{1}{2} \mathbf{x}' \mathbf{R}^{-1} \mathbf{x}\right\}. \tag{5.2}$$

5.1. *Monte Carlo integrálás*

Az (5.1) integrál *Monte Carlo módszerrel* történő kiszámítását az integrálközép módszerével lehet elvégezni. Csonkoljuk az (5.1) integrál D integrálási tartományát a $D^* = \{\mathbf{x} | -\mathbf{b} \leqq \mathbf{x} \leqq \mathbf{h}\}$ téglatestre valamely konstans $\mathbf{b} > \mathbf{0}$ vektorral. D^* területét T^*-gal jelölve az integrál egy becslése

$$\theta_1 = \frac{1}{N} T^* \sum_{i=1}^{N} \varphi(\mathbf{v}^{(i)}), \tag{5.3}$$

ahol $\mathbf{v}^{(i)}$ egyenletes D^*-ban. Erre a becslésre építhető egy *V1 algoritmus* a valószínűség kiszámítására.

A becslésnek egy, a gyakorlatban előforduló feladatok esetén jobb változata az a módszer, amikor közvetve integrálunk. Pontosabban, legyen $K = \{\mathbf{x} | -\mathbf{b} \leqq \mathbf{x} \leqq \mathbf{b}\}$ és az integrálást a T^{**} területű $K - D^*$ tartomány felett végezzük

$$\theta_2 = \frac{1}{N} T^{**} \sum_{i=1}^{N} \varphi(\mathbf{v}^{(i)}), \tag{5.4}$$

ahol most $\mathbf{v}^{(i)}$ a $K - D^*$-ban egyenletes. Erre építhető a *V2 algoritmus*. Ez a *V1 algoritmusnál* jobb — kisebb szórású eredményt ad, de a gyakorlatban $n = 6$ dimenzió felett ez a módszer is használhatatlan [31].

Monte Carlo módszerek hibáját kétféleképpen szokták megadni. A becslés szórását adják meg, amelyet a módszer standard hibájának neveznek, vagy pedig a központi határeloszlástétel alapján konfidenciaintervallumot adnak a becslésre. A szórást általában a generált mintából számítjuk. Jelöljük σ-val a $T^* \varphi(\mathbf{v}^{(i)})$ valószínűségi változó szórását, akkor a θ_1 becslés standard hibája $\sigma/\sqrt{N}$. A θ_1 valószínűségi változó közelítőleg normális eloszlású, így a becslés hibáját a

$$P\{|\theta_1 - I| < 1{,}97\sigma/\sqrt{N}\} \geqq 0{,}95$$

valószínűséggel is jellemezhetjük.

5.2. *Elfogadás — elvetés*

A keresett (5.1) érték felírható

$$p = P\{\xi \in D\} \tag{5.5}$$

alakban is, ahol ξ egy $\varphi(\mathbf{x})$ sűrűségfüggvényű normális eloszlású valószínűségi változó. Legyen

$$f(\mathbf{x}) = \begin{cases} 1, & \text{ha } \mathbf{x} \leqq \mathbf{h}, \\ 0, & \text{különben.} \end{cases}$$

Így a keresett érték

$$p = \int_{R^n} f(\mathbf{x})\, d\Phi(\mathbf{x}) \tag{5.6}$$

alakba írható. A megfelelő becslés

$$\theta_3 = \frac{1}{N} \sum_{i=1}^{N} f(\xi^{(i)}), \tag{5.7}$$

az erre épülő *V3 algoritmus* pedig a következő: N darab $\varphi(\mathbf{x})$ sűrűségfüggvényű ξ vektort generálunk, ha ezek közül N_1 a D tartományba esik akkor a $p \sim N_1/N$ értéket fogadjuk el. Az $f(\xi^i)$ valószínűségi változó binomiális eloszlású p paraméterrel, tehát a θ_3 becslés standard hibája

$$D(\theta_3) = \sqrt{p(1-p)/N}.$$

Ennek a becslésnek két jó tulajdonsága van. Egyrészt a $\sigma = \sqrt{p(1-p)}$ szórás 0,5-nél mindig kisebb, másrészt a hiba a dimenziószámtól nem függ. Egy $n = 100$ dimenziós, normális eloszlású vektor generálásának ideje 0,13 sec, így a θ_3 becsléssel 0,1 standard hibájú eredményt 3 sec alatt lehet kapni, 0,01 hibájút pedig 300 sec alatt $n = 100$ esetén.

A *Csebisev egyenlőtlenséget* alkalmazva a

$$P\{|p-\theta_3| \geqq \varepsilon\} \leqq \frac{p(1-p)}{N\varepsilon^2}$$

egyenlőtlenségből adott megbízhatósághoz N értéke kiszámítható. Kisebb mintaszámot kapunk, ha a *Csebisev egyenlőtlenség* BERNSTEJN által élesített formáját használjuk, amely szerint

$$P\{|p-\theta_3| \geqq \varepsilon\} \leqq 2 \exp\left[- \frac{N\varepsilon^2}{2p(1-p)\left[1 + \frac{\varepsilon}{2p(1-p)}\right]^2} \right].$$

A *V3 algoritmus* módosítható úgy, hogy kis valószínűségekre ($p < 0{,}3$) az $f(\xi^{(i)})$ függvényérték kiszámítása kevesebb munkát igényeljen. Az alapgondolatunk az, hogy nem számítjuk ki minden egyes i index esetén a $\xi^{(i)}$ vektor minden komponensét. [32]

5.3. *A polár módszer felhasználása*

A normális sűrűségfüggvény egy ellipszoid felett kiintegrálható. Ezt felhasználva konstruálható egy *V5 algoritmus* a valószínűség kiszámítására. A módszer a nagy valószínűségek esetén működik igen gyorsan [33].

5.4. *Iránymenti integrálás*

Az előbbi módszerek közül az elfogadás—elvetés θ_3 becslése az egyetlen, amely általánosan alkalmazható $n = 50$ dimenzióban is, minden p valószínűségre. A továbbiakban ezt a becslést javítjuk azáltal, hogy a szórását csökkentjük.

A θ_3 becslés az $f(\xi^{(i)})$ valószínűségi változók számtani közepe. Az $f(\xi^{(i)})$ változó csak a 0 vagy 1 értéket veheti fel, attól függően, hogy a $\xi^{(i)}$ vektor a D tartományon kívül esik, vagy benne van. Célszerű az $f(\xi^{(i)})$ helyett olyan valószínűségi változót keresni, amely nem csak a 0 és 1 értékeket veheti fel.

Használjuk a polár módszert a $\xi^{(i)}$ vektor generálására. Ha egy irányt már előállítottunk, akkor meghatározhatjuk, hogy az adott irány esetén mi a valószínűsége a D tartományba esésnek, vagyis az adott irány mentén integrálhatunk.

Jelöljük $V(y)$-nal az E ellipszoid térfogatára nézve egyenletes eloszlású η valószínűségi változó eloszlásfüggvényét és F_n-nel az n szabadságfokú χ eloszlású k valószínűségi változó eloszlásfüggvényét. Ekkor

$$\int_{R^n} f(\mathbf{x})\,d\Phi(\mathbf{x}) = \int_{E_f}\int_0^{\infty} f(k\mathbf{y})\,dF_n(k)\,dV(\mathbf{y}) = \int_{E_f} e(\mathbf{y})\,dV(\mathbf{y}), \tag{5.8}$$

ahol az $e(\mathbf{y})$ függvény az E ellipszoid $E_f=\{\mathbf{y}|\mathbf{y}'\mathbf{R}^{-1}\mathbf{y}=1\}$ felületén értelmezzük $\mathbf{h}\geqq\mathbf{0}$ esetére (a többi esetre az értelmezés könnyen kiterjeszthető)

$$e(\mathbf{y}) = F_n\Big(\max_{\substack{r\mathbf{y}\in D\\ r\geqq 0}} r\Big). \tag{5.9}$$

Az F_n függvény argumentumában szereplő maximum könnyen kiszámítható adott y-ra, ugyanis azt az r konstanst kell kiszámítani, amelyre az $r\mathbf{y}$ vektor a D tartomány határán van. Ehhez az

$$r = \begin{cases} \min\limits_{y_i>0} \dfrac{h_i}{y_i}, & \text{ha} \quad \mathbf{y} \not\leqq \mathbf{0}, \\ +\infty, & \text{ha} \quad \mathbf{y} \leqq \mathbf{0} \end{cases} \tag{5.10}$$

minimumot kell kiszámítani. Ugyanis, ha a minimum a j indexre teljesül, akkor az $r\mathbf{y}$ vektor a $h_i=x_i$, $i=1,\ldots,n$ hipersíkok közül a j-ediket metszi el először.

Az (5.8) integrálás 1-gyel csökkenti a szabad változók számát; az integrálás az n-dimenziós térben elhelyezkedő $(n-1)$-dimenziós felületen végzendő.

A *Monte Carlo módszerekben* az ellentétes változók — *(antithetic variates)* — szóráscsökkentő eljárása úgy használható, hogy a becslésben negatívan korrelált valószínűségi változók összegét számítjuk ki. Most ez az eljárás úgy alkalmazható, hogy az $\mathbf{y}$ vektor mellett a $-\mathbf{y}$ vektort is felhasználjuk a becslésben. Használjuk a

$$g(\mathbf{y}) = \frac{1}{2}\left(e(\mathbf{y})+e(-\mathbf{y})\right) \tag{5.11}$$

jelölést, ennek segítségével a becslés

$$\theta_4 = \frac{1}{N}\sum_{i=1}^{N} g(\mathbf{y}^{(i)}), \tag{5.12}$$

ahol az $\mathbf{y}^{(i)}$, $i=1,\ldots,N$ vektorok $V(\mathbf{y})$ eloszlásfüggvényűek, azaz az E ellipszoid térfogatára nézve egyenletes eloszlásúak.

Az ortáns esetben ($\mathbf{h}=\mathbf{0}$) a $g(\mathbf{y})$ valószínűségi változó csak a 0 vagy 0,5 értéket veheti fel; $P\{g(\mathbf{y})=0{,}5\}=2p$, így

$$D^2(\theta_4) = \frac{1}{N}\left[2p\,\frac{1}{4}-p^2\right] = \frac{p(0{,}5-p)}{N} < D^2(\theta_3) = \frac{p(1-p)}{N}.$$

Nem ortáns esetben a θ_3 valószínűségi változó szórása általában kétszerese a θ_4 becslés szórásának. Mivel $\mathbf{h}=\mathbf{0}$, $p=0{,}5$-re $D^2(\theta_4)=0$, ezért a θ_4 becslés $n=1$ dimenzióban a pontos valószínűséget adja.

Végül egy algoritmust adunk a $g(\mathbf{y})$ függvény kiszámítására.

A $g(\mathbf{y})$ függvény (5.11) kiszámítása

1. Legyen $c_1 \leftarrow K$, $c_2 \leftarrow K$ és $i \leftarrow 0$ (K egy nagy pozitív konstans, $K=10^{10}$).
2. Növeljük az indexet $i \leftarrow i+1$. Ha $i>n$, akkor menjünk az 5. lépésre.
3. Legyen $s \leftarrow h_i/y_i$, ha $s<0$, akkor menjünk 4-re. Ha $s<c_1$, akkor legyen $c_1 \leftarrow s$. Menjünk vissza 2-re.
4. Ha $-s<c_2$, akkor legyen $c_2 \leftarrow s$. Menjünk vissza 2-re.
5. Legyen $s \leftarrow 0{,}5\left(F_n(c_1)+F_n(c_2)\right)$ és adjuk át az s-t, mint a $g(\mathbf{y})$ függvény értékét.

5.5. *Ortonormált alappontok*

Az egydimenziós kvadratúra képletekhez hasonlóan a g függvény integrálására is adhatók „ekvidisztans" alappontok.

Legyen $B=(\mathbf{b}^{(1)}, \ldots, \mathbf{b}^{(n)})$ egy ortonormált, véletlen elhelyezkedésű vektorrendszer az n-dimenziós térben. Tekintsük ennek $\mathbf{U}=\mathbf{TB}$ transzformáltját, ahol $\mathbf{R}=\mathbf{TT}'$, a kapott $\mathbf{u}^{(i)}=\mathbf{T}\mathbf{b}^{(i)}$, $i=1, \ldots, n$ vektorrendszer $\mathbf{R}^{-1}$ ortonormált, vagyis $\mathbf{u}^{(i)\prime}\mathbf{R}^{-1}\mathbf{u}^{(j)}=$ $=\delta_{ij}$, ahol $\delta_{ij}=0$, ha $i \neq j$ és $\delta_{ij}=1$, ha $i=j$. Alkalmazzuk az $\mathbf{u}^{(i)}$ vektorok mindegyikére a θ_4 becslést, vagyis legyen ϱ a következő valószínűségi változó.

$$\varrho = \frac{1}{n} \sum_{i=1}^{n} g(\mathbf{u}^{(i)}),$$

akkor ennek a $\varrho^{(j)}$, $j=1, \ldots, N$ realizációit felhasználva kapjuk a

$$\theta_5 = \frac{1}{N} \sum_{j=1}^{N} \varrho^{(j)} \tag{5.12}$$

becslést a valószínűségre.

Egy ortonormált $\mathbf{B}$ bázis ennél jobban is kihasználható. Tekintsük a $\mathbf{b}^{(i)}$, $i=$ $=1, \ldots, n$ vektorokat, ezek közül k darabot $\binom{n}{k}$ különböző módon tudunk kiválasztani, legyen egy ilyen kiválasztás eredménye a $\mathbf{b}^{(i_1)}, \ldots, \mathbf{b}^{(i_k)}$ vektorokból álló rendszer. Ezeknek a felhasználásával képezzük a

$$\mathbf{b}^{(\mathbf{s}'\mathbf{i})} = \frac{1}{\sqrt{k}} [s_1 \mathbf{b}^{(i_1)} + s_2 \mathbf{b}^{(i_2)} + \ldots + s_k \mathbf{b}^{(i_k)}]$$

vektorokat, ahol $s_i=+1$ vagy -1, $i=1, \ldots, k$. A $\mathbf{b}^{(\mathbf{s}'\mathbf{i})}$ vektorok egységnyi hosszúságúak. Az $\mathbf{s}$ előjelvektort egy $(i_1, \ldots, i_k)$ indexrendszerhez 2^{k-1} különböző módon adhatjuk meg úgy, hogy a $\mathbf{b}^{(\mathbf{s}'\mathbf{i})}$ vektorok mind különböző egyenest határozzanak

meg. Tehát a **B** ortonormált bázis vektorjai közül k darabot kiválasztva és ezek minden lehetséges előjellel képzett összegét tekintve összesen $2^{k-1}\binom{n}{k}$ különböző irányt állíthatunk elő. Jelöljük O_k-val azt a becslést, amelyet a θ_4 becslésnek erre a $2^{k-1}\binom{n}{k}$ számú vektorra való alkalmazásával kapunk. Természetesen O_1 azonos a θ_5 becsléssel.

Ha a **B** bázis vektorjainak $\mathbf{u}^{(i)}=\mathbf{T}\mathbf{b}^{(i)}$ transzformáltjait tároljuk, akkor az O_k becslésnél szereplő $\mathbf{T}\mathbf{b}^{(s,i)}$ vektort nem mátrixszorzással, hanem összeadással és kivonással lehet megkapni:

$$\mathbf{T}\mathbf{b}^{(s,i)} = \mathbf{T}\frac{1}{\sqrt{k}}\sum_{j=1}^{k} s_j \mathbf{b}^{(i_j)} = \frac{1}{\sqrt{k}}\sum_{j=1}^{k} s_j \mathbf{u}^{(i_j)}.$$

Felvetődik a kérdés, hogy az $O_1, \ldots, O_n$ becslések közül adott dimenziószám esetén melyiket használjuk. Erre a kérdésre csak az adott számítógépre megírt programmal való kísérletezés adhat végleges választ, de néhány megfontolás tehető.

Egy becslés a másiknál hatékonyabbnak tekinthető, ha felhasználásával ugyanannyi idő alatt kisebb szórású becslést kapunk. Az idő helyett a végzendő műveletek számát vizsgálhatjuk, a szórás helyett pedig a vektorok számát (feltéve, hogy a független esethez hasonlóan több vektor felhasználása kisebb szórású eredményt ad). Az O_k becslésben felhasználandó vektorok közül egynek az előállításához $n(k-1)$ kivonás vagy összeadás szükséges, valamint k számú index növelése és ellenőrzése (programozási fogásokkal ez némileg csökkenthető). Jelöljük t_1-gyel egy összeadás idejét, t_2-vel egy **B** bázis előállításának és transzformálásának idejét. Az O_k becslés egy vektorának előállításához szükséges idő durván

$$n(k-1)t_1 + \frac{t_2}{2^{k-1}\binom{n}{k}}. \tag{5.13}$$

Így az O_{k-1} becslés O_k-nál hatékonyabb lehet, ha az U rendszer előállításának ideje a vektorok számához képest kicsi — annak ellenére, hogy az O_k becslésben több vektort használunk fel mint O_{k-1}-ben — mivel az egy vektor kiszámításához szükséges idő lényegesen megnő (5.13) szerint.

Természetesen adott n-re azok az O_k becslések, amelyek az O_{k-1} becslésnél kevesebb vektort állítanak elő (de több munkával), az O_{k-1} becslésnél kevésbé hatékonyak. A jellemző k_0 küszöbszám az $N_k/N_{k-1}<1$ feltételből meghatározható, ahol N_k adott n-re az O_k becslésben levő vektorok száma ($n=6$-ra $k_0=5$, így O_5 már nem jobb O_4-nél).

Ezeknél a becsléseknél nagyon sok vektor esetében kell a g függvény értékét kiszámítani, ezért módosítjuk az előző pont végén leírt eljárást. Megmutatjuk, hogy a 3. lépésben szereplő osztás elkerülhető.

A módosított függvényérték kiszámítását egyszerűség kedvéért az O_2 becslés felhasználása során előállított vektorokra írjuk le. Az $\mathbf{u}^{(i)}=\mathbf{T}\mathbf{b}^{(i)}$ vektorok kiszámításával egyidejűleg számítsuk ki és tároljuk a

$$v_{ij} = \frac{u_j^{(i)}}{h_j}$$

mennyiségeket. A g függvény kiszámításához az $\frac{1}{\sqrt{2}}(\mathbf{u}^{(i)}+\mathbf{u}^{(j)})$ vektor esetén a következő minimumot kell meghatározni:

$$r=\min_k\left\{\frac{h_k}{\frac{1}{\sqrt{2}}(u_k^{(i)}+u_k^{(j)})}\,\middle|\,\frac{h_k}{u_k^{(i)}+u_k^{(j)}}>0\right\}=\min_k\left\{\frac{\sqrt{2}}{\frac{u_k^{(i)}}{h_k}+\frac{u_k^{(j)}}{h_k}}\,\middle|\,\frac{u_k^{(i)}+u_k^{(j)}}{h_k}>0\right\}=$$

(5.14)

$$=\min_k\left\{\frac{\sqrt{2}}{v_{ik}+v_{jk}}\,\middle|\,v_{ik}+v_{jk}>0\right\}=\frac{\sqrt{2}}{\max_k\{v_{ik}+v_{jk}|v_{ik}+v_{jk}>0\}}.$$

Természetesen hasonlóan számítható a minimum a $\frac{h_k}{u_k^{(i)}+u_k^{(j)}}<0$ feltétellel adott k-indexekre.

Ezzel a módosítással az n számú osztás helyett egy vektor esetén csak egy osztásra van szükség a függvényérték kiszámításához. Az előző kiszámítási módnál egyetlen **B** bázisból kiindulva az O_2 becslésnél $n^2(n-1)$ osztásra lenne szükség, míg a fentebbi módosítással csak $3n^2-2n$ osztás kell (n^2 osztás a v_{ij} értékek meghatározásához, $2n(n-1)$ osztás a $\sqrt{2}$ és a maximum hányadosának előállításához).

A fentebbi gondolatokat összefoglalva leírjuk az O_2 becslés kiszámításához szükséges részletes algoritmust, vagyis a

$$\varrho=\frac{1}{n(n-1)}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}g\left(\frac{1}{\sqrt{2}}(\mathbf{u}^{(i)}+\mathbf{u}^{(j)})\right)+g\left(\frac{1}{\sqrt{2}}(\mathbf{u}^{(i)}-\mathbf{u}^{(j)})\right)$$

valószínűségi változó $\varrho^{(i)}$, $i=1, \ldots, N$ realizációból számítható

$$\theta_6=O_2=\frac{1}{N}\sum_{i=1}^{N}\varrho^{(i)} \tag{5.15}$$

közelítést. Az $O_3, O_4, \ldots$ stb. becslések algoritmusa az alábbihoz hasonló.

V6 Valószínűség ortonormált 2 becsléssel

1. Állítsuk be a $p\leftarrow 0$, $q\leftarrow 0$ kezdeti értékeket.
2. Legyen $q\leftarrow q+1$, ha $q>N$, akkor menjünk a 14. lépésre. Generáljuk a $\mathbf{b}^{(i)}$, $i=1, \ldots, n$ vektorokat, amelyek komponensei függetlenek, standard normális eloszlásúak. Legyen $s\leftarrow\sum_{i=1}^{n} b_i^{(1)2}$, $\mathbf{b}^{(1)}\leftarrow\mathbf{b}^{(1)}/\sqrt{s}$ és $i\leftarrow 1$.
3. [Ortogonalizáljuk a **B** vektorrendszert.] Legyen $i\leftarrow i+1$ és képezzük az $x_j\leftarrow\mathbf{b}^{(i)'}\mathbf{b}^{(j)}$, $j=1, \ldots, i-1$ skalárszorzatokat. Módosítsuk a $\mathbf{b}^{(i)}$ vektort: $\mathbf{b}^{(i)}\leftarrow\mathbf{b}^{(i)}-\sum_{j=1}^{i-1}x_j\mathbf{b}^{(j)}$, és végül normáljuk $s\leftarrow\sum_{j=1}^{n}b_j^{(i)2}$, $\mathbf{b}^{(i)}\leftarrow\mathbf{b}^{(i)}/\sqrt{s}$. Ha $i\neq n$, akkor ismételjük meg ezt a lépést előlről.

4. [Előállítjuk a transzformált **U** rendszert.] Legyen $\mathbf{u}^{(i)} \leftarrow \mathbf{T}\mathbf{b}^{(i)}$, $i=1, \ldots, n$, ekkor $\mathbf{u}^{(i)}$-k egy véletlen $\mathbf{R}^{-1}$ ortonormált rendszert alkotnak. Számítsuk ki a $v_{ij} \leftarrow u_j^{(i)}/h_j$ értékeket minden $i, j=1, \ldots, n$ indexre és legyen $i \leftarrow 0$, $s \leftarrow 0$.
5. [A vektorok képzéséhez index beállítás.] Legyen $i \leftarrow i+1$, ha $i>n-1$, akkor menjünk a 13. lépésre. Legyen $j \leftarrow i$.
6. Legyen $j \leftarrow j+1$, ha $j>n$, akkor menjünk vissza az 5. lépésre. Állítsuk be a $k \leftarrow 0$, $c_1 \leftarrow \varepsilon$, $c_2 \leftarrow \varepsilon$, $c_3 \leftarrow \varepsilon$, $c_4 \leftarrow \varepsilon$, $(\varepsilon = 10^{-10})$ értékeket.
7. [Az (5.14) kifejezés kiszámítása.] Növeljük az indexeket, $k \leftarrow k+1$, ha $k>n$, akkor menjünk 12. lépésre. Számítsuk ki az $x \leftarrow v_{ik}+v_{jk}$ és az $y \leftarrow v_{ik}-v_{jk}$ értékeket.
8. Ha $x<0$, akkor menjünk a 9. lépésre. Ha $c_1<x$, akkor $c_1 \leftarrow x$. Menjünk a 10. lépésre.
9. Ha $c_2<-x$, akkor $c_2 \leftarrow -x$.
10. Ha $y<0$, akkor menjünk a 11. lépésre. Ha $c_3<y$, akkor $c_3 \leftarrow y$. Menjünk vissza a 7. lépésre.
11. Ha $c_4<-y$, akkor $c_4 \leftarrow -y$. Menjünk vissza a 7. lépésre.
12. [Számítsuk ki a g függvény értékeit.] A g függvény értékeit hozzáadjuk az eddig kiszámítottakhoz $s \leftarrow s+F_n(\sqrt{2}/c_1)+F_n(\sqrt{2}/c_2)+F_n(\sqrt{2}/c_3)+F_n(\sqrt{2}/c_4)$ és menjünk vissza a 6. lépésre.
13. Legyen $p \leftarrow p+s/(2n(n-1))$ és menjünk a 2. lépésre új bázis generálásához.
14. Adjuk át a $p \leftarrow p/N$ értéket mint a valószínűség közelítő értékét.

Az O_3 becslés algoritmusa a fentiből kis módosítással kapható. A 7. lépésben x és y helyett az $x_1 \leftarrow v_{ik}+v_{jk}+v_{lk}$, $x_2 \leftarrow v_{ik}+v_{jk}-v_{lk}$, ... értékeket kell számítani, a 8—11. lépések megkétszerezendők stb.

Végül néhány olyan becslést sorolunk fel, amelyeket kipróbáltunk, de nem bizonyultak jónak: vagy a szórás nem csökkent a kiindulásként használt becslés szórásához képest, vagy a kiszámítási idő növekedett meg igen nagy mértékben.

Tekintsünk egy **B** ortonormált véletlen vektorrendszert és egy $\mathbf{e}^{(i)}$ i-edik egységvektor által meghatározott $\mathbf{e}^{(i)'}\mathbf{x}=0$ hipersíkot. A **B** bázisnak a hipersíkra való tükörképe is bázis. A tükrözés végrehajtása csak az egyes vektorok i-edik előjelének megváltoztatásából áll, valamint az is könnyen belátható, hogy a **T** mátrixszal való transzformálást sem kell teljes egészében újra elvégezni a tükrözött bázis esetében (az $\mathbf{U}=\mathbf{TB}$ rendszert felhasználjuk a tükrözött bázis transzformáltjának kiszámításához). Több előjelet egyszerre is változtathatunk. A becslésben **B**-t és **B** tükörképeit használtuk O_k becslések kiindulásaként.

Az E hiperellipszoidnak az **R** mátrix $\mathbf{v}^{(i)}$ sajátvektorjai által meghatározott egyenesek szimmetriatengelyei. Egy $\mathbf{U}=\mathbf{TB}$ rendszernek a $\mathbf{v}^{(i)}$ vektor egyenesére való tükörképe könnyen megkapható, és szintén $\mathbf{R}^{-1}$ ortonormált lesz (lásd a 3.3. pontban az *M2 módosítást*).

Azok a becslések, amelyek az O_k becslésekből oly módon keletkeznek, hogy az egyes vektorokra nem a g függvényt számítjuk ki, hanem egy véletlen vektorhosszat generálunk és a beesési gyakoriságot számítjuk — mind lényegesen rosszabbak.

Megpróbálkoztunk regresszióval, mind a közölt $\theta_3-\theta_6$ becslések, mind a fentebb megemlített becslések között; nem sikerült olyan változatot találni, amely a példák nagy részében jobb eredményre vezetett volna.

A legkisebb négyzetek módszerével a g függvény egy kvadratikus közelítését is próbáltuk meghatározni; az approximálás csak alacsony, $n<5$ dimenzióban végezhető el, egyébként a megoldandó egyenletrendszer determinánsa közel szinguláris.

5.6. Számítógépes tapasztalatok

A valószínűségeket kiszámító szubrutinok működésének helyességét nagyszámú példán ellenőriztük. A példák részletes leírása a Függelék 4. pontjában található. A példákban a valószínűség vagy ismert, vagy pedig kiszámítható, így az ellenőrzés segítségükkel végrehajtható.

Az algoritmusok futási eredményeit a 11—19. példákon az 5.1 táblázatban foglaltuk össze. A példákra vonatkozólag a példaszámot, az n dimenziószámot és az ismert p valószínűséget adjuk meg. A következő becsléseket futtattuk le.

1. Az elfogadás-elvetés θ_3 becslése független, normális eloszlású vektorokkal.
2. SZÁNTAI [127] módszerének alkalmazása normális valószínűségek kiszámítására. SZÁNTAI az eredeti módszert regresszióval egészítette ki, ezt az újabb változatot futtattuk.
3. Irány menti integrálás θ_4 becslése, független normális vektorok által meghatározott egyenesek mentén.
4. Irány menti integrálás θ_4 becslése, az *A2 alapmódszer* által generált stacionárius vektorok által meghatározott egyenesek mentén.
5. Irány menti integrálás θ_4 becslése az *A3 alapmódszer* által generált stacionárius vektorok felhasználásával.
6. Véletlen ortonormált bázis vektorjaira a θ_4 becslés alkalmazása, vagyis az O_1 becslés.
7. Véletlen ortonormált bázis O_2 becslés: a bázisbeli vektorok összege és különbsége.
8. Véletlen ortonormált bázis O_3 becslés: három vektor minden lehetséges módon vett összege és különbsége.
9. Véletlen ortonormált bázis O_4 becslés: négy vektor összege és különbsége.

Numerikus integrálási formuláknál a módszer hatékonyságát a hibával és a függvényértékek kiszámításának számával szokták jelelmezni. Itt, mivel nem a $\varphi(\mathbf{x})$ függvény integrálját számítjuk ki közvetlenül, így egy *Monte Carlo módszereknél* szokásos mutatót fogunk az algoritmusok rangsorolásához használni. Ha az egyik módszerrel egy σ_1 standard hibájú becslést t_1 idő alatt kapunk, egy másik módszerrel σ_2 hibájú becslést t_2 idő alatt kapunk, akkor a második módszernek az elsőre vonatkoztatott hatékonyságát $\frac{\sigma_1^2 t_1}{\sigma_2^2 t_2}$ hányadossal mérjük. Ez megmutatja, hogy a második módszer hányszor gyorsabb az elsőnél; ugyanakkora standard hibát menynyiszer kevesebb idő alatt lehet a második módszer felhasználásával elérni. Becsléseink hatékonyságát a θ_3 becslésre vonatkoztatva adjuk meg.

Az 5.1. táblázatban az $N=100$ mintaszám esetén a fentebb felsorolt becsléseknek három jellemzőjét adjuk meg: a mintából számított szórást, a futáshoz szükséges időt másodpercekben, valamint a θ_3 becsléshez viszonyított hatékonyságot. Az első becslésnél azért kapunk 1-től eltérő hatékonyságokat, mert a módszerek szórását az N elemű mintából becsültük, míg a hatékonyság kiszámításában a θ_3 becslés szórását a pontos $\sqrt{p(1-p)/N}$ képlettel számítottuk. A 18. és 19. példán az O_4 becslést csak $N=25$ mintaszámmal futtattuk és ebből számítottuk ki a közölt értékeket, a *Szántai algoritmus* eredményeit pedig $N=400$-as mintából számítottuk.

5.1 TÁBLÁZAT Becslések összehasonlítása $n = 4-20$ dimenzióban ($N = 100$)

Példák \ Becslések		Elfogadás–elvetés	Szántai algor.	Irány m., független	Irány m., stac., $A2$	Irány m., stac., $A3$	Ortonorm. O_1	Ortonorm. O_2	Ortonorm. O_3	Ortonorm. O_4
11.	σ	0,049	0,014	0,018	0,0092	0,0063	0,0066	0,0017	0,0020	0,0030
$n=4$	t	0,076	0,10	0,16	2,6	1,6	0,69	1,0	1,1	0,82
$p=0,6406$	hat.	0,9	8,3	3,0	0,7	2,7	5,7	55,0	35,0	22,4
12.	σ	0,043	0,029	0,022	0,012	0,0077	0,0074	0,0022	0,0015	0,0029
$n=5$	t	0,096	0,11	0,24	2,4	1,8	1,15	1,8	2,5	2,4
$p=0,2324$	hat.	0,9	1,8	1,4	0,4	1,5	2,7	19,2	30,0	7,8
13.	σ	0,0098	0,00054	0,0016	0,0017	0,0010	0,00074	0,00014	0,000058	0,00015
$n=6$	t	0,12	0,13	0,22	2,6	2,0	1,7	2,7	4,8	5,9
$p=0,9904$	hat.	1,0	303,0	19,0	1,5	6,0	12,8	198,0	748,0	82,6
14.	σ	0,050	0,017	0,024	0,018	0,0097	0,0043	0,0014	0,00049	0,00021
$n=10$	t	0,25	0,27	0,40	3,5	2,9	6,1	10,6	32,0	90,2
$p=0,5830$	hat.	0,9	7,5	2,7	0,5	2,2	5,3	27,0	79,0	140,0
15.	σ	0,035	0,021	0,015	0,013	0,0076	0,0024	0,00079	0,00032	0,000084
$n=10$	t	0,25	0,26	0,35	3,4	2,8	6,2	10,3	30,0	83,8
$p=0,8527$	hat.	1,0	2,5	3,8	0,5	1,9	9,0	49,4	104,0	518,0
16.	σ	0,040	0,012	0,024	0,020	0,0099	0,0024	0,00064	0,00034	0,00014
$n=15$	t	0,47	0,49	0,69	4,7	4,0	18,1	31,3	144,0	704,4
$p=0,8012$	hat.	0,9	9,3	1,7	0,4	1,8	6,8	57,4	45,0	52,0
17.	σ	0,027	0,023	0,017	0,014	0,0075	0,0033	0,0010	0,00046	0,00024
$n=15$	t	0,45	0,48	0,69	4,7	4,0	17,9	31,0	140,0	691,0
$p=0,1105$	hat.	1,3	1,7	2,0	0,4	1,9	2,2	12,5	15,0	11,0
18.	σ	0,025	0,0086	0,010	0,011	0,0059	0,0019	0,00061	0,00040	0,00016
$n=20$	t	0,75	0,76	0,95	5,6	4,9	39,5	68,3	420,0	2935,0
$p=0,9229$	hat.	1,0	9,0	5,0	0,7	3,1	3,5	20,8	8,1	7,1
19.	σ	0,049	0,040	0,020	0,016	0,0080	0,0034	0,00097	0,00047	0,00027
$n=20$	t	0,73	0,77	0,95	5,8	5,4	39,8	69,2	423,0	2991,0
$p=0,4180$	hat.	1,0	1,5	4,5	1,0	5,0	3,7	27,2	19,2	8,3

5.2 TÁBLÁZAT Nagy pontosságú eredmények $n=4$—10 dimenzióban

Példák \ Becslések		Elfogadás—elvetés	Irány m., független	Ortonorm. O_1	Ortonorm. O_2	Ortonorm. O_3	Ortonorm. O_4
11.	p_c	0,63720	0,64259	0,63940	0,64080	0,64056	
$n=4$	σ	0,00473	0,00194	0,00052	0,00016	0,00020	—
$p=0,64064$	h	0,00344	0,00195	0,00124	0,00016	0,00008	
	t	7,6	16,5	70,2	104,0	116,0	
12.	p_c	0,22970	0,22963	0,23227	0,23232	0,23241	
$n=5$	σ	0,00421	0,00224	0,00066	0,00020	0,00015	—
$p=0,23241$	h	0,00271	0,00278	0,00013	0,00009	0,00000	
	t	9,7	23,9	115,3	180,5	257,1	
13.	p_c	0,990400	0,990516	0,990370	0,990415	0,990433	
$n=6$	σ	0,001210	0,000256	0,000076	0,000016	0,000006	—
$p=0,990436$	h	0,000030	0,000080	0,000066	0,000021	0,000002	
	t	12,8	22,3	172,9	277,3	476	
14.	p_c	0,57860	0,58186	0,58269	0,58300		0,58299
$n=10$	σ	0,00498	0,00249	0,00043	0,00013	—	0,00010
$p=0,58300$	h	0,00440	0,00114	0,00031	0,00000		0,00001
	t	25,2	39,3	616,0	1065,8		369,8
15.	p_c	0,85300	0,85351	0,85252	0,85286		0,85270
$n=10$	σ	0,00344	0,00169	0,00024	0,00008	—	0,00003
$p=0,85276$	h	0,00024	0,00075	0,00023	0,00010		0,00005
	t	25,8	35,8	615,0	1030,9		337,1

Az egyes példák esetén legjobban működő becslések eredményét a táblázatban bekereteztük. A közölt futási eredmények megegyeznek a többi, nem közölt számítógépes tapasztalattal. A számítógépes kísérletezésből a következő általános megjegyzéseket vonhatjuk le.

1. Ha egy példában a kiszámítandó valószínűség 1-hez közeli, akkor nemcsak a 0_3 becslés ad a kis valószínűségű példáknál kapható eredménynél jobb eredményt, hanem az egyes módszerek hatékonyságának a növekedése is nagyobb lesz, mint a kis valószínűségi példák esetén tapasztalható hatékonyságnövekedés.

2. A legjobb eredményeket az ortonormált becslések adják; $n=5$ dimenzióban O_2 vagy O_3, $n=10$ dimenzióban O_4, $n=15$—20 dimenzióban pedig az O_2 (esetleg az O_3) becslés.

3. Két tizedesjegyre pontos eredményt $n=10$ dimenzióig 0,1 sec alatt, 20 dimenzióig 1 sec alatt lehet kapni (0,01 standard hibájú becslést). Az operációkutatás nemlineáris algoritmusai közül a *megengedett irányok módszerének* még ekkora pontosságra sincs szüksége [102], a VEINOTT-féle *metszősík algoritmus* [128] és a *redukált gradiens módszere* [83] viszont kisebb hibájú eredményt igényel. A közölt *Monte Carlo módszerek* segítségével ezeket a kisebb hibájú becsléseket is ki tudjuk számítani. Ezt mutatjuk be az 5.2 táblázatban, ahol a 11—15. példákon elért eredmények láthatók. A becsléseket az $N=10\,000$-es mintaszámmal számítottuk ki, kivételt képeznek az O_4 becslés futásai, itt $N=400$ volt a mintaszám. Minden módszernél négy számmal adjuk meg az adott példán végzett futást: a valószínűség számított p_c értékét, a mintából számított σ szórást, a tényleges h hibát és a másodpercekben mért t futási időt közöltük.

Az 5.2 táblázatban közölt 25 valószínűség közül csak hét esetben volt nagyobb a tényleges hiba a szórásnál és mindössze egyetlen esetben haladta meg a hiba a szórás kétszeresét, jó összhangban az elmélettel.

Az algoritmusokat a magasabb dimenziójú példákon futtatva kapott eredményeket közöljük a következő két táblázatban. Az 5.3 táblázat tartalmazza a 20—26. példákon, az 5.4 táblázat pedig a 27—30. példákon kapott eredményeket. A becsléseket az $N=100$ mintaszámra futtattuk, kivéve a *-gal jelölt eseteket; itt $N=25$-re számított értékből határoztuk meg az $N=100$-nak megfelelő, táblázatban is közölt értékeket. Az egyes esetekben az 5.1 táblázatban közöltekhez hasonlóan a σ szórást, a t időt másodpercekben és az elfogadás—elvetés becsléshez viszonyított hatékonyságot adtuk meg.

A futási eredmények alapján $n=20$ dimenzióig az algoritmusok hatékonysága az 5.1 táblázatban mutatott képpel azonos. Az $n=20$—50 dimenziókban az O_2 becslés, vagy pedig a θ_4 becslés — stacionárius vektorok ($A3$) felhasználásával — adja a legjobb eredményt. Ezek a legjobb algoritmusok csak 10—20-szor gyorsabbak az elfogadás—elvetés becslés algoritmusánál (szemben az $n=5$—15 dimenzióban tapasztalt 50—500-szoros hatékonysággal), mégis elég ahhoz, hogy két tizedesjegyre pontos eredményt kapjunk 50 dimenzióig 10 másodpercnél kevesebb idő alatt.

Alacsony valószínűségekre (0,1 és még ennél is kevesebb) a becslések hatékonysága nem kielégítő. Ezekben az esetekben az elfogadás—elvetés becslés alacsony valószínűségekre kidolgozott $V4$ algoritmusa (a ξ és a $-\xi$ vektor egyidejű számításával) biztosítja a legjobb eredményt.

A módszerek erejét kívánjuk szemléltetni az utolsó táblázatban közölt adatokkal. Itt hat darab ismeretlen valószínűségű példán (a 31—37. példákon) kapott eredményeket közlünk. A mintaszám $N=1600$ volt. Az első három példa esetében a

5.3 TÁBLÁZAT Becslések összehasonlítása $n=5-50$ dimenzióban

Példák \ Becslések		Elfogadás–elvetés	Irány m., független	Irány m., stac., $A3$	Ortonorm. O_1	Ortonorm. O_2	Ortonorm. O_3	Ortonorm. O_4
20.	σ	0,019	0,0046	0,0015	0,0014	0,00024	0,00022	* 0,00030
$n=5$	t	0,10	0,22	1,7	1,1	1,7	2,5	2,4
$p=0,973$	hat.	0,7	5,1	5,9	11,0	239,0	192,0	117,0
21.	σ	0,027	0,0074	0,0028	0,0022	0,00064	0,00024	* 0,000054
$n=10$	t	0,25	0,39	2,9	6,1	10,5	31,6	88,0
$p=0,938$	hat.	0,8	6,7	6,4	4,7	34,0	77,6	577,0
22.	σ	0,043	0,017	0,0077	0,0031	0,00097	0,00048	* 0,00017
$n=15$	t	0,47	0,70	4,1	18,1	31,4	144,9	716,0
$p=0,773$	hat.	0,9	4,0	3,4	4,5	28,0	24,3	36,0
23.	σ	0,040	0,021	0,0080	0,0030	0,0010	0,00052	* 0,00037
$n=20$	t	0,76	0,97	5,4	40,0	69,3	427,5	3025,0
$p=0,704$	hat.	1,2	3,6	4,5	4,3	21,1	13,4	3,7
24.	σ	0,049	0,025	0,0086	0,0035	0,00091	* 0,00054	
$n=30$	t	1,5	1,7	7,6	124,0	214,0	1964,0	—
$p=0,566$	hat.	1,0	3,2	6,6	2,4	21,0	6,4	
25.	σ	0,050	0,024	0,0093	0,0035	0,00084	* 0,00052	
$n=40$	t	2,5	2,8	10,7	277,0	486,0	5921,9	—
$p=0,443$	hat.	1,0	3,6	6,6	1,8	18,0	3,8	
26.	σ	0,048	0,024	0,0088	0,0021	0,00094		
$n=50$	t	3,7	4,2	13,5	530,4	914,0	—	—
$p=0,365$	hat.	1,0	3,5	8,3	3,7	10,2		

5.4 TÁBLÁZAT

Becslések összehasonlítása $n=20$—50 dimenzióban

Példák \ Becslések		Elfogadás—elvetés	Iránym. független	Iránym. stac. $A3$	Ortonorm. O_1	Ortonorm. O_2	Ortonorm. O_3	Ortonorm. O_4
27.	σ	0,014	0,0050	0,0018	0,00087	0,0004	* 0,00016	* 0,00007
$n=20$	t	0,75	0,96	5,2	39,5	67,8	100,5	714,0
$p=0,9859$	hat.	0,8	4,2	5,6	3,4	9,6	9,1	7,4
28.	σ	0,019	0,0050	0,0028	0,0012	0,00052		
$n=30$	t	1,5	1,7	7,9	123,6	209,1	—	—
$p=0,9717$	hat.	0,8	9,0	6,2	2,3	7,1		
29.	σ	0,017	0,0051	0,0028	0,0011	0,00048		
$n=40$	t	2,5	2,8	10,8	281,3	480,0	—	—
$p=0,9690$	hat.	1,0	9,9	8,2	2,1	6,8		
30.	σ	0,021	0,011	0,0028	0,00097	0,00053		
$n=50$	t	3,7	4,1	13,7	541,0	919,8	—	—
$p=0,9618$	hat.	0,8	2,5	12,7	2,6	5,2		

5.5 TÁBLÁZAT

Ismeretlen valószínűségű példák eredményei

Példa 31—37.	Valószínűség	Standard hiba (szórás)	Becslés	Idő (sec)	Hatékonyság
$n=5$	0,719411	$\pm 0{,}000196$	O_3	40,8	128
$n=10$	0,662662	$\pm 0{,}000043$	O_4	1444	206
$n=20$	0,492468	$\pm 0{,}000301$	O_2	1106	18
$n=5$	0,995469	$\pm 0{,}000016$	O_3	39,4	455
$n=10$	0,976035	$\pm 0{,}000008$	O_4	1389	659
$n=20$	0,966064	$\pm 0{,}000124$	O_2	1084	14

valószínűség közepes, míg a második háromnál igen magas volt. Megjegyezzük, hogy a *Szántai-féle algoritmus* hatékonysága ezeken a példákon rendre 12, 7, 3, 78, 62 és 84 volt egy $N=2500$-as mintából számítva.

5.7. Más módszerek és összehasonlítások

Azokat az irodalomban közölt eredményeket foglaljuk össze ebben a pontban, amelyek a többdimenziós normális eloszlás eloszlásfüggvényének kiszámítására vagy többdimenziós integrálásra vonatkoznak. Ezeket a módszereket az általunk adott módszerekkel összehasonlítjuk, bár számítógépes tapasztalatok hiánya miatt ez nem végezhető el minden eljárás esetében.

Az 1963-ig elért eredményeket GUPTA [56] foglalta össze; tetszőleges korreláció mátrix és integrálási határ esetén nem létezik megfelelő eljárás. JOHNSON és KOTZ [67] 1972-ben megjelent könyvükben hasonló véleményt fejeznek ki, bár itt *Monte Carlo integrálással* történt próbálkozásra utalnak. A $\varphi(\mathbf{x})$ függvényre vonatkozó *Monte Carlo integrálási módszer* — a θ_1 vagy a θ_2 becslés — alkalmazásával $n=6$-nál nagyobb dimenzióban még egy értékes jegyet sem lehet minden esetben biztosítani néhány perc gépidő alatt [31].

A továbbiakban az alacsony $n \leqq 4$ dimenziós eloszlásfüggvények kiszámítására vonatkozó eredményeket tekintjük át. Az $n=2$ dimenziós eloszlásfüggvényre más kétváltozós függvények bevezetésével, illetőleg ezek approximációjával lehet 5—6 tizedes pontosságú eredményt kapni [67]. A DONELLY [41] által közölt algoritmus egy ilyen átalakítás felhasználásával 15 értékes jegyre adja meg a valószínűséget.

MILTON [85] egy adaptív *Simpson kvadratúra* felhasználásával adott $n=6$ dimenzióig használható algoritmust, az eredmény 10^{-3}—10^{-4} hibájú. Ennél az eljárásnál jobb a DUTT [43] által közölt módszer, amely a tetrachorikus kifejtést transzformálja többdimenziós *Fourier transzformáltak* összegévé és ezeket *Gauss kvadratúrával* számítja ki. A megfelelő program $n=4$ dimenzióban nyolc tizedes pontosságú eredményt ad DUTT szerint. Három és négy dimenzióban az irodalom alapján ez a legjobb eredmény. Számítási eredményt $n>4$ dimenzióban DUTT nem közölt; az elvégzendő munka n növekedésével rohamosan nő. Az eljárás, mint minden, a $\varphi(\mathbf{x})$ függvény kiintegrálásával foglalkozó módszer, érzékeny a nagy korrelációkra; ha $|\mathbf{R}|$ a zérushoz közeli, akkor a tetrachorikus sor lassan konvergál.

Megemlítjük még ANDEL [6] munkáját, amelyben a szerző a sűrűségfüggvény sorbafejtését használja fel és ESCOUFIER [45] cikkét; ezeket számítógépes tapasztalatok hiánya miatt nem lehet értékelni.

A nem kifejezetten a normális eloszlásfüggvény kiszámítására felépített többdimenziós integrálási módszerek közül GOOD és GASKINS [54] *Centroid integrálási módszerét* említjük először. A cikk a *Centroid módszer* segítségével példaként normális sűrűségfüggvény integrálját számítják ki egy szimplex felett független komponensek esetén ($\mathbf{R}=\mathbf{I}$). A módszert néhány példán futtatva 10^{-5}—10^{-10} hibát adnak meg a szerzők. A módszer használhatóságának problémájához két gondolatot vetünk fel: 1. Egy eloszlásfüggvény értékének kiszámítása esetén kérdéses, hogy az integrálási tartomány szimplexekre, vagy más tartományokra történő felbontásánál $n \geqq 5$ dimenzióban a keletkező nagy számú feladat numerikusan kezelhető-e. 2. Ha az eloszlás korrelált, akkor a függvényértékek kiszámítási idejének megnövekedése miatt kérdéses, hogy akár egy szimplex felett is kiszámítható-e az integrál.

Az n-dimenziós egységkockában értelmezett függvények integrálására CRANLEY és PATTERSON [28] adtak regressziós módszereket $n=6$ dimenzióig (10^{-4} pontosság). Megemlítjük még HASELGROVE [60] és TSUDA [133] munkáit, ezek legfeljebb 8 dimenzióig alkalmazhatók.

A θ_3 becslésnek egy ötletes változata a SZÁNTAI [127] által közölt *Monte Carlo módszer*. Ez bizonyos esetekben — nagy valószínűségekre — jól működik, de használata az általános esetben a θ_3 becsléshez képest nem jelent tízszeresnél nagyobb hatékonyságot (lásd az 5.1 és 5.5 Táblázatot).

A közölt módszerek a nagy korrelációra egyáltalán nem érzékenyek. A dimenziószámot tekintve lényeges javulást jelentenek az eddigi eredményekhez képest. A legjobb ismert módszerekkel is összehasonlítható pontosságú eredményt adnak már $n=5$ dimenzióban, magasabb dimenzióban pedig módszereink pontosabbak. A következő táblázatban összefoglaljuk, hogy a legjobb becslések alkalmazása esetén n dimenzióban 10^{-2}, illetőleg 10^{-3} szórású (standard hibájú) eredmény eléréséhez hány másodperc idő szükséges. A táblázat a „legrosszabb" feladatokra kapott időket tartalmazza, magas valószínűségű példákra a táblázatban közölt időknek csak egy tized- vagy huszadrészére van szükség.

5.6 TÁBLÁZAT

Adott pontossághoz szükséges idő (sec)

Hiba \ $n=$	5	10	15	20	30	50
0,01	0,05	0,04	0,3	0,6	2,2	9,0
0,001	5,2	4,0	30,0	60,0	214,0	900,0

Módszereink a DAVIS és RABINOWITZ ([29], p. 314) véleménye szerint elérhető négy tizedes pontosságot $n=12$ dimenzióig biztosítják, az $n=12$—50 dimenziós esetekben még egy tizedesjegyet tudnak adni az elérhetőnek gondolt két tizedesjegyhez.

Két hátrányát említjük meg a közölt módszereknek. A konvergencia sebessége, mint minden *Monte Carlo módszernél* csak $O(N^{-1/2})$. Ha nem eloszlásfüggvényt kell kiszámítani, hanem egy általános halmaz valószínűségét, akkor a leghatékonyabb becsléseink (θ_4 és az ortonormált becslések) gyorsasága a g függvény gyors kiszámíthatóságán múlik.

FÜGGELÉK

Ebben a részben az algoritmusok számítógépes kipróbálásának néhány részletével foglalkozunk. Áttekintjük a felhasznált függvényeljárásokat, megadjuk a gépi kódban megírt véletlenszám generátorok teljes listáját és a véletlenszám generátorok által előállított számokra elvégzett statisztikai próbák eredményeit. Az utolsó két pontban numerikus példákat közlünk.

1. Függvényeljárások a normális és az χ^2 eloszlás eloszlásfüggvényének kiszámítására

Az egydimenziós normális eloszlásfüggvény kiszámítására több közelítő formula létezik (lásd pl. [67], II. vol. p. 43). Mi általában az MTA CDC 3300 számítógépén meglevő NDTR könyvtári szubrutint használtuk, amely a következő formulákon alapszik:

$$p = 1-\varphi(x)\cdot t\cdot(a_1+a_2t+a_3t^2+a_4t^3+a_5t^4),$$

ahol

$$t = 1{,}0/(1{,}0-0{,}2316419\,|x|),$$

$$\varphi(x) = 0{,}3989423\exp(-0{,}5x^2),$$

$$a_1 = 0{,}3193815, \quad a_2 = 0{,}3565638, \quad a_3 = 1{,}781478,$$

$$a_4 = -1{,}821256, \quad a_5 = 1{,}330274$$

és

$$\Phi(x) \doteqdot \begin{cases} p, & \text{ha } x \geqq 0, \\ 1-p, & \text{ha } x < 0. \end{cases}$$

A $\Phi(x)$ függvényérték kiszámítási ideje 210 μsec, a képlet hibája 10^{-7} nagyságrendű.

A példákban a Φ valószínűségek kiszámítására a HILL és JOYCE [62] által közölt algoritmus FORTRAN-ba átírt dupla pontosságú (15 értékes jegy) változatát használtuk. A szerzők szerint a pontosság csak a számítógép szóhosszúságától függ, így a Honeywell 66/60-as gépen a hiba 10^{-15}, egy függvényérték kiszámítási ideje 720 μsec.

A χ^2_n eloszlásfüggvénye kiszámítására több szubrutint kipróbáltunk. A Honeywell XINGAM elnevezésű szubrutinja, amely *Gauss-integrálással* számítja ki a függvényértékeket, 10^{-4} nagyságrendű hibákat is megenged. Az eloszlásfüggvény közelítő kiszámítására a *Fischer-féle közelítés* [67]

$$P\{\chi^2_n < x\} \doteqdot \Phi(\sqrt{2x}-\sqrt{2n-1}),$$

vagy az ennél pontosabb *Wilson—Hilferty approximációs képlet*

$$P\{\chi^2_n < x\} \doteqdot \Phi([(x/n)^{1/3}-1+2/(9n)]\sqrt{9n/2})$$

használható, ahol Φ a standard normális eloszlásfüggvény. Az utóbbi közelítés három tizedesre pontos eredményt ad 1000—1200 μsec alatt.

Ennél pontosabb eredmény érhető el az IBM SSP-ben található PRCHI2 függvény (futási ideje 40 msec körül), vagy pedig a CDC 3300 könyvtárában levő PRCHI2 függvény használatával. Ez utóbbi a HILL és PIKE [63] cikkében közölt algoritmus

FORTRAN változata; az IBM programhoz hasonló pontosságot ad (kb. 10^{-7}), de futási ideje csak 1000 μsec körül van. Mivel a χ_n^2 eloszlás eloszlásfüggvényét, illetőleg az n szabadságfokú χ eloszlás F_n eloszlásfüggvényét sokszor kellett a programban használni, ezért a következő eljárást alkalmaztuk. Meghatároztuk azt a c_1 és c_2 számot, melyre $F_n(x) \leqq 10^{-6}$, ha $x \leqq c_1$ és $F_n(x) \geqq 1 - 10^{-6}$, ha $x \geqq c_2$ teljesült. A $[c_1, c_2]$ intervallumot 1000 egyenlő részre osztottuk, az osztási pontokban felvett függvényértékeket egy 1000 elemű vektorban tároltuk. Az $F_n(x)$, $x \in [c_1, c_2]$ függvényérték meghatározását ezek után a vektorban tárolt értékek közti lineáris interpolációval végeztük el. Az így kapott függvényeljárás végrehajtási ideje átlagosan 90 μsec, az általa adott értékek 10^{-6}-nál kisebb hibájúak.

2. *A véletlenszám generátorok és gépi kódos programjuk*

Egyenletes eloszlású számok generálását egy ARATÓ MÁTYÁSTÓL származó ötlet alapján a következő algoritmus felhasználásával valósítottuk meg.

Legyenek m_1 és m_2 két 36 bites egész szám, amelyeknek tetszőleges kezdeti érték adandó (IRANU és IRANV).

1. Balra forgatjuk m_2-t 7 bittel.
2. Legyen $m_3 \leftarrow m_1 + m_2$, $m_1 \leftarrow m_2$, $m_2 \leftarrow m_3$.
3. Balra forgatjuk m_1-t és m_2-t 4 bittel.
4. [Normalizáljuk m_2-t.] Adjuk át az $u \leftarrow m_2/2^{36}$ számot, mint [0, 1)-ben egyenletes eloszlásút.

```
 1          SYMDEF    UNGE,
                        NAHDI,
                        EAHSI
 2 UNGE     NULL
 3          STI       .E.L..
 4          STX1      .E.L..
 5          LDA       IRANU
 6          ALR       7
 7          ADLA      IRANV
 8          LDQ       IRANV
 9          LLR       4
10          STA       IRANV
11          STQ       IRANU
12          ARL       1
13          LDE       0,DU
14          FNO
15          FST       FRANU
16          RET       .E.L..
17 NAHDI    NULL
18          STI       .E.L..
19          STX1      .E.L..
20          LDA       IRANU
21          ALR       10
22          ADLA      IRANV
23          ALR       7
24          LDQ       IRANV
25          STA       IRANV
26          STQ       IRANU
27          STA       SIGN
28          ANA       MOD
29          TZE       TAIL
30          STA       IND
31          LXL1      IND
32          FLD       A—1,1
33          FST       AM
34          LDA       IRANU
35          ARL       1
36          LDE       0,DU
37          FNO
38          FST       US
39 S4       FSB       T—1,1
40          TMI       S5
41          FMP       H—1,1
42 S17      FAD       AM
43          SZN       SIGN
44          TMI       OUT
45          FNEG
46 OUT      FST       XNOR
47          RET       .E.L..
48 S5       FLD       A,1
49          FSB       AM
50          FST       TEM2
51          LDA       IRANU
52          ALR       4
53          ADLA      IRANV
54          ALR       3
```

```
 55          LDQ     IRANV      112          STA     IND
 56          STQ     IRANU      113          LXL2    IND
 57          STA     IRANV      114          FLD     T—1
 58          ARL     1          115          FST     AM
 59          LDE     0, DU      116 S10      LDA     IRANU
 60          FNO                117          ALR     1

 61          FST     TEMP       118          TMI     S12
 62          FMP     TEM2       119          STA     IRANU
 63          FST     W          120          FLD     AM
 64          ADE     MOD4       121          FAD     D, 2
 65          FAD     AM         122          FST     AM
 66          FMP     W          123          ADX2    1, DU
 67          FNEG               124          TRA     S10
 68          FAD     US         125 S12      ALR     1
 69 S6       FST     TEM4       126          ARL     1
 70          TMI     S7         127          LDE     0, DU
 71          FLD     W          128          FNO
 72          TRA     S17        129          FST     TEM2
 73 S7       LDA     IRANU      130 S13      FMP     D, 2
 74          ALR     4          131          FST     W
 75          ADLA    IRANV      132          ADE     MOD4
 76          ALR     3          133          FAD     AM
 77          LDQ     IRANV      134          FMP     W
 78          STQ     IRANU      135          FST     TEM3
 79          STA     IRANV      136 S14      LDA     IRANU

 80          ARL     1          137          ALR     7
 81          LDE     0, DU      138          ADLA    IRANV
 82          FNO                139          LDQ     IRANV
 83          FST     TEM3       140          ALR     2
 84          FSB     US         141          STA     IRANV
 85          TMI     S8         142          STQ     IRANU
 86          LDA     IRANU      143          ARL     /
 87          ALR     4          144          LDE     0, DU
 88          ADLA    IRANV      145          FNO
 89          ALR     3          146          FST     US
 90          LDQ     IRANV      147          FSB     TEM3
 91          STQ     IRANU      148          TMI     S15
 92          STA     IRANV      149          FLD     W
 93          ARL     1          150          TRA     S17
 94          LDE     0, DU      151 S15      LDA     IRANU
 95          FNO                152          ALR     5
 96          FST     US         153          ADLA    IRANV
 97          TRA     S4         154          ALR     2
 98 S8       LDA     IRANU      155          LDQ     IRANV

 99          ALR     4          156          STA     IRANV
100          ADLA    IRANV      157          STQ     IRANU
101          ALR     3          158          ARL     1
102          LDQ     IRANV      159          LDE     0, DU
103          STQ     IRANU      160          FNO
104          STA     IRANV      161          FST     TEMP
105          ARL     1          162          FSB     US
106          LDE     0, DU      163          TMI     S16
107          FNO                164          LDA     IRANU
108          FST     US         165          ALR     4
109          FSB     TEM3       166          ADLA    IRANV
110          TRA     S6         167          LDQ     IRANV
111 TAIL     LDA     =5         168          ALR     3
```

```
169          STA     IRANV
170          STQ     IRANU
171          ARL     1
172          LDE     0,DU
173          FNO
174          FST     TEM3

175          TRA     S13
176 S16      FLD     TEMP
177          FST     TEM3
178          TRA     S14
179 MOD      OCT     000000000037
180 MOD2     OCT     377777777777
181 MOD3     OCT     000001000000
182 MOD4     OCT     777000000000
183 IND      OCT
184 SIGN     OCT
185 AM       DEC
186 US       DEC
187 USMT     DEC
188 TEM2     DEC
189 TEM3     DEC
190 TEM4     DEC
191 W        DEC
192 EAHSI    NULL
193          STI     .E.L..
194          STX1    .E.L..

195          STZ     C
196          LDA     IRANU
197          ALR     7
198          ADLA    IRANV
199          LDQ     IRANV
200          LLR     4
201          STA     IRANV
202          STQ     IRANU
203 SI2      TMI     SI4
204          STA     IDEIG
205          FLD     QSA
206          FAD     C
207          FST     C
208          LDA     IDEIG
209          ALR     1
210          TRA     SI2
211 SI4      ALR     1
212          ARL     1
213          LDE     0, DU

214          FST     URAN
215          FSB     QSA
216          TPL     SI6
217          FLD     URAN
218          FAD     C
219          FST     XEXP
220          RET     .E.L..
221 SI6      LXL1    =1, DL
222          LDA     IRANU
223          ALR     7
224          ADLA    IRANV
225          LDQ     IRANV
226          LLR     4

227          STA     IRANV
228          STQ     IRANU
229          ARL     1
230          LDE     0, DU
231          FNO
232          FST     UMIN

233 SI7      LDA     IRANU
234          ALR     7
235          ADLA    IRANV
236          LDQ     IRANV
237          LLR     4
238          STA     IRANV
239          STQ     IRANU
240          ARL     1
241          LDE     0, DU
242          FNO
243          FST     USTAR
244          FSB     UMIN
245          TPL     SI8
246          FLD     USTAR
247          FST     UMIN
248 SI8      FLD     URAN
249          FSB     QSA, 1
250          TMI     SI9
251          ADX1    1, DU

252          TRA     SI7
253 SI9      FLD     QSA
254          FMP     UMIN
255          FAD     C
256          FST     XEXP
257          RET     .E.L..
258 TEMP     BSS     1
259 C        BSS     1
260 URAN     BSS     1
261 UMIN     BSS     1
262 USTAR    BSS     1
263  IDEIG   OCT
264          BLOCK   ADAT
265 A        BSS     32
266 T        BSS     31
267 H        BSS     31
268 D        BSS     47
269 QSA      BSS     10
270          BLOCK   RACO
271 IRANU    OCT
272 IRANV    OCT
273 FRANU    DEC
274 FRANV    DEC
275 FRANZ    DEC
276 XNOR     DEC
277 XEXP     DEC
278 XCHI     DEC
279 XBET     BSS     1
280 YS       BSS     20
281 XI       BSS     20
282          BLOCK   ELCO
283 ND       BSS     1

284          END
```

A fenti algoritmus gépi kódban (GMAP) írt programja (UNGE) egy véletlen számot 24 μsec alatt állított elő (másodpercenként ~40 ezer szám).

Normális eloszlású számok generálását végzi a NAHDI szubrutin, amely AHRENS és DIETER [2], [3] *FL_5 algoritmusa* gépi kódban. A szükséges konstansokat az idézett cikkekben meg lehet találni, az ADAT elnevezésű COMMON mezőbe töltendők. Az adatok jelölése a programban azonos a cikkekben használt jelölésekkel. Egy normális eloszlású szám előállításának ideje 60 μsec (másodpercenként ~17 ezer).

Exponenciális eloszlású számok generálását az EAHSI szubrutin végzi AHRENS és SIBUYA [1] módszere alapján, a konstansok a QSA tömbben vannak tárolva. Egy exponenciális eloszlású szám előállításához szükséges idő 63 μsec (másodpercenként ~16 ezer szám).

Néhány gépi kódos művelet idejét adjuk meg:

Floating store	1,0 μsec
Floating add	1,7 μsec
Floating multiply	3,1 μsec

A Honeywell 66/60 adott gépi konfigurációja az irodalomban megadott fentebbi futási időknél több időt igényel egy művelet végrehajtására: tapasztalataink szerint a fenti időket 1,4-gyel kell szorozni ahhoz, hogy a tényleges futási időket megkapjuk. Az MTA CDC 3300-as gépénél mintegy kétszer gyorsabb.

3. A véletlenszám generátorok statisztikai próbái

Az elvégzett próbák lényegében χ^2 próbák voltak. A [0, 1) intervallumot $K=1000$ egyenlő részre osztottuk, a vizsgált χ^2_N változót pedig a [0, 1)-ben egyenletes eloszlású véletlen számok $N=100\,000$ realizációjára a

$$\chi^2_N = \sum_{j=1}^{K} \frac{(v_j - Np_j)^2}{Np_j}$$

összefüggésből számítottuk, ahol $p_j=1/K=0{,}001$, v_j pedig a j-edik intervallumba való beesés gyakorisága. Összesen 11-féle próbát végeztünk, minden próbát tízszer hajtottunk végre, így az első próba esetén összesen 10^6 egyenletes eloszlású számot generáltunk, a második próba esetén $2 \cdot 10^6$ számot stb.

Az egyenletes eloszlású számok próbáit a MACLAREN és MARSAGLIA [76] által végzett próbákhoz hasonlóan konstruáltuk meg.

a) Egyenletesség. A [0, 1) intervallumot 1000 részre osztva az UNGE szubrutin által előállított u_i, $i=1, \ldots, N$ számok beesési gyakoriságát ellenőriztük.

b) n szám maximuma. Ha v_i, $i=1, \ldots, n$ a [0, 1) intervallumban egyenletes eloszlású, független valószínűségi változók, akkor a $v_{\max} = \max_{i=1,\ldots,n} v_i$ valószínűségi változó eloszlásfüggvénye $F(a)=a^n$. Az $r=F(v_{\max})=v^n_{\max}$ valószínűségi változó [0, 1)-ben egyenletes eloszlású, a próbát az r valószínűségi változó realizációira végeztük, vagyis az UNGE generátor által előállított u_i, $i=1, \ldots, nN$ értékekből számított $r_j=[\max(u_{jn+1}, \ldots, u_{jn+n})]^n$, $j=0, \ldots, N-1$ számokra. A próbát az $n=2, 3, 10, 20$ esetére számítottuk.

c) n szám minimuma. A próba felépítése az előzőhöz hasonló, most azonban az $r_j = 1 - [1 - \min(u_{jn+1}, \ldots, u_{jn+n})]^n$, $j = 0, \ldots, N-1$ számok egyenletességét ellenőriztük χ^2 próba segítségével. A próbát az $n = 2, 3, 10, 20$ értékekre végeztük el.

d) Exponenciális és normális számok. Az EAHSI és NAHDI generátorok által előállított x_i és y_i számokat transzformáltuk az $r_i = 1 - e^{-x_i}$ és az $s_i = \Phi(x_i)$, $i = 1, \ldots, N$ kifejezések segítségével, ahol Φ az egydimenziós normális eloszlásfüggvény. A próbával az r_i és s_i számok egyenletességét ellenőriztük. (Ez a két próba közvetve az UNGE generátor ellenőrzését is adja.)

A számítógépes futások eredményét az 1. táblázatban foglaltuk össze. A táblázatban felsorolt számok a $P\{\chi_N^2 > \chi_0^2\}$ valószínűségek százalékban kifejezve, ahol χ_0^2 egy $K-1$ szabadságfokú χ^2 eloszlású valószínűségi változó, χ_N^2 pedig a véletlenszám generátorok által előállított számokra, illetőleg ezek transzformáltjaira a pont elején levő képlettel számított érték.

1. TÁBLÁZAT

Véletlen számok statisztikai próbáinak eredményei

Próba \ Próbafutás	1	2	3	4	5	6	7	8	9	10
Egyenletesség	99	57	76	5,6	76	41	9,4	47	71	57
2 szám max	66	0,8	88	81	61	67	99	11	90	97
2 szám min	98	46	17	94	14	48	29	77	13	72
3 szám max	28	44	73	49	90	85	2,3	65	25	71
3 szám min	24	52	19	88	69	6,9	18	78	94	84
10 szám max	1,7	34	19	17	74	38	41	24	38	99
10 szám min	18	8,1	44	86	59	20	67	71	30	56
20 szám max	0,2	27	38	65	29	96	48	74	19	32
20 szám min	46	96	15	69	7,2	11	3,3	21	14	32
Exponenciális	19	99	34	85	35	78	6,7	95	2,9	28
Normális	25	31	31	37	22	0,03	4,6	75	15	28

A számítógépes programok által generált számok eloszlásának helyességét mutatja, hogy az elvégzett 110 χ^2 próba közül mindössze három esetben kaptunk 1 százaléknál, vagyis 0,01-nél kisebb valószínűséget (2 szám max próba 2. futás, 20 szám max 1. futás és a normális 6. futás). Ezek közül a legkisebb érték a 0,03%, amely nagyobb a MACLAREN által kritikusnak vett $\varepsilon = 0{,}02\%$-nál.

Az exponenciális és a normális számok eloszlását még a $P\{\xi < c\}$ valószínűségek kiszámításával is ellenőriztük. A $c = 0, 0{,}2, \ldots, 3{,}0$ értékek mindegyikére kiszámítottuk a valószínűséget elfogadás—elvetéssel. A tényleges hiba sehol sem volt nagyobb h_{95}-nél, a 95%-os biztonsággal a *Csebisev—Bernstejn egyenlőtlenségből* $N = 100\,000$-es mintaszám mellett számított hibahatárnál. A részletes futási eredményeket terjedelmességük miatt nem közöljük.

A *Monte Carlo módszerekkel* egyidős a vita, hogy mit kell vagy lehet véletlen számnak tekinteni és mi az értékük a fentiekhez hasonló statisztikai próbáknak. Véleményünk szerint a véletlenszám generátorok igazi próbája a gyakorlat. Olyan feladatok megoldására kell felhasználni a generált számokat, amelyek eredménye ismert; ha a véletlen számok felhasználásával kapott érték az egzakt eredménytől

csak a megengedett hibával tér el, akkor ilyen típusú feladatok megoldására az adott véletlenszám generátorok használhatók. Tehát a felhasznált generátorok jóságát tulajdonképpen az 5.2 táblázat eredményei mutatják; több dimenziós normális valószínűségek ezekkel a generátorokkal számíthatók.

4. Ismert valószínűségű példák

Felsoroljuk azokat a példákat, amelyek kiszámításával a számítógépes programokat ellenőriztük. A példák a következő eseteket, vagy azok párosítását tartalmazzák: kétdimenziós korrelált, három vagy négydimenziós korrelált ortáns, azonos korrelációjú és független esetek. A példákat szándékosan úgy válogattuk össze, hogy kis és nagy valószínűségek egyaránt szerepeljenek.

A kétdimenziós korrelált esetek valószínűségét a BIVNOR szubrutin [41] segítségével számítottuk ki. Az ortáns esetekben a valószínűségek kiszámíthatók, ezeket a példákat DUTT [43] cikkéből vettük. Az azonos korrelációjú esetek valószínűségeit GUPTA [56] adta meg. Független esetben az egyes komponensek valószínűségének szorzata adja az együttes valószínűséget (az egydimenziós valószínűségeket a [62] algoritmussal számítottuk).

A példákat a p valószínűség, az n dimenziószám, a $\mathbf{h}$ integrálási határ és az r_{ij}, $i=1, \ldots, n-1$, $j=i+1, \ldots, n$ korrelációs együtthatók segítségével adjuk meg. A példában explicite meg nem adott r_{ij}, $i \neq j$ együtthatók mindig zérusnak veendők. Az $\mathbf{R}$ korreláció mátrix szimmetrikussága miatt a fenti adatok a példát teljesen leírják.

1. Példa. $p=0{,}73519173$, $n=2$, $\mathbf{h}=(0{,}9\ 1{,}4)$, $r_{12}=-0{,}8$.
2. Példa. $p=0{,}88257314$, $n=2$, $\mathbf{h}=(1{,}2\ 1{,}8)$, $r_{12}=0{,}9$.
3. Példa. $p=0{,}34371247$, $n=2$, $\mathbf{h}=(-0{,}4\ 2{,}2)$, $r_{12}=0{,}4$.
4. Példa. $p=0{,}17401540$, $n=3$, $\mathbf{h}=(0\ 0\ 0)$, $r_{12}=0{,}8$, $r_{13}=-0{,}4$, $r_{23}=0{,}1$.
5. Példa. $p=0{,}12159722$, $n=4$, $\mathbf{h}=(0\ 0\ 0\ 0)$, $r_{12}=0{,}5$, $r_{23}=0{,}5$, $r_{34}=0{,}309$.
6. Példa. $p=0{,}15$, $n=4$, $\mathbf{h}=(0\ 0\ 0\ 0)$, $r_{12}=0{,}5$, $r_{13}=0{,}5$, $r_{23}=0{,}5$, $r_{34}=0{,}5$.
7. Példa. $p=0{,}20$, $n=4$, $\mathbf{h}=(0\ 0\ 0\ 0)$, $r_{12}=0{,}612$, $r_{13}=0{,}25$, $r_{14}=0{,}408$, $r_{23}=0{,}666$, $r_{24}=0{,}406$, $r_{34}=0{,}612$.
8. Példa. $p=0{,}1666666$, $n=4$, $\mathbf{h}=(0\ 0\ 0\ 0)$, $r_{12}=0{,}707$, $r_{23}=0{,}5$, $r_{34}=0{,}707$.
9. Példa. $p=0{,}04797920$, $n=3$, $\mathbf{h}=(0\ 0\ 0)$, $r_{12}=-0{,}9$, $r_{13}=-0{,}7$, $r_{23}=0{,}8$.
10. Példa. $p=0{,}24319006$, $n=3$, $\mathbf{h}=(0\ 0\ 0)$, $r_{12}=0{,}95$, $r_{13}=-0{,}6$, $r_{23}=-0{,}4$.
11. Példa. $p=0{,}64064279$, $n=4$, $\mathbf{h}=(2{,}1\ 1{,}2\ 0{,}6\ 3{,}8)$, $r_{12}=0{,}8$, $r_{34}=-0{,}9$.
12. Példa. $p=0{,}23241$, $n=5$, $h_1=\ldots=h_5=0{,}3$, $r_{ij}=0{,}4$, $i=1, \ldots, 4$, $j=i+1, \ldots, 5$.
13. Példa. $p=0{,}99043613$, $n=6$, $\mathbf{h}=(3{,}8\ 4{,}1\ 2{,}5\ 3{,}2\ 3{,}9\ 2{,}8)$, $r_{12}=-0{,}5$ $r_{34}=0{,}1$, $r_{56}=-0{,}7$.
14. Példa. $p=0{,}58300606$, $n=10$, $\mathbf{h}=(1{,}7\ 0{,}8\ 5{,}1\ 3{,}2\ 2{,}4\ 1{,}8\ 2{,}7\ 1{,}5\ 1{,}2\ 2{,}6)$, $r_{12}=-0{,}6$, $r_{34}=0{,}9$, $r_{56}=0{,}4$, $r_{78}=0{,}2$, $r_{9,10}=-0{,}8$.
15. Példa. $p=0{,}85276$, $n=10$, $h_1=\ldots=h_{10}=1{,}5$, $r_{ij}=0{,}9$, $i=1, \ldots, 9$, $j=i+1, \ldots, 10$.
16. Példa. $p=0{,}80128$, $n=15$, $h_1=\ldots=h_{11}=2{,}9$, $\mathbf{h}=(2{,}9\ \ldots\ 2{,}9\ 2{,}7\ 1{,}6\ 1{,}2\ 2{,}1)$, $r_{ij}=0{,}2$, $i=1, \ldots, 10$, $j=i+1, \ldots, 11$, $r_{12,13}=0{,}3$, $r_{14,15}=-0{,}95$.
17. Példa. $p=0{,}11054$, $n=15$, $h_1=\ldots=h_7=0{,}5$, $\mathbf{h}=(0{,}5\ \ldots\ 0{,}5\ 1{,}2\ 2{,}7\ 1{,}8\ 0{,}4\ 2{,}4\ 0{,}6\ 1{,}5\ 0{,}9)$, $r_{ij}=0{,}625$, $i=1, \ldots, 6$, $j=i+1, \ldots, 7$, $r_{89}=0{,}8$, $r_{10,11}=0{,}7$, $r_{12,13}=-0{,}8$, $r_{14,15}=-0{,}9$.

18. Példa. $p=0{,}92295$, $n=20$, $h_1=\ldots=h_8=3{,}1$, $\mathbf{h}=(3{,}1\ \ldots\ 3{,}1\ 2{,}8\ 3{,}2\ 4{,}3\ 3{,}8\ 2{,}3\ 1{,}9\ 2{,}6\ 2{,}2\ 3{,}1\ 2{,}3\ 2{,}7\ 3{,}4)$, $r_{ij}=0{,}875$, $i=1, \ldots, 7$, $j=i+1, \ldots, 8$, $r_{9,10}=0{,}1$, $r_{11,12}=0{,}8$, $r_{13,14}=-0{,}2$, $r_{15,16}=0{,}5$, $r_{17,18}=0{,}4$, $r_{19,20}=-0{,}9$.

19. Példa. $p=0{,}41805$, $n=20$, $h_1=\ldots=h_{12}=0{,}9$, $\mathbf{h}=(0{,}9\ \ldots\ 0{,}9\ 2{,}8\ 1{,}5\ 3{,}2\ 1{,}2\ 2{,}4\ 2{,}7\ 1{,}9\ 1{,}4)$, $r_{ij}=0{,}8$, $i=1, \ldots, 11$, $j=i+1, \ldots, 12$, $r_{13,14}=0{,}7$, $r_{15,16}=-0{,}2$, $r_{17,18}=-0{,}9$, $r_{19,20}=0{,}1$.

A következő 20—26. példákban a korreláció mátrix az egységmátrix volt, a $\mathbf{h}$ integrálási határ pedig a $\mathbf{h}^{(1)}$ vektor első n komponense volt, ahol $\mathbf{h}^{(1)}=(2{,}84\ 2{,}06,\ 2{,}64\ 3{,}44\ 3{,}73\ 2{,}55\ 2{,}99\ 2{,}24\ 2{,}16\ 2{,}97\ 1{,}73\ 1{,}43\ 2{,}39\ 2{,}02\ 1{,}74\ 2{,}51\ 2{,}54\ 1{,}95\ 1{,}66\ 2{,}53\ 1{,}53\ 2{,}34\ 2{,}46\ 2{,}69\ 1{,}88\ 1{,}99\ 2{,}21\ 2{,}70\ 2{,}88\ 1{,}57\ 2{,}15\ 2{,}31\ 2{,}02\ 1{,}93\ 1{,}55\ 3{,}20\ 2{,}33\ 1{,}60\ 1{,}77\ 3{,}20\ 3{,}10\ 2{,}37\ 1{,}50\ 1{,}43\ 2{,}84\ 3{,}34\ 2{,}17\ 2{,}21\ 3{,}30\ 2{,}98)$.

20. Példa. $p=0{,}97365838$, $n=5$.
21. Példa. $p=0{,}93883813$, $n=10$.
22. Példa. $p=0{,}77302876$, $n=15$.
23. Példa. $p=0{,}70443223$, $n=20$.
24. Példa. $p=0{,}56617742$, $n=30$.
25. Példa. $p=0{,}44324325$, $n=40$.
26. Példa. $p=0{,}36586192$, $n=50$.

A 27—30. példákban a korreláció mátrix az egységmátrix, a $\mathbf{h}$ vektor pedig a $\mathbf{h}^{(2)}$ vektor első n komponense, ahol $\mathbf{h}^{(2)}=(2{,}86\ 4{,}35\ 3{,}17\ 4{,}32\ 3{,}79\ 3{,}74\ 2{,}66\ 4{,}04\ 3{,}21\ 3{,}56\ 4{,}35\ 3{,}21\ 3{,}93\ 4{,}94\ 3{,}06\ 2{,}97\ 3{,}81\ 2{,}80\ 4{,}67\ 3{,}36\ 3{,}59\ 2{,}84\ 3{,}35\ 4{,}48\ 4{,}14\ 2{,}50\ 2{,}93\ 3{,}11\ 2{,}93\ 3{,}06\ 4{,}98\ 3{,}21\ 4{,}67\ 4{,}58\ 4{,}83\ 3{,}65\ 3{,}43\ 3{,}14\ 4{,}78\ 3{,}17\ 3{,}98\ 3{,}23\ 4{,}23\ 3{,}01\ 4{,}31\ 3{,}33\ 3{,}79\ 3{,}11\ 2{,}65\ 3{,}81)$.

27. Példa. $p=0{,}98591377$, $n=20$.
28. Példa. $p=0{,}97170781$, $n=30$.
29. Példa. $p=0{,}96907774$, $n=40$.
30. Példa. $p=0{,}96180948$, $n=50$.

5. Feladatok

Hat numerikus feladatot adunk meg, ezek esetében a valószínűség ismeretlen. A feladatokat véletlenszám generátorok segítségével állítottuk össze. Az $\mathbf{R}^{(1)}$ vagy $\mathbf{R}^{(2)}$ mátrixok bal felső $n\times n$-es sarokmátrixa adja meg az adott példa korreláció mátrixát, a határt pedig a $\mathbf{h}^{(3)}$ vagy $\mathbf{h}^{(4)}$ vektorok első n komponense.

31. Példa. $n=5$, $\mathbf{h}^{(3)}$, $\mathbf{R}^{(1)}$.
32. Példa. $n=10$, $\mathbf{h}^{(3)}$, $\mathbf{R}^{(1)}$.
33. Példa. $n=20$, $\mathbf{h}^{(3)}$, $\mathbf{R}^{(1)}$.
34. Példa. $n=5$, $\mathbf{h}^{(4)}$, $\mathbf{R}^{(2)}$.
35. Példa. $n=10$, $\mathbf{h}^{(4)}$, $\mathbf{R}^{(2)}$.
36. Példa. $n=20$, $\mathbf{h}^{(4)}$, $\mathbf{R}^{(2)}$.

Itt $\mathbf{h}^{(3)}=(1{,}03\ 2{,}59\ 2{,}20\ 2{,}71\ 1{,}09\ 2{,}67\ 1{,}54\ 2{,}85\ 2{,}22\ 3{,}42\ 1{,}19\ 3{,}77\ 2{,}62\ 2{,}55\ 2{,}45\ 3{,}86\ 2{,}05\ 1{,}30\ 1{,}65\ 3{,}40)$, a másik határvektor pedig $\mathbf{h}^{(4)}=(2{,}95\ 2{,}97\ 3{,}78\ 3{,}00\ 4{,}02\ 2{,}85\ 4{,}34\ 2{,}17\ 2{,}81\ 4{,}03\ 4{,}37\ 2{,}59\ 3{,}10\ 4{,}18\ 4{,}34\ 4{,}03\ 2{,}72\ 3{,}89\ 3{,}11\ 3{,}01)$.

Az $\mathbf{R}^{(1)}$ és $\mathbf{R}^{(2)}$ mátrixot a főátló feletti elemekkel adjuk meg.

Az $\mathbf{R}^{(1)}$ *mátrix soronként:*

1. sor 0,039 0,097 −0,041 0,040 0,144 0,095 0,553 −0,223 −0,035 −0,215 0,113 −0,409 −0,435 0,519 −0,024 −0,416 0,209 0,129 −0,013
2. sor 0,474 0,398 0,221 0,344 0,134 0,385 −0,009 0,030 −0,257 −0,268 0,069 0,037 0,072 0,155 −0,070 0,002 −0,192 0,257
3. sor 0,054 −0,048 −0,080 −0,145 0,446 −0,043 −0,288 −0,230 −0,177 0,210 −0,054 −0,145 −0,092 −0,120 −0,050 0,161 0,307
4. sor −0,099, −0,272 −0,003 0,232 −0,013 0,138 −0,099 −0,036 0,103 −0,002 0,115 0,484 −0,277 0,274 −0,239 0,062
5. sor 0,087 −0,145 −0,241 −0,141 −0,173 0,119 0,024 −0,016 0,051 0,490 0,335 −0,102 0,258 0,022 0,225
6. sor 0,397 0,002 −0,282 0,022 −0,202 0,144 −0,293 −0,091 0,096 −0,087 0,264 −0,228 0,018 −0,020
7. sor −0,132 0,179 0,226 −0,120 −0,024 −0,173 −0,142 −0,140 −0,121 0,180 −0,404 0,080 0,052
8. sor −0,186 −0,085 −0,219 −0,257 −0,138 −0,134 0,067 0,113 −0,310 0,147 −0,219 0,079
9. sor 0,072 0,195 −0,468 −0,035 0,362 −0,354 −0,224 0,114 −0,113 0,065 −0,009
10. sor −0,218 0,148 −0,446 −0,102 0,056 −0,012 0,281 −0,243 −0,176 −0,191
11. sor −0,016 0,113 0,185 −0,213 0,259 0,146 −0,130 −0,208 −0,188
12. sor 0,133 −0,427 0,410 −0,189 0,106 −0,299 0,488 −0,046
13. sor 0,0 −0,003 −0,203 −0,123 −0,159 0,194 0,129
14. sor −0,337 0,127 0,139 0,200 0,110 −0,107
15. sor 0,193 −0,259 0,070 0,224 0,024
16. sor −0,167 0,257 −0,412 −0,180
17. sor −0,475 0,094 −0,117
18. sor −0,092 0,242
19. sor 0,137

Az $\mathbf{R}^{(2)}$ *mátrix soronként*

1. sor −0,006 −0,198 −0,412 −0,352 −0,169 0,251 0,003 −0,213 0,012 0,094 0,115 0,017 0,129 −0,195 0,123 −0,010 0,247 0,149 0,554
2. sor −0,140 0,115 0,246 0,290 0,209 −0,077 0,197 −0,106 0,039 −0,195 0,401 −0,220 0,158 −0,296 0,108 0,259 −0,212 −0,098
3. sor 0,295 0,409 −0,163 −0,210 −0,141 −0,471 0,203 0,288 0,045 0,212 −0,020 −0,026 0,436 −0,164 −0,178 0,272 −0,058
4. sor 0,387 −0,115 −0,233 0,127 −0,289 −0,068 0,235 −0,055 −0,039 −0,006 0,194 −0,203 0,379 0,146 0,017 0,183
5. sor −0,314 −0,106 −0,075 0,088 0,210 0,218 −0,438 0,090 0,078 0,222 −0,128 −0,168 0,107 0,097 0,130
6. sor 0,367 0,043 0,065 0,237 −0,204 0,252 0,056 −0,279 0,086 −0,343 0,298 0,238 0,140 −0,402
7. sor 0,133 0,287 0,225 −0,076 0,168 −0,185 0,134 −0,056 −0,168 −0,048 0,064 0,191 −0,055
8. sor 0,076 0,006 −0,025 0,266 −0,572 0,235 0,240 0,102 0,404 0,017 0,180 0,103

9. sor 0,071 −0,146 −0,197 −0,213 −0,287 0,286 −0,146 −0,070 −0,262 −0,392 −0,367
10. sor 0,169 0,031 −0,200 0,001 0,091 0,230 0,016 0,212 0,125 −0,159
11. sor −0,143 −0,195 0,285 −0,024 0,261 −0,092 −0,235 −0,041 0,236
12. sor 0,001 0,016 −0,065 0,116 0,222 0,053 −0,121 −0,118
13. sor −0,245 0,031 0,022 −0,391 0,297 −0,043 −0,196
14. sor −0,005 0,143 −0,177 −0,112 0,366 0,440
15. sor −0,239 0,002 0,374 −0,214 −0,140
16. sor −0,236 −0,321 0,045 −0,132
17. sor 0,164 −0,123 0,085
18. sor −0,033 0,158
19. sor 0,371

IRODALOM

[1] AHRENS, J. H., DIETER, U., "Computer methods for sampling from the exponential and normal distributions", *Comm. ACM* **15** (1972) 873—882.
[2] AHRENS, J. H., DIETER, U., "Extensions of Forsythe's method for random sampling from the normal distributions", *Math. Comp.* **27** (1973) 927—937.
[3] AHRENS, J. H., DIETER, U., "Computer methods for sampling from gamma, beta, Poisson and binomial distributions", *Computing* **12** (1974) 223—246.
[4] AHRENS, J. H., DIETER, U., "Non-uniform random numbers", Preprint 1974, 192 pages.
[5] ALAGAR, V. S., "The distribution of the distance between random points", *J. Appl. Prob.* **13** (1976) 558—566.
[6] ANDEL, J., "On multiple normal probabilities of rectangles", *Aplikace Mat.* **16** (1971) 172—181.
[7] ATKINSON, A. C., "An easily programmed algorithm for generating gamma random variables', *J. Roy. Stat. Soc. A* **140** (1977) 232—234.
[8] ATKINSON, A. C., PEARCE, M. C., "The computer generation of beta, gamma and normal random variables", *J. Roy. Stat. Soc. A* **139** (1976) 431—461.
[9] ATKINSON, A. C., WHITTAKER, J., "A switching algorithm for the generation of beta random variables with at least one parameter less than one", *J. Roy. Stat. Soc. A* **139** (1976) 462—467.
[10] BÁNKÖVI, GY., "A note on the generation of beta distributed and gamma distributed random variables", *Publ. Math. Inst. Hung. Acad. Sci.* **9** (1964) 555—563.
[11] BÁNKÖVI, GY., "A decomposition-rejection technique for generating exponential random variables", *Publ. Math. Inst. Hung. Acad. Sci.* **9** (1964) 573—581.
[12] BARR, D. R., SLEZAK, N. L., "A comparison of multivariate normal generators", *Comm. ACM* **15** (1972) 1048—1049.
[13] BÉKÉSSY, A., "Remarks on beta distributed random numbers", *Publ. Math. Inst. Hung. Acad. Sci.* **9** (1964) 565—571.
[14] BELL, J. R., "Normal random deviates", Algorithm 334, *Comm. ACM* **11** (1968) 498.
[15] BENE, B., "Reichentransformationen zur gemeinsamen Normalverteilung von familiären Erkrangungsneigungen", Compstat, 1974 Vienna (előadás).
[16] BOHRER, R., "A multivariate t probability integral", *Biometrika* **60** (1973) 647—654.
[17] BRENT, R. P., "A Gaussian pseudo-random number generator", *Algorithm* 488, *Comm. ACM* **17** (1974) 704—706.
[18] BURR, I. W., "A useful approximation to the normal distribution function, with application to simulation", *Technometrics* **9** (1967) 647—651.
[19] BUSZLENKO, N. P., GOLENKO, D. I., SREJGYER, J. A., SZOBOL, I. M., SZRAGOVICS, V. G *Monte Carlo módszerek,* (Műszaki Könyvkiadó, Budapest, 1965).
[20] BUSZLENKO, N. P., *Bonyolult rendszerek szimulációja,* (Műszaki Könyvkiadó, Budapest, 1972).
[21] BUTCHER, J. C., "Random sampling from the normal distribution", *Computer J.* **3** (1961) 251—253.
[22] BUTLER, E. L., "General random number generator", *Algorithm* 370, *Comm. ACM* **13** (1970) 49—52.
[23] BUTLER, J. W., "Machine sampling from given probability distributions", in: Symp. on Monte Carlo methods, Ed. *H. A. Meyer* (J. Wiley, New York, 1956) 249—264.

[24] CHAMAYOU, J. M. F., "On a direct algorithm for the generation of log-normal pseudo-random numbers", *Computing* **16** (1976) 69—76.
[25] COOK, J. M., "Rational formulae for the production of spherically symmetric probability distribution", *MTAC* **11** (1957) 81—82.
[26] COOK, J. M., "Remarks on a recent paper", *Comm. ACM* **2** (1959) 26.
[27] COVEYOU, R. R., MACPHERSON, R. D., "Fourier analysis of uniform random number generators", *J. ACM* **14** (1967) 100—119.
[28] CRANLEY, R., PATTERSON, T. N. L., "A regression method for the Monte Carlo evaluation of multidimensional integrals", *Num. Math.* **16** (1970) 58—72.
[29] DAVIS, P. J., RABINOWITZ, P., *Methods of Numerical Integration* (Academic Press, New York, 1975).
[30] DEÁK, I., „Egy sztochasztikus programozási modell számítógépes kiértékelése", *MTA Számítástechnikai Központ Közleményei* **9** (1972) 33—49.
[31] DEÁK, I., „A többdimenziós normális eloszlásfüggvény Monte Carlo integrálással történő kiszámításának számítógépes tapasztalatai", *MTA Számítástechnikai és Automatizálási Kutató Intézet Közleményei*, **19** (1978) 47—60.
[32] DEÁK, I., „A többdimenziós tér halmazai valószínűségeinek kiszámítása normális eloszlás esetén", *Alk. Mat. Lapok* **2** (1976) 17—26.
[33] DEÁK, I., „A többdimenziós normális eloszlásfüggvény Monte Carlo kiszámítása az Ellipszoid módszer segítségével", *Alk. Mat. Lapok* **2** (1976) 341—349.
[34] DEÁK, I., "Monte Carlo evaluation of the multidimensional normal distribution function by the Ellipsoid method", *Problems of Contr. Inf. Theo.* 7 (1978) 203—212.
[35] DEÁK, I., "Comparison of methods for generating uniformly distributed random points in and on a hypersphere", *Problems of Contr. Inf. Theo.* **8** (1979)105—113.
[36] DEÁK, I., "The Ellipsoid method for generating normally distributed random vectors", *Zastosowania Mat.* **17** 1979.
[37] DIETER, U., "Pseudo-random numbers: the exact distribution of pairs", *Math. Comp.* **25** (1971) 855—883.
[38] DIETER, U., "How to calculate shortest vectors in a lattice", *Math. Comp.* **29** (1975) 827—833.
[39] DIETER, U., AHRENS, J. H., "A combinatorial method for the generation of normally distributed random numbers", *Computing* **11** (1973) 137—146.
[40] DIETER, U., AHRENS, J. H., "Pseudo-random numbers", Preprint (1974) 216 pages.
[41] DONELLY, T. G., "Bivariate normal distribution", *Algorithm* 462, *Comm. ACM* **16** (1973) 638.
[42] DUTKIEWICZ, S., „Generowanie n-wymiarowych sympleksów skierowanych losowo", *Algorytmy* **XI** (1974) 29—47.
[43] DUTT, J. E., "A representation of multivariate normal probability integrals by integral transforms", *Biometrika* **60** (1973) 637—645.
[44] DUTT, J. E., "On computing the probability integral of a general multivariate *t*", *Biometrika* **62** (1975) 201—205.
[45] ESCOUFIER, Y., "Calculs de probabilités par une méthode de Monte-Carlo pour une variable *p*-normale", *Revue de Statistique Appl.* **15** (1967) 5—15.
[46] FELLEN, B. M., "An implementation of the Tausworthe generator", *Comm. ACM* **12** (1969) 413.
[47] FORSYTHE, G. E., "Von Neumann's comparison method for random sampling from the normal and other distributions", *Math. Comp.* **26** (1972) 817—826.
[48] FOX, L., *An Introduction to Numerical Linear Algebra* (Clarendon Press, Oxford, 1964).
[49] FRANKLIN, J. N., "Numerical simulation of stationary and nonstationary Gaussian random processes", *SIAM Review* 7 (1965) 68—80.
[50] FRANKLIN, M. A., SEN, A., "Comparison of exact and approximate variate generation methods for the Erlang distribution", *J. Stat. Comp. Simul.* **4** (1975) 1—18.
[51] FULLERTON, W., "Modified incomplete gamma function", *Algorithm* 435. *Comm. ACM* **15** (1972) 993—995.
[52] GICHMAN, I. I., SKOROCHOD, A. W., *Wstep do teorii procesów stochastycznych* (Pans. Wyd. Naukowe, Warszawa, 1968).
[53] GOLDSTEIN, R. B., "Chi-Square quantiles", *Algorithm* 451, *Comm. ACM* **16** (1973) 483—485.
[54] *Good, I. J., Gaskins, R. A.*, "The centroid method of numerical integration", *Num. Math.* **16** (1971) 343—359.
[55] GREENWOOD, A. J., "A fast generator for gamma-distributed random variables", Compstat 1974, ed. *G. Bruckmann, F. Ferschl, L. Schmetterer* (Physica Verlag, Wien, 1974) 19—27.

[56] GUPTA, S. S., "Probability integrals of multivariate normal and multivariate t", *Ann. Math. Stat.* **34** (1963) 792—828.
[57] HABER, S., "Numerical evaluation of multiple integrals", *SIAM Review*, **12** (1970) 481—526.
[58] HALTON, J. H., "A retrospective and prospective survey of the Monte Carlo method", *SIAM Review* **12** (1970) 1—63.
[59] HAMMERSLEY, J. M., HANDSCOMB, D. C., *Monte Carlo Methods* (Methuen, London, 1964).
[60] HASELGROVE, C. B., "A method for numerical integration", *Math. Comp.* **15** (1961) 323—337.
[61] HICKS, J. S., WHEELING, R. F., "An efficient method for generating uniformly distributed points on the surface of an *n*-dimensional sphere", *Comm. ACM* **2** (1959) 17—19.
[62] HILL, I. D., JOYCE, S. A., "Normal curve integral", *Algorithm* 304, *Comm. ACM* **10** (1967) 374—375.
[63] HILL, I. D., PIKE, M. C., "Chi-squared integral", *Algorithm* 299, *Comm. ACM* **10** (1967) 243—244.
[64] HULL, T. E., DOBELL, A. R., "Random number generators", *SIAM Review* **4** (1962) 230—254.
[65] HURST, R. L., KNOP, R. E., "Generation of random correlated normal variables", *Algorithm* 425, *Comm. ACM* **15** (1972) 355—357.
[66] JANSSON, B., *Random Number Generators* (Almqvist and Wiksell, Stockholm. 1966).
[67] JOHNSON, N. L., KOTZ, S., *Distributions in Statistics, I—IV. vol.* (J. Wiley, New York, 1972).
[68] JÖHNK, M. D., "Erzeugung von Betaverteilten und Gammaverteilten Zuffälszahlen", *Metrika* **8** (1964) 5—15.
[69] KAMINSKY, F. C., RUMPF, D. L., "Simulating nonstationary Poisson processes: a comparison of alternatives including the correct approach", *Simulation* **28** (1977) July, 17—20.
[70] Kendall, M. G., *A Course in the Geometry of n Dimensions*, in: Griffin's stat. monogr. and courses, ed. *M. G. Kendall*, (Griffin, London, 1961).
[71] KENDALL, M. G., MORAN, P. A. P., *Geometrical Probability*, in: Griffin's stat. monogr. and courses, ed. *M. G. Kendall*, (Griffin, London, 1963).
[72] KNOP, R. E., "Random vectors uniform in a solid angle", *Algorithm* 381, *Comm. ACM* **13** (1970) 326.
[73] KNUTH, D. E., *The Art of Computer Programming, II. vol. Seminumerical algorithms*, (Reading Mass., Addison Wesley, 1969).
[74] LEWIS, T. G., PAYNE, W. H., "Generalized feedback shift register pseudo-random number algorithm", *J. ACM* **20** (1973) 456—468.
[75] LUKASZEWSKA, L., PLESZCZYNSKA, E., „Generowanie realizacji procesu Poissona ze zmienna intensywnoscia", *Algorytmy* 7 (1967) 49—60.
[76] MACLAREN, M. D., MARSAGLIA, G., "Uniform random number generators", *J. ACM* **12** (1965) 83—89.
[77] MAGHSOODLOO, S., "Eccentricities for which ellipsoidal probabilities are good approximations to spherical probabilities", *J. Stat. Comp. Simul.* **3** (1975) 369—378.
[78] MARSAGLIA, G., "Experessing a random variable in terms of uniform random variables", *Ann. Math. Stat.* **32** (1961) 894—898.
[79] MARSAGLIA, G., "Generating discrete random variables in a computer", *Comm. ACM* **6** (1963) 37—38.
[80] MARSAGLIA, G., "Generating a variable from the tail of the normal distribution", *Technometrics* **6** (1964) 101—102.
[81] MARSAGLIA, G., "Random numbers fall mainly in the planes", *Proc. Nat. Acad. Sci. USA* **61** (1968) 25—28.
[82] MARSAGLIA, G., MACLAREN, M. D., BRAY, T. A., "A fast procedure for generating normal random variables", *Comm. ACM* **7** (1964) 4—10.
[83] MAYER, J., "Computational experiences with the reduced gradient method", in: Coll, Math. Soc. J. Bolyai 12 ed. *A. Prékopa* (Progress in Op. Res., held at Eger, Hungary, 1974) 613—624.
[84] MCGRAW, D. K., WAGNER, J. F., „Elliptically symmetric distributions", *IEEE Trans. on Inf. Theory* **IT—14** (1968) 110—120.
[85] MILTON, R. C., "Computer evaluation of the multivariate normal integral", *Technometrics* **14** (1972) 881—889.
[86] MIYATAKE, O., INOUE, H., YOSHIZAWA, Y., "Generation of physical random numbers", *Math. Japonica* **20** (1975) 207—217.
[87] MULLER, M. E., "A comparison of methods for generating normal deviates on digital computers", *J. ACM* **6** (1959) 376—383.

[88] MULLER, M. E., "A note on a method for generating points uniformly on n-dimensional spheres", *Comm. ACM* **2** (1959) 19—20.
[89] MURRY, H. F., "A general approach for generating natural random variables", *IEEE Trans. on Comp.* **C—19** (1970) 1210—1213.
[90] NANCE, R. E., OVERSTREET, C. JR., "A bibliography on random number generation", *Computing Reviews* **13** (1972) 495—508.
[91] NEUMANN, J., "Various techniques used in connection with random digits", *J. Res. NBS. Appl. Math. Series* **3** (1951) 36—38.
[92] NEWMANN, T. G., ODELL, P. L., *The Generation of Random Variates* in: Griffin's stat. monogr. and courses, ed. *M. G. Kendall* (Griffin, London, 1971).
[93] PAYNE, W. H., "Fortran Tausworthe pseudo-random number generator", *Comm. ACM* **13** (1970) 57.
[94] PAYNE, W. H., "Normal random numbers: using machine analysis to choose the best algorithm", *ACM Trans. on Math. Software* **3** (1977) 346—358.
[95] PHILLIPS, D. T., BEIGHTLER, C. S., "Procedure for generating gamma variates with non-integer parameter sets", *J. Stat. Comp. Simul.* **1** (1972) 197—208.
[96] POLGE, R. J., HOLLIDAY, E. M., BHAGAVAN, B. K., "Generation of a pseudo-random set with desired correlation and probability distribution", *Simulation* **20** (1973) 153—158.
[97] PRÉKOPA, A., *Valószínűségelmélet* (Műszaki Könyvkiadó, Budapest, 1974).
[98] PRÉKOPA, A., "On probabilistic constrained programming", in: Proc. Princeton Symp. on Math. Prog. (Princeton Univ. Press, Princeton N. J., 1970) 113—138.
[99] PRÉKOPA, A.: „Sztochasztikus rendszerek optimalizálási problémáiról", doktori értekezés, Magyar Tudományos Akadémia, Budapest, 1970.
[100] PRÉKOPA, A., "Logarithmic concave measures with application to stochastic programming", *Acta Scientiarum Mathematicarum* **32** (1971) 301—316.
[101] PRÉKOPA, A., "Contributions to the theory of stochastic programming", *Mathematical Programming* **4** (1973) 202—221.
[102] PRÉKOPA, A., GANCZER, S., DEÁK, I., PATYI, K., „A STABIL sztochasztikus programozási modell és annak kísérleti alkalmazása a magyar villamosenergia-iparra", *Alk. Mat. Lapok* **1** (1975) 3—22.
[103] PRÉKOPA, A., GANCZER, S., DEÁK, I., PATYI, K., "The STABIL stochastic programming modell and its experimental application to the electrical energy sector of the Hungarian economy", in: Proc. of the Int. Symp. on Stochastic Programming, ed. *M. Dempster* (sajtó alatt) és *Studies in applied stochastic programming I., ed. A. Prékopa, MTA SZTAKI Tanulmányok* **80** (1978) 5—30.
[104] PRÉKOPA, A., "Application of stochastic programming to engineering design", in: Proc. of the IX. Int. Symp. on Mathematical Programming, ed. *A. Prékopa* (Akadémiai Kiadó, Budapest, 1978).
[105] PRÉKOPA, A., RAPCSÁK, T., ZSUFFA, I., „Egy új módszer sorbakapcsolt tározórendszerek tervezésére", *Alk. Mat. Lapok* (sajtó alatt).
[106] PRÉKOPA, A., RAPCSÁK, T., ZSUFFA, I., "Energiatermelő tározó optimális méretezése", (előkészületben).
[107] PROLL, L. G., "Remark on algorithm 370", *Comm. ACM* **15** (1972) 467—468.
[108] RAMBERG, J. S., SCHMEISER, B. W., "An approximate method for generating symmetric random variables", *Comm. ACM* **15** (1972) 987—990.
[109] RAMBERG, J. S., SCHMEISER, B. W., "An approximate method for generating asymmetric random variables", *Comm. ACM* **17** (1974) 78—82.
[110] RAMBERG, J. S., TADIKAMALLA, P. R., "An algorithm for generating gamma variates based on the Weibull distribution", *AIEE Trans.* **6** (1974) 257—260.
[111] RAPCSÁK, T., „Egy tározási modell számítástechnikai megoldása", egyetemi doktori disszertáció, ELTE TTK, Budapest, 1974.
[112] RAPCSÁK, T., „Egy külső pont eljárás konvex nemlineáris programozási feladatok megoldására", *Alk. Mat. Lapok* **1** (1975) 357—364.
[113] RÉNYI, A., *Valószínűségszámítás* (Tankönyvkiadó, Budapest, 1966).
[114] RONNING, G., "A simple scheme for generating multivariate gamma distributions with non-negative covariance matrix", *Technometrics* **19** (1977) 179—183.
[115] ROSENBLATT, M., "Multiply schemes and shuffling", *Math. Comp.* **29** (1975) 929—934.
[116] RUDOLPH, E., HAWKINS, D. M., "Random number generators in cyclic queuing applications", *J. Stat. Comp. Simul.* **5** (1976) 65—71.

[117] SALFI, R., "A long-period random number generator with application to permutations", Compstat 1974, ed. *G. Bruckmann, F. Ferschl, L. Schmetterer* (Physica Verlag, Wien, 1974) 28—35.
[118] SCHAFFER, H. E., "Generator of random numbers satisfying the *Poisson distribution*", *Algorithm* 369, *Comm. ACM* **13** (1970) 49.
[119] SCHEUER, E. M., STOLLER, D. S., "On the generation of normal random vectors", *Technometrics* **4** (1962) 278—281.
[120] SCHRACK, G. F., "Remark on algorithm 381", *Comm. ACM* **15** (1972) 468.
[121] SERAPHIN, D. S., "A fast random number generator for IBM 360", *Comm. ACM* **12** (1969) 695.
[122] SIBUYA, M., "A method for generating uniformly distributed points on n-dimensional speheres", *Ann. Inst. Stat. Math.* **14** (1962—63) 81—85.
[123] SIBUYA, M., "Generating doubly exponential random numbers", *Ann. Inst. Stat. Math. Suppl.* **5** (1968) 1—7.
[124] SIDÁK, Z., "Remarks on Andel's paper "On multiple normal probabilities of rectangles"", *Aplikace Mat.* **16** (1971) 182—187.
[125] STRAZICZKY, B., "On an algorithm for solution of the two stage stochastic programming problem", *Operation Research Verfahren* **19** (1973) 142—156.
[126] STROUD, A. H., *Approximate Calculation of Multiple Integrals,* (Prentice Hall, Englewood Cliffs, 1971).
[127] SZÁNTAI, T., „Egy eljárás a többdimenziós normális eloszlásfüggvény és gradiense értékeinek meghatározására", *Alk. Mat. Lapok* **2** (1976) 27—39.
[128] SZÁNTAI, T., „A Prékopa-féle STABIL sztochasztikus programozási modell numerikus megoldásáról", *Alk. Mat. Lapok* **2** (1976) 93—101.
[129] TADIKAMALLA, P. R., RAMBERG, J. S., "An approximate method for generating gamma and other variates", *J. Stat. Comp. Simul.* **3** (1975) 275—282.
[130] TAUSWORTHE, R. C., "Random numbers generated by linear recurrence modulo two", *Math. Comp.* **19** (1965) 201—209.
[131] TOOTILL, J. P. R., ROBINSON, W. D., ADAMS, A. G., "The runs up-and-down performance of Tausworthe pseudo-random number generators", *J. ACM* **18** (1971) 381—399.
[132] TOOTILL, J. P. R., ROBINSON, W. D., EAGLE, D. J., "An asymptotically random Tausworthe sequence", *J. ACM* **20** (1973) 469—481.
[133] TSUDA, T., "Numerical integration of functions of very many variables", *Num. Math.* **20** (1973) 377—391.
[134] WALKER, A. J., "An efficient method for generating discrete random variables with general distributions", *ACM Trans. on. Math. Software* **3** (1977) 253—256.
[135] WALLACE, N. D., "Computer generation of gamma random variates with non-integral shape parameters", *Comm. ACM* **17** (1974) 691—695.
[136] WESTLAKE, W. J., "A uniform random number generator based on the combination of two congruential generators", *J. ACM* **14** (1967) 337—340.
[137] WHEELER, D. J., "An approximation for simulation of gamma distributions", *J. Stat. Comp. Simul.* **3** (1975) 225—232.
[138] WHITLESEY, J. RB., "A comparison of the correlational behavior of random number generators for the IBM 360", *Comm. ACM* **11** (1968) 641—644.
[139] WHITLESEY, J. RB., "On the multidimensional uniformity of pseudo-random generators", *Comm. ACM* **12** (1969) 247.
[140] ZAREMBA, S. K., "The mathematical basis of Monte Carlo and quasi-Monte Carlo methods", *SIAM Review* **10** (1968) 303—314.
[141] ZIELINSKI, R., *Generatory liczb losowych* (Wyd. Naukowo-Techniczne, Warszawa, 1972).
[142] ZIELINSKI, R., "A Monte Carlo estimator of the gradient", Compstat 1974, ed. *G. Bruckmann, F. Ferschl, L. Schmetterer* (Physica Verlag, Wien, 1974) 61—69.
[143] YUEN, C. K., "Testing random number generators by *Walsh transform*", *IEEE Trans. on Comp.* **C—26** (1977) 329—333.
[144] Андерсон, Т., *Введение в многомерный статистический анализ,* (Изд. Физ. Мат. Лит., Москва, 1963).
[145] Ермаков, С. М., *Метод Монте Карло и смежные вопросы,* (Изд. Наука, Москва, 1971).
[146] Миркин, Л. И., Рабинович, М. А., Ярославский, Л. П., «Метод генерирования коррелированных гауссовских псевдослучайных чисел на ЭВМ», *Ж. Выч. Мат. и Мат. Физ.* **12**(1972) 1353—1357.

[147] Михайлов, Г. А., «О методе 'повторения' для моделирования случайных векторов и процессов», *Теор. вер. и её прим.* **19**(1974) 873—878.
[148] Ососков, Г. А., «Быстрый способ получения случайной последовательности с пуассоновым законом распределения», *Ж. Выч. Мат. и Мат. Физ.* **16**(1976) 1052—1057.
[149] Смирнов, Н. В., Большев, Л. Н., *Таблицы для вычисления функции двухмерного нормального распределения* (Изд. Ак. Наук, Москва, 1962).
[150] Соболь, И. М., *Численные методы Монте Карло*, (Изд. Наука, Москва, 1973).

(Beérkezett: 1978. május 17.)

DEÁK ISTVÁN
MTA SZÁMÍTÁSTECHNIKAI ÉS AUTOMATIZÁLÁSI KUTATÓ INTÉZET
1250 BUDAPEST, ÚRI U. 49.

MONTE CARLO METHODS FOR COMPUTING PROBABILITIES OF SETS IN HIGHER DIMENSIONAL SPACES IN CASE OF NORMAL DISTRIBUTION

I. DEÁK

The evaluation of the distribution function of the multidimensional normal distribution is often required in stochastic programming models and in multivariate statistical problems. In the paper for one thing we made a summary of the author's results in this field for another several new algorithms are presented. In the first chapter preliminary results are cited. In the second third and fourth chapters fast algorithms are presented for generating stationary normal vectors. In the fifth chapter (i) previously published results concerning the computation of multiple normal probabilities are outlined (ii) algorithms based on antithetic varietes techniques are presented for computing the normal distribution function and (iii) the methods are compared computing a large number of examples (thirty examples in dimensions $n=2$—50 are reproduced in Appendix 4 and six examples with unknown probabilities in dimensions $n=5$—20 in Appendix 5). As a result of the presented algorithms the three digit accurate computation of the multiple normal distribution function is possible in 5 seconds up to $n=10$ dimensions, in 1 min. up to $n=20$ dimensions and in 15 min. up to $n=50$ dimensions. For probabilities near to 1 which are especially interesting values from the point of view of stochastic programming the computation may be carried out in one tenth of the times given above.

Alkalmazott Matematikai Lapok 4 (1978) 95—108

AZ ÁLTALÁNOS ENTRÓPIA FÜGGVÉNY KONVEX HALMAZON VALÓ MINIMALIZÁLÁSÁRÓL

KÉRI GERZSON
Budapest

A célfüggvény egyenesek mentén való aszimptotikus viselkedését vizsgáljuk, majd ennek alapján két különböző típusú transzformációt vezetünk be azoknak az optimalizálási feladatoknak a vizsgálatára, melyeknél a címben szereplő függvényt minimalizáljuk konvex zárt halmazon. Az első fajta transzformáció a feltételi halmaz meghagyása mellett a célfüggvény nemlineáris részét rövidíti le, vagy legrosszabb esetben a célfüggvényt is változatlanul hagyja. A második fajta transzformáció lényegét tekintve egy vetítés, amely a célfüggvény lineáris részét húzza össze oly módon, hogy a csak lineárisan tartalmazott változókat egyetlen változóval helyettesíti, a feltételi halmazt pedig a vetületével. A két transzformáció együtt — illetve lineáris feltételrendszer esetén az első fajta transzformáció egyedül is — eszközül szolgál a szóban forgó optimalizálási feladatok korlátosságára, valamint a minimum elérésére vonatkozó kérdés algoritmikus eldöntésére. Az erre a célra alkalmas algoritmus menetének leírását mellőzzük, mivel az magától értetődően adódik a 4—7. szakaszok tételei, valamint az ezeket szemléltető két táblázat (7. szakasz) alapján. Ehelyett — most már lineáris feltételrendszer esetére szorítkozva — a 8. szakaszban az algoritmusnak zártabb formában történő tételszerű megfogalmazását adjuk.

1. Bevezetés

A geometriai programozás tárgya szűkebb értelemben az úgynevezett primál, illetve duál geometriai program megoldása, tágabb értelemben ezenkívül a primál, illetve duál geometriai programmal kapcsolatos mindennemű matematikai elmélet és gyakorlati felhasználás.

A primál geometriai program alatt a következő feladat értendő: Maximalizálandó a

$$\sum_{j=1}^{n} \beta_j \eta_j \tag{1.1}$$

függvény a

$$\sum_{i \in I_k} e^{\sum_{j=1}^{n} \alpha_{ij}\eta_j - \gamma_i} \leqq 1 \quad (k = 1, 2, \ldots, p) \tag{1.2}$$

$$\sum_{j=1}^{n} \alpha_{ij}\eta_j \leqq \gamma_i \quad (i \in I_0)$$

feltételek mellett.

A duál geometriai program alatt a következő feladat értendő: Minimalizálandó a

$$\sum_{k=0}^{p} \sum_{i \in I_k} \gamma_i \xi_i + \sum_{k=1}^{p} \log \frac{\prod_{i \in I_k} \xi_i^{\xi_i}}{\left(\sum_{i \in I_k} \xi_i\right)^{\sum_{i \in I_k} \xi_i}} \tag{1.3}$$

függvény a

$$\sum_{k=0}^{p} \sum_{i \in I_k} \alpha_{ij} \xi_i = \beta_j \quad (j = 1, 2, \ldots, n)$$

(1.4)

$$\xi_i \geqq 0 \quad \left(i \in \bigcup_{k=0}^{p} I_k\right)$$

feltételek mellett.

KLAFSZKY EMIL javaslatára — akinek egy algoritmus kiterjesztéséhez kívántam segítséget nyújtani — kezdtem foglalkozni a duál geometriai programra vonatkozó egzisztencia-problémával: azzal a problémával, hogy a duál geometriai programok mely osztályaira igaz az az állítás, hogy a célfüggvény felveszi a minimumát a feltételi halmazon[1]. Bár jól kezelhető kritériumot eddig még nem sikerült kidolgoznom, mégis úgy gondoltam, hogy a kapott eredményeket érdemes egy dolgozat formájában összegyűjteni.

2. Jelölések és alapfeltevések

c_k-val, illetve x_k-val fogjuk jelölni az I_k halmaz elemeivel indexezett γ_i-k illetve ξ_i-k által alkotott vektort, φ betűvel a vizsgált függvényt, φ_k-val pedig a benne szereplő nemlineáris tagokat. Egymás mellé írt két vektor ezek skalár szorzatát jelenti. (Nem teszünk különbséget sor- és oszlopvektor között.) Az eddigiek szerint tehát

$$\varphi(x) = \varphi(x_0, x_1, \ldots, x_p) = \sum_{k=0}^{p} c_k x_k + \sum_{k=1}^{p} \varphi_k(x_k). \tag{2.1}$$

Jelöljük m_k-val az I_k indexhalmaz legnagyobb elemét:

$$I_0 = \{1, 2, \ldots, m_0\},$$
$$I_k = \{m_{k-1}+1, m_{k-1}+2, \ldots, m_k\} \quad (k = 1, 2, \ldots, p). \tag{2.2}$$

Ekkor a $\varphi(x)$ függvényt írhatjuk úgy is, hogy

$$\varphi(x) = \varphi(\xi_1, \xi_2, \ldots, \xi_{m_p}) = \sum_{j=1}^{m_p} \gamma_j \xi_j + \sum_{k=1}^{p} \log \frac{\prod_{j=m_{k-1}+1}^{m_k} \xi_j^{\xi_j}}{\left(\sum_{j=m_{k-1}+1}^{m_k} \xi_j\right)^{\sum_{j=m_{k-1}+1}^{m_k} \xi_j}}. \tag{2.3}$$

A $\varphi(x)$ függvény értelmezési tartományának az $R^{(m_p)}$ tér azon $x=(\xi_1, \xi_2, \ldots, \xi_{m_p})$ vektorainak a halmazát tekintjük, melyekre

$$\xi_j \in [0, +\infty), \quad \text{ha} \quad m_0 < j \leqq m_p \tag{2.4}$$

és

$$\xi_j \in (-\infty, +\infty), \quad \text{ha} \quad j \leqq m_0. \tag{2.5}$$

A (2.4)—(2.5) feltételekkel meghatározott halmazt D-vel fogjuk jelölni.

[1] A célfüggvény alulról való korlátossága nyilvánvalóan nem elég a minimum eléréséhez.

A továbbiakban az (1.4)—(1.5) feladatnál általánosabb alábbi feladatot tekintjük: Minimalizálandó a $\varphi(x)$ függvény az $x \in K$ feltétel mellett, ahol K konvex zárt halmaz. A K feltételi halmazról végig megköveteljük, hogy fennálljon

$$K \subset D, \tag{2.6}$$

és hogy legyen K-ban olyan x vektor, melyre teljesül[2]

$$\xi_j > 0, \quad \text{ha} \quad m_0 < j \leqq m_p. \tag{2.7}$$

A dolgozat hátralevő részében K mindig e tulajdonságokkal rendelkező konvex zárt halmazt[3] fog jelenteni, bár e követelmények előírását nem fogjuk még a tételek kimondásánál sem külön elismételni.

Az a kikötés, hogy létezzen (2.7)-et kielégítő vektor K-ban, nem szűkíti az általánosságot, csak a tárgyalás egyszerűbbé tételét szolgálja. A közlésre kerülő tételek és egyéb állítások elvben egyszerűen — de jelöléstechnikailag bonyolultan — átvihetők az általánosabb esetre, amikor a szóban forgó kikötéssel nem élünk.[4]

A K halmaz recessziós kúpjának nevezzük és K^2-vel jelöljük a

$$\{y \mid y \neq 0, \text{ tetszőleges } \lambda \geqq 0, \ x \in K \text{ esetén } x + \lambda y \in K\} \cup \{0\} \tag{2.8}$$

konvex zárt kúpot. Ha K poliéder, akkor *Motzkin tétele* értelmében K előállítható

$$K = K^1 + K^2 = \{y + z \mid y \in K^1, \ z \in K^2\} \tag{2.9}$$

formában, ahol K^1 egy konvex politóp.

3. A $\varphi(x)$ függvény alaptulajdonságai

A geometriai programozás irodalmából ismeretes, hogy a $\varphi_k(x_k)$ függvények rendelkeznek az alábbi tulajdonságokkal:

a) $\varphi_k(x_k) \leqq 0$ minden $x_k \geqq 0$ esetén.

b) $\varphi_k(x_k) = 0$ akkor és csak akkor teljesül, ha a $\xi_i (i \in I_k)$ változók közül legfeljebb egy vesz fel zérótól különböző értéket.

c) $\varphi_k(\lambda x_k) = \lambda \varphi_k(x_k)$, ha $\lambda \geqq 0$ és $x_k \geqq 0$.

d) $\varphi_k(x_k + s_k) \leqq \varphi_k(x_k) + \varphi_k(s_k)$, ha $x_k \geqq 0$
és $s_k \geqq 0$. Egyenlőség itt akkor és csak akkor teljesül, ha

$$\xi_j \sum_{i \in I_k} \sigma_i = \sigma_j \sum_{i \in I_k} \xi_i \tag{3.1}$$

fennáll minden $j \in I_k$ esetén, ahol σ_j az s_k vektor valamely komponensét jelenti.

[2] A geometriai programozás irodalmában az ehhez hasonló feltevést kanonikussági feltételnek nevezik. A kanonikussághoz általában azt szokták megkövetelni, hogy K csak nemnegatív vektorokat tartalmazzon, és létezzen minden koordinátájában pozitív vektor K-ban.

[3] Az eddigi kikötésekből már következik, hogy K nem lehet üres halmaz.

[4] Azzal a kérdéssel, hogy egyenlőségek és egyenlőtlenségek formájában megadott K konvex zárt halmazhoz miként találhatjuk meg azokat a j indexeket, amelyekhez tartozó ξ_j komponensek csak 0 értéket tudnak felvenni minden $x \in K$ esetén, itt nem kívánunk foglalkozni. Ez ugyan egyáltalán nem elhanyagolható kérdés, azonban e dolgozatban a továbbiakban a K halmazt soha nem egyenlőtlenségrendszerrel fogjuk definiálni.

A φ_k függvényekre felsorolt tulajdonságok közül a c) és d) tulajdonságokat a $\varphi(x)$ függvény az alábbi módon örökli:

$$\varphi(\lambda x) = \lambda\varphi(x), \quad \text{ha} \quad \lambda \geqq 0 \quad \text{és} \quad x \in D. \tag{3.2}$$

$$\varphi(x+s) \leqq \varphi(x)+\varphi(s), \quad \text{ha} \quad x \in D \quad \text{és} \quad s \in D. \tag{3.3}$$

Egyenlőség (3.3)-ban akkor és csak akkor teljesül, ha (3.1) fennáll minden $k \in \{1, 2, \ldots, p\}$ és $j \in I_k$ esetén.

Egy további tulajdonságként a $\varphi(x)$ függvény aszimptotikus viselkedését írjuk fel egyenesek mentén. Az elemi analízis eszközeivel bizonyítható, hogy tetszőleges $x \in D$, $s \in D$, $s \neq 0$ vektorok esetén

$$\lim_{\tau \to +\infty} \varphi(x+\tau s) = \begin{cases} -\infty, \text{ ha } \varphi(s) < 0, \\ +\infty, \text{ ha } \varphi(s) > 0, \\ -\infty, \text{ ha } \varphi(s) = 0 \text{ és található olyan } k, i \text{ indexpár,} \\ \qquad \text{melyekre } k \in \{1, 2, \ldots, p\},\ i \in I_k,\ s_k \neq 0,\ \sigma_i = 0 \\ \qquad \text{és } \xi_i \neq 0, \\ \sum_{k=0}^{p} c_k x_k + \sum_0 \varphi_k(x_k) + \sum_1 x_k \operatorname{grad} \varphi_k(s_k), \\ \qquad \text{ha } \varphi(s) = 0 \text{ és tetszőleges } k \in \{1, 2, \ldots, p\}, \\ \qquad i \in I_k \text{ indexekre } s_k \neq 0,\ \sigma_i = 0 \text{ esetén } \xi_i = 0. \end{cases} \tag{3.4}$$

$\sum_0$-nál, illetve $\sum_1$-nél itt azokra a $k \in \{1, 2, \ldots, p\}$ indexekre összegzünk, melyekre $s_k = 0$, illetve $s_k \neq 0$.

A (3.2)—(3.4) formulák alapján megállapíthatók a $\varphi(x)$ függvény következő tulajdonságai: A $\varphi(x+\tau s)$ kifejezés értéke akkor és csak akkor nem függ τ-tól, ha $\varphi(s) = 0$ és (3.1) teljesül minden $k \in \{1, 2, \ldots, p\}$ és $j \in I_k$ esetén. Ha pedig $\varphi(s) = 0$, de (3.1) nem áll fenn minden ilyen k, j indexpárra, akkor $\varphi(x+\tau s)$ szigorúan monoton csökkenő függvénye τ-nak.

4. Egy szükséges feltétel a $\varphi(x)$ függvény alulról való korlátosságára

4.1. TÉTEL. Ha a $\varphi(x)$ függvény alulról korlátos a K halmazon, akkor fennállnak a következők:

a) $\varphi(s) \geqq 0$ minden $s \in K^2$ esetén.

b) Ha s tetszőleges olyan vektor, melyre $s \in K^2$ és $\varphi(s) = 0$, akkor minden $k \in \{1, 2, \ldots, p\}$ esetén $s_k = 0$ vagy $s_k \gg 0$.

Bizonyítás. Legyen x a K halmaznak egy olyan pontja, melyre $x_k \gg 0$ teljesül $k \in \{1, 2, \ldots, p\}$ esetén. A 2. szakaszban tett általános feltevés szerint ilyen x pont létezik. A recessziós kúp definíciója szerint (illetve $s = 0$ esetén triviálisan) teljesül, hogy tetszőleges $s \in K^2$ és tetszőleges $\tau \geqq 0$ esetén $x + \tau s \in K$. Ennélfogva a $\varphi(x)$ függvénynek a K halmazon alulról való korlátosságából következik, hogy

$$\lim_{\tau \to +\infty} \varphi(x+\tau s) > -\infty \tag{4.1}$$

tetszőleges $s \in K^2$ esetén. A $\varphi(x)$ függvény aszimptotikus viselkedését leíró (3.4) formula szerint ez csak úgy lehetséges, ha $\varphi(s) \geqq 0$ és ha $\varphi(s) = 0$ esetén tetszőleges $k \in \{1, 2, \ldots, p\}$, $i \in I_k$ indexekre $s_k \neq 0$-ból és $\sigma_i = 0$-ból következik, hogy $\xi_i = 0$. Ez a követelmény pedig $\xi_i \neq 0$ miatt az a) és b) feltételek formájában is megfogalmazható.

5. Egy elégséges feltétel a $\varphi(x)$ függvény minimumának elérésére

Az elégséges feltételt általánosabban mondjuk ki a következő állításban.

5.1. LEMMA. Ha K nem üres konvex zárt halmaz, $f(x)$ folytonos homogén konvex függvény[5] egy K-t tartalmazó konvex zárt kúpon[6] és $f(s)>0$ teljesül minden $s\in K^2-\{0\}$ esetén, akkor az $f(x)$ függvény felveszi a minimumát a K halmazon.

A bizonyítást a soron következő 5.2—5.5 lemmák felhasználásával fogjuk megadni. Ezek, egy kivétellel a [11] könyvben bizonyított tételek között megtalálhatók.

5.2. LEMMA. (l. [11]-ben a 8.3.3. következményt). Ha K és L konvex zárt halmazok, K^2 és L^2 ezek recessziós kúpja és $K\cap L$ nem üres, akkor $K\cap L$ recessziós kúpja azonos a $K^2\cap L^2$ halmazzal.

5.3. LEMMA. (l. [11]-ben a 8.4. tételt). Egy K nem üres konvex zárt halmaz akkor és csak akkor korlátos, ha K^2 egyedül a 0 vektort tartalmazza.

5.4. LEMMA. (l. [11]-ben a 9.6. tételt). Ha K nem üres konvex zárt halmaz, amely nem tartalmazza a 0 vektort, akkor

$$\text{cl cone } K = \text{cone } K\cup K^2. \tag{5.1}$$

5.5. LEMMA. Ha K konvex zárt halmaz, $f(x)$ homogén konvex függvény a K halmazon és $f(y)>0$ teljesül minden $y\in\text{cl cone } K-\{0\}$ vektor esetén, akkor az $f(x)$ függvény felveszi a minimumát a K halmazon.

Bizonyítás. A feltételekből nyilvánvalóan következik, hogy $f(x)$ alulról korlátos a K halmazon. Tegyük fel indirekten, hogy $f(x)$ nem veszi fel a minimumát K-n. Ekkor van olyan $x^{(i)}\in K-\{0\}$ vektorsorozat, melyre $\lambda^{(i)}=|x^{(i)}|\to+\infty$ és az $f(x^{(i)})$ sorozat konvergens. Jelentse $y^{(i)}$ azt az egységvektort, melyre

$$x^{(i)} = \lambda^{(i)} y^{(i)}. \tag{5.2}$$

Ekkor

$$y^{(i)}\in\text{cone } K, \tag{5.3}$$

a $\lambda^{(i)}f(y^{(i)})$ sorozat pedig konvergens. $\lambda^{(i)}\to+\infty$ miatt $f(y^{(i)})\to 0$, az $y^{(i)}$ sorozat valamely $y^{(0)}$ torlódási pontjára tehát $|y^{(0)}|=1, f(y^{(0)})=0$ és $y^{(0)}\in$cl cone K, ellentétben a feltételekkel.

Az 5.1. lemma bizonyítása:

Legyen C egy K-t tartalmazó olyan konvex zárt kúp, melyen az $f(x)$ függvény folytonos és homogén konvex. Legyen

$$L = \{x | x\in C,\ f(x)\leqq 0\},$$

ekkor nyilvánvalóan L is konvex zárt kúp.

[5] Az $f(x)$ függvény homogén konvex a H halmazon, ha teljesül a) $f(\lambda x)=\lambda f(x)$ minden $\lambda\geqq 0$ és minden $x\in H$ esetén, és b) $f(x+y)\leqq f(x)+f(y)$ minden $x, y\in H$ esetén.

[6] A $\varphi(x)$ függvény esetén a D halmaz ilyen, mivel ez minden szóba jövő K-t tartalmaz.

Tekintsük először azt az esetet, amikor létezik olyan $x \in K$ vektor, melyre $f(x) \leqq \leqq 0$. Ugyanezt a kikötést úgy is mondhatjuk, hogy a $K \cap L$ halmaz nem üres. Mivel $f(s) > 0$ teljesül minden $s \in K^2 - \{0\}$ esetén, ezért a $K^2 \cap L$ halmaz csak a 0 vektort tartalmazza. Az 5.2. lemma szerint a $K^2 \cap L$ halmaz a $K \cap L$ halmaznak a recessziós kúpja. (L nem üres konvex zárt kúp, ezért nyilvánvalóan $L^2 = L$.) Most az 5.3. lemma alkalmazásával adódik, hogy a $K \cap L$ halmaz korlátos. A folytonos $f(x)$ függvény felveszi a minimumát a $K \cap L$ nem üres korlátos konvex zárt halmaz valamely $x^{(0)}$ pontjában, s ez az $x^{(0)}$ nyilvánvalóan minimumhely az egész K halmazra nézve is.

A még le nem tárgyalt esetben $f(x) > 0$ minden $x \in K$ esetén, következésképpen $f(y) > 0$ minden $y \in \operatorname{cone} K$ esetén. Az 5.1. lemma feltevését is figyelembe véve azt kapjuk, hogy $f(y) > 0$ teljesül minden $y \in \operatorname{cone} K \cup K^2 - \{0\}$ esetén is. A homogén konvex $f(x)$ függvényre $f(0) = 0$, így K nem tartalmazza a 0 vektort, tehát az 5.4. lemma alkalmazható. Ennélfogva az előző állítást úgy is mondhatjuk, hogy $f(y) > 0$ teljesül minden $y \in \operatorname{cl} \operatorname{cone} K - \{0\}$ esetén, azaz teljesülnek az 5.5. lemma feltételei, és e lemma értelmében $f(x)$ felveszi K-n a minimumát.

6. A $\varphi(x)$ függvény nemlineáris részének a lerövidítése

Ebben a szakaszban a $\varphi(x)$ függvényből egy másik ugyanolyan típusú függvényt származtatunk, amely általában kevesebb nemlineáris tagot tartalmaz, mint az eredeti függvény.

6.1. TÉTEL. Tegyük fel, hogy létezik olyan $s \in K^2 - \{0\}$ vektor, melyre $\varphi(s) = 0$ és tetszőleges $k \in \{1, 2, \ldots, p\}$ esetén vagy $s_k = 0$, vagy pedig $s_k \gg 0$. Ekkor a következőket állítjuk:

a) $\varphi(x)$ akkor és csak akkor alulról korlátos a K halmazon, ha a

$$(6.1) \qquad \varphi^{(1)}(x) = \lim_{\tau \to +\infty} \varphi(x + \tau s) = \sum_{k=0}^{p} c_k x_k + \sum_0 \varphi_k(x_k) + \sum_1 x_k \operatorname{grad} \varphi_k(s_k)$$

függvény alulról korlátos K-n, ahol $\sum_0$-ban az $s_k = 0$ kikötés mellett, $\sum_1$-ben az $s_k \neq 0$ kikötés mellett összegzünk, továbbá fennáll

$$(6.2) \qquad \inf \{\varphi(x) | x \in K\} = \inf \{\varphi^{(1)}(x) | x \in K\}.$$

b) $\varphi(x)$ akkor és csak akkor veszi fel K-n a minimumát, ha egyrészt $\varphi^{(1)}(x)$ felveszi K-n a minimumát, és másrészt $\varphi^{(1)}(x)$ minimumhelyei közül legalább egy $\tilde{x}$ vektorra teljesül, hogy a $\varphi(\tilde{x} + \tau s)$ kifejezés értéke nem függ τ-tól.

Bizonyítás. A feltevések szerint a (3.4) egyenlőség a legutolsó jobb oldallal érvényes minden $x \in D$ esetén, s ekkor a 3. szakasz végén tett megjegyzés szerint tetszőleges $x \in D$ esetén $\varphi(x + \tau s)$ értéke vagy nem függ τ-tól, vagy pedig τ-nak szigorúan monoton csökkenő függvénye, tehát mindenképpen $\varphi^{(1)}(x) \leqq \varphi(x)$. Ennélfogva

$$(6.3) \qquad \inf \{\varphi^{(1)}(x) | x \in K\} = \inf \{\varphi(x) | x \in K\}.$$

A recessziós kúp definíciója szerint $x\in K$-ból következik, hogy $x+\tau s\in K$, s így

$$\inf\{\varphi(y)|y\in K\} \leqq \lim_{\tau\to+\infty} \varphi(x+\tau s) \tag{6.4}$$

fennáll minden $x\in K$ esetén, tehát (6.3) fordított irányú egyenlőtlenséggel is érvényes.

A b) állítás egyszerűen adódik a)-ból és abból a tényből, hogy $\varphi(x+\tau s)$ monoton nemnövekvő függvénye τ-nak minden $x\in D$ esetén.

6.2. TÉTEL. Tegyük fel, hogy létezik olyan $s\in K^2-\{0\}$ vektor, melyre $\varphi(s)=0$ és $s_k\gg 0$ valamennyi $k\in\{1, 2, \ldots, p\}$ esetén. Ebben az esetben ahhoz, hogy $\varphi(x)$ alulról korlátos legyen K-n szükséges, poliedrikus K halmaz esetén pedig elégséges is, hogy a

$$\varphi^{(1)}(x) = \sum_{k=0}^{p} c_k x_k + \sum_{k=1}^{p} x_k \operatorname{grad} \varphi_k(s_k) = x \operatorname{grad} \varphi(s) \tag{6.5}$$

függvényre teljesüljön $\varphi^{(1)}(x)\geqq 0$ minden $x\in K^2$ vektor esetén.

Bizonyítás. A 6.2. tételben szereplő $\varphi^{(1)}(x)$ függvény, amely nem más mint a 6.1. tétel szerint is származtatott $\varphi^{(1)}(x)$ függvény, most x-nek lineáris függvénye. A lineáris programozás egyik ismert tétele szerint egy lineáris $f(x)$ függvény akkor és csak akkor alulról korlátos egy K poliéderen, ha $f(x)\geqq 0$ a K^2 recessziós kúpon. Ebből és a 6.1. tételből egyszerűen adódik a 6.2. tételnek a poliedrikus K halmazra vonatkozó állítása. Az általánosabb esetre kimondott szükségesség a 4.1 és 6.1. tételekből adódik.

7. A $\varphi(x)$ függvény lineáris részének összehúzása

Ebben a szakaszban meg fogjuk mutatni, hogy ha a $\varphi(x)$ függvényre és a K halmazra a 4—6. szakaszokban szereplő tételek nem alkalmazhatók, akkor a $\varphi(x)$ függvényt egy olyan, hasonló típusú

$$\psi(\xi_0, \xi_{m_0+1}, \xi_{m_0+2}, \ldots, \xi_{m_p}) \tag{7.1}$$

függvényre tudjuk redukálni, amely $\varphi(x)$-nél kevesebb változót tartalmaz. Most néhány felhasználásra kerülő lemmát sorolunk fel.

7.1. LEMMA. (l. [11]-ben a 8.1. tételt). Ha K nem üres konvex halmaz, K^2 pedig ennek recessziós kúpja, akkor

$$K^2 = \{y|x+y\in K \text{ minden } x\in K \text{ esetén}\} \tag{7.2}$$

7.2. LEMMA. Legyen $M: R^{(m)}\to R^{(n)}$ egy lineáris leképezés, amely értelmezve van valamely $D\subset R^{(m)}$ halmazon, és az $x\in D$ vektorhoz az $\bar{x}=Mx\in R^{(n)}$ vektort rendeli. Legyen K egy nem üres konvex zárt halmaz, melyre $K\subset D$.

Jelöljük K^2-vel a K halmaz recessziós kúpját, $M(K)^2$-vel pedig az

$$M(K) = \{Mx|x\in K\} \tag{7.3}$$

halmaz recessziós kúpját, és legyen

$$M(K^2) = \{My|y\in K^2\}. \tag{7.4}$$

Azt állítjuk, hogy

$$M(K)^2 \supset M(K^2), \tag{7.5}$$

poliedrikus K halmaz esetén pedig

$$M(K)^2 = M(K^2). \tag{7.6}$$

Bizonyítás. A lineáris leképezések elméletéből ismeretes, hogy konvex zárt halmaz képe is konvex zárt halmaz, konvex zárt kúp képe is konvex zárt kúp, poliedrikus konvex zárt halmaz képe is poliedrikus konvex zárt halmaz.

Tegyük fel, hogy $\bar{y} \in M(K^2)$ és $\bar{x} \in M(K)$. Ekkor léteznek olyan $y \in K^2$ és $x \in K$ vektorok, melyekre $\bar{y} = My$ és $\bar{x} = Mx$. A 7.1. lemma szerint $x + y \in K$, tehát $\bar{x} + \bar{y} \in M(K)$. Most a képhalmazokra alkalmazva a 7.1. lemmát, azt kapjuk, hogy $\bar{y} \in M(K)^2$, tehát a (7.5) tartalmazás valóban fennáll.

Ha K poliedrikus konvex zárt halmaz, akkor a *Motzkin-féle felbontási tétel* szerint létezik olyan K^1 konvex politóp, melyre

$$K = K^1 + K^2, \tag{7.7}$$

és ennélfogva

$$M(K) = M(K^1) + M(K^2). \tag{7.8}$$

Mivel az $M(K^1)$ halmaz korlátos, ezért az $M(K)$ halmaz recessziós kúpja azonos az $M(K^2)$ halmaz recessziós kúpjával, az utóbbi pedig magával az $M(K^2)$ kúppal. Éppen ezt fejezi ki a (7.6) egyenlőség.

7.3. TÉTEL. Tekintsük azt az

$$M: R^{(m_p)} \to R^{(m_p - m_0 + 1)} \tag{7.9}$$

lineáris leképezést, amely az

$$x = (\xi_1, \xi_2, \ldots, \xi_{m_p}) \in R^{(m_p)} \tag{7.10}$$

vektorhoz az

$$\bar{x} = (\bar{\xi}_0, \bar{\xi}_{m_0+1}, \bar{\xi}_{m_0+2}, \ldots, \bar{\xi}_{m_p}) = \left(\sum_{j=1}^{m_0} \gamma_j \xi_j, \xi_{m_0+1}, \xi_{m_0+2}, \ldots, \xi_{m_p} \right) \tag{7.11}$$

vektort rendeli. Jelöljük $M(K)^2$-vel az $M(K)$ halmaz recessziós kúpját és legyen

$$\psi(\bar{x}) = \psi(\bar{\xi}_0, \bar{\xi}_{m_0+1}, \bar{\xi}_{m_0+2}, \ldots, \bar{\xi}_{m_p}) = \bar{\xi}_0 + \sum_{k=1}^{p} c_k \bar{x}_k + \sum_{k=1}^{p} \varphi_k(\bar{x}_k). \tag{7.12}$$

Ekkor a következőket állítjuk:

a) $\varphi(x)$ akkor és csak akkor alulról korlátos K-n, ha $\psi(\bar{x})$ alulról korlátos $M(K)$-n, és ebben az esetben

$$\inf \{\varphi(x) | x \in K\} = \inf \{\psi\{\bar{x}\} | \bar{x} \in M(K)\}. \tag{7.13}$$

b) $\varphi(x)$ akkor és csak akkor veszi fel a minimumát K-n, ha $\psi(\bar{x})$ felveszi a minimumát $M(K)$-n.

c) Létezik olyan $\bar{x}\in M(K)$ vektor, melyre $\bar{\xi}_j>0$ teljesül $j=m_0+1, m_0+2, \ldots, m_p$ esetén.

d) Ha az $\{\bar{s}|\bar{s}\in M(D), \psi(\bar{s})=0\}$ halmazban van zérótól különböző vektor, akkor minden ilyen $\bar{s}=(\bar{\sigma}_0, \bar{\sigma}_{m_0+1}, \bar{\sigma}_{m_0+2}, \ldots, \bar{\sigma}_{m_p})$ vektorra $\bar{\sigma}_j\neq 0$ teljesül legalább egy $j\in\{m_0+1, m_0+2, \ldots, m_p\}$ index esetén.

Bizonyítás. A $\psi(\bar{x})$ függvény értelmezése szerint fennáll $\varphi(x)=\psi(Mx)$ minden $x\in D$ esetén, tehát a $\varphi(x)$ függvény értékkészlete a K halmazon azonos a $\psi(\bar{x})$ függvény értékkészletével az $M(K)$ halmazon, ebből pedig a 7.3. tétel a) és b) állítása nyilvánvalóan adódik.

A c) állítás igazsága következik abból, hogy hasonló állítást feltételeztünk a K halmazra vonatkozóan.

A d) állítás igazsága indirekt okoskodással látható be. Ha valamely $\bar{s}\in M(D)$ vektorra $\bar{\sigma}_j=0$ minden $j\in\{m_0+1, m_0+2, \ldots, m_p\}$ esetén, akkor $\psi(\bar{s})=\bar{\sigma}_0$, és ennélfogva vagy $\bar{s}=0$, vagy pedig $\psi(\bar{s})\neq 0$.

7.4. TÉTEL. Ha K poliéder, a $\varphi(x)$ függvényre pedig teljesül $\varphi(s)>0$ minden $s\in K^2-Z$ esetén, ahol

$$Z=\{x|x\in D,\ c_0x_0=0,\ x_k=0,\ \text{ha } k\in\{1, 2, \ldots, p\}\} \tag{7.14}$$

akkor a 7.3. tételben definiált $\psi(\bar{x})$ függvényre $\psi(\bar{s})>0$ teljesül minden $\bar{s}\in M(K)^2-\{0\}$ esetén.

Bizonyítás. $s\in Z$ esetén $\varphi(s)=0$, és így $\varphi(s)\geqq 0$ teljesül minden $s\in K^2$ esetén. Ebből következően $\psi(\bar{s})\geqq 0$ minden $\bar{s}\in M(K^2)=M(K)^2$ esetén. (Ehhez tudnunk kell még a 7.2. lemma állítását.) Ha valamely $\bar{s}\in M(K^2)$ esetén $\psi(\bar{s})=0$, akkor a 7.3. tétel d) állítása szerint vagy $\bar{s}=0$, vagy pedig $\bar{s}_k\neq 0$ teljesül legalább egy $k\in\{1, 2, \ldots, p\}$ index esetén.

Tegyük most fel indirekten, hogy valamely $\bar{s}\in M(K)^2-\{0\}$ vektor esetén $\psi(\bar{s})=0$. Tekintsünk egy olyan $s\in K^2$ vektort, melyre $M(s)=\bar{s}$. Erre a vektorra $\varphi(s)=\psi(\bar{s})=0$, ennélfogva a tétel feltevése szerint $s\in Z$-nek kell teljesülnie. A Z halmaz definíciója szerint viszont $s_k=\bar{s}_k=0$ fennáll $k=1, 2, \ldots, p$ esetén. Így azonban ellentmondásba kerültünk a 7.3. tétel d) állításával.

7.5. TÉTEL. Ha K poliéder, a $\varphi(x)$ függvényre pedig $\varphi(s)>0$ teljesül minden $s\in K^2-Z$ esetén, ahol Z a (7.14) alatt definiált halmaz, akkor a $\varphi(x)$ függvény felveszi K-n a minimumát.

Bizonyítás. A 7.4. tétel szerint a 7.3. tételben definiált $\psi(\bar{x})$ függvény és az $M(K)$ halmaz teljesíti az 5.1. lemma feltételeit. Az 5.1. lemma alkalmazásával adódik tehát, hogy $\psi(\bar{x})$ felveszi a minimumát az $M(K)$ halmazon. A ψ függvény és az M leképezés értelmezése alapján nyilvánvaló, hogy ekkor a $\varphi(x)$ függvény is felveszi a minimumát a K halmazon.

A továbbiakban megmutatjuk, hogy a 4—7. szakaszokban bizonyított tételek együttvéve minden lehetséges esetben közölnek valamilyen használható információt a $\varphi(x)$ függvénynek a K halmazon való minimalizálásával kapcsolatosan. A tételek alkalmazhatóságának a terjedelmét a levonható konklúzióval együtt az 1. és 2. táblázatban foglaljuk össze. A K halmazról továbbra is feltesszük a következőket:

1. TÁBLÁZAT

Eset	Feltevés	Konklúzió	Igazolás
I.	$\varphi(s)<0$ valamely $s\in K^2$ esetén vagy $\varphi(s)=0$ valamely $s\in K^2$ esetén és legalább egy $k\in\{1,2,\ldots,p\}$ indexre s_k komponensei között előfordul zéró és pozitív értékű is.	$\varphi(x)$ alulról nem korlátos a K halmazon	4.1. tétel
II.	$\varphi(s)>0$ minden $s\in K^2-\{0\}$ esetén.	$\varphi(x)$ felveszi a minimumát a K halmazon	5.1. lemma
III.	Valamely $s\in K^2-\{0\}$ vektor esetén teljesülnek a következők: *a)* $\varphi(s)=0$, *b)* minden $k\in\{1,2,\ldots,p\}$ indexre vagy $s_k\gg 0$, vagy pedig $s_k=0$, *c)* legalább egy $k\in\{1,2,\ldots,p\}$ indexre $s_k\gg 0$.	$\varphi(x)$ nemlineáris része rövidíthető	6.1. tétel
IV.	$m_0\geqq 2$	$\varphi(x)$ lineáris része összehúzható	7.3. tétel

2. TÁBLÁZAT

Eset	Feltevés	Konklúzió	Igazolás
I.	$\varphi(s)<0$ valamely $s\in K^2$ esetén vagy $\varphi(s)=0$ valamely $s\in K^2$ esetén és legalább egy $k\in\{1,2,\ldots,p\}$ indexre s_k komponensei között előfordul zéró és pozitív értékű is.	$\varphi(x)$ alulról nem korlátos a K poliéderen	4.1. tétel
II.	$\varphi(s)>0$ minden $s\in K^2-Z$ esetén, ahol $Z=\{x\mid x\in D,\ c_0x_0=0,\ x_k=0,$ ha $k\in\{1,2,\ldots,p\}\}$	$\varphi(x)$ felveszi a minimumát a K poliéderen	7.5. tétel
III.	Valamely $s\in K^2-\{0\}$ vektor esetén teljesülnek a következők: *a)* $\varphi(s)=0$, *b)* minden $k\in\{1,2,\ldots,p\}$ indexre $s_k\gg 0$, vagy $s_k=0$, *c)* legalább egy $k\in\{1,2,\ldots,p\}$ indexre $s_k\gg 0$.	$\varphi(x)$ nemlineáris része rövidíthető	6.1. tétel
IV.	Valamely $s\in K^2-\{0\}$ vektor esetén teljesülnek a következők: *a)* $\varphi(s)=0$, *b)* minden $k\in\{1,2,\ldots,p\}$ indexre $s_k\gg 0$.	$\varphi(x)$ lineáris függvényre redukálható	6.2. tétel

a) K konvex zárt halmaz, b) tetszőleges $x\in K$ vektorra teljesül $\xi_j\geqq 0$ minden $j\in\{m_0+1, m_0+2, \ldots, m_p\}$ esetén, végül c) K-ban található olyan x vektor, melyre teljesül $\xi_j>0$ minden $j\in\{m_0+1, m_0+2, \ldots, m_p\}$ esetén. E feltevések mellett az 1. táblázatnak megfelelően osztályozhatjuk a lehetséges eseteket. Ha viszont még azt is feltesszük, hogy a K halmaz poliéder, akkor az esetek osztályozását a 2. táblázat mutatja.

A két táblázat eseteihez tartozó feltevéseket lehetőség szerint úgy fogalmaztuk meg, hogy könnyen felismerhető legyen a kapcsolat az esetek és a tételek között. A felsorolt esetek nem zárják ki egymást, például a 2. táblázatban szereplő IV. eset lényegében a III. eset részének tekinthető[7], viszont az 1. táblázatba foglalt I—IV. esetek minden lehetőséget kimerítenek, és ugyanez mondható a 2. táblázatban foglalt I—III. esetekre is. A 2. táblázatra vonatkozóan ez az állítás elemi logikai műveletekkel igazolható.

Az 1. táblázatra vonatkozó állítás bizonyítása céljából nézzük meg, mit mondhatunk, ha az I—III. esetek lehetőségét kizárjuk. Ekkor teljesülniük kell a következőknek:

a) $\varphi(s) \geqq 0$ minden $s \in K^2 - \{0\}$ esetén.
b) Ha $s \in K^2 - \{0\}$ és $\varphi(s) = 0$, akkor $s_k = 0$ minden $k \in \{1, 2, \ldots, p\}$ esetén.
c) Létezik olyan $s \in K^2 - \{0\}$ vektor, melyre $\varphi(s) = 0$.

Ha $m_0 \leqq 1$, akkor b) és c) nem állhat fenn egyidejűleg, $m_0 = 0$ esetén ugyanis közöttük logikai ellentmondás jön létre, $m_0 = 1$ esetén pedig — minthogy ekkor a φ és ψ függvények azonosak — ezért a 7.3. tétel d) állítása éppen azt állítja, hogy b) és c) nem állhat fenn egyidejűleg. Ezek szerint csak $m_0 \geqq 2$ lehetséges, azaz az I—III. esetek lehetőségének kizárása maga után vonja a IV. eset fennállását.

8. Egy szükséges és elégséges feltétel a $\varphi(x)$ függvény alulról való korlátosságára konvex poliéderen

Ebben a szakaszban arra az esetre korlátozódunk, amikor K poliéder.

Az előbb beláttuk, hogy a 2. táblázatban szereplő I—III. esetek valamelyike okvetlenül fennáll. Ha tehát nem található a III. esetnek megfelelő s vektor, akkor a II. esethez tartozás szükséges és elégséges ahhoz, hogy a $\varphi(x)$ függvény alulról korlátos legyen a K poliéderen. Ha ily módon nem dönthető el a korlátosság, azaz a III. eset áll fenn, akkor jelöljük $s^{(1)}$-gyel (majd az eljárás esetleges ismétlései során $s^{(2)}$-vel, $s^{(3)}$-mal stb.) a III. esethez tartozó feltevésben szereplő s vektort, S_1-gyel (az ismétlés során S_2-vel, S_3-mal stb.) azon k indexek halmazát, melyekre $s_k^{(1)} \gg 0$ (az ismétlés során $s_k^{(2)} \gg 0$, $s_k^{(3)} \gg 0$ stb.), $\varphi^{(1)}$-gyel (az ismétlés során $\varphi^{(2)}$, $\varphi^{(3)}$-mal stb.) a 6.1. tétel szerint a nemlineáris rész lerövidítésével származtatott függvényt. A származtatott függvénnyel megismételve az eljárást, véges sorszámú[8] — mondjuk a q-adik — lépésben már csak az I. vagy a II. eset állhat fenn. A vázolt gondolatmenetben elmondottakat a következő tételben fogalmazhatjuk meg:

8.1. TÉTEL. A $\varphi(x)$ függvény akkor és csak akkor alulról korlátos a K poliéderen, ha valamely $q \in \{0, 1, 2, \ldots, p\}$ esetén léteznek olyan $s^{(1)}, s^{(2)}, \ldots, s^{(q)} \in K^2$ vek-

[7] Ha teljesen precízek akarunk lenni, akkor a IV. eset kissé „kilóg" a III. esetből, amennyiben a IV. esetben $\varphi(x)$ lehet lineáris függvény is, a III. esetben viszont nem. Azt azonban mindenképpen kimondhatjuk, hogy a III. eset és a lineáris $\varphi(x)$ esete együttesen tartalmazza a IV. esetet.

[8] Az I_k indexhalmazokról az általánosság csorbítása nélkül kiköthettünk volna, hogy minden I_k tartalmaz legalább két elemet, és ekkor az 1. és 2. táblázat III. esetben mindig ténylegesen, s nem pedig esetleg csak formálisan rövidülne a $\varphi(x)$ függvény nemlineáris része. A ciklizálás lehetőségét azonban kizárja a formális rövidülés is, az I_k indexhalmazok minimális elemszámára tett megkötés viszont bonyodalmakat okozna általánosabb feltételek mellett, amikor is később eltekintünk attól a követelménytől, hogy legyen K-ban a (2.7) egyenlőtlenségnek eleget tevő x vektor.

torok és olyan $S_1, S_2, \ldots, S_q \subset \{1, 2, \ldots, p\}$ páronként diszjunkt, nem üres indexhalmazok, melyekre teljesülnek a következők:

$$\left.\begin{array}{ll} \text{a)} & s_k^{(i)} \gg 0, \quad \text{ha} \quad k \in S_i \\ & s_k^{(i)} = 0, \quad \text{ha} \quad k \in \tilde{S}_i = \{1, 2, \ldots, p\} - \bigcup_{r=1}^{i} S_r \end{array}\right\} \quad (i = 1, 2, \ldots, q) \tag{8.1}$$

$$\text{b)} \quad \varphi^{(i)}(x) = \sum_{k=0}^{p} c_k x_k + \sum_{r=1}^{i} \sum_{k \in S_r} x_k \operatorname{grad} \varphi_k(s_k^{(r)}) + \sum_{k \in \tilde{S}_i} \varphi_k(x_k) \tag{8.2}$$

jelöléssel[9] fennáll

$$\varphi^{(i)}(s^{(i+1)}) = 0, \quad (i = 0, 1, 2, \ldots, q-1). \tag{8.3}$$

$$\text{c)} \quad \varphi^{(q)}(s) > 0 \quad \text{minden} \quad s \in K^2 - Z_q \tag{8.4}$$

esetén, ahol

$$Z_q = \left\{x \mid x \in D,\ c_0 x_0 + \sum_{r=1}^{q} \sum_{k \in S_r} (c_k + \operatorname{grad} \varphi_k(s_k^{(r)})) x_k = 0,\ x_k = 0, \quad \text{ha} \quad k \in \tilde{S}_q\right\}. \tag{8.5}$$

A 8.1. tétel feltevéseinek a teljesülése esetén vezessük be a $\min \{\varphi(x) | x \in K\}$ feladatra a q-reguláris (q itt változó) elnevezést. E terminológiában 0-regularitás a 7.5. tétel feltevéseinek a teljesülését, vagyis a primálra a *Slater-feltétel* teljesülését jelenti. A 8.1. tételt most úgy is fogalmazhatjuk, hogy a $\varphi(x)$ függvény akkor és csak akkor korlátos alulról a K poliéderen, ha a $\min \{\varphi(x) | x \in K\}$ feladat q-reguláris valamilyen nemnegatív egész q esetén.

A 7.5. tétel szerint a 8.1. tétel feltevései azt is garantálják, hogy a $\varphi^{(q)}(x)$ függvény felveszi a minimumát a K poliéderen. Ennek a ténynek az ismeretében a 6.1. tétel felhasználásával — elsősorban annak b) állítása segítségével — igazolható a következő tételek állítása:

8.2. TÉTEL. A 8.1. tétel feltételeinek fennállása esetén $\varphi(x)$ akkor és csak akkor veszi fel a minimumát a K poliéderen, ha teljesülnek a következők:

a) $\varphi^{(q)}(x)$ felveszi a minimumát a

$$\tilde{K} = \left\{x \mid x \in K,\ \xi_j \sum_{i \in I_k} \sigma_i^{(r)} = \sigma_j^{(r)} \sum_{i \in I_k} \xi_j, \quad \text{ha} \quad r \in \{1, 2, \ldots, q\}, \quad k \in S_r \text{ és } j \in I_k\right\} \tag{8.6}$$

poliéderen,

b) fennáll a

$$\min \{\varphi^{(q)}(x) | x \in K\} = \min \{\varphi^{(q)}(x) | x \in \tilde{K}\} \tag{8.7}$$

egyenlőség.

8.3. TÉTEL. Ha fennállnak a 8.1. tétel feltételei és teljesül a (8.7) egyenlőség, akkor a

$$\min \{\varphi^{(q)}(x) | x \in \tilde{K}\} \tag{8.8}$$

feladat optimális megoldásainak a halmaza azonos a

$$\min \{\varphi(x) | x \in K\} \tag{8.9}$$

feladat optimális megoldásainak a halmazával.

[9] Azaz $\varphi^{(i)}(x) = \lim_{\tau_i \to +\infty} \lim_{\tau_{i-1} \to +\infty} \ldots \lim_{\tau_1 \to +\infty} \varphi\left(x + \sum_{r=1}^{i} \tau_r s^{(r)}\right)$.

9. Általánosítási lehetőségek

A 8. szakaszban szereplő tételekhez hasonló tételek megfogalmazhatók akkor is, ha nem korlátozódunk poliéder feltételi halmaz esetére. Ebben az esetben, a 8. szakasz elején vázolt algoritmust úgy módosítjuk, hogy minden $\varphi^{(i)}(x)$ függvényre alkalmazzuk a 7.3. tételben definiált transzformációt, és ezután az 1. táblázat I—III. eseteit vesszük figyelembe. A gondolatmenet azonban ekkor is kifejezhető a korábbitól kevésbé eltérő módon, csak a $\varphi^{(i)}$ függvények használatával, a 7.3. tételben definiált transzformáció explicit alkalmazása nélkül és megfogalmazható a 8.1. tételnek megfelelő következő tétel:

9.1. TÉTEL. A $\varphi(x)$ függvény akkor és csak akkor alulról korlátos a K halmazon, ha valamely $q \in \{0, 1, 2, \ldots, p\}$ esetén léteznek olyan $s^{(1)}, s^{(2)}, \ldots, s^{(q)} \in D$ vektorok és olyan $S_1, S_2, \ldots, S_q \subset \{1, 2, \ldots, p\}$ páronként diszjunkt, nem üres indexhalmazok, melyekre teljesülnek a következők:

$$\left.\begin{array}{l} \text{a) } s_k^{(i)} \gg 0, \quad \text{ha} \quad k \in S_i \\ \quad\; s_k^{(i)} = 0, \quad \text{ha} \quad k \in \tilde{S}_i = \{1, 2, \ldots, p\} - \bigcup_{r=1}^{i} S_r \end{array}\right\} \quad (i = 1, 2, \ldots, q)$$

b) Jelöljük M_i-vel azt a lineáris leképezést, amely az x vektorhoz azt a vektort rendeli, melynek komponensei:

$$\bar{\xi}_0 = c_0 x_0 + \sum_{r=1}^{i} \sum_{k \in S_r} x_k [c_k + \operatorname{grad} \varphi_k(s_k^{(r)})],$$

$$\bar{\xi}_j = \xi_j, \quad \text{ha} \quad j \in I_k \quad \text{és} \quad k \in \tilde{S}_i.$$

Legyen $\varphi^{(i)}(x)$ formailag ugyanaz, mint a (8.2) alatt deflniált függvény:

$$\varphi^{(i)}(x) = \sum_{k=0}^{p} c_k x_k + \sum_{r=1}^{i} \sum_{k \in S_r} x_k \operatorname{grad} \varphi_k(s_k^{(r)}) + \sum_{k \in \tilde{S}_i} \varphi_k(x_k).$$

E jelölésekkel álljon fenn

$$\left.\begin{array}{l} s^{(i)} \in M_{i-1}^{-1}[M_{i-1}(K)^2] \\ \varphi^{(i-1)}(s^{(i)}) = 0 \end{array}\right\} \quad (i = 1, 2, \ldots, q \quad \text{esetén}).$$

és

c) $$\varphi^{(q)}(s) > 0 \quad \text{minden} \quad s \in M_q^{-1}[M_q(K)^2 - \{0\}]$$

esetén.

A 8.2. és 8.3. tétel hasonló jellegű általánosításához elég ezek szövegében a „8.1. tétel" kifejezést a „9.1. tétel" kifejezéssel, a „poliéder" szót pedig a „halmaz" szóval helyettesíteni.

A dolgozatnak a 4. szakasztól kezdődő, és az előző mondattal végződő részére teljes egészében érvényes általánosítást, helyesebben egy kényelmi szempontból tett megkötés feloldását jelenti az alábbi módosítás:

Tekintsünk el a kanonikussági feltételtől, azaz attól a követelménytől, hogy létezzen K-ban a (2.7) egyenlőtlenségnek eleget tevő x vektor, és legyen $k = 1, 2, \ldots$

..., p-re $I_k^* = \{j | j \in I_k, \xi_j > 0$ legalább egy $x \in K$ vektorra$\}$. Ekkor a dolgozat fent meghatározott része teljes egészében érvényben marad, ha az $s_k \gg 0$ előírást minden előfordulása esetén a $\sigma_j > 0$ $(j \in I_k^*)$ előírással helyettesítjük, hasonlóan az $x_k \gg 0$ előírást minden előfordulása esetén a $\xi_j > 0$ $(j \in I_k^*)$ előírással helyettesítjük.

IRODALOM

[1] AVRIEL, M. AND WILLIAMS, A. C., "On the primal and dual constraint sets in geometric programming", *Journal of Mathematical Analysis and Applications* **32** (1970) 684—688.

[2] AVRIEL, M. AND WILLIAMS, A. C., "Complementary geometric programming", *SIAM Journal on Applied Mathematics* **19** (1970) 125—141.

[3] DUFFIN, R. J. AND PETERSON, E. L., "Duality theory for geometric programming", *SIAM Journal on Applied Mathematics* **14** (1966) 1307—1349.

[4] DUFFIN, R. J. AND PETERSON, E. L., "The proximity of (algebraic) geometric programming to linear programming", *Mathematical Programming* **3** (1972) 250—253.

[5] DUFFIN, R. J., PETERSON, E. L. AND ZENER, C., *Geometric Programming* (John Wiley, New York, 1966).

[6] GALE, D., *The Theory of Linear Economic Models* (McGraw-Hill, New York, Toronto and London, 1960).

[7] GOCHET, W. AND SMEERS, Y., "Constraint sets of geometric programs characterized by auxiliary problems", *SIAM Journal on Applied Mathematics* **29** (1975) 708—718.

[8] KLAFSZKY, E., Geometriai programozás (*Magyar Tudományos Akadémia Számítástechnikai Központja, Közlemények*, **8** szám, Budapest, 1972) 41—65.

[9] KLAFSZKY, E., "Geometriai programozás és néhány alkalmazása", kandidátusi értekezés. Magyar Tudományos Akadémia, Budapest, 1973.

[10] KLAFSZKY, E., *Geometric Programming* (Hungarian Committee for Systems Analysis, Seminary Notes, Mathematics, No. **11**., Budapest, 1976.)

[11] ROCKAFELLAR, R. T., *Convex Analysis* (Princeton University Press, Princeton, 1970).

(Beérkezett: 1978. október 9.)

KÉRI GERZSON
MTA SZÁMÍTÁSTECHNIKAI ÉS AUTOMATIZÁLÁSI KUTATÓ INTÉZET
1502 BUDAPEST, KENDE U. 13—17.

ON THE MINIMIZATION OF THE GENERAL ENTROPY FUNCTION ON A CONVEX SET

G. KÉRI

Some classes of dual geometrical programming problems are specified on the score of the asymptotic behaviour of the objective function along straight lines. The analysis and description of this asymptotic behaviour serves as a tool for the determination of the following:

Whether the objective function of a concrete geometrical programming problem

a) is bounded (from below) on the constraint set, or not;

b) assumes its minimum, or not?

Our results can be formulated in more general

a) for the non-canonical problem;

b) for the case of the minimization over a non-polyhedral convex set.

Alkalmazott Matematikai Lapok 4 (1978) 109—116

AZ OPTIMALITÁS MÁSODRENDŰ FELTÉTELEIRŐL

RAPCSÁK TAMÁS
Budapest

A dolgozatban egy másodrendű elegendő feltételt fogalmazunk meg, mely a nemlineáris programozási feladat optimumpontjában teljesül. Ezután az itt bevezetett feltétel differenciálgeometriai tartalmát vizsgáljuk. Megadjuk e feltétel differenciálgeometriai megfelelőjét, s ennek segítségével ellenőrzését egy olyan mátrix legnagyobb sajátértékének kiszámítására vezetjük vissza, amelyet a gradiens vektor és a *Hesse-mátrix* segítségével explicit alakban is előállítunk.

1. Bevezetés

A nemlineáris programozási feladatok elméleti kérdéseinek megoldásához nagy segítséget adnak az optimalitás másodrendű szükséges és elegendő feltételei. Mégis viszonylag kevéssé ismert e feltételek természete és tulajdonságai. Ezeket az optimalitási kritériumokat részletesen tárgyalja könyvében A. V. FIACCO és G. P. MCCORMICK [8], D. G. LUENBERGER [12] és M. AVRIEL [2].

A dolgozat első részében a definíciókat és a felhasználásra kerülő lemmákat ismertetjük. A következő részben az ismert tételeken kívül két tétel található. Az elsőben lokális kvázikonvexitási tulajdonság felhasználásával egy másodrendű szükséges, a másodikban pedig az előzőekben már említett másodrendű elegendő feltételt fogalmazzuk meg.

E feltételek vizsgálatához kvázikonvexitási kérdések vizsgálata kapcsolódott [22]. Ez utóbbi témakörben lényeges eredmények születtek nálunk is az elmúlt években, melyek közül többet felhasználtunk.

A dolgozat következő részében az itt bevezetett másodrendű elegendőségi feltétel differenciálgeometriai tartalmát vizsgáljuk. Megadjuk e feltételek differenciálgeometriai megfelelőjét, s ennek segítségével ellenőrzésüket egy mátrix legnagyobb sajátértékének kiszámítására vezetjük vissza.

Ez a számítás numerikusan elvégezhető, ugyanis a mátrix explicit alakja a gradiens vektor és a *Hesse mátrix* felhasználásával megadható.

2. Definíciók, előzetes lemmák

2.1. DEFINÍCIÓ. [18] Legyen $f(\mathbf{x})$ egy valós értékű függvény definiálva egy nyílt $\Gamma \subseteq R^n$ halmazon. Akkor mondjuk, hogy $f(\mathbf{x})$ lokálisan kvázikonvex egy $\mathbf{x}_0 \in \Gamma$ pontban (a Γ halmazra vonatkoztatva), ha $f(\mathbf{x})$ $\mathbf{x}_0$-ban differenciálható és

$$\left.\begin{array}{l}\mathbf{x} \in \Gamma \\ f(\mathbf{x}) \leqq f(\mathbf{x}_0)\end{array}\right\} \Rightarrow \nabla f(\mathbf{x}_0)(\mathbf{x}-\mathbf{x}_0) \leqq 0.$$

(A $\nabla f(\mathbf{x}_0)$ sorvektor $f(\mathbf{x})$ $\mathbf{x}_0$ pontbeli gradiensét jelöli.)

2.2. DEFINÍCIÓ. [14] Legyen $f(\mathbf{x})$ egy valós értékű függvény definiálva egy nyílt $\Gamma \subseteq R^n$ halmazon. Akkor mondjuk, hogy $f(\mathbf{x})$ pszeudokonvex egy $\mathbf{x}_0 \in \Gamma$ pontban (a Γ halmazra vonatkoztatva), ha $f(\mathbf{x})$ differenciálható az $\mathbf{x}_0$ pontban és

$$\left.\begin{array}{l} \mathbf{x} \in \Gamma \\ \nabla f(\mathbf{x}_0)(\mathbf{x}-\mathbf{x}_0) \geqq 0 \end{array}\right\} \Rightarrow f(\mathbf{x}) \geqq f(\mathbf{x}_0).$$

Ezek a definíciók megtalálhatók MARTOS B. [18] és O. L. MANGASARIAN [14] könyveiben. Ugyanott megtalálhatók azok az állítások is, amelyek a különböző kvázikonvexitási és pszeudokonvexitási definíciók közötti kapcsolatokat és eltéréseket mutatják.

2.3. LEMMA. Ha $f(\mathbf{x})$ folytonos a Γ halmazon, az $\mathbf{x}_0 \in \Gamma$ pontban differenciálható és lokálisan kvázikonvex, $\nabla f(\mathbf{x}_0) \neq 0$, akkor $f(\mathbf{x})$ az $\mathbf{x}_0$ pontban pszeudokonvex.

A 2.3. lemma bizonyításával együtt [22]-ben megtalálható. Hasonló állításokat találunk [5], [7]-ben.

2.4. LEMMA. Ha $f(\mathbf{x})$ kétszer folytonosan differenciálható a Γ halmazon és egy $\mathbf{x}_0 \in \Gamma$ pontban lokálisan kvázikonvex, $\nabla f(\mathbf{x}_0) \neq \mathbf{0}$, akkor

$$\mathbf{y}^T \nabla^2 f(\mathbf{x}_0)\mathbf{y} \geqq 0, \quad \text{ha} \quad \nabla f(\mathbf{x}_0)\mathbf{y} = 0.$$

(Az $f(\mathbf{x})$ függvény *Hesse-mátrixát* az $\mathbf{x}_0$ pontban a $\nabla^2 f(\mathbf{x}_0)$ szimbólum jelöli.)

A 2.4. lemma más formában [3]-ban és [9]-ben is megtalálható. [22]-ben a lemma egy egyszerű bizonyítását adtuk.

3. Az optimalitás másodrendű szükséges és elegendő feltételei

Tekintsük a következő feladatot:

$$\min f(\mathbf{x}) \tag{3.1}$$

$$g_i(\mathbf{x}) \geqq 0, \quad i = 1, \ldots, m,$$

ahol $f(\mathbf{x})$, $g_i(\mathbf{x})$, $i=1, \ldots, m$ egy $\Gamma \subseteq R^n$ nyílt halmazon értelmezett kétszer folytonosan differenciálható függvények. Legyen az $\mathbf{x}^*$ pont a (3.1) feladat optimális megoldása és legyen

$$R = \{\mathbf{x} | g_i(\mathbf{x}) \geqq 0, \ i = 1, \ldots, m, \ \mathbf{x} \in \Gamma\}. \tag{3.2}$$

Az R tartomány egy $\mathbf{x}$ pontjában $B(\mathbf{x})$ az aktív indexek halmaza, ha

$$g_i(\mathbf{x}) = 0, \quad i \in B(\mathbf{x}); \quad g_i(\mathbf{x}) > 0, \quad i \notin B(\mathbf{x}).$$

Vezessük be az alábbi jelöléseket.

$$L(\mathbf{x}, \lambda) = f(\mathbf{x}) - \sum_{i=1}^{m} \lambda_i g_i(\mathbf{x}),$$

$$Z_1 = \{\mathbf{y} | \nabla g_i(\mathbf{x}^*)\mathbf{y} = 0, \ i \in B(\mathbf{x}^*)\},$$

$$Z_2 = \{\mathbf{y} | \nabla g_i(\mathbf{x}^*)\mathbf{y} = 0, \ i \in B(\mathbf{x}^*) \quad \text{és} \quad \lambda_i^* > 0\},$$

ahol λ_i^* az optimális *Lagrange szorzókat* jelenti,

$$Z_3 = \{\mathbf{y} | \nabla f(\mathbf{x}^*)\mathbf{y} = 0\}.$$

3.1. TÉTEL. Tegyük fel, hogy az $\mathbf{x}^*$ pontban az aktív feltételek gradiensei függetlenek. Ha $\mathbf{x}^*$ egy lokális optimuma a (3.1) problémának, akkor létezik olyan $\lambda^* \in R^m$ vektor, hogy $\lambda^* \geqq \mathbf{0}$,

$$\nabla f(\mathbf{x}^*) - \sum_{i=1}^{m} \lambda_i^* \nabla g_i(\mathbf{x}^*) = 0,$$

(3.3)

$$\lambda_i^* g_i(\mathbf{x}^*) = 0, \quad i = 1, \ldots, m$$

(3.4) $$\mathbf{y}^T \nabla_{\mathbf{x}}^2 L(\mathbf{x}^*, \lambda^*)\mathbf{y} \geqq 0, \quad \text{ha} \quad \mathbf{y} \in Z_1.$$

A 3.1 tétel bizonyításával együtt megtalálható [2], [8], [12]-ben. A (3.3) feltételeket *Kuhn—Tucker*, a (3.4) feltételt pedig *az optimalitás másodrendű szükséges feltételének* nevezzük. Megjegyezzük, hogy az aktív gradiensek függetlensége mind az első, mind a másodrendű szükséges feltételhez tartozó regularitási feltétel is.

3.2. TÉTEL. Ha az $\mathbf{x}^*$ pontban teljesülnek a *Kuhn—Tucker feltételek,* és ott az $f(\mathbf{x})$, $-\sum_{i \in B(\mathbf{x}^*)} \lambda_i^* g_i(\mathbf{x})$ függvények lokálisan kvázikonvexek, $\nabla f(x^*) \neq \mathbf{0}$, akkor

(3.5) $$\mathbf{y}^T \nabla_{\mathbf{x}}^2 L(\mathbf{x}^*, \lambda^*)\mathbf{y} \geqq 0, \quad \text{ha} \quad \mathbf{y} \in Z_3.$$

Bizonyítás. Tudjuk azt, hogy

$$\nabla_{\mathbf{x}}^2 L(\mathbf{x}^*, \lambda^*) = \nabla^2 f(\mathbf{x}^*) - \sum_{i \in B(\mathbf{x}^*)} \lambda_i^* \nabla^2 g_i(\mathbf{x}^*).$$

Mivel $\nabla f(\mathbf{x}^*) = \sum_{i \in B\mathbf{x}^*)} \lambda_i^* \nabla g_i(\mathbf{x}^*)$, így az $f(\mathbf{x})$, $-\sum_{i \in B(\mathbf{x}^*)} \lambda_i^* g_i(\mathbf{x})$ függvényekre alkalmazva a 2.4. lemmát kapjuk az állítást. Megjegyezzük, hogy a (3.5) egyenlőtlenség erősebb állítást tartalmaz, mint a (3.4) ugyanis $Z_1 \subset Z_3$. Ha a 3.2. tételben azt feltételeztük volna, hogy maguk a $-g_i(\mathbf{x})$, $i \in B(\mathbf{x}^*)$ függvények lokálisan kvázikonvexek az $\mathbf{x}^*$ pontban, akkor a (3.5) egyenlőtlenségek csak a Z_1 halmazra teljesülnek ([22]).

3.3. TÉTEL. Ha az $\mathbf{x}^*$ pontban teljesülnek a *Kuhn—Tucker feltételek* és $\mathbf{y} \in Z_2$ esetén

(3.6) $$\mathbf{y}^T \nabla_{\mathbf{x}}^2 L(\mathbf{x}^*, \lambda^*)\mathbf{y} > 0,$$

akkor $\mathbf{x}^*$ szigorú lokális minimuma a (3.1) problémának. A tétel bizonyításával együtt megtalálható [2], [12]-ben.

3.4. TÉTEL. Ha a megengedett tartomány egy $\mathbf{x}^*$ pont egy környezetében konvex, az $\mathbf{x}^*$ pontban teljesülnek a *Kuhn—Tucker feltételek* és tetszőleges $\mathbf{y} \in Z_3$ esetén

(3.7) $$\mathbf{y}^T \nabla^2 f(\mathbf{x}^*)\mathbf{y} > 0,$$

akkor $\mathbf{x}^*$ szigorú, lokális minimuma a (3.1) problémának.

Bizonyítás: A 3.3. tétel segítségével beláthatjuk, hogy a

$$\min f(\mathbf{x})$$

(3.8)

$$\nabla f(\mathbf{x}^*)(\mathbf{x} - \mathbf{x}^*) \geqq 0$$

feladatnak $\mathbf{x}^*$ szigorú, lokális minimuma. Másrészt a *Kuhn—Tucker feltételek* teljesüléséből következik, hogy

$$\{\mathbf{y}|\nabla g_i(\mathbf{x}^*)\mathbf{y} \geqq 0,\ i \in B(\mathbf{x}^*)\} \subset \{\mathbf{y}|\nabla f(\mathbf{x}^*)\mathbf{y} \geqq 0\}. \tag{3.9}$$

Azonban a megengedett tartomány az optimum pont egy környezetében konvex, tehát ez a környezet benne van az aktív gradiensek által kifeszített kúpban, ami a tétel állítását igazolja.

4. A másodrendű feltételek differenciálgeometriai vizsgálata

Ebben a részben a (3.7) feltételt vizsgáljuk. Elég csak azzal az esettel foglalkozni, mikor $\nabla f(\mathbf{x}_0) \neq 0$. Megmutatjuk, hogy ennek és (3.7)-nek a teljesülése az $f(\mathbf{x})$ függvény $f(\mathbf{x})=f(\mathbf{x}_0)$ nívófelületének $\mathbf{x}_0$ pontján átmenő felületi sík görbék maximálgörbületének negativitásával egyenértékű. Megmutatjuk azt is, hogy ennek a feltételnek az ellenőrzése elvégezhető egy mátrix legnagyobb sajátértékének meghatározásával. Az utolsó részben ezt a mátrixot adjuk meg az $f(\mathbf{x})$ függvény $\mathbf{x}_0$ pontbeli gradiensének és *Hesse mátrixának* segítségével, így a számítás numerikusan is elvégezhetővé válik. A továbbiakban alkalmazott jelöléseket és elnevezéseket illetően lásd a [21] és [23] irodalmat.

Az $f(\mathbf{x})=f(\mathbf{x}_0)$ nívófelületet vizsgáljuk az $\mathbf{x}_0$ pont közelében. Legyen ez egy $n-1$ dimenziós elemi felületként megadva a következő alakban:

$$\mathbf{x}(\mathbf{u}) = \begin{pmatrix} x_1(u_1, \ldots, u_{n-1}) \\ \vdots \\ x_n(u_1, \ldots, u_{n-1}) \end{pmatrix}, \quad (u_1, \ldots, u_{n-1}) \in U. \tag{4.1}$$

Tekintsünk ezen a felületen egy az $\mathbf{x}_0$ ponton áthaladó $\mathbf{x}(t)$ görbét.

$$\mathbf{x}(t) = \mathbf{x}(\mathbf{u}(t)) = \begin{pmatrix} x_1(u_1(t), \ldots, u_{n-1}(t)) \\ \vdots \\ x_n(u_1(t), \ldots, u_{n-1}(t)) \end{pmatrix}, \quad t \in I. \tag{4.2}$$

Legyen ennek a görbének az érintővektora

$$\mathbf{v} = \dot{\mathbf{x}}(t) = \frac{d\mathbf{x}(t)}{dt} = \begin{pmatrix} \dfrac{dx_1(t)}{dt} \\ \vdots \\ \dfrac{dx_n(t)}{dt} \end{pmatrix}.$$

4.1. TÉTEL. $\dfrac{1}{|\nabla f(\mathbf{x})|}\mathbf{v}^T \nabla^2 f(\mathbf{x})\mathbf{v} = -\dot{\mathbf{u}}^T \mathbf{B}\dot{\mathbf{u}}$, ha $\nabla f(\mathbf{x})\mathbf{v} = 0$, $\nabla f(\mathbf{x}) \neq \mathbf{0}$, $\mathbf{v} \neq \mathbf{0}$

ahol $\mathbf{B}$ jelenti a második alapmennyiségek mátrixát.

Bizonyítás. Legyen $\mathbf{x}(\mathbf{u}(t))$ egy tetszőleges felületi görbe, ezért

$$f(\mathbf{x}(\mathbf{u}(t))) = f(\mathbf{x}_0). \tag{4.1}$$

Differenciáljuk t szerint a (4.1) egyenlőség mindkét oldalát. Így azt kapjuk, hogy

$$\nabla f(\mathbf{x})\mathbf{v} = 0. \tag{4.2}$$

Ezt újból differenciálva kapjuk, hogy

$$\mathbf{v}^T \nabla^2 f(\mathbf{x})\mathbf{v} + \nabla f(\mathbf{x})\ddot{\mathbf{x}}(t) = 0. \tag{4.3}$$

Tudjuk azt, hogy

$$\dot{x}_i(t) = \frac{\partial x_i}{\partial u_1}\dot{u}_1(t) + \frac{\partial x_i}{\partial u_2}\dot{u}_2(t) + \ldots + \frac{\partial x_i}{\partial u_{n-1}}\dot{u}_{n-1}(t) = \nabla_{\mathbf{u}} x_i \dot{\mathbf{u}}, \quad i = 1, \ldots, n, \tag{4.4}$$

$$\ddot{x}_i(t) = \dot{\mathbf{u}}^T \nabla_{\mathbf{u}}^2 x_i \dot{\mathbf{u}} + \nabla_{\mathbf{u}} x_i \ddot{\mathbf{u}}, \quad i = 1, \ldots, n. \tag{4.5}$$

Mivel nívófelületet vizsgálunk, ezért a $\dfrac{\partial \mathbf{x}}{\partial u_i}$, $i = 1, \ldots, n-1$ paramétervonalérintők és a gradiens vektor ortogonálisak, tehát

$$\nabla f(\mathbf{x}) \begin{pmatrix} \dfrac{\partial x_1}{\partial u_1} & \cdots & \dfrac{\partial x_1}{\partial u_{n-1}} \\ \vdots & & \vdots \\ \dfrac{\partial x_n}{\partial u_1} & \cdots & \dfrac{\partial x_n}{\partial u_{n-1}} \end{pmatrix} \ddot{\mathbf{u}} = 0, \tag{4.6}$$

vagy más formában írva a

$$\nabla f(\mathbf{x})(\nabla_{\mathbf{u}} x_1 \ddot{\mathbf{u}}, \ldots, \nabla_{\mathbf{u}} x_n \ddot{\mathbf{u}})^T = 0 \tag{4.7}$$

egyenlőség teljesül. Ez viszont azt jelenti, hogy (4.3) helyett a

$$\mathbf{v}^T \nabla^2 f(\mathbf{x})\mathbf{v} + \nabla f(\mathbf{x})(\dot{\mathbf{u}}^T \nabla_{\mathbf{u}}^2 x_1 \dot{\mathbf{u}}, \ldots, \dot{\mathbf{u}}^T \nabla_{\mathbf{u}}^2 x_n \dot{\mathbf{u}})^T = 0 \tag{4.8}$$

egyenlőséget írhatjuk. Most belátjuk, hogy

$$\nabla f(\mathbf{x})(\dot{\mathbf{u}}^T \nabla_{\mathbf{u}}^2 x_1 \dot{\mathbf{u}}, \ldots, \dot{\mathbf{u}}^T \nabla_{\mathbf{u}}^2 x_n \dot{\mathbf{u}})^T = |\nabla f(\mathbf{x})| \dot{\mathbf{u}}^T \mathbf{B} \dot{\mathbf{u}}. \tag{4.9}$$

A bal oldalon a kijelölt szorzást elvégezve kapjuk, hogy

$$\nabla f(\mathbf{x})(\dot{\mathbf{u}}^T \nabla_{\mathbf{u}}^2 x_1 \dot{\mathbf{u}}, \ldots, \dot{\mathbf{u}}^T \nabla_{\mathbf{u}}^2 x_n \dot{\mathbf{u}})^T = \tag{4.10}$$

$$= \sum_{i=1}^{n} \dot{\mathbf{u}}^T \begin{pmatrix} \dfrac{\partial f}{\partial x_i}\dfrac{\partial^2 x_i}{\partial u_1^2} & \cdots & \dfrac{\partial f}{\partial x_i}\dfrac{\partial^2 x_i}{\partial u_1 \partial u_{n-1}} \\ \vdots & & \vdots \\ \dfrac{\partial f}{\partial x_i}\dfrac{\partial^2 x_i}{\partial u_{n-1}\partial u_1} & \cdots & \dfrac{\partial f}{\partial x_i}\dfrac{\partial^2 x_i}{\partial u_{n-1}^2} \end{pmatrix} \dot{\mathbf{u}}.$$

Mivel esetünkben az érintősík $\mathbf{n}$ normálisának és a $\nabla f(\mathbf{x})$ vektornak az iránya megegyezik, valamint

$$b_{ij} = \mathbf{n}\left(\frac{\partial^2 x_1}{\partial u_i \partial u_j}, \ldots, \frac{\partial^2 x_n}{\partial u_i \partial u_j}\right), \quad i, j = 1, \ldots, n-1, \tag{4.11}$$

így valóban látszik, hogy a (4.9) egyenlőség igaz. Ez viszont éppen a tétel állítását igazolja.

A 4.1 tételből következik, hogy a (3.7) feltétel akkor teljesül, ha a **B** mátrix negatív definit, azaz ha a **B** mátrix legnagyobb sajátértéke is negatív.

Most megmutatjuk, hogy ez az $\mathbf{x}_0$ ponton átmenő felületi síkgörbék maximálgörbületének negativitását jelenti. A G maximálgörbületre teljesülni kell a $\det|\mathbf{B}-\mathbf{GH}|=0$ egyenletnek, ahol **H** az első alapmennyiségek mátrixa. Ez a vizsgálatok szempontjából elég kellemetlen forma. Azonban **H** pozitív definit, így alkalmazható a következő tétel:

4.2. TÉTEL. [4] Legyen **B**, **H** két valós, $(n-1)\times(n-1)$-es szimmetrikus mátrix, ahol **H** pozitív definit. Akkor létezik olyan nem szinguláris **T** mátrix, hogy

$$\mathbf{H} = \mathbf{T}\mathbf{T}^T$$

(4.12)

$$\mathbf{B} = \mathbf{T}\begin{pmatrix} \mu_1 & & & \\ & \mu_2 & & \\ & & \ddots & \\ & & & \mu_{n-1} \end{pmatrix}\mathbf{T}^T,$$

ahol μ_i, $i=1, \ldots, n-1$ a **B** mátrix sajátértékeit jelöli.

Ha ezt a tételt felhasználjuk, akkor azt kapjuk, hogy

$$0 = \det|\mathbf{B}-\mathbf{GH}| = \det\left|\mathbf{T}\begin{pmatrix} \mu_1 & & & \\ & \mu_2 & & \\ & & \ddots & \\ & & & \mu_{n-1} \end{pmatrix}\mathbf{T}^T-\mathbf{GTT}^T\right| = \tag{4.13}$$

$$= \det|\mathbf{T}|\det\left|\begin{pmatrix} \mu_1 & & & \\ & \mu_2 & & \\ & & \ddots & \\ & & & \mu_{n-1} \end{pmatrix}-\mathbf{GI}\right|\det|\mathbf{T}^T|.$$

De $\det|\mathbf{T}|=\det|\mathbf{T}^T|$, így

$$0 = (\mu_1-\mathrm{G})(\mu_2-\mathrm{G})\ldots(\mu_{n-1}-\mathrm{G}), \tag{4.14}$$

ami állításunkat igazolja.

Az eddig elmondottak tetszőleges paraméterek mellett igazak. Most olyan paramétereket választunk, hogy a számítások könnyen elvégezhetőek legyenek.

Mivel a gradiens nem nulla, az általánosság megszorítása nélkül feltehetjük, hogy az n-edik komponens különbözik nullától.

Legyen $x_1=u_1, \ldots, x_{n-1}=u_{n-1}$, így az implicit-függvény tételt alkalmazva kapjuk az $f(\mathbf{x})=f(\mathbf{x}_0)$ felület

$$\mathbf{x}(\mathbf{u}) = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_{n-1} \\ x_n(x_1, \ldots, x_{n-1}) \end{pmatrix} \tag{4.15}$$

előállítását. Ezt a felületet explicit módon nem ismerjük, azonban a számításokhoz szükséges első és második alapmennyiségeket az implicit függvények differenciálási szabálya alapján a gradiens vektor és a *Hesse-mátrix* segítségével elő tudjuk állítani a következő formában:

$$h_{ij} = \frac{\dfrac{\partial f}{\partial x_i}\dfrac{\partial f}{\partial x_j}}{\left(\dfrac{\partial f}{\partial x_n}\right)^2}, \quad i, j = 1, \ldots, n-1; \; i \neq j \tag{4.16}$$

$$h_{ii} = 1 + \left(\frac{\dfrac{\partial f}{\partial x_i}}{\dfrac{\partial f}{\partial x_n}}\right)^2, \quad i = 1, \ldots, n-1 \tag{4.17}$$

$$b_{ij} = -\frac{1}{|\nabla f|}\frac{\partial^2 f}{\partial x_i \partial x_j} + \frac{1}{|\nabla f|}\frac{\dfrac{\partial^2 f}{\partial x_i \partial x_n}\dfrac{\partial f}{\partial x_i}}{\dfrac{\partial f}{\partial x_n}} + \tag{4.18}$$

$$+ \frac{1}{|\nabla f|}\frac{\dfrac{\partial^2 f}{\partial x_j \partial x_n}\dfrac{\partial f}{\partial x_i}}{\dfrac{\partial f}{\partial x_n}} - \frac{1}{|\nabla f|}\frac{\dfrac{\partial^2 f}{\partial x_n^2}\dfrac{\partial f}{\partial x_i}\dfrac{\partial f}{\partial x_j}}{\left(\dfrac{\partial f}{\partial x_n}\right)^2}, \quad i, j = 1, \ldots, n-1.$$

IRODALOM

[1] ARROW, K. J. AND ENTHOVEN, A. C., "Quasi concave programming", *Econometria* **29** (1961) 778—800.

[2] AVRIEL, M., *Nonlinear Programming, Analysis and Methods* (Prentice-Hall, Inc. Englewood Cliffs, New Jersey, 1976).

[3] AVRIEL, M., "*r*-convex functions", *Mathematical Programming* **2** (1972) 309—323.

[4] BELLMANN, R., *Introduction to Matrix Analysis* (McGraw-Hill Book Company, New York, Toronto, London, 1960).

[5] COTTLE, R. W., AND FERLAND, J. A., "On pseudo-convex functions of nonnegative variables", *Mathematical Programming* **1** (1971) 95—101.

[6] FARKAS, M., „A feltételes szélsőértékről", *Mat. Lapok* **24** (1973) 1—2.

[7] FERLAND, J. A., "Mathematical programming problems with quasi-convex objective functions", *Mathematical Programming* **3** (1972) 296—301.

[8] FIACCO, A. V., AND MCCORMICK, G. P., *Nonlinear Programming, Sequential Unconstrained Minimization Techniques* (Wiley and Sons, New York, 1968).

[9] GERENCSÉR, L., "On a close relation between quasiconvex and convex functions and related investigations", *Math. Operationsforschung und Statistik* **4** (1973) Heft 3, 201—211.

[10] GERENCSÉR, L., „Nemlineráis programozási feladatok megoldása szekvenciális módszerekkel" *MTA SZTAKI Tanulmányok* **49** (1976).

[11] KÉRI, G., "An examination of nonnegativity and quasiconvexity conditions of quadratic forms on the non-negativ orthant", *Studia Scientiarum Mathematicarum Hungaricae* **6** (1971) 193—196.

[12] LUENBERGER, D. G., *Introduction to Linear and Nonlinear Programming* (Addison-Wesley Publishing Company Inc., 1973).

[13] MANGASARIAN, O. L., "Pseudo-convex functions", *SIAM Journal on Control* **3** (1965) 281—290.

[14] MANGASARIAN, O. L., *Nonlinear Programming* (McGraw-Hill Book Company, 1969).

[15] MANN, H. B., "Quadratic forms with linear constraints", *Amer. Math. Monthly* **50** (1943) 430—433.
[16] MARTOS, B., "The direct power of adjacent vertex programming methods", *Management Science* **12** (1965) 241—252.
[17] MARTOS, B., "Quasi-convexity and quasi-monotonicity in nonlinear programming", *Studia Sci. Math. Hungaricae* **2** (1967) 265—273.
[18] MARTOS, B., *Nonlinear Programming Theory and Methods* (Akadémiai Kiadó, Bp. 1975).
[19] PRÉKOPA, A., „Sztochasztikus rendszerek optimalizálási problémáiról", Akadémiai doktori disszertáció, Bp. 1970.
[20] PRÉKOPA, A., „Eine Erweiterung der sogenannter Methode der „zulässige Richtungen" der nichtlinearen Optimierung auf den Fall quasikonkaver Restriktionsfunctionen", *Math. Operationsforschung und Stat.*, (1973).
[21] RAPCSÁK, A. ÉS TAMÁSSY, L., *Differenciálgeometria* (Tankönyvkiadó, Bp. 1967)
[22] RAPCSÁK, T., „A SUMT-módszer alkalmazása nem konvex programozási feladatok esetén", *Alk. Mat. Lapok* **2** (1976) 427—437.
[23] SZOLCSÁNYI, E., *Differenciálgeometria és vektoranalízis* (Tankönyvkiadó, Bp. 1973).
[24] ZANGWILL, W. I., *Nonlinear Programming: A Unified Approach* (Prentice-Hall, Inc. Englewood Cliffs, 1969).
[25] WETTERLING, W., „Über Minimalbedingungen und Newton-Iteration bei nichtlinearen Optimierungsaufgaben", *International Series of Numerical Mathematics* **15** (1970) 93—99.
[26] WETTERLING, W., AND COLLATZ, L., *Optimization Problems* (Springer-Verlag, New York, Heidelberg, Berlin, 1975).

(Beérkezett: 1978. április 21.)

RAPCSÁK TAMÁS
MTA SZÁMÍTÁSTECHNIKAI ÉS AUTOMATIZÁLÁSI KUTATÓ INTÉZET
1250 BUDAPEST, I., ÚRI U. 49.

ON SECOND ORDER OPTIMALITY CONDITIONS

T. RAPCSÁK

In this paper a second order optimality condition is given. We show that this condition is sufficient and a similar condition — using the pseudo-convex property in the optimum point — is necessary. The differential geometric interpretation of these conditions is given too.

On this basis the conditions can be checked calculating the greatest eigenvalue of a matrix. Using the gradient vector and the *Hessian-matrix* the matrix can be given explicitly.

Alkalmazott Matematikai Lapok 4 (1978) 117—134

HIPERGRÁF ELMÉLETEN ALAPULÓ ÚJ CLUSTER DEFINÍCIÓ ÉS TECHNIKA, II.*

FUTÓ PÉTER
Budapest

3. A hipergráf kvázi komponenseinek meghatározása

A kvázi komponensek megkeresésének alapgondolata

A hipergráf kvázi komponenseit a hipergráf fokozatos szétdarabolásával keressük meg, hasonlóan az LS halmazok LAWLER [21] által vázolt meghatározásához.

A hipergráf szétdarabolását (azaz a minimális értékű két részre vágást) egyszerű szerkezetű kereslet-kínálat feladatra vezetjük vissza, ennek következtében a *vágás* praktikusan *kevés lépésszámmal* végezhető el, egy komponens-kereső rutin minden egyes vágást követő alkalmazása pedig *a végrehajtandó vágások számának jelentős csökkenését* eredményezheti.

A $H=(X;\mathcal{E})$ hipergráf nem triviális kvázi komponensei kifeszítő ponthalmazok (2.9. lemma), és az általuk kifeszített rész hipergráfok összefüggők (2.17. megjegyzés). Tehát a nem triviális kvázi komponenseket a $H=(X;\mathcal{E})$ hipregráf egy és csak egy komponense tartalmazza, sőt ezek kvázi komponensei a megfelelő komponens által kifeszített rész hipergráfoknak is (2.16. megjegyzés). (Természetesen ugyanez teljesül a triviális kvázi komponensekre is.) Ugyanakkor a komponensek által kifeszített rész hipergráfok valamennyi kvázi komponense a $H=(X;\mathcal{E})$ hipergráfnak is kvázi komponense (2.7. lemma).

Tehát a $H=(X;\mathcal{E})$ hipergráf kvázi komponenseit úgy határozhatjuk meg, hogy *rendre megkeressük* a *komponensei* által kifeszített rész hipergráfok *kvázi komponenseit*. Ehhez először meghatározzuk $H=(X;\mathcal{E})$ komponenseit *(R1 rutin)*.

Legyen P a $H=(X;\mathcal{E})$ hipergráf egy tetszőleges komponense. Akkor P kvázi komponense H_P-nek (2.12. megjegyzés). Ha $|P|=1$, akkor nem tartalmaz több kvázi komponenst. Ha $|P|=2$, akkor $S=P$ választás mellett *meghatározzuk azt a* $T^*\in P_S$ *ponthalmazt, amelyre a* $\bar{w}_P(T^*)$ *vágásérték minimális (R2 rutin)*.

A H_P hipergráf P-től különböző kvázi komponenseit T^* vagy $P-T^*$ (2.11. tétel), sőt H_{T^*} vagy H_{P-T^*} komponensei is (2.14. tétel) tartalmazzák.

Tehát *meghatározzuk* H_{T^*} és H_{P-T^*} *komponenseit (R1 rutin)*.

Ezt követően az így kapott komponensekkel úgy járunk el, mint P-vel, szem előtt tartva, hogy a minimális vágást mindig a teljes H_P hipergráfon keressük.

Az eljárást addig folytatjuk, amíg valamennyi komponens egy elemű lesz. Közben a 2.22. megjegyzésen alapuló *próbával kiszűrjük* a rész hipergráfok komponensei közül azokat, amelyek H_p-nek *kvázi komponensei*. Azt, hogy ily módon H_P összes kvázi komponensét megkapjuk, és hogy az eljárás véges, a 3.5. tétel bizonyítja.

* A dolgozat első része az *Alkalmazott Matematikai Lapok* 3 (1977) számában jelent meg. Az ott ismertetett definíciókra és tételekre a szövegben több ízben hivatkozunk.

Tehát a kvázi komponensek megkeresése két rutinra: a hipergráf komponenseinek meghatározása *(R1 rutin)* és a hipergráf minimális értékű két részre vágásának meghatározására *(R2 rutin)* épül.

A komponensek megkeresésére (R1 rutin) a szakirodalomban több egyszerű és hatékony eljárás található (KNUTH [18], KLAFSZKY [17]). Ezek közül az indexelési technikán alapuló (KLAFSZKY [17]) jól ismert módszert használjuk; ennek ismertetésétől itt eltekintünk.

A következő részben az *R2 rutinnal* foglalkozunk részletesen.

Hipergráf minimális értékű két részre vágásának meghatározása (R2 rutin)

A kvázi komponensek megkeresésének *alapvető fontosságú rutinja* az *R2 rutin*, amely a következő feladat megoldására szolgál:

F1: Adott a $H=(X;\mathscr{E})$ összefüggő hipergráf ($|X|\geqq 2$) és az élein értelmezett $v(E_j)>0$ ($j=1,\ldots,m$) függvény. Legyen adott X-nek egy legalább két elemű S részhalmaza. ($S\subseteqq X$, $|S|\geqq 2$). Határozzuk meg azt a $P^*\in\mathscr{P}_S=\{P|P\neq\emptyset, P\subset S\}$ ponthalmazt, amelyre teljesül az, hogy bármely $P\in\mathscr{P}_S$ esetén $\overline{w}(P^*)\leqq\overline{w}(P)$.

Tehát az *F1* feladatnál azzal a pótlólagos *feltétellel* keressük a *minimális értékű vágást* generáló ponthalmazt, hogy az egy előre kijelölt $X-S$ ponthalmaz egy pontját se tartalmazza.

Az *F1* feladatot úgy fogjuk megoldani, hogy *több lépésen keresztül visszavezetjük* igen egyszerű szerkezetű irányított gráfok (két kijelölt pontot szeparáló) minimális vágásainak meghatározására. Ehhez első lépésben definiáljuk az *F2* feladatot.

F2: Adott a $H=(X;\mathscr{E})$ összefüggő hipergráf és az élein értelmezett $v(E_j)>0$ ($j=1,\ldots,m$) függvény. Legyen $|X|\geqq 2$. Legyen adott az $\mathscr{E}$ élhalmaznak két kijelölt részhalmaza $\mathscr{F}^0$ és $\mathscr{G}^0$ amelyek a következő tulajdonságokkal rendelkeznek:

(I) $\emptyset\neq\mathscr{F}^0\subset\mathscr{E}$, $\emptyset\neq\mathscr{G}^0\subset\mathscr{E}$

(II) $\mathscr{H}(\mathscr{F}^0)\cap\mathscr{H}(\mathscr{G}^0)=\emptyset$.

Határozzuk meg azokat a $\mathscr{F}^*$ és $\mathscr{G}^*$ élhalmazokat, amelyek a következő tulajdonságokkal rendelkeznek:

(1) $\mathscr{F}^0\subseteqq\mathscr{F}^*\subset\mathscr{E}$, $\mathscr{G}^0\subseteqq\mathscr{G}^*\subset\mathscr{E}$

(2) $\mathscr{H}(\mathscr{F}^*)\cap\mathscr{H}(\mathscr{G}^*)=\emptyset$

(3) tetszőleges az (1) és (2) követelményeket kielégítő $\mathscr{F}$ és $\mathscr{G}$ halmazok választása esetén $w(\mathscr{F}^*)+w(\mathscr{G}^*)\geqq w(\mathscr{F})+w(\mathscr{G})$.

3.1. Megjegyzés: A $w: 2^{\mathscr{E}}\to R^+$ függvény pozitivitásából és a (2) tulajdonságából következik, hogy az *F2* feladat megfogalmazható úgy is, hogy a (3) tulajdonságot a következő tulajdonsággal helyettesítjük: (3′) tetszőleges az (1) és (2) követelményeket kielégítő F és G halmazok választása esetén $w(\mathscr{E}-(\mathscr{F}^*\cup\mathscr{G}^*))\leqq w(\mathscr{E}-(\mathscr{F}\cup\mathscr{G}))$.

3.1. TÉTEL. Az *F1* feladat megoldása legfeljebb $2(|S|-1)$ számú *F2* feladat elvégzésével meghatározható.

Bizonyítás: Legyen $x_0 (x_0 \in S)$ előre rögzített, $x_j (x_j \in S, x_j \neq x_0)$ pedig tetszőleges pont.

A $H=(X;\mathscr{E})$ hipergráf felhasználásával konstruáljuk meg a $H_{0j}=(X;\mathscr{E}_{0j})$ hipergráfot, ahol $\mathscr{E}_{0j}=\mathscr{E}\cup\{x_0\}\cup\{\{x_j\}\cup(X-S)\}$. Legyen $v(\{x_0\})=v(\{x_j\}\cup(X-S))=1$. Jelölje a H_{0j} hipergráfban értelmezett leképezéseket $\mathscr{H}_{0j}: 2^{\mathscr{E}_{0j}}\to 2^X$, $\mathfrak{E}_{0j}: 2^X\to 2^{\mathscr{E}_{0j}}$, $\mathfrak{E}'_{0j}: 2^X\otimes 2^X\to 2^{\mathscr{E}_{0j}}$. Legyen $\mathscr{F}^0_{0j}=\{x_0\}$, $\mathscr{G}^0_{0j}=\{\{x_j\}\cup(X-S)\}$.

Nyilvánvaló, hogy

(I) $\emptyset\neq\mathscr{F}^0_{0j}\subset\mathscr{E}_{0j}$, $0\neq\mathscr{G}^0_{0j}\subset\mathscr{E}_{0j}$

(II) $\mathscr{H}_{0j}(\mathscr{F}^0_{0j})\cap\mathscr{H}_{0j}(\mathscr{G}^0_{0j})\neq\emptyset$.

Végezzük el az *F2* feladatot a $\mathscr{H}_{0j}$ hipergráf, az $\mathscr{F}^0_{0j}$ és $\mathscr{G}^0_{0j}$ élhalmazok felhasználásával.

Jelölje annak egy megoldását $\mathscr{F}^*_{0j}$ és $\mathscr{G}^*_{0j}$.

Hasonlóképpen a $\mathscr{E}_{j0}=\mathscr{E}\cup\{x_j\}\cup\{\{x_0\}\cup(X-S)\}$ választás mellett megszerkesztjük a $\mathscr{H}_{j0}$ hipergráfot, kijelöljük az $\mathscr{F}^0_{j0}=\{x_j\}$ és a $\mathscr{G}^0_{j0}=\{\{x_0\}\cup(X-S)\}$ élhalmazokat, és megoldjuk a megfelelő *F2* feladatot. Jelölje a feladat egy megoldását $\mathscr{F}^*_{j0}$ és $\mathscr{G}^*_{j0}$.

Legyen P^* az *F1* feladatnak egy megoldása. $P^*\in P_S$ miatt vagy $x_0\in P^*$ és $\exists x_j\in S-P^*$ vagy $x_0\in S-P^*$ és $\exists x_j\in P^*$. A bizonyítás további gondolatmenetéből adódik, hogy az általánosság megszorítása nélkül feltehetjük azt, hogy $x_0\in P^*$ és $x_j\in S-P^*$.

Definiáljuk P^* felhasználásával egy most meghatározott x_j esetére az $\mathscr{F}_{0j}=\mathfrak{E}'_{0j}(P^*|P^*)$ és a $\mathscr{G}_{0j}=\mathfrak{E}'_{0j}((X-P^*)|(X-P^*))$ élhalmazokat. Nyilvánvaló, hogy az $\mathscr{F}_{0j}, \mathscr{G}_{0j}$ élhalmaz párra az *F2* feladat (1) és (2) feltétele teljesül, azaz $\mathscr{F}^0_{0j}\subseteq\mathscr{F}_{0j}$ és $\mathscr{G}^0_{0j}\subseteq\mathscr{G}_{0j}$, valamint, hogy $\mathscr{H}_{0j}(\mathscr{F}_{0j})\cap\mathscr{H}_{0j}(\mathscr{G}_{0j})=\emptyset$.

Tehát $w(\mathscr{E}_{0j}-(\mathscr{F}^*_{0j}\cup\mathscr{G}^*_{0j}))\leq w(\mathscr{E}_{0j}-(\mathscr{F}_{0j}\cup\mathscr{G}_{0j}))$.

Mivel az $\mathscr{F}^*_{0j}, \mathscr{G}^*_{0j}$ és az $\mathscr{F}_{0j}, \mathscr{G}_{0j}$ élhalmaz párokra egyaránt teljesül az *F2* feladat (1) feltétele ezért

(i) $w(\mathscr{E}-(\mathscr{F}^*_{0j}\cup\mathscr{G}^*_{0j}))\leq w(\mathscr{E}-(\mathscr{F}_{0j}\cup\mathscr{G}_{0j}))$ is igaz.

Jelölje az $\mathscr{F}^*, \mathscr{G}^*$ élhalmaz pár a fenti módon felírható legfeljebb $2(|S|-1)$ számú *F2* feladat megoldásai közül egy olyat, amelyre $w(\mathscr{E}-(\mathscr{F}^*\cup\mathscr{G}^*))$ minimális, azaz $w(\mathscr{E}-(\mathscr{F}^*\cup\mathscr{G}^*))\leq w(\mathscr{E}-(\mathscr{F}^*_{0j}\cup\mathscr{G}^*_{0j}))$ tetszőleges $x_j\in S$ $(x_j\neq x_0)$ választása mellett.

Mivel $\mathscr{H}(\mathscr{G}^*)\supset X-S$, ezért az *F2* feladat (2) feltétele miatt $\mathscr{H}(\mathscr{F}^*)\subset S$. $\mathscr{H}(\mathscr{F}^*)\neq\emptyset$, mivel legalább egy, egy pontot tartalmazó élt tartalmaz. Tehát $\mathscr{H}(\mathscr{F}^*)\in P_S$ és így

(ii) $$\bar{w}(P^*)\leq\bar{w}(\mathscr{H}(\mathscr{F}^*)).$$

A 2.8. lemma gondolatmenetének felhasználásával azonnal adódik, hogy $C(P^*)=\mathscr{E}-(\mathscr{F}_{0j}\cup\mathscr{G}_{0j})$ tehát

(iii) $$\bar{w}(P^*)=w(\mathscr{E}-(\mathscr{F}_{0j}\cup\mathscr{G}_{0j})).$$

Tegyük fel, hogy $C(\mathscr{H}(\mathscr{F}^*))\subseteq\mathscr{E}-(\mathscr{F}^*\cup\mathscr{G}^*)$ nem teljesül, azaz $\exists E_j\in C(\mathscr{H}(\mathscr{F}^*))$, amelyre $E_j\in\mathscr{F}^*\cup\mathscr{G}^*$. Mivel $\mathscr{H}(\mathscr{F}^*)$ kifeszítő (2.1. lemma), ezért ha $E_j\in C(\mathscr{H}(\mathscr{F}^*))$, akkor $E_j\notin\mathscr{F}^*$ tehát $E_j\in\mathscr{G}^*$.

Ez viszont ellentmond az *F2* feladat (2) feltételének. Tehát $C(\mathcal{H}(\mathcal{F}^*)) \subseteq \mathcal{E} - (\mathcal{F}^* \cup \mathcal{G}^*)$, azaz (i), (ii) és (iii) felhasználásával $w(\mathcal{E}-(\mathcal{F}^*_{0j} \cup \mathcal{G}^*_{0j})) = \overline{w}(P^*) = w(\mathcal{E}-(\mathcal{F}^* \cup \mathcal{G}^*))$ egyenlőség adódik.

A tétel *konstruktív jellegű,* ugyanis láttuk, hogy a szóbajövő $2(|S|-1)$ számú *F2* feladat megoldásai között, ha $\mathcal{F}^*, \mathcal{G}^*$ megoldásra teljesül az, hogy $w(\mathcal{E}-(\mathcal{F}^* \cup \mathcal{G}^*))$ minimális, akkor $\mathcal{H}(\mathcal{F}^*)$ megoldása az *F1* feladatnak.

3.2. Megjegyzés: A 3.1. tétel bizonyításából egyszerű számítással adódik, hogy ha $x_0 \in P^*$, akkor tetszőleges $x_j \in S - P^*$ választása esetén a megfelelő *F2* feladatot elvégezve $w(\mathcal{E}-(\mathcal{F}^*_{0j} \cup \mathcal{G}^*_{0j})) = \overline{w}(P^*)$, ha $x_0 \in S - P^*$, akkor tetszőleges $x_j \in P^*$ választása esetén a megfelelő *F2* feladatot elvégezve $w(\mathcal{E}-\mathcal{F}^*_{j0} \cup \mathcal{G}^*_{j0})) = \overline{w}(P^*)$.

A következőkben az *F2* feladat kis módosításával definiáljuk az *F3* feladatot, amelyről a 3.2. tétel mondja ki, hogy ekvivalens *F2*-vel (megoldásaik halmaza azonos).

F3: Adott a $H=(X;\mathcal{E})$ összefüggő hipergráf ($|X| \geqq 2$) és az élein értelmezett $v(E_j) > 0$ ($j=1, \ldots, m$) függvény. Legyen adott az $\mathcal{E}$ élhalmaznak két részhalmaza, $\mathcal{F}^0$ és $\mathcal{G}^0$, amelyek a következő tulajdonságokkal rendelkeznek:

(I′) $\emptyset \neq \mathcal{F}^0 \subset \mathcal{E}, \quad \emptyset \neq \mathcal{G}^0 \subset \mathcal{E},$

(II′) $\mathcal{H}(\mathcal{F}^0) \cap \mathcal{H}(\mathcal{G}^0) = \emptyset.$

Szerkesszük meg a következő két élhalmazt:

$$\hat{\mathcal{F}} = \{E_j | E_j \in \mathcal{E}, \ \mathcal{H}(\{\mathcal{E}_j\}) \cap \mathcal{H}(\mathcal{G}^0) = \emptyset\}$$

$$\hat{\mathcal{G}} = \{E_j | E_j \in \mathcal{E}, \ \mathcal{H}(\{\mathcal{E}_j\}) \cap \mathcal{H}(\mathcal{F}^0) = \emptyset\}.$$

Határozzuk meg azokat az $\mathcal{F}^{*\prime}$ és $\mathcal{G}^{*\prime}$ élhalmazokat, amelyek a következő tulajdonságokkal rendelkeznek:

(1′) $\mathcal{F}^{*\prime} \subseteq \hat{\mathcal{F}} \quad \mathcal{G}^{*\prime} \subseteq \hat{\mathcal{G}}$

(2′) $\mathcal{H}(\mathcal{F}^{*\prime}) \cap \mathcal{H}(\mathcal{G}^{*\prime}) = \emptyset$

(3′) tetszőleges az (1′) és (2′) követelményeket kielégítő $\mathcal{F}$ és $\mathcal{G}$ halmazok választása esetén $w(\mathcal{F}) + w(\mathcal{G}) \leqq w(\mathcal{F}^*) + w(\mathcal{G}^*)$.

3.2. TÉTEL. Az *F2* és az *F3* feladat megoldásainak halmaza megegyezik.

Bizonyítás: Legyen $\mathcal{F}^*, \mathcal{G}^*$ megoldása *F2*-nek. $\mathcal{H}(\mathcal{F}^0) \cap \mathcal{H}(\mathcal{G}^0) = \emptyset$ miatt $\mathcal{F}^0 \subseteq \hat{\mathcal{F}}$, $\mathcal{G}^0 \subseteq \hat{\mathcal{G}}$. Nyilvánvaló, hogy már az *F2* feladat (1) és (2) feltételének teljesülése esetén is teljesül az *F3* (1′) és (2′) feltétele. Tehát $\mathcal{F}^* \subseteq \hat{\mathcal{F}}$, $\mathcal{G}^* \subseteq \hat{\mathcal{G}}$ és $\mathcal{H}(\mathcal{F}^*) \cap \mathcal{H}(\mathcal{G}^*) = \emptyset$. Ezzel szemben (1) és (2) teljesülése nem vezethető le csak (1′) és (2′) fennállásából. Fel kell tenni (3′) fennállását is.

Legyen $\mathcal{F}^{*\prime}, \mathcal{G}^{*\prime}$ megoldása az *F3* feladatnak.

(1′)miatt $\mathcal{H}(\mathcal{F}^{*\prime}) \cap \mathcal{H}(\mathcal{G}^0) = \emptyset$, $\mathcal{H}(\mathcal{G}^{*\prime}) \cap \mathcal{H}(\mathcal{F}^0) = \emptyset$. Ha vagy $\mathcal{F}^0 \nsubseteq \mathcal{F}^{*\prime}$, vagy $\mathcal{G}^0 \nsubseteq \mathcal{G}^{*\prime}$, akkor $\mathcal{F}^{*\prime} \cup \mathcal{F}^0 \subseteq \hat{\mathcal{F}}$ és $\mathcal{G}^{*\prime} \cup \mathcal{G}^0 \subseteq \hat{\mathcal{G}}$, és hasonlóképpen $\mathcal{H}(\mathcal{F}^{*\prime} \cup \mathcal{F}^0) \cap \mathcal{H}(\mathcal{G}^{*\prime} \cup \mathcal{G}^0) = \emptyset$ is teljesül, és így a 2.11. megjegyzés miatt $w(\mathcal{F}^{*\prime} \cup \mathcal{F}^0) + w(\mathcal{G}^{*\prime} \cup \mathcal{G}^0) > w(\mathcal{F}^{*\prime}) + w(\mathcal{G}^{*\prime})$, amely ellentmond (3′)-nek.

Tehát az $\mathcal{F}^{*\prime}, \mathcal{G}^{*\prime}$ élhalmazpár kielégíti az (1) feltételt, és (2′) fennállása miatt a (2) feltételt is.

Mivel az $\mathcal{F}^*, \mathcal{G}^*$ élhalmazpár megoldása *F2*-nek és az $\mathcal{F}^{*\prime}, \mathcal{G}^{*\prime}$ élhalmazpár eleget tesz (1)-nek és (2)-nek, ezért $w(\mathcal{F}^{*\prime})+w(\mathcal{G}^{*\prime}) \leqq w(\mathcal{F}^*)+w(\mathcal{G}^*)$.

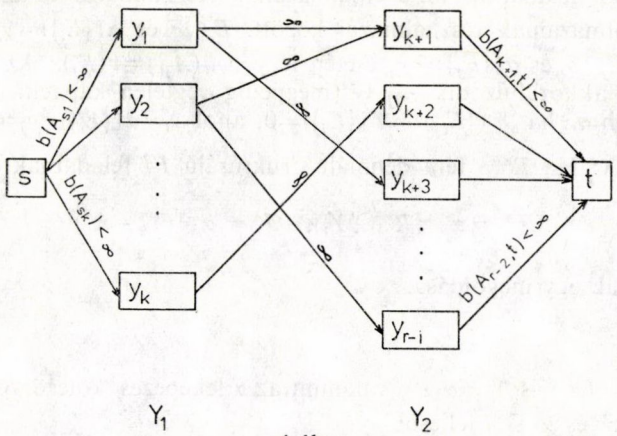
Mivel az $\mathcal{F}^{*\prime}, \mathcal{G}^{*\prime}$ élhalmazpár megoldása *F3*-nak, és az $\mathcal{F}^*, \mathcal{G}^*$ élhalmazpár kielégíti (1′)-t és (2′)-t, ezért $w(\mathcal{F}^*)+w(\mathcal{G}^{*\prime}) \geqq w(\mathcal{F}^*)+w(\mathcal{G}^*)$.

Tehát $w(\mathcal{F}^*)+w(\mathcal{G}^{*\prime})=w(\mathcal{F}^*)+w(\mathcal{G}^*)$.

Az *F3* feladat megoldását a következő igen egyszerű szerkezetű hálózat minimális vágásának meghatározásával kaphatjuk meg. (Vö. 3.3. tétel)

F4: Adott a $G=(Y; \mathcal{A})$ hálózat. Az $Y=\{y_1, \ldots, y_r\}$ halmaz elemei a hálózat pontjai. Az $A_{ij} \in \mathcal{A}$ ív az y_i pontból az y_j pontba fut.
Legyen $b(A_{ij})>0$ $(A_{ij} \in \mathcal{A})$ a hálózat ívein értelmezett kapacitás függvény.
A hálózat pontjai 4 diszjunkt csoportba sorolhatók: $Y=\{s\} \cup Y_1 \cup Y_2 \cup \{t\}$
(Az s pontot forrásnak, a t pontot nyelőnek nevezzük.)

A hálózat ívei 3 diszjunkt csoportba sorolhatók: $\mathcal{A}=\mathcal{A}_1 \cup \mathcal{A}_2 \cup \mathcal{A}_3$. Az $\mathcal{A}_1$ halmaz ívei véges kapacitásúak és az s forrásból húzódnak az összes $y_i \in Y_1$ ponthoz. Az $\mathcal{A}_2$ halmaz ívei végtelen kapacitásúak és bizonyos $y \in Y_1$ pontokból húzódnak bizonyos $y_j \in Y_2$ pontokba. Az $\mathcal{A}_3$ halmaz ívei véges kapacitásúak és az összes $y_j \in Y_2$ pontból húzódnak a t nyelőbe.



4. ábra

Jelölje $\mathcal{R}$ azon ponthalmazok osztályát, amelyek az s forrást tartalmazzák, de a t nyelőt nem tartalmazzák.

$$\mathcal{R} = \{R | R \subseteq Y,\ s \in R,\ t \notin R\}.$$

Legyen $\mathcal{V}(R)$ az R halmaz által generált *irányított vágás* és $d(R)$ az irányított vágás értéke.

$$\mathcal{V}(R) = \{A_{ij} | A_{ij} \in \mathcal{A},\ Y_i \in R,\ Y_j \in Y-R,\ R \in \mathcal{R}\},$$

$$d(R) = \sum_{A_{ij} \in \mathcal{V}(R)} b(A_{ij}).$$

Határozzuk meg azt az $R^* \in \mathcal{R}$ ponthalmazt, amely által generált irányított vágás értéke minimális, azaz $d(R^*) \leqq d(R)$, ha $R \in \mathcal{R}$.

3.3. Megjegyzés: Az F4 feladat az irodalomban elterjedt *kereslet-kínálat feladat*nak egy *speciális esete,* ahol az $y_i \in Y_1$ termelő helyről tetszőleges sok árut szállíthatunk az $y_j \in Y_2$ fogyasztó helyre, de a termelő helyek között valamint a fogyasztó helyek között áru nem szállítható. $b(A_{si})$ jelöli az $y_i \in Y_1$ termelő hely kínálatát, $b(A_{jt})$ pedig az $y_j \in Y_2$ fogyasztó hely keresletét. (KLAFSZKY [17]).

3.4. Megjegyzés: Az *F4* feladatnál bevezetett irányított vágás csak az R halmaz pontjaiból az $Y-R$ halmaz pontjaiba húzódó íveket tartalmazza, a fordított irányú íveket nem. Tehát nem egyezik meg a 2.11. definícióban bevezetett vágással.

3.3. TÉTEL. Az *F3* feladathoz szerkeszthető olyan *F4* feladat, amelynek elvégzésével *F3* megoldása meghatározható.

Bizonyítás: Szerkesszünk meg egy az *F4* feladathoz leírt struktúrájú hálózatot, az *F3* feladat felhasználásával.

Értelmezzünk egy kölcsönösen egyértelmű

$$(\alpha\colon 2^{\hat{\mathscr{F}}} \to 2^{Y_1};\quad \alpha\colon 2^{\hat{\mathscr{G}}} \to 2^{Y_2})$$

leképezést az *F3* feladat $\hat{\mathscr{F}}$, ill. $\hat{\mathscr{G}}$ élhalmazának részhalmazai és az *F4* feladat Y_1, ill. Y_2 pont-halmazainak részhalmazai között. $E_i \in \hat{\mathscr{F}}$ és $\alpha(\{E_i\}) = y_i$ esetén legyen $b(A_{si}) = v(E_i)$ $E_j \in \hat{\mathscr{G}}$ és $\alpha(\{E_j\}) = y_j$ esetén legyen $b(A_{jt}) = v(E_j)$. Az $y_i \in Y_1$ pontból akkor és csak akkor húzódik A_{ij} ív (mégpedig végtelen kapacitású $(b(A_{ij}) = \infty)$) az $y_j \in Y_2$ ponthoz, ha $\mathscr{H}(\{E_i\}) \cap \mathscr{H}(\{E_j\}) \neq \emptyset$, ahol $y_i = \alpha(\{E_i\})$, $y_j = \alpha(\{E_j\})$.

3.4. LEMMA: Ha R^* a fent definiált struktúrájú *F4* feladatnak egy megoldása, akkor

$$\mathscr{F}^* = \alpha^{-1}(R^* \cap Y_1), \quad \mathscr{G}^* = \alpha^{-1}(Y_2 - R^*)$$

az *F3* feladatnak egy megoldása.

Bizonyítás:

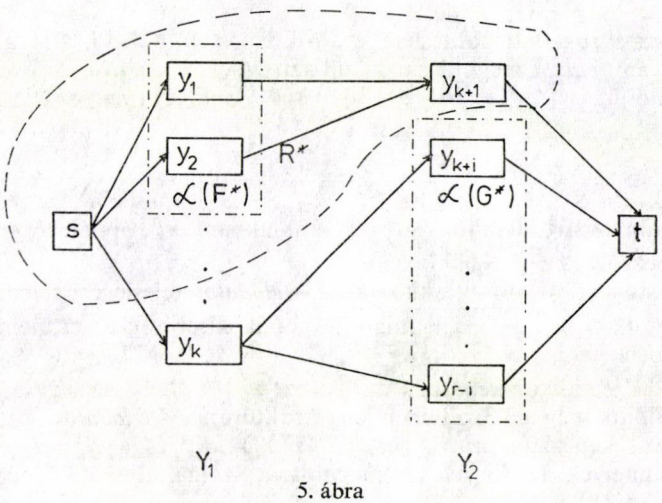
(1) Mivel $Y_1 = \alpha(\hat{\mathscr{F}})$ és $Y_2 = \alpha(\hat{\mathscr{G}})$, valamint az α leképezés kölcsönösen egyértelmű, ezért $\hat{\mathscr{F}}^* \subseteq \mathscr{F}$ és $\mathscr{G}^* \subseteq \hat{\mathscr{G}}$ teljesül.

(2) Az $\{s\} = R$ halmaz eleme az $\mathscr{R}$ halmazosztálynak és $d(\{s\})$ véges, tehát $d(R^*)$ is véges. Tehát ha $y_i \in R^* \cap Y_1$ és $y_j \in Y_2 - R^*$, akkor $b(A_{ij}) = \infty$ miatt $A_{ij} \notin \mathscr{V}(R^*)$, azaz ha $E_i \in \mathscr{F}^*$ és $E_j \in \mathscr{G}^*$, akkor $\mathscr{H}(\{E_i\}) \cap \mathscr{H}(\{E_j\}) = \emptyset$. Így $\mathscr{H}(\mathscr{F}^*) \cap \mathscr{H}(\mathscr{G}^*) = \emptyset$ is teljesül.

(3) Legyenek $\mathscr{F}$ és $\mathscr{G}$ olyan élhalmazok, amelyekre az *F3* feladat (1) és (2) feltétele teljesül, azaz $\mathscr{F} \subseteq \hat{\mathscr{F}}$, $\mathscr{G} \subseteq \hat{\mathscr{G}}$, és $\mathscr{H}(\mathscr{F}) \cap \mathscr{H}(\mathscr{G}) = \emptyset$. Legyen $R = \{s\} \cup \alpha(\mathscr{F}) \cup \alpha(\hat{\mathscr{G}} - \mathscr{G})$.

Mivel $\mathscr{H}(\mathscr{F}) \cap \mathscr{H}(\mathscr{G}) = \emptyset$ és az α leképezés kölcsönösen egyértelmű, ezért $\mathscr{V}(R)$ nem tartalmazhat ∞ kapacitású élt, tehát $d(R^*) \leqq d(R) < \infty$. Ezt részletesebben kifejtve kapjuk, hogy

$$\sum_{y_i \in Y_1 - R^*} b(A_{si}) + \sum_{y_j \in Y_2 \cap R^*} b(A_{jt}) \leqq \sum_{y_i \in Y_1 - R} b(A_{si}) + \sum_{y_j \in Y_2 \cap R} b(A_{jt}).$$



5. ábra

Ez az egyenlőtlenség *F3* feladat jelöléseivel is leírható:

$$w(\alpha^{-1}(Y_1-R^*))+w(\alpha^{-1}(Y_2\cap R^*))\leqq w(\alpha^{-1}(Y_1-R))+w(\alpha^{-1}(Y_2\cap R)),$$

azaz

$$w(\mathscr{F}-\mathscr{F}^*)+w(\hat{\mathscr{G}}-\mathscr{G}^*)\leqq w(\hat{\mathscr{F}}-\mathscr{F})+w(\hat{\mathscr{G}}-\mathscr{G}).$$

Ezt az egyenlőtlenséget a $w(\hat{\mathscr{F}})+w(\hat{\mathscr{G}})=w(\mathscr{F})+w(\hat{\mathscr{G}})$ triviális egyenlőségből kivonva, a 2.9. megjegyzés felhasználásával $w(\mathscr{F}^*)+w(\mathscr{G}^*)\geqq w(\mathscr{F})+w(\mathscr{G})$ adódik.

Tehát az *F3 feladat* megoldásával megkaptuk az *F4 feladatnak* egy megoldását is.

3.5. Megjegyzés: Egyszerűen belátható, hogy $R^*\cap Y_1\neq\emptyset$ és $Y_2-R^*\neq\emptyset$, ugyanis az $\alpha(\mathscr{F}^0)$ és az $\alpha(\mathscr{G}^0)$ halmazbeli pontok csak a forrással, illetve csak a nyelővel vannak ívvel összekötve, tehát a minimális irányított vágásra $\alpha(\mathscr{F}^0)\subseteqq R^*$ és $R^*\cap\alpha(\mathscr{G}^0)=\emptyset$ teljesül.

3.6. Megjegyzés: Az *F3 feladathoz* szerkesztett *F4 feladat mérete csökkenthető* annak a ténynek a felhasználásával, hogy az $\alpha(\mathscr{F}^0)$ halmaz pontjai csak az *s* nyelővel, ill. az $\alpha(\mathscr{G}^0)$ pontjai csak a *t* forrással vannak összekötve.

Legyen $Y_1'=\alpha(\hat{\mathscr{F}}-\mathscr{F}^0)$ és $Y_2'=\alpha(\hat{\mathscr{G}}-\mathscr{G}^0)$. Az íveket definiáljuk a 3.3. tételnek megfelelő módon. Legyen $R^{*\prime}$ a redukált méretű *F4 feladat* megoldása. Ekkor az *F3 feladat* megoldását nyilvánvalóan az $\mathscr{F}^*=\alpha^{-1}(R^{*\prime}\cap Y_1)\cup F^0$ és a $G^*=\alpha^{-1}(Y_2'-R^{*\prime})\cup G^0$ halmazok adják.

Az *F4 feladat* megoldására *Edmonds és Karp* egyik *algoritmusát* használjuk. (EDMONDS, KARP [9]) Választásunkat indokolja egyrészt az, hogy ez az algoritmus tetszőleges pozitív kapacitás függvény választása esetén is *konvergens,* másrészt az, hogy valóban *hatékony* és lépésszáma egyszerű eszközökkel felülről becsülhető.

A fentiek alapján már megszerkeszthetjük az *F1 feladat megoldását szolgáltató R2 rutint:*

R2 rutin:

Legyen $x_0 \in S$ rögzített pont. Jelölje a j-ik lépés előtt $\mathcal{F}^{*(j)}, \mathcal{G}^{*(j)}$ a $j-1$ számú, már elvégzett *F3 feladat* megoldásai közül azt, amelyre a $w[\mathcal{E}-(\mathcal{F}^{*(j)} \cup \mathcal{G}^{*(j)})]$ érték minimális. Jelöljük $w[\mathcal{E}-(\mathcal{F}^{*(j)} \cup \mathcal{G}^{*(j)})]$-ot $w^{*(j)}$-vel. Legyen $\mathcal{F}^{*(1)}=\emptyset$, $\mathcal{G}^{*(1)}=\emptyset$, $w^{*(1)}=w(\mathcal{E})$.

j-ik lépés $(1 \leqq j \leqq 2(|S|-1))$:

Képezzük $j \leqq |S|-1$ esetén az $\mathcal{F}^0_{0j}, \mathcal{G}^0_{0j}$, $j > |S|-1$ esetén az $\mathcal{F}^0_{i0}, \mathcal{G}^0_{i0}$ $(i=j-(|S|-1))$ halmazpárt. Jelöljük ezt a továbbiakban $\mathcal{F}^0_k, \mathcal{G}^0_k$-lal. $\mathcal{F}^0_k$ és $\mathcal{G}^0_k$ felhasználásával képezzük az $\hat{\mathcal{F}}_k, \hat{\mathcal{G}}_k$ halmazokat.

Ha $w[\mathcal{E}-(\hat{F}_k \cup \hat{G}_k)] \geqq w^{*(j)}$, akkor az *F3 feladatot* felesleges elvégezni, ugyanis $w[\mathcal{E}-(\mathcal{F}^*_k \cup \mathcal{G}^*_k)] \geqq w[\mathcal{E}-(\hat{\mathcal{F}}_k \cup \hat{\mathcal{G}}_k)]$ miatt $w^{*(j)}$-nál kisebb értéket nem kaphatunk. Ez esetben tehát $\mathcal{F}^{*(j+1)}=\mathcal{F}^{*(j)}$, $\mathcal{G}^{*(j+1)}=\mathcal{G}^{*(j)}$, $w^{*(j+1)}=w^{*(j)}$. Ha $w[\mathcal{E}-(\hat{\mathcal{F}}_k \cup \hat{\mathcal{G}}_k)] < w^{*(j)}$, akkor képezzük az $Y'_1=\alpha(\hat{\mathcal{F}}_k-\mathcal{F}^0_k)$ és az $Y'_2=\alpha(\hat{\mathcal{G}}_k-\mathcal{G}^0_k)$ halmazokat, és oldjuk meg a 3.3. tételben leírt struktúrájú *F4 feladatot.* Az $R^{*\prime}$ megoldás felhasználásával képezzük az $\mathcal{F}^*_k=\alpha^{-1}(R^{*\prime} \cap Y'_1) \cup \mathcal{F}^0_k$ és a $\mathcal{G}^*_k=\alpha^{-1}(Y'_2-R^*) \cup \mathcal{G}^0_k$ halmazokat, amelyek az *F3 feladat* megoldását szolgáltatják (3.4. tétel, 3.6. megjegyzés).

Ha $w[\mathcal{E}-(\mathcal{F}^*_k \cup \mathcal{G}^*_k)] \geqq w^{*(j)}$, akkor legyen $\mathcal{F}^{*(j+1)}=\mathcal{F}^{*(j)}$, $\mathcal{G}^{*(j+1)}=\mathcal{G}^{*(j)}$, $w^{*(j+1)}=w^{*(j)}$.

Ha $w[\mathcal{E}-(\mathcal{F}^*_k \cup \mathcal{G}^*_k)] < w^{*(j)}$, akkor legyen $\mathcal{F}^{*(j+1)}=\mathcal{F}^*_k$, $\mathcal{G}^{*(j+1)}=\mathcal{G}^*_k$, $w^{*(j+1)}=w[\mathcal{E}-(\mathcal{F}^*_k \cup \mathcal{G}^*_k)]$.

A 3.1. tétel felhasználásával adódik, hogy az *F1 feladat* megoldását $P^*=\mathcal{H}(\mathcal{F}^{*(2|S|-1)})$ szolgáltatja.

3.7. Megjegyzés: $S=X$ esetén $\mathcal{F}^0_{j0}=\mathcal{G}^0_{0j}$, $\mathcal{F}^0_{0j}=\mathcal{G}^0_{j0}$ miatt elegendő $|S|-1$ számú *F2 feladatot* elvégezni, azaz $P^*=\mathcal{H}(\mathcal{F}^{*(|S|-1)})$ már szolgáltatja az *F1 feladat* megoldását.

3.8. Megjegyzés: Az *R2 rutin* apró módosításával meghatározhatjuk az *F1 feladat összes megoldását* is. Az összes megoldás ismerete azonban nem gyorsítja olyan mértékben a kvázi komponensek megkeresését, mint amilyen mértékben növeli az egész eljárás tárolási helyigényét.

A kvázi komponensek meghatározása

A hipergráf kvázi komponensei meghatározásának feladata a következőképpen fogalmazható meg:

F5: Adott a $H=(X; \mathcal{E})$ $(|X| \geqq 1)$ (nem feltétlenül összefüggő) hipergráf és $v(E_j)>0$ $(j=1, \ldots, m)$ a hipergráf élein értelmezett pozitív függvény. Határozzuk meg a $H=(X; \mathcal{E})$ hipergráf összes kvázi komponensét, azaz azokat a Q $(\emptyset \neq Q \subseteq X)$ ponthalmazokat, amelyekre teljesül az, hogy bármely $T \in \mathcal{P}_Q$ esetén $\bar{w}(Q) < \bar{w}(T)$.

3.9. Megjegyzés: Az $|X| \geqq 1$ feltétel nem jelent megszorítást, ugyanis, ha $|X|=0$ akkor $H=(X; \mathcal{E})$ az üres hipergráf, amelynek nincs kvázi komponense, hiszen a kvázi komponens nem lehet üres halmaz.

3.10. Megjegyzés: Az egy pontot tartalmazó élek egyik vágásnak sem lehetnek elemei, ezért a kvázi komponensek meghatározásánál a $H=(X;\mathscr{E})$ hipergráf élei közül *elhagyhatók*. Az elhagyás után adódhatnak olyan pontok, amelyeket egyetlen él sem tartalmaz. Ezek nyilván más ponthoz nem kapcsolódtak, ezért egyetlen nem triviális kvázi komponensnek sem elemei (2.14. megjegyzés).

Tehát a kvázi komponensek meghatározása előtt elhagyhatók, tárolva azt az információt, hogy triviális kvázi komponensek.

3.11. Megjegyzés: A $H=(X;\mathscr{E})$ hipergráfban — a 3.10. megjegyzésben leírt redukció után — *összevonhatjuk* azokat *a pontokat, amelyeket ugyanazon élek tartalmazzák*. Vezessük be a $T(\mathscr{F})=\{x|x\in X, \mathfrak{E}(\{x\})=\mathscr{F}\}$ jelölést. Nyilvánvaló, hogy $|T(\mathscr{F})|\geqq 2$. A kvázi komponens definíciójából következik, hogy — tetszőleges $v(E_j)>0$ $(j=1, \ldots, m)$ súlyfüggvény esetén — a $T(\mathscr{F})$ ponthalmaz akkor és csak akkor kvázi komponens, ha $\exists E_j\in\mathscr{F}$ amelyre $\mathscr{H}(E_j)=T(\mathscr{F})$ teljesül. A 2.14. megjegyzés alapján evidens, hogy $P\in\mathscr{P}_{T(\mathscr{F})}$ halmaz nem lehet kvázi komponens.

A $T(\mathscr{F})$ halmaz pontjait egy pontba vonjuk össze, tárolva azt az információt, hogy valamennyien triviális kvázi komponensek voltak és esetleg $T(\mathscr{F})$ maga is kvázi komponens volt.

3.12. Megjegyzés: Korábban megmutattuk, hogy egy nem összefüggő *hipergráf kvázi komponenseit úgy határozhatjuk meg,* hogy rendre *megkeressük a komponensei által kifeszített rész hipergráfok kvázi komponenseit.*

Tehát az *F5 feladat* megoldása — célszerűen a 3.10. megjegyzésben és a 3.11. megjegyzésben leírt redukciók elvégzése után — az *R1 komponenskereső rutin* alkalmazásával a komponensek számával megegyező számú *F6* feladat megoldására redukálódott.

F6: Adott a $H=(X;\mathscr{E})$ $(|X|\geqq 2)$ *összefüggő* hipergráf és $v(E_j)>0$ $(j=1, \ldots, m)$ a hipergráf élein értelmezett pozitív függvény. Határozzuk meg a $H(X;\mathscr{E})$ hipergráf összes kvázi komponensét, azaz azokat a Q $(\emptyset\neq Q\subseteq X)$ ponthalmazokat, amelyekre teljesül az, hogy bármely $T\in\mathscr{P}_Q$ esetén $\overline{w}(Q)<\overline{w}(T)$.

Az F6 feladat megoldására szolgáló algoritmus

Jelölje $\mathscr{Q}_j=\{Q_1, \ldots, Q_{t_j}\}$ az algoritmus j-edik lépése előtt meghatározott kvázi komponensek halmazát.

Jelölje $\mathscr{K}_j=\{K_j^{(l_j,h_j)}, \ldots, K_{v_{j-1}}^{(l_{v_{j-1}},h_{v_{j-1}})}\}$ az eljárás j-edik lépése előtt azon ponthalmazok rendezett osztályát, amelyekről az eljárás j-edik és további lépéseiben döntjük el, hogy kvázi komponensek-e vagy sem.

Az eljárás akkor ér véget, ha a j-edik lépésben konstruált $\mathscr{K}_{j+1}$ halmaz üres.

1. lépés: $\mathscr{Q}_1=\emptyset$, $\mathscr{K}_1=\{x\}$. A $H=(X;\mathscr{E})$ hipergráf összefüggő, tehát komponens. Így a 2.12. megjegyzés alapján kvázi komponens is.

Ha $|X|=2$ azaz $X=\{x_1, x_2\}$, akkor a 2.13. megjegyzés alapján $\{x_1\}$ és $\{x_2\}$ is triviális kvázi komponens. A 2.20. megjegyzés alapján a hipergráfnak legfeljebb 3 kvázi komponense lehet. Ezeket már ismerjük. Tehát $\mathscr{Q}_2=\{X, \{x_1\}, \{x_2\}\}$, $\mathscr{K}_2=\emptyset$. Az eljárás véget ért.

Ha $|X|>2$, akkor $S_1=X$ választás mellett az *R2 rutin* felhasználásával meghatározzuk azt a $T_1^*\in\mathscr{P}_{S_1}$ halmazt, amelyre tetszőleges $T\in\mathscr{P}_{S_1}$ halmaz választása esetén $\bar{w}(T_1^*)\leqq\bar{w}(T)$ teljesül.

A 2.12. következmény alapján $H=(X;\mathscr{E})$ összes — X-től különböző — kvázi komponensét T_1^* vagy $X-T_1^*$ tartalmazza.

A nem triviális kvázi komponensek kifeszítő ponthalmazok (2.9. lemma) ezért ezeket vagy az $\mathscr{H}(\mathfrak{E}'((X-T_1^*)|(X-T_1^*)))$ vagy az $\mathscr{H}(\mathfrak{E}'(T_1^*|T_1^*))$ kifeszítő ponthalmazok is tartalmazzák (2.2. lemma).

Tehát az $\mathscr{R}_{11}=X-\{\mathscr{H}(\mathfrak{E}'(T_1^*|T_1^*))\cup\mathscr{H}(\mathfrak{E}'((X-T_1^*)|(X-T_1^*)))\}$ ponthalmaz csak triviális kvázi komponenseket tartalmaz. A 2.13. megjegyzés alapján $\mathscr{R}_{11}$ valamennyi pontja kvázi komponens.

Ha $\mathscr{R}_{11}=X$, akkor $H=(X;\mathscr{E})$ az X halmazon kívül nem tartalmaz másik nem triviális kvázi komponenst. Tehát $\mathscr{Q}_2=\{X\}\cup\mathscr{R}_{11}$, $\mathscr{K}_2=\emptyset$. Az eljárás véget ért.

Ha $\mathscr{R}_{11}\subset X$, akkor vagy $\mathscr{H}(\mathfrak{E}'(T_1^*|T_1^*))\neq\emptyset$ és/vagy $\mathscr{H}(\mathfrak{E}'((X-T_1^*)|(X-T_1^*)))\neq\emptyset$.

Mivel a nem triviális kvázi komponensek által kifeszített rész hipergráfok összefüggőek, (2.17. megjegyzés) ezért ezeket $H_{T_1^*}$ vagy $H_{X-T_1^*}$ komponensei is tartalmazzák. Határozzuk meg az *R1* rutin segítségével $H_{T_1^*}$ és $H_{X-T_1^*}$ komponenseit.

Az így kapott komponensek közül az 1 eleműek halmazát jelölje $\mathscr{R}_{12}$. Ezek nyilván csak triviális kvázi komponenst tartalmaznak. A 2.13. megjegyzés alapján az $\mathscr{R}_{12}$ halmaz valamennyi pontja kvázi komponens.

Ha $\mathscr{R}_{11}\cup\mathscr{R}_{12}=X$, akkor a $H=(X;\mathscr{E})$ hipergráf az X halmazon kívül nem tartalmaz másik nem triviális kvázi komponenst. Tehát $\mathscr{Q}_2=\{X\}\cup\mathscr{R}_{11}\cup\mathscr{R}_{12}$, $\mathscr{K}_2=\emptyset$. Az eljárás véget ért.

Ha $\mathscr{R}_{11}\cup\mathscr{R}_{12}\subset X$, akkor jelölje $H_{T_1^*}$ legalább két pontot tartalmazó komponenseit rendre $K_1^{(1,0)}, K_2^{(1,0)}, \ldots, K_{i_1}^{(1,0)}$, $H_{X-T_1^*}$ legalább két pontot tartalmazó komponenseit rendre $K_{i_1+1}^{(1,1)}, \ldots, K_{i_1+l}^{(1,l)}, \ldots, K_{v_1}^{(1,v_1-i_1)}$. Jelölje a $K^{(1,l)}$ komponens által generált vágás értékét — a komponens sorszámától (alsó index) függetlenül w_{1l} ($l=0, \ldots, \ldots, v_1-i_1$).

Ez a jelölés korrekt, mivel a $H_{T_1^*}$ hipergráf komponensei által generált vágások értéke megegyezik $\bar{w}(T_1^*)$-gal (2.19. tétel) azaz $\bar{w}(T_1^*)=\bar{w}(K_1^{(1,0)})=\ldots=\bar{w}(K_{i_1}^{(1,0)})=w_{1,0}$.

A ponthalmaz által generált vágás definíciója értelmében $\bar{w}(T_1^*)=\bar{w}(X-T_1^*)$.

Tehát az $X-T_1^*$ halmazra is alkalmazhatjuk a 2.19. tételt: $\bar{w}(X-T_1^*)=\bar{w}(K_{i_1+1}^{(1,1)})=\ldots=\bar{w}(K_{v_1}^{(1,v_1-i_1)})$. A $w(T_1^*)=\bar{w}(X-T_1^*)$ egyenlőség felhasználásával $w_{1,0}=w_{1,1}=\ldots=w_{1,v_1-i_1}$ adódik.

Legyen

$$\mathscr{K}_2=\{K_1^{(1,0)}, \ldots, K_{v_1}^{(1,v_1-i_1)}\},$$

$$\mathscr{Q}_2=\{x\}\cup\mathscr{R}_{11}\cup\mathscr{R}_{12}.$$

j. lépés: ($j\geqq 2$) $\mathscr{Q}_j=\{Q_1, \ldots, Q_{t_j}\}$ $\mathscr{K}_j=\{K_j^{(l_j,h_j)}, \ldots, K_{v_{j-1}}^{(l_{v_{j-1}},h_{v_{j-1}})}\}$. Nyilvánvaló, hogy $\mathscr{K}_j\neq\emptyset$, ugyanis ellenkező esetben az eljárás már a j-ik lépés előtt véget ért volna.

Legyen $K_j^{(l_j,h_j)}$ a K_j rendezett halmaz első eleme.

Az előző lépésekből evidens, hogy $|K_j^{(l_j,h_j)}|\geqq 2$. A $K_j^{(l_j,h_j)}$ halmaz által generált vágás értékét, w_{l_j,h_j}-t, már az l_j-edik lépésben meghatároztuk ($l_j<j$).

Ha $|K_j^{(l_j,h_j)}|=2$, azaz $K_j^{(l_j,h_j)}=\{x_{j_1}, x_{j_2}\}$, akkor $\mathscr{P}_{K_j^{(l_j,h_j)}}=\{x_{j_1}, x_{j_2}\}$. Határ-

rozzuk meg a $\overline{w}(\{x_{j_1}\})$ és a $\overline{w}(\{x_{j_2}\})$ értékeket. A kvázi komponens definíciója szerint

— ha $w_{l_j,h_j}<\overline{w}(\{x_{j_1}\})$ és $w_{l_j,h_j}<\overline{w}(\{x_{j_2}\})$, akkor $K_j^{(l_j,h_j)}$ kvázi komponens. Tehát $\mathcal{Q}_{j+1}=\mathcal{Q}_j\cup\{K_j^{(l_j,h_j)}\}\cup\{x_{j_1}\}\cup\{x_{j_2}\}$,

— ha a fenti két egyenlőtlenség legalább egyike nem teljesül, akkor $K_j^{(l_j,h_j)}$ nem kvázi komponens, tehát $\mathcal{Q}_{j+1}=\mathcal{Q}_j\cup\{x_{j_1}\}\cup\{x_{j_2}\}$.

Mivel $K_j^{(l_j,h_j)}$ más nem triviális kvázi komponenst nem tartalmazhat, ezért mindkét esetben $\mathcal{K}_{j+1}=\mathcal{K}_j-\{K_j^{(l_j,h_j)}\}$. Ha $\mathcal{K}_{j+1}=\emptyset$, akkor az eljárás véget ért.

Ha $|K_j^{(l_j,h_j)}|>2$, akkor $S_j=K_j^{(l_j,h_j)}$ választás mellett az *R2 rutin* felhasználásával meghatározzuk azt a $T_j^*\in\mathcal{P}_{S_j}$ halmazt, amelyre tetszőleges $T\in\mathcal{P}_{S_j}$ halmaz választása esetén $\overline{w}(T_j^*)\leqq\overline{w}(T)$ teljesül.

Azt, hogy a $K_j^{(l_j,h_j)}$ halmaz kvázi komponens-e, a 2.24. megjegyzés alapján, az l_j-edik lépésben meghatározott w_{l_j,h_j} érték és $\overline{w}(T_j^*)$ összehasonlításával döntjük el.

Ha $w_{l_j,h_j}<\overline{w}(T_j^*)$, akkor $K_j^{(l_j,h_j)}$ kvázi komponens. Legyen ez esetben $\mathcal{R}_{j_0}=\{K_j^{(l_j,h_j)}\}$.

Ha $w_{l_j,h_j}=\overline{w}(T_j^*)$, akkor $K_j^{(l_j,h_j)}$ nem kvázi komponens. Legyen ez esetben $\mathcal{R}_{j_0}=\emptyset$.

A 2.12. következmény alapján a $H=(X;\mathcal{E})$ hipergráfnak az $S_j=K_j^{(l_j,h_j)}$ halmaz által valódi részként tartalmazott összes kvázi komponensét T_j^* vagy $S_j-T_j^*$ tartalmazza.

Ezek közül a nem triviális kvázi komponensek kifeszítő ponthalmazok (2.9. lemma) ezért ezeket vagy az $\mathcal{H}(\mathfrak{E}'(T_j^*|T_j^*))$ vagy az $\mathcal{H}(\mathfrak{E}'((S_j-T_j^*)|(S_j-T_j^*)))$ kifeszítő ponthalmazok is tartalmazzák (2.2. lemma). Tehát az $\mathcal{R}_{j_1}=K_j^{(l_j,h_j)}-\mathcal{H}(\mathfrak{E}'(T_j^*|T_j^*))\cup\mathcal{H}(\mathfrak{E}'((S_j-T_j^*)|(S_j-T_j^*))))$ ponthalmaz csak triviális kvázi komponenseket tartalmaz.

A 2.13. megjegyzés alapján $\mathcal{R}_{j_1}$ valamennyi pontja kvázi komponens.

Ha $\mathcal{R}_{j_1}=K_j^{(l_j,h_j)}$, akkor $K_j^{(l_j,h_j)}$ valódi részként csak triviális kvázi komponenst tartalmaz, tehát $\mathcal{Q}_{j+1}=\mathcal{Q}_j\cup\mathcal{R}_{j_0}\cup\mathcal{R}_{j_1}$ $\mathcal{K}_{j+1}=\mathcal{K}_j-\{K_j^{(l_j,h_j)}\}$. Ha $\mathcal{K}_{j+1}=\emptyset$, akkor az eljárás véget ért.

Ha $\mathcal{R}_{j_1}\subset K_j^{(l_j,h_j)}$, akkor vagy $\mathcal{H}(\mathfrak{E}'(T_j^*|T_j^*))\neq\emptyset$ és/vagy $\mathcal{H}(\mathfrak{E}'((S_j-T_j^*)|(S_j-T_j^*)))\neq\emptyset$.

Mivel a nem triviális kvázi komponensek által kifeszített rész hipergráfok összefüggőek, (2.17. megjegyzés), ezért ezeket $H_{T_j^*}$ vagy $H_{S_j-T_j^*}$ komponensei is tartalmazzák.

Határozzuk meg az *R1 rutin* segítségével $H_{T_j^*}$ és $H_{S_j-T_j^*}$ komponenseit.

Az így kapott komponensek közül az 1 eleműek halmazát jelölje $\mathcal{R}_{j_2}$. Ezek nyilván nem tartalmaznak nem triviális kvázi komponenst, és maguk triviális kvázi komponensek (2.13. megjegyzés).

Ha $\mathcal{R}_{j_1}\cup\mathcal{R}_{j_2}=K_j^{(l_j,h_j)}$, akkor $K_j^{(l_j,h_j)}$ valódi részként csak triviális kvázi komponenst tartalmaz. Tehát $\mathcal{Q}_{j+1}=\mathcal{Q}_j\cup\mathcal{R}_{j_0}\cup\mathcal{R}_{j_1}\cup\mathcal{R}_{j_2}$ $\mathcal{K}_{j+1}=\mathcal{K}_j-\{K_j^{(l_j,h_j)}\}$. Ha $\mathcal{K}_{j+1}=\emptyset$, akkor az eljárás véget ért.

Ha $\mathcal{R}_{j_1}\cup\mathcal{R}_{j_2}\subset K_j^{(l_j,h_j)}$, akkor jelölje $H_{T_j^*}$ legalább két pontot tartalmazó komponenseit rendre $K_{v_{j-1}+1}^{(j,0)},\ldots,K_{v_{j-1}+i_j}^{(j,0)}$; $H_{S_j-T_j^*}$ legalább két pontot tartalmazó komponenseit rendre $K_{v_{j-1}+i_j+1}^{(j,1)},\ldots,K_{v_j}^{(j,v_j-i_j)}$.

Jelölje a $K^{(j,l)}$ komponens által generált vágás értékét — a komponens sorszámától (alsó index) függetlenül — $w_{j,l}$. $(l=0,\ldots,v_j-i_j)$

Ez a jelölés korrekt, mivel a $H_{T_j^*}$ hipergráf komponensei által generált vágások értéke megegyezik $\bar{w}(T_j^*)$-gal (2.19. tétel) azaz $\bar{w}(T_j^*)=\bar{w}(K^{(j,0)}_{v_{j-1}+1}=\ldots=\bar{w}(K^{(j,0)}_{v_{j-1}+i_j})=$ $=w_{j,0}$.

A $H_{S_j-T_j^*}$ hipergráf komponensei által generált vágások értékei nem ismertek, ezért ezeket meghatározzuk.

Legyen

$$\mathscr{K}_{j+1}=\mathscr{K}_j-\{K_j^{(l_j,h_j)}\}\cup\{K^{(j,0)}_{v_{j-1}+1},\ldots,K^{(j,v_j-i_j)}_{v_j}\}=\{K_{j+1}^{(l_{j+1},h_{j+1})},\ldots,K_{v_j}^{(l_{v_j},h_{v_j})}\}$$

$$(l_{v_j}=j),\ (h_{v_j}=v_j-i_j)$$

$$\mathscr{Q}_{j+1}=\mathscr{Q}_j\cup\mathscr{R}_{j0}\cup\mathscr{R}_{j1}\cup\mathscr{R}_{j2}.$$

3.5. TÉTEL. Az *Fő Feladat* megoldására, azaz az összefüggő $H=(X;\mathscr{E})$ hipergráf összes kvázi komponensének meghatározására szolgáló algoritmus legfeljebb $|X|-1$ számú lépésben véget ér. ($|X|\geqq 2$.) Az utolsó, r-edik lépésben kapott $\mathscr{Q}_{r+1}$ halmaz tartalmazza a $H=(X;\mathscr{E})$ hipergráf összes kvázi komponensét.

Bizonyítás: $|X|=n$ szerinti teljes indukcióval

a) Ha $n=2$, akkor az eljárás az 1. lépésben véget ér, tehát $r\leqq n-1$ igaz.

b) Ha $n\geqq 3$, akkor elvégezzük az algoritmus 1. lépését.

Ha $\mathscr{R}_{11}=X$ vagy $\mathscr{R}_{11}\cup\mathscr{R}_{12}=X$, akkor az algoritmus már az 1. lépésben véget ér, tehát $r\leqq n-1$ igaz. Ha $\mathscr{R}_{11}\cup\mathscr{R}_{12}\subset X$, akkor jelölje a kapott legalább 2 elemű komponensek elemszámát rendre $n_1,\ldots,n_{v_1}$. Nyilván $\sum_{i=1}^{v_1} n_i\leqq n$.

Az X által tartalmazott, az első lépésben nem meghatározott kvázi komponensek a fenti komponensek által kifeszített rész hipergráfoknak is kvázi komponensei. Indukciós feltevésünk értelmében (ez alkalmazható, mivel a kapott komponensek legalább két eleműek) a komponensek által kifeszített hipergráfok kvázi komponenseit legfeljebb $n_1-1,\ldots,n_{v_1}-1$ számú lépésben határozhatjuk meg.

Tehát az algoritmus lépésszáma legfeljebb $\sum_{i=1}^{v_1}(n_i-1)+1$, ugyanis az 1. lépést már elvégeztük.

Ha $\sum_{i=1}^{v_1} n_i=n$, akkor $v_1\geqq 2$, mivel ez esetben $\mathscr{H}(\mathfrak{E}'(T_1^*|T_1^*))\neq\emptyset$ és $\mathscr{H}(\mathfrak{E}'((X-T_1^*)|(X-T_1^*)))\neq\emptyset$ kellett legyen. Tehát $\sum_{i=1}^{v_1}(n_i-1)+1\leqq\sum_{i=1}^{v_1} n_i-1\leqq n-1$.

Ha $v_1=1$, akkor $n_1<n$, mivel ez esetben vagy $\mathscr{H}(\mathfrak{E}'(T_1^*|T_1^*))=\emptyset$ vagy $\mathscr{H}(\mathfrak{E}'((X-T_1^*)|(X-T_1^*)))=\emptyset$ teljesült. Tehát $(n_1-1)+1\leqq n-1$. Tehát $r\leqq n-1$ igaz.

Legyen a $Q(\emptyset\neq Q\subseteq X)$ halmaz a $H=(X;\mathscr{E})$ hipergráfnak tetszőleges kvázi komponense.

Legyen $K_j^{(l_j,h_j)}$ az $\bigcup_{i=1}^{r}\mathscr{K}_i$ halmaz elemei közül az a legszűkebb halmaz, amely Q-t tartalmazza.

Ilyen legszűkebb halmaz biztos van, mert $X\in\mathscr{K}_1$ miatt $X\in\bigcup_{i=1}^{r}\mathscr{K}_i$.

Jelölje $\mathscr{K}_j$ azt a rendezett halmazt, amelynek $K_j^{(l_j,h_j)}$ az első eleme.

Ha $Q \subseteq K_j^{(l_j,h_j)}$, akkor vagy $Q \in \mathcal{Q}_{j+1}$ vagy van olyan $K \in \{K_{v_{j-1}+1}^{(j,0)}, \ldots, K_{v_j}^{(j,v_j-i_j)}\}$ halmaz, amely Q-t tartalmazza. Mivel $K \subset K_j^{(l_j,h_j)}$ és $K \in \bigcup_{i=1}^{r} \mathcal{K}_i$, ezért az utóbbi lehetőség ellentmondásra vezet. Tehát $Q \in \mathcal{Q}_{j+1}$ és $j \leqq r$. $\mathcal{Q}_j \subseteq \mathcal{Q}_{j+1}$ $(j=1, \ldots, r)$ miatt $Q \in \mathcal{Q}_{r+1}$.

4. Az új cluster technika és alkalmazása

A cluster technika és programcsomagja

A cluster analízis gráf és hipergráf modelljei, a hipergráf kvázi komponensének fogalma, és a kvázi komponensek megkeresésére szolgáló eljárás együttesen alkotja az új cluster technikát.

Ennek alapköve a kvázi komponens fogalmán alapuló cluster definíció.

4.1. DEFINÍCIÓ. Az objektumok *cluster*jei az *1. modell* gráfjának, illetve a *3. modell* hipergráfjának kvázi komponensei. A deszkriptorok *cluster*jei a *2. modell* hipergráfjának kvázi komponensei.

Az így definiált clusterek legfontosabb tulajdonságai a következők:

— ha a Q halmaz clustere a H objektum vagy deszkriptor halmaznak, akkor Q-nak bármely valódi részhalmaza „erősebben kapcsolódik" a Q halmazhoz, mint annak környezetéhez. (2.10. tétel),
— ha a Q halmaz clustere a H objektum vagy deszkriptor halmaznak, akkor Q clustere H minden olyan részhalmazának is, amely a Q halmazt tartalmazza (2.16. megjegyzés),
— a clusterek vagy diszjunktak vagy tartalmazzák egymást (2.13. következmény),
— a H halmaz valamennyi egy elemű részhalmaza cluster (2.13. megjegyzés),
— a H halmaz clustereinek száma kisebb, mint a H halmaz elemei számának kétszerese (2.20. megjegyzés).

A cluster definíció egyenes következménye, hogy a clusterek meghatározása a kvázi komponensek megkeresésére szolgáló algoritmuson alapul.

Ennek ismeretében már vázolhatók az új cluster eljárás legfontosabb jellemzői: *hierarchikus* jellegű, *divizív, konvergens.*

A fentiek alapján az új cluster technika jellemző vonásai a következőkben foglalhatók össze:

a) Lehetővé teszi *kettőnél több* objektum vagy deszkriptor hasonlóságának *közvetlen modellezését,* sőt a *kapcsolatok súlyozás*át is.

b) *Elkerülhető* a *hasonlósági mérőszámok* megszerkesztése. (A statisztikai cluster analízisnél ezt a deszkriptorok definiálása váltja fel.)

c) *Az explicit cluster definíció* módot nyújt a clusterek tulajdonságainak előzetes vizsgálatára.

d) A cluster kereső eljárás *hierarchikus, divizív jellegű, konvergens és hatékony.* Az összes clustert és csak azokat szolgáltatja az objektumok és deszkriptorok sorrendjétől függetlenül.

e) Az objektumok és deszkriptorok modellezése és a clusterek megkeresése is ugyanazzal a módszerrel történik; az alkalmazott *modellek egymás duálisai.*

Az új cluster technika számítógépes programcsomagja assembler és FORTRAN modulokból épül fel és IBM 370, valamint R20 számítógépen fut.

A programcsomag feladata egyrészt a cluster analízis hipergráf modelljeinek megszerkesztése, másrészt a kvázi komponensek, azaz a clusterek meghatározása.

A programcsomag vagy a teljes objektum halmazra, vagy annak *tetszőleges részhalmaz*ára fogja elvégezni a clusterelemzést. Szükség esetén a felhasználó a *deszkriptorok*at tetszőlegesen súlyozhatja. A programcsomag jelenlegi kiépítettségében legfeljebb 500 objektum és 5000 deszkriptor (legfeljebb 1000 különböző) feldolgozására képes.

A feldolgozás eredményeként a felhasználó egyrészt magát a teljes *cluster hierarchiát* és a clusterekbe rendezett objektumok nevét, másrészt a clusterek *megkeresésének folyamatát* kapja kézhez.

A programcsomag gyakorlati alkalmazásánál (4.2 rész) szerzett számítástechnikai tapasztalatokat összegezi a következő táblázat.

Jelölések:

D = objektumok száma
N_d = átlagos deszkriptor szám objektumonként
T^x = legalább két objektumot jellemző deszkriptorok száma
N_k = komponensek száma
$P_{\max}$ = a nem egy elemű komponens(ek) elemszáma
$F4$ = az elvégzett $F4$ kereslet-kínálat feladatok száma
$F1$ = az elvégzett $F1$ minimális kettévágási feladatok száma
N_{cl} = a nem triviális clusterek száma.

Feladat jele	D	N_d	T^x	N_k	P_{max}	$F4$	$F1$	N_{cl}	Cluster keresés IBM 370 gépidő	Virtuális memória
20	63	6	54	3	61	714	38	9	3′ 15.17″	318 K
20*	63	6	54	3	61	635	39	13	2′ 25.78″	318 K
30	82	6	86	1	82	990	56	6	3′ 19.61″	318 K
30*	82	6	86	1	82	1007	58	12	3′ 24.34″	318 K
40	56	5—6	63	4	53	421	37	11	1′ 09.89″	318 K
50	90	6	109	5	86	849	66	17	5′ 38.26″	320 K

A fenti feladatoknál az adatkezelő programok futási ideje 30″ alatt volt, tehát a cluster keresés időigényéhez képest elenyésző.

A 20 és 20* valamint a 30 és 30* feladat modellje csak abban különbözik, hogy a *-gal nem jelölt feladatoknál minden deszkriptorhoz 1 súlyt, míg a *-gal jelölt feladatoknál minden deszkriptorhoz 1-től 5-ig terjedő súlyokat rendeltünk. Látható, hogy ezeknél a feladatoknál a deszkriptorok súlyozása nem módosította lényegesen az algoritmus futási idejét.

A $P_{\max}$ és az $F1$ jelű oszlopok számadatainak egybevetéséből kitűnik, hogy az *R1 komponens kereső rutin* használata eredményes volt (ugyanis nélküle $P_{\max} = F1 + 1$ lenne). A táblázat adatainak részletesebb elemzése azt mutatja, hogy az algoritmus futtatási ideje a feladat méretein kívül erősen függ a hipergráf modell szerkezetétől is.

A cluster technika gyakorlati alkalmazása

A hipergráf modellen alapuló cluster technika lehetővé teszi olyan objektumrendszerek szerkezetének feltárását, amelyekben az *objektumok*at elsősorban a *kvalitatív jellemzők* és/vagy a *deszkriptorok* jellemzik. (A cluster analízissel vizsgált — közgazdasági, orvosbiológiai, információs stb. rendszerek jelentős hányada ilyen tulajdonságú.)

A következőkben az új technika alkalmazását mutatjuk be a kutatásirányítás területén.

Az *Építéstudományi Intézetben* 1977. január 1. óta üzemel a TPA/i *számítógéppel segített kutatásirányítási rendszer,* amely felhasználja az új cluster technikát.

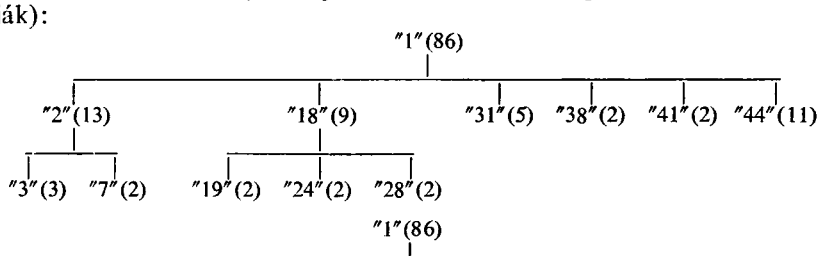
A sikeres kutatásirányítás alapvető feltétele a *kutatói szervezetek* tevékenységi körének és a nemrég befejezett, folyamatban levő és tervezett *kutatások tematikájának* naprakész ismerete. Ez utóbbi célból (1976. január 1. óta) a kutatási témák (objektumok) többségéről a TPA/i *adatbankja* hétre kész információt tárol. Minden egyes téma tárgyát, célját, módszerét az *Intézetben* működő építéstudományi és információtudományi szakemberekből álló team képezi le tárgyszavakra — az állandóan bővülő, jelenleg 4000 tárgyszót tartalmazó építéstudományi tezaurusz felhasználásával.

A szervezeti egységek tényleges kutató tevékenysége feltérképezésének, a helyzetelemzésnek kulcskérdése a *tematikai centrumok* ismerete, amelyeket az új cluster technika számítógépes programjai segítségével határoztunk meg.

A tematikai centrumok, azaz a clusterek jellemzését megkönnyítette, hogy az új *cluster technika* nemcsak a clustereket és azok objektumait, hanem a *clustert* egyértelműen *leíró tárgyszavakat is megadja,* sőt a szóban forgó clustert a környezetével (a többi clusterrel, illetve objektummal) összekapcsoló deszkriptorokat is.

A továbbiakban bemutatjuk az *Intézet* egyik főosztályán az *Építésgépesítési és Technológiai Tagozaton* végzett kutatómunka struktúráját:

A *Tagozat* témáinak száma 90 volt, jellemzésükre átlagosan 6 tárgyszót használtak. A legalább két témát jellemző tárgyszavak száma 109 volt. A hipergráf modell négy 1 elemű komponenst és egy 86 eleműt tartalmazott, amelynek nem triviális kvázi komponenseinek (a feladat legalább két elemű clustereinek) hierarchiáját mutatja a következő ábra (a zárójelbe írt számok a megfelelő cluster elemszámát mutatják):



A „2" cluster fogja össze a vasbetonnal, vasbeton szerkezetekkel (ezen belül elsősorban az UNIVÁZ szerkezettel) és ezek szereléstechnológiájával kapcsolatos kutatásokat.

A „3" clusterbe kerültek a közműalagutakkal, közműfolyosókkal és általában a közművesítéssel kapcsolatos témák.
A „7" cluster témái a vasbeton fúrásával kapcsolatos kísérletekkel foglalkoznak.
A „18" cluster fogja össze az építőipari és építőanyagipari szállítással, tárolással, a központi telepek létesítésével kapcsolatos témákat.
Ezen belül:
a „19" cluster témái a konténeres szállítással,
a „28" cluster témái anyagmozgatási tanácsadással,
a „24" cluster témái nagy panelok, födémelemek szállításával foglalkoznak.
A „31" cluster gipsz felhasználásával készülő válaszfallapok és pallók gyártásának beindításával foglalkozó kutatásokat fogja össze.
A „38" cluster témái a hulladék kezelés és a főzés energiaigényét vizsgálják.
A „41" cluster a lakótelepek energiaellátásával és közművesítésével kapcsolatos kutatásokat fogja össze.
A „44" cluster témái a különböző építésmódok, technológiák fejlesztésével foglalkoznak. Vizsgálják az öntött, a vázpanelos, a zsaluemeléses, a könnyűszerkezetes építésmódot és a melegbeton és habbeton technológiát.
Az „56" cluster a lézer technológia alkalmazásával foglalkozik a földalatti vasútépítésnél, bányászatban és útépítésnél. Ide tartoznak a vasútépítés, útépítés és a bányászat technológiai fejlesztését szolgáló új gépek és berendezések kidolgozásával kapcsolatos témák is.
A „68" cluster témái a távfűtővezetékek vizsgálatával és elsősorban bitumoperlites hőszigetelésével foglalkoznak.
A „92" cluster fogja össze a befejező munkák gépesítésével foglalkozó kutatásokat.
A „98" clusterbe kerültek a csatornacső gyártási technológiák vizságlatával foglalkozó témák.
A „101" cluster témái acélhuzalok vizsgálatával foglalkoznak.

A témák jellemzésére általában jól megválasztott, találó tárgyszavakat használtak. Ezért a *cluster analízis* a deszkriptorok súlyozása nélkül is jól *feltárta* a kutatási rendszer *tematikai centrum*ait, annak ellenére, hogy ennek a tagozatnak igen *szerteágazó a tevékenysége.*

A homonimák okozta *veszély*ekre hívja fel a figyelmet a „2" cluster. Ugyanis a „vasbeton" és a „fúrás" tárgyszóval összekapcsolja a „3" clusterbe tartozó közművesítési és a „7" clusterbe tartozó vasbeton fúrási témákat. Ennek oka, hogy a „fúrás" szót „3" clusternél és a „7" clusternél más-más értelemben használják. A „18" cluster nemcsak a szállítással kapcsolatos kutatásokat fogja össze, hanem megmutatja annak *közvetlen kapcsolat*át a komplex építőipari telepek létesítésével. Érdekes a „44" cluster, amely egy *új technológia* (lézer) *révén fog össze* viszonylag távol esőnek vélt területeket. Ki kell emelni még az „56" clustert, amely az építésmódok és technológiák fejlesztési feladatai közötti *kapcsolatok*ra hívja fel a figyelmet.

A kutatásirányítási alkalmazás eredményeit összefoglalva megállapíthatjuk, hogy a hipergráf modellen alapuló *cluster technika lehetővé teszi* nagy kutatási témahalmazok *tematikai centrum*ainak, és azok *hierarchiájának feltárását.* Ennek révén valóban *átfogó képet* nyerhetünk az adott kutatási területekről. Természetesen ez a kép nem egyezik meg teljesen a kutatási egységek hosszú távú feladataival. A *cluster analízis* alkalmazásának további *előnye,* hogy *képes feltárni* az egymástól *távol*

esőnek képzelt *kutatási területek kapcsolatait* és a *kialakuló új kutatási területeket* is. *Felhívja a figyelmet a tárgyszavazás hibáira,* a nem releváns tárgyszavak, és a homonimák alkalmazásának veszélyeire.

IRODALOM

[1] ADAMSON, G. W. AND BOREHAM, J., "The use of an association measure based on character structure to identify semantically related pairs of words and document titles". *Inform. Stor. Retr.* **10** (1974) 253—260.

[2] ANDERBERG, M. R., *Cluster Analysis for Applications* (Academic Press, 1973).

[3] AUGUSTSON, J. C. AND MINKER, J., "An analysis of some graph theoretical cluster techniques", *Journal of A. C. M.* **17** (1970) 571—588.

[4] BALAS, E. AND PADBERG, M., "On the set covering problem: II. An algorithm for set partitioning", *Op Res.* **23** (1975) 74—90.

[5] BENEDIKT, V., KELEMEN, K., PINTÉR, ZS. ÉS VÁRI, P.-NÉ, „Cluster analízis és lényegkiemelő eljárás-rendszer terve", *SZÁMKI* (1976).

[6] BERGE, C., *Graphs and Hypergraphs* (North Holland/American Elsevier, 1973).

[7] BOULTON, P. M. AND WALLACE, C. S., "An information measure for single link classification", *Comp. Journ.* **18** (1975) 236—238.

[8] DIDAY, E. AND SCHROEDER, A., "A new approach in mixed distributions detection", *IRIA, Rapport de Recherche,* No. 52., 1974.

[9] EDMONDS, J. AND KARP, R., "Theoretical improvements in algorithmic efficiency for network flow problems"; *Journal of A. C. M.* **19** (1972) 248—264.

[10] FORGY, E. W., "Classification so as to relate to outside variables", *Final Report, Conf. Cluster Analysis of Multivariate Data,* (13.01—13.12.), Cath. Univ. America, 1966.

[11] FRITZ, J., „Tanuló algoritmusok alkalmazása az alakfelismerésben", *MTA MKI,* 1975.

[12] FUTÓ, P., "Computer aided management of industrial research — the method LOGEL", 12. Progr. Op. Res., North Holland, 1976, 353—371.

[13] FUTÓ, P., „Hipergráf elméleten alapuló új cluster definíció és technika, I"., *Alk. Mat. Lapok* **3** (1977) 111—129.

[14] FUTÓ, P., „A cluster analízis egy új modellje és algoritmusa", *SZIGMA,* 1977/4 199—220.

[15] GOWER, J. C., "A comparison of some methods of cluster analysis", *Biometrika* **23** (1967) 623—637.

[16] HU, T. C., *Integer Programming and Network Flows,* (Addison-Wesley, Reading, Massachusetts, 1969).

[17] KLAFSZKY, E., *Hálózati folyamatok,* (Bólyai J. Mat. Társ., 1969).

[18] KNUTH, D. E., *The Art of Computer Programming; Vol.* **1**.*: Fundamental Algorithms,* (Addison-Wesley, 1968).

[19] KOVÁCS, L. B., *A diszkrét programozás kombinatorikus módszerei.* (Bólyai J. Mat. Társ., 1969).

[20] KUNSZT, GY., *A tudományos kutatás logikai modellezése és tematikai irányítása* (Akadémiai Kiadó, 1975).

[21] LAWLER, E. L., "Cutsets and partitions of hypergraphs"; *Networks* **3** (1973) 275—286.

[22] LAWLER, E. L., "Algorithms, graphs and complexity", *Networks* **5** (1975) 89—92.

[23] LUCCIO, F. AND SAMI, M., "On the decomposition of networks into minimally interconnected networks", *IEEE Trans. Circuit Theory,* **CT 16** (1969) 184—188.

[24] MACQUEEN, J. B., "Some methods for classification and analysis of multivariate observations," *Proc. Symp. Math. Stat. and Prob.,* **1** (1967) 281—297.

[25] MULLIGAN, G. B. AND CORNEIL, P. G., "Corrections to Bierstone's algorithm for generating clique"; *Journal of A. C. M.* **19** (1972) 244—247.

[26] OSTEEN, R. E., Clique detection algorithms based on line addition and linc removal", *SIAM Juornal Appl. Math.* **26** (1974) 126—135.

[27] SIBSON, R., "SLINK — An optimally efficient algorithm for the single-link cluster method", *Comp. Journ.* **16** (1973) 30—34.

[28] SPARCK-JONES, K., *Automatic Indexing "74"* (Comp. Lab., Univ. of Cambridge, 1974).

[29] SREJDER, JU. A., *Egyenlőség, hasonlóság, rendezés* (Gondolat, 1975).

[30] TANIMOTO, T. T., *An Elementary Mathematical Theory of Classification and Prediction;* (IBM, 1958).

[31] WARD, J. H. "Hierarchical grouping to optimize an objective function", *Journ. Amer. Statist. Assoc.* **58** (1963) 236—244.
[32] WISHART, D., "An algorithm for hierarchical classifications"; *Biometr.* **22** (1969) 165—170.
[33] ZADEH, N., "Theoretical efficiency of the Edmonds—Karp algorithm for computing maximal flows", *Journal of A. C. M.* **19** (1972) 184—192.

(Beérkezett: 1978. január 3.)

FUTÓ PÉTER
ÉPÍTÉSTUDOMÁNYI INTÉZET
1502 BUDAPEST, PF. 71.

NEW CLUSTER DEFINITION AND TECHNIQUE BASED ON HYPERGRAPH THEORY

P. FUTÓ

In a previous paper (*Alkalmazott Matematikai Lapok* 3 (1977)) hypergraph models of cluster analysis and new cluster definition based on the concept of quasi component of a hypergraph have been introduced.

The first part of this paper presents a non-agglomerative convergent polynomially bounded hierarchic cluster algorithm based on the determination of the quasi components of the hypergraph models. Computational experiences and practical applications are described in the second part of the paper.

Alkalmazott Matematikai Lapok 4 (1978) 135—141

NEMLINEÁRIS EGYENLETRENDSZEREK MEGOLDÁSA RENDSZÁMNÖVELÉSSEL

GERGELY JÓZSEF
Budapest

Az [1] dolgozatban egy iterációs eljárást ismertettünk nemlineáris egyenletrendszerek megoldására. A [2] dolgozatban vizsgáltuk a módszer használatát lineáris esetben. A [3] dolgozatban megmutattuk, hogy a vizsgált módszer a *Newton módszer* egy általánosításának tekinthető. Jelen dolgozatban bebizonyítjuk a módszer konvergenciáját.

1. A megoldási módszer

Tekintsük az

$$\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), \ldots, f_n(\mathbf{x}))^T = 0, \quad \mathbf{x} \in R^n, \quad \mathbf{x} = (x_1, \ldots, x_n)^T \tag{1.1}$$

nemlineáris egyenletrendszert. Legyen az (1.1) egyenletrendszer egy megoldása $\mathbf{x}^*$ és legyen S az $\mathbf{x}^*$-nak egy olyan környezete, ahol az $\mathbf{f}(\mathbf{x})$ függvény folytonosan differenciálható. Jelölje $\mathbf{A}(\mathbf{x})$ az $\mathbf{f}(\mathbf{x})$ függvény *Jacobi mátrixát*

$$\mathbf{A}(\mathbf{x}) = \left\| \frac{\partial f_i}{\partial x_j} \right\| = \|a_{ij}(\mathbf{x})\|, \quad i, j = 1, \ldots, n. \tag{1.2}$$

Jelöljük az $\mathbf{A}(\mathbf{x})$ mátrix bal felső k-adik minormátrixát $\mathbf{A}_k(\mathbf{x})$-szel, azaz $\mathbf{A}_k(\mathbf{x}) = \|a_{ij}(\mathbf{x})\|$, $i, j = 1, \ldots, k$. Tegyük fel, hogy a gyök S környezetében az összes minormátrix nemszinguláris. Tegyük fel, hogy ismeretes az $\mathbf{x}^*$ gyök egy kiinduló $\mathbf{x}^0 \in S$ közelítése és legyen

$$\mathbf{A} = \mathbf{A}(\mathbf{x}^0), \quad \mathbf{f} = \mathbf{f}(\mathbf{x}^0), \quad (a_{ij} = a_{ij}(\mathbf{x}^0),\ f_i = f_i(\mathbf{x}^0),\ i, j = 1, \ldots, n).$$

A gyök egy újabb $\mathbf{x}^1$ közelítését $\mathbf{x}^1 = x^0 + \mathbf{v}^n$ alakban keressük, ahol $\mathbf{v}^n$-et a következő módon számítjuk (lásd [3]). Legyen $\mathbf{v}^0 = 0$ és

$$\left.\begin{aligned} \mathbf{v}^k &= (v_1^k, \ldots, v_k^k, 0, \ldots, 0)^T, \\ \mathbf{c}^k &= (c_1^{k-1}, \ldots, c_{k-1}^{k-1}, 1, 0, \ldots, 0)^T \end{aligned}\right\} k = 1, \ldots, n.$$

$k = 1, \ldots, n$-re végrehajtjuk a következő algoritmusban kijelölt számításokat.

I. Algoritmus (a rendszámnövelés algoritmusa)

I/a. Kiszámítjuk a

$$\hat{\mathbf{c}}^{k-1} = -\mathbf{A}_{k-1}^{-1}\,\mathbf{u}^{k-1}$$

vektort, ahol

$$\hat{\mathbf{c}}^{k-1} = (c_1^{k-1}, \ldots, c_{k-1}^{k-1})^T,$$

$$\mathbf{u}^{k-1} = (a_{1k}, \ldots, a_{k-1,k})^T$$

$k-1$ elemű vektorok ($k \geqq 2$);

I/b. Megoldjuk az

$$f_k(\mathbf{x}^0 + \mathbf{v}^{k-1} + \delta v_k \mathbf{c}^k) = 0$$

nemlineáris egyenletet a δv_k ismeretlenre. (A későbbi (4. pontban való) felhasználás céljából legyen $\bar{\mathbf{x}} = \mathbf{x}^0 + \mathbf{v}^{k-1} + \delta v_k \mathbf{c}^k$);

I/c. Legyen

$$\mathbf{v}^k = \mathbf{v}^{k-1} + \delta v_k \mathbf{c}^k.$$

2. Az eltérés függvény minimalizálása

Tegyük fel, hogy az $f_i(\mathbf{x})$ függvények az $\mathbf{x}^*$ pont S környezetében kétszer differenciálhatók és a második parciális differenciálhányadosaik korlátosak. Ekkor az S környezet bármely $\mathbf{x}^0$ pontjára vonatkozó *Taylor soruk* felírható a következő alakban

$$f_i(\mathbf{x}) = f_i(\mathbf{x}^0) + \sum_{j=1}^{n} a_{ij} v_j + O(\|\mathbf{v}\|^2), \quad \mathbf{v} = \mathbf{x} - \mathbf{x}^0. \tag{2.1}$$

Definiáljuk a következő függvényeket

$$F_i(\mathbf{x}) = [f_i(\mathbf{x})]^2, \quad i = 1, \ldots, n,$$

$$F(\mathbf{x}) = \sum_{i=1}^{n} F_i(\mathbf{x}). \tag{2.2}$$

A (2.2) kifejezések alapján az S környezetben

$$F(\mathbf{x}) > 0, \quad \text{ha} \quad \mathbf{x} \neq \mathbf{x}^* \quad \text{és} \quad F(\mathbf{x}^*) = 0.$$

Ezért az $F(\mathbf{x})$ függvény a minimumát az $\mathbf{x} = \mathbf{x}^*$ gyökhelyen veszi fel és az $F(\mathbf{x})$ függvény minimumának közelítő meghatározása egybeesik az $\mathbf{f}(\mathbf{x}) = 0$ egyenletrendszer közelítő megoldásának megkeresésével. Megfogalmazunk egy algoritmust az $F(\mathbf{x})$ függvény minimalizálási feladatára. Az algoritmus n lépésből áll. A k-adik lépésben minimalizáljuk az $F_k(\mathbf{x})$ függvényt azon feltétel mellett, hogy a már minimalizált $F_i(\mathbf{x})$, $i<k$ függvények értéke ne változzék. Az 1. pont jelöléseit használva az algoritmus k-adik lépése a következőképpen fogalmazható meg:

II. Algoritmus

II/a. Keressük a $\mathbf{v}^k$ vektort $\mathbf{v}^k=\mathbf{v}^{k-1}+\delta v_k\mathbf{c}^k$ alakban, ahol a $\mathbf{c}^k$ vektornak (mint az 1. pontból látható) $k-1$ ismeretlen összetevője van. Számítsuk ki ezt a $k-1$ ismeretlent a

$$dF_i = 0, \quad i = 1, \ldots, k-1 \tag{2.3}$$

feltételi egyenletekből. A (2.3) egyenletrendszer $k-1$ homogén egyenletéből a c_i^{k-1}, $i=1, \ldots, k-1$ összetevők kiszámíthatók (a k-adik ismeretlen δv_k kiesik az egyenletekből);

II/b. Minimalizáljuk az $F_k(\mathbf{x}^0+\mathbf{v}^{k-1}+\delta v_k\mathbf{c}^k)$ függvényt a δv_k változó szerint. (Ez ugyanazt jelenti mint az

$$f_k(\mathbf{x}^0+\mathbf{v}^{k-1}+\delta v_k\mathbf{c}^k) = 0$$

nemlineáris egyenlet megoldása a δv_k változóra nézve);

II/c. Legyen

$$\mathbf{v}^k = \mathbf{v}^{k-1}+\delta v_k\mathbf{c}^k.$$

Megmutatjuk, hogy a II. algoritmus lépései ugyanazt végzik, mint az I. algoritmuséi.

Az F_i függvény (2.2)-es képlet szerinti definícióját figyelembe véve a II/a lépés pontosabban a

$$\frac{\partial F_i(\mathbf{x}^0+\mathbf{v}^{k-1}+\delta v_k\mathbf{c}^k)}{\partial(\delta v_k)} = 2\left[\sum_{j=1}^{n} a_{ij}(\mathbf{v}^{k-1}+\delta v_k c_j^k)+f_i\right]\left[\sum_{j=1}^{n} a_{ij}c_j^k\right] = 0$$

$i=1, 2, \ldots, k-1$ egyenletrendszer megoldását jelenti. Minthogy $v_j^{k-1}=0$, ha $j\geqq k$ és $c_j^k=0$, ha $j>k$, azt kapjuk, hogy

$$\left(\sum_{j=1}^{k-1} a_{ij}v_j^{k-1}+f_i\right)\left(\sum_{j=1}^{k} a_{ij}c_j^k\right)+\left(\sum_{j=1}^{k} a_{ij}c_j^k\right)\delta v_k = 0, \quad i = 1, \ldots, k-1. \tag{2.4}$$

A (2.4) egyenletrendszer mindegyik egyenlőségének első tagja zérus, mivel az első tényező teljesíti a $k+1$-edik lépés minimum feltételét. Ezért a (2.4) egyenlőségek második tagjából kapjuk a feltételi egyenleteket a c_j^k komponensek meghatározására, vagyis a

$$\sum_{j=1}^{k} a_{ij}c_j^k = 0, \quad i = 1, \ldots, k-1$$

egyenletrendszert. Ebből a $\hat{\mathbf{c}}^{k-1}=(c_1^{k-1}, \ldots, c_{k-1}^{k-1})^T$ vektor kiszámítható:

$$\hat{\mathbf{c}}^{k-1} = -\mathbf{A}_{k-1}^{-1}\mathbf{u}^{k-1}$$

ugyanúgy mint a I/a lépésben.

A II/b és II/c lépés azonos az I/b, illetve az I/c lépéssel.

Ezzel beláttuk, hogy az I. algoritmus és a II. algoritmus ugyanazt az eredményt szolgáltatja.

3. A módszer konvergenciája

A 2. pont k-adik lépésében minimalizáltuk az F_k függvényt és ugyanakkor a már minimalizált F_i függvények változatlanok maradtak, azaz $dF_i=0$ maradt $i<k$-ra. Ezért a (2.2) definíciót figyelembe véve $k=1, \ldots, n$-re igaz, hogy

$$\sum_{i=1}^{k} F_i(\mathbf{x}^0+\mathbf{v}^k) < \sum_{i=1}^{k} F_i(\mathbf{x}^0+\mathbf{v}^{k-1}),$$

így $k=n$-re az $\mathbf{x}^1=x^0+\mathbf{v}^n$ jelölést használva

$$F(\mathbf{x}^1) < F(\mathbf{x}^0).$$

Ez pedig azt jelenti, hogy $\mathbf{x}^1$ jobb megoldása az (1.1) egyenletrendszernek mint $\mathbf{x}^0$ volt, vagyis az iterációs eljárásunk konvergál.

Foglaljuk össze a konvergencia feltételeit.

F_1. Legyen az $\mathbf{f}(\mathbf{x})$ függvény a gyök közelében kétszer differenciálható és minden második parciális differenciálhányados korlátos;

F_2. Az $\mathbf{f}(\mathbf{x})$ függvény *Jacobi mátrixának* minormátrixai a gyök közelében legyenek nem szingulárisak;

F_3. Minden k-ra legyen megoldható az $f_k(\mathbf{x}^0+\mathbf{v}^{k-1}+\delta v_k \mathbf{c}^k)=0$ egyenlet a δv_k ismeretlenben. A fentiekben bebizonyítottuk a következő tételt.

3.1. TÉTEL. Az F_1, F_2 és az F_3 feltétel mellett az 1. pontban leírt rendszámnöveléses módszer konvergens.

Az F_3 feltétel teljesüléséhez utalunk arra, hogy a $\mathbf{c}^k$ vektor egy irányt jelöl ki a k dimenziós altérben és ennek az iránynak a mentén kell megoldanunk a k-adik egyenletet. Vagyis az F_3 feltétel szemléletesen úgy fogalmazható, hogy a $\mathbf{c}^k$ irány messe az $f_k(\mathbf{x})=0$ egyenlettel adott felületet.

Ha az (1.1) rendszer *Jacobi mátrixáról* feltesszük, hogy az pozitív definit, akkor egyrészt teljesül az F_2 feltétel is, de teljesül az F_3 feltétel is. Ugyanis ismeretes, hogy pozitív definit mátrix bal felső minormátrixai nemszingulárisak, ami éppen az F_2 feltétel követelménye.

Vizsgáljuk meg az F_3 feltétel teljesülését pozitív definit mátrix esetén. Nézzük meg az f_k függvénynek a $\mathbf{c}^k$ irány mentén δv_k-tól függő változását. Ezért differenciáljuk az f_k függvényt δv_k szerint

$$\frac{\partial f_k}{\partial(\delta v_k)} = \sum_{j=1}^{k-1} \frac{\partial f_k}{\partial x_j} c_j^{k-1} \delta v_k + \frac{\partial f_k}{\partial x_k} \delta v_k = \left(\sum_{j=1}^{k-1} a_{kj} c_j^{k-1} + a_{kk} \right) \delta v_k. \tag{3.1}$$

A [2]-es dolgozatban beláttuk, hogy a (3.1) jobb oldalán a zárójelben szereplő mennyiség éppen az eliminációs k-adik lépésében a mátrix főátlójában keletkező mennyiség lesz. Ha a mátrix pozitív definit, akkor ez a mennyiség mindig pozitív. Ebben az esetben viszont mindig megoldható az $f_k=0$ egyenlet a δv_k változó szerint.

Ezek után érvényes a

3.2'. TÉTEL. Ha teljesül az F_1, feltétel és az $\mathbf{f}$ függvény *Jacobi mátrixa* pozitív definit, akkor az 1. pontban leírt algoritmussal számolt iterációs eljárás konvergens.

4. A konvergencia gyorsasága

Módosítsuk az I. algoritmusunk I/b lépését úgy, hogy az

$$f_k(\mathbf{x}^0+\mathbf{v}^{k-1}+\delta v_k \mathbf{c}^k) = 0$$

nemlineáris egyenlet helyett oldjuk meg az f_k függvény lineáris közelítéséből nyert

$$\sum_{j=1}^{k} a_{kj}(v_j^{k-1}+c_j^{k-1}\delta v_k)+a_{kk}\delta v_k = -f_k(\mathbf{x}^0)$$

lineáris egyenletet a δv_k változóra nézve. És legyen ezután $\bar{\bar{\mathbf{x}}}=\mathbf{x}^0+\mathbf{v}^{k-1}+\delta v_k \mathbf{c}^k$.

Vizsgáljuk meg, hogy az I. algoritmus fent leírt módosítása (továbbiakban módosított algoritmus) milyen hatással lesz az F_k függvény minimalizálására. Számítsuk ki az F_k függvény értékét az I/b pontban definiált $\bar{\mathbf{x}}$ és a fenti $\bar{\bar{\mathbf{x}}}$ helyen, azt kapjuk, hogy

$$0 = F_k(\bar{\mathbf{x}}) \leqq F_k(\bar{\bar{\mathbf{x}}}). \tag{4.1}$$

A (4.1) egyenlőtlenségben az egyenlőség csak akkor állhat fenn, ha az $f_k=0$ egyenlet megoldása megegyezik az f_k függvény linearizált közelítése zérushelyével. Ez a helyzet, ha például az f_k függvény lineáris. Ha az f_k függvény nemlineáris és a gyök közelében konvex vagy konkáv, ez elég ahhoz, hogy a szigorú egyenlőtlenség álljon fenn a (4.1) képletben.

Tegyük fel tehát, hogy az f_k függvények közt van legalább egy, olyan nemlineáris, amelyik a gyök közelében konvex vagy konkáv. (Ha minden f_k lineáris, akkor az iterációs eljárás egy lépésben a pontos megoldást adja, lásd [2].) Akkor ebből következik, hogy F függvény értéke kisebb lesz az I. algoritmussal számolt megoldáshelyen, mint a módosított algoritmussal számolt megoldáshelyen. Ez azt jelenti, hogy az I. algoritmus jobb megoldást szolgáltat az (1.1) egyenletrendszer megoldására, mint a módosított algoritmus.

A [3] dolgozatban beláttuk, hogy a módosított algoritmusban leírt iterációs eljárás pontosan megegyezik a *Newton iterációs eljárással,* amely négyzetesen konvergens. Ezzel bizonyítottuk, hogy:

4.1. TÉTEL. Az I. algoritmusban leírt iterációs eljárás legalább négyzetesen konvergens.

5. Megjegyzés, példák

Az előzőekben bizonyított tételek szerint az $\mathbf{f}(\mathbf{x})$ függvény *Jacobi mátrixának* pozitív definitsége (és az F_1 feltétel teljesülése) esetén az I. algoritmusban megfogalmazott rendszámnöveléses módszer a nemlineáris egyenletrendszerek megoldására gyorsabban konvergál mint a *Newton módszer.* Ha azonban a pozitív definitség nem teljesül, előfordulhat az is, hogy a rendszámnöveléses módszer nem mindig alkalmazható még a *Newton módszer* konvergenciája esetén sem. Mindez persze fordítva is igaz: lehet, hogy a rendszámnöveléses módszer konvergál, de *Newton módszer* nem is alkalmazható. Mindkét esetre mutatunk példát.

1. Példa. Oldjuk meg az

$$f_1(x, y) = x^2+y-1 = 0$$
$$f_2(x, y) = 4x+y^2-1 = 0 \tag{5.1}$$

nemlineáris egyenletrendszert az $x_0=1$, $y_0=1$ közelítésből kiindulva először *Newton módszerrel.* A bal oldal *Jacobi mátrixa*

$$\begin{Vmatrix} 2x & 1 \\ 4 & 2y \end{Vmatrix}, \tag{5.2}$$

ami az (x_0, y_0) pontban szinguláris, így a *Newton módszer* nem alkalmazható. A rendszámnöveléses módszer alkalmazása a következőket jelenti: Első lépésben megoldjuk az

$$f_1(x_0+\delta x, y_0) \equiv (1+\delta x)^2 = 0$$

egyenletet: a megoldás $\delta x=-1$ és $x^1=x_0+\delta x=0$. A második lépésben $c=-0{,}5$ és megoldjuk az

$$f_2(x^1+c\delta y, y_0+\delta y) \equiv 4(-0{,}5\delta y)+(1+\delta y)^2-1 = 0$$

egyenletet, amiből $\delta y=0$ és így a megoldás

$$x_1 = x^1+c\delta y = 0$$
$$y_1 = y_0+\delta y = 1.$$

Tehát egy iterációs lépésben a pontos megoldáshoz jutottunk.

2. Példa. Oldjuk meg az (5.1) egyenletrendszert az $x_0=0$, $y_0=0$ közelítésből kiindulva a rendszámnöveléses módszerrel. Az (5.2) *Jacobi mátrix* az (x_0, y_0) pontban

$$\begin{Vmatrix} 0 & 1 \\ 4 & 0 \end{Vmatrix},$$

így a rendszámnöveléses módszer nem alkalmazható. Viszont mint azt a [3] dolgozatban kimutattuk a rendszámnöveléses módszer alkalmazásához lehetőségünk van sor vagy oszlopcserékre. Cseréljük fel a két egyenletet az (5.1) egyenletrendszerben. Akkor az első lépésben megoldandó a

$$4\delta x-1 = 0$$

egyenlet, amiből $\delta x=0{,}25$ és $x^1=0{,}25$. A második lépésben pedig a megoldandó egyenlet ($c=0$)

$$(0{,}25)^2+\delta y-1 = 0,$$

aminek a megoldása $\delta y=0{,}9375$. Így az első lépésben kapjuk, hogy

$$x_1 = x^1+c\delta y = 0{,}25, \quad y_1 = y_0+\delta y = 0{,}9375.$$

A 2. példa esetén a *Newton módszer* viszont változtatás nélkül alkalmazható és az első iteráció után

$$x_1 = 0{,}25, \quad y_1 = 1$$

adódik.
A minimalizálandó függvény értékei pedig

$$F(0{,}25; 1) = 1{,}00390625 > F(0{,}25; 0{,}9375) = 0{,}772476.$$

IRODALOM

[1] GERGELY, J., „Numerikus módszer nemlineáris egyenletrendszerek megoldására", *Alk. Mat. Lapok* **2** (1976) 127—134.
[2] GERGELY, J., „Lineáris egyenletrendszerek megoldása rendszámnöveléssel", *Alk. Mat. Lapok*, **3** (1977) 193—197.
[3] GERGELY, J., „Newton-módszer módosítása nemlineáris egyenletrendszerek megoldására", *Alk. Mat. Lapok*, **3** (1977) 199—205.

(Beérkezett: 1978. július 6.)

DR. GERGELY JÓZSEF
MTA SZÁMÍTÁSTECHNIKAI ÉS AUTOMATIZÁLÁSI KUTATÓ INTÉZET
1250 BUDAPEST, ÚRI U. 49.

NUMERICAL SOLUTION OF THE NONLINEAR SYSTEMS BY METHOD OF BORDERING

J. GERGELY

In the paper the convergence of the method of bordering for solution of the nonlinear systems is proved in the case of the positive definite *Jacobian matrix*. It can be seen that the convergence of the method of bordering is better then the convergence of the *Newton's method*.

Alkalmazott Matematikai Lapok **4** (1978) 143—149

NAGYMÉRETŰ RITKA MÁTRIXOK INVERTÁLÁSA

GERGELY JÓZSEF
Budapest

A nagyméretű mátrixok invertálása a számítógépekben nagy memóriakapacitást igényel és csak valamilyen külső memória segítségével végezhető el. A nagy memóriára ritka mátrixok invertálása esetén is szükség van, hiszen az invertálás közben egyre telítettebb lesz a mátrix. Az invertálás közben nagymennyiségű adatátvitelre van szükség a belső és a külső memóriák közt, ami nagyon időigényes.

A dolgozatban módszereket tárgyalunk és hasonlítunk össze a számítási és adatátviteli idők szempontjából.

1. Bevezetés

Mint ismeretes a ritka mátrixok inverze lehet teljes. Ezért az inverz számolása közben a mátrix ritkaságából adódó előnyök érvényesítése nehézkes, az inverz teljes tárolásánál pedig a mátrix ritkaságából származó előnyök megszűnnek.

Az alábbiakban olyan nagyméretű ritka mátrixok invertálásának gépi módszereivel foglalkozunk, amelyek még a ritkaság kihasználásával sem férnek el a számítógép gyors (operatív) memóriájában. Az invertálandó mátrixot olyan részekre bontjuk, amelyek már kezelhetők a gyors memóriában. A teljes mátrixot a gép egy külső (háttér) tárolójában helyezzük el és itt építjük fel az inverzet is. Az inverz számolása közben a külső tárolóból részletekben hozzuk a gyors memóriába a mátrixot, majd a részeredményeket visszaírjuk a külső tárolóba. A számolási idő és a gépen lekötött tömbök és tárolók mennyisége erősen függ attól, hogy milyen invertálási módszert alkalmazunk.

Az alábbiakban két különböző módszert vizsgálunk. Kiszámítjuk a külső tárolóhoz való fordulások számát. Adott gépi konfigurációt feltételezve összehasonlítjuk a vizsgált módszereket.

Feltevések, jelölések:

1) Tegyük fel, hogy az $n \times n$-es **A** ritka mátrix olyan sok nem 0 elemet tartalmaz, illetve a számolás közben olyan sok nem 0 elem képződik, hogy a vizsgált módszerek használatához egyidejűleg nem fér el a számítógépünk gyors memóriájában.
2) Az **A** mátrix nemszinguláris és az invertálás megkezdése előtt a mátrix úgy van rendezve, hogy a számolás közben sehol se lép fel szingularitás, továbbá főelem választást, oszlop vagy sorcserét nem hajtunk végre. (Ha például az **A** mátrix pozitív definit, akkor ez a feltevés jogos).
3) A gép külső tárolója mágneslemez vagy mágnesszalag, amit KM-mel jelölünk. A gyors vagy operatív memóriát GM-mel jelöljük. A memóriák közti átvitelre a $\Rightarrow$ jelet használjuk.

2. A Doolittle-féle módszer

Az **A** mátrix invertálása *Doolittle-féle módszerrel* (ami a *Gauss eliminációnak* egy variánsa), a következőképpen fogalmazható meg (lásd [4]). Bontsuk fel az **A** mátrixot $\mathbf{A}=\mathbf{FG}$ szorzat alakra *Gauss elimináció* segítségével, ahol **F** alsó háromszögmátrix, **G** felső háromszögmátrix. Ezután $\mathbf{A}^{-1}=\mathbf{G}^{-1}\mathbf{F}^{-1}$. Az $\mathbf{F}^{-1}$ és a $\mathbf{G}^{-1}$ mátrixokat pedig úgy számíthatjuk ki legegyszerűbben, hogy rendre megoldjuk az $\mathbf{FY}=\mathbf{E}$, majd a $\mathbf{GX}=\mathbf{Y}$ mátrix egyenleteket (ahol **E** az egységmátrix) és $\mathbf{X}=\mathbf{A}^{-1}$ a keresett inverz mátrix. Ebben a pontban mindezt abban az esetben valósítjuk meg, amikor **A** ritka mátrix. Az [1] dolgozat jelöléseit használva az

$$\mathbf{A}=\begin{Vmatrix} a_{11} & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \dots & a_{nn} \end{Vmatrix}$$

mátrixból készítsük el a

$$\begin{matrix} d_{11} & u_{12} & u_{13} & \dots & u_{1n} \\ l_{21} & d_{22} & u_{23} & \dots & u_{2n} \\ \vdots & \vdots & \vdots & & \vdots \\ l_{n1} & l_{n2} & l_{n3} & \dots & d_{nn} \end{matrix}$$

úgynevezett faktortáblázatot, ahol

$$d_{ii}=1/a_{ii}^{(i-1)};\quad u_{ij}=a_{ij}^{(i)},\ i<j;\quad l_{ij}=a_{ij}^{(j-1)},\ i>j.$$

($a_{ij}^{(k)}$ a *Gauss elimináció* k-adik lépése után az i, j indexű mátrixelem).

Bevezetjük a következő mátrixokat: legyenek

$$\mathbf{D}_i,\ \mathbf{L}_i \quad \text{és} \quad \mathbf{U}_i$$

olyan mátrixok, amelyek annyiban különbözzenek az egységmátrixtól, hogy legyen

$\mathbf{D}_i$ i-edik sora $(0, \dots, d_{ii}, 0, \dots, 0)$
$\mathbf{L}_i$ i-edik oszlopa $(0, \dots, 1, -l_{i+1,i}, \dots, -l_{ni})$
$\mathbf{U}_i$ i-edik sora $(0, \dots, 0, 1, -u_{i,i+1}, \dots, -u_{in})$.

Ezen mátrixok segítségével az inverz a következőképpen írható fel:

$$\mathbf{A}^{-1}=\mathbf{U}_1\mathbf{U}_2\dots\mathbf{U}_{n-2}\mathbf{U}_{n-1}\mathbf{D}_n\mathbf{L}_{n-1}\dots\mathbf{L}_2\mathbf{D}_2\mathbf{L}_1\mathbf{D}_1\mathbf{E}=\mathbf{ULE}, \tag{2.1}$$

ahol E az egységmátrix, $\mathbf{U}=\mathbf{U}_1\dots\mathbf{U}_{n-1}$, $\mathbf{L}=\mathbf{D}_n\mathbf{L}_{n-1}\dots\mathbf{L}_1\mathbf{D}_1$.

A feltevésünk 2. pontja értelmében az **A** mátrixban előre ki lehet jelölni azokat a helyeket, ahol nem 0 elem léphet fel, (lásd [2, 28. old.]).

Ugyanezeken a helyeken léphet fel nem 0 a faktortáblázatban is. A továbbiakban a nem 0 mátrixelem azt jelenti, hogy az vagy már eredetileg $\neq 0$, vagy a számolás közben lett $\neq 0$.

Az **A** mátrixot sorfolytonosan tároljuk, és soronként úgy bontjuk r részre, ($\mathbf{A}_1, \dots, \mathbf{A}_r$-re) hogy a részek külön-külön kezelhetők a GM-ben. A részek tartalmazzanak $s_1, \dots, s_r$ sort, $s_1+\dots+s_r=n$.

A faktortáblázat kialakításához az eliminációt a sorcsoportok közt hajtjuk végre, mégpedig

1. lépés: elvégezzük az eliminációt az $\mathbf{A}_1$-en belül;
2. lépés: az $\mathbf{A}_1$-gyel eliminálju az $\mathbf{A}_2$ sorait, majd az $\mathbf{A}_2$-n belül eliminálunk;
3. lépés: $\mathbf{A}_1$-gyel, majd $\mathbf{A}_2$-vel eliminálju $\mathbf{A}_3$ sorait, majd $\mathbf{A}_3$-on belül eliminálunk stb.

Az eliminálás végrehajtásához a GM-ben két (azonos méretű) tömbre, $\mathbf{T}_1$ és $\mathbf{T}_2$-re van szükségünk.

A $\mathbf{T}_1$ és $\mathbf{T}_2$, valamint a KM közötti átvitelek:
1. lépésben: KM$\Rightarrow\mathbf{T}_1$ és $\mathbf{T}_1\Rightarrow$KM
2. lépésben: KM$\Rightarrow\mathbf{T}_2$, KM$\Rightarrow\mathbf{T}_1$ és $\mathbf{T}_2\Rightarrow$KM
3. lépésben: KM$\Rightarrow\mathbf{T}_3$ kétszer KM$\Rightarrow\mathbf{T}_1$ és $\mathbf{T}_3\Rightarrow$KM
k. lépésben: k-szor KM$\Rightarrow$GM és egy GM$\Rightarrow$KM átvitel, vagyis $k+1$ átvitel a memóriák közt. Így az elimináció r lépésében összesen

$$2+3+\ldots+r+1=\frac{(r+1)(r+2)}{2}-1$$

memóriák közti átvitelre van szükség.

$\mathbf{A}^{-1}$-et a (2.1) képlet segítségével oszlop csoportonként állítjuk elő. Legyen $\mathbf{E}_1, \mathbf{E}_2, \ldots, \mathbf{E}_p$ az egységmátrix $q_1, q_2, \ldots, q_p$ darabszámú oszlopát tartalmazó része ($q_1+q_2+\ldots+q_p=n$). A (2.1) képletben minden $\mathbf{LE}$ szorzás a faktortáblázat r részletének egyszeri végigolvasását jelenti. Ez összesen pr KM$\Rightarrow$GM átvitelt jelent. Valójában erre mindre nincs szükség, ugyanis $\mathbf{E}_i$ esetében a faktortáblázat i-nél kisebb indexű soraival nem kell számolni. Emiatt az átvitelek mintegy fele szükségtelen, vagyis valójában $\frac{1}{2}pr$ átvitel kell. A (2.1) képlet szerinti $\mathbf{U}(\mathbf{LE}_i)$ szorzás is a faktortáblázat egyszerű végigolvasását jelenti, de ezt a file végén kell kezdeni, vagyis az r-edik rész olvasása, majd az $r-1$-edik rész olvasása stb. Így az inverz teljes felépítéséhez a faktortáblázat $\frac{3}{2}p$-szeri végigolvasására van szükség.

A *Doolittle-féle módszerrel* való invertálásnál tehát összesen

$$\frac{(r+1)(r+2)}{2}+\frac{3}{2}rp-1$$

darab $\mathbf{T}_1=\mathbf{T}_2$ hosszúságú tömbök átvitelére van szükség. Az inverz az $\mathbf{E}_1, \mathbf{E}_2, \ldots, \mathbf{E}_p$ bontásnak megfelelően oszlopfolytonosan képződik.

3. Módosított mátrix inverze

Legyen ismert az $n\times n$-es $\mathbf{B}$ mátrix inverze $\mathbf{D}=\mathbf{B}^{-1}$. Módosítsuk a $\mathbf{B}$ mátrixot egy $\mathbf{uv}^T$ diáddal. Ismeretes, hogy ekkor

$$(\mathbf{B}+\mathbf{uv}^T)^{-1}=\mathbf{D}-\frac{1}{1+\mathbf{v}^T\mathbf{Du}}\mathbf{Duv}^T\mathbf{D}. \tag{3.1}$$

A (3.1) képlet ritka mátrixok esetén célszerűen használható, lásd [3]. Itt a (3.1) képlet használhatóságát abban az esetben vizsgáljuk, amikor az 1. feltevésünk teljesül, vagyis az invertálandó mátrix nem helyezhető el a gyors memóriában. Tegyük fel, hogy $a_{ii} \neq 0$, $i=1, \ldots, n$.

Legyen a kiindulási mátrixunk ($\mathbf{A}_0$) az $\mathbf{A}$ mátrix fődiagonálisában álló elemekből álló diagonálmátrix. $\mathbf{A}_0^{-1}=\mathbf{D}_0$ ugyancsak diagonálmátrix. Az inverzet sorfolyamatosan építjük fel a KM-ben. Legyen

$$\mathbf{A}_k = \mathbf{A}_{k-1}+\mathbf{e}_k \mathbf{a}_k^T, \quad \mathbf{A}_k^{-1} = \mathbf{D}_k,$$

ahol $\mathbf{e}_k$ a k-adik egységvektor, $\mathbf{a}_k^T$ az $\mathbf{A}$ mátrix k-adik sora kivéve a k-adik elemet, ami legyen 0. Jelölje a $\mathbf{D}_k$ mátrix elemeit $d_{ij}^{(k)}$, i-edik oszlopát pedig $\mathbf{d}_i^{(k)}$. Legyenek az $\mathbf{a}_k^T$ mátrixsor nem 0 elemeinek oszlopindexei $j_1, j_2, \ldots, j_s$, azaz

$$a_{kj_1} \neq 0, \; a_{kj_2} \neq 0, \ldots, a_{kj_s} \neq 0,$$

a többi eleme 0.

A $\mathbf{B}=\mathbf{A}_{k-1}$, az $\mathbf{u}=\mathbf{e}_k$ és a $\mathbf{v}^T=\mathbf{a}_k^T$ helyettesítéssel a (3.1) képlet a következő alakba megy át

$$\mathbf{A}_k^{-1} = (\mathbf{A}_{k-1}+\mathbf{e}_k \mathbf{a}_k^T)^{-1} =$$

(3.2)

$$= \mathbf{D}_{k-1} - \frac{1}{1+\sum_{i=1}^{s} a_{kj_i} d_{j_i k}^{(k-1)}} \mathbf{d}_k^{(k-1)} \left[\sum_{i=1}^{s} a_{kj_i} d_{j_i 1}^{(k-1)}, \ldots, \sum_{i=1}^{s} a_{kj_i} d_{j_i n}^{(k-1)} \right].$$

A (3.2) kifejezést kiszámítva $k=1, 2, \ldots, n$-re, $\mathbf{A}_n^{-1}=\mathbf{A}^{-1}$. A számolás megkezdése előtt az $\mathbf{A}$ mátrix nem 0 elemeit és azok oszlopindexeit elhelyezzük egy külső tárolón, ahonnan soronként olvassuk be a GM-be, a $\mathbf{D}_0$ inverz d_{ii} főátlóelemeit pedig a GM egy tömbjében tároljuk. A (3.2) képlet számolása $k=1$-re a következőket jelenti: beolvassuk az $\mathbf{A}$ első sorának nem 0 elemeit és azok oszlopindexeit. A (3.2) képletben szereplő összegek $k=1$ esetben mindössze legfeljebb egy tagból állnak

$$\sum_{i=1}^{s} a_{1j_i} d_{j_i} = a_{1m} d_{mm},$$

amik csak akkor különböznek 0-tól, ha $a_{1m} \neq 0$. Minthogy $\mathbf{a}_1^T$ első komponense $\mathbf{a}_{11}=0$ és $\mathbf{d}_1^{(0)}$-nek csak az első komponense $\neq 0$, ezért $\mathbf{D}_1$ első sora lesz 0-tól különböző azoknál a j indexű elemeknél, amelyekre $a_{1j} \neq 0$ volt, és ezek szorzódnak d_{11}-gyel. $k=1$-re a (3.2) képletből számolt $\mathbf{D}_1$ első sorát d_{11}-gyel kiegészítve elhelyezzük a KM-ben.

A (3.2) képlet számolása $k=2$-re a következőket jelenti: Beolvassuk az $\mathbf{A}$ második sorának nem 0 elemeit és azok oszlopindexeit. A (3.2) képletben szereplő összegek állhatnak most zéró, egy vagy két tagból, attól függően, hogy $a_{21}=0$, vagy $a_{21} \neq 0$. Ha $a_{21} \neq 0$, akkor be kell olvassuk a $\mathbf{D}_1$ első sorát, mert ennek elemei az összegekben tényezőként szerepelnek. Ha $a_{21}=0$, akkor erre nincs szükség. Kiszámítjuk a (3.2) képletben szereplő összegeket és felépítjük a

$$\mathbf{q}^T = \left\{ p \sum_{i=1}^{s} a_{2j_i} d_{j_i m} \right\}, \quad m = 1, \ldots, n$$

sorvektort, ahol $p=\frac{1}{1+a_{21}d_{12}^{(1)}}$. Ezután ha $d_{12}^{(1)}\neq 0$, a $d_{12}^{(1)}\cdot\mathbf{q}$ vektort, valamint a $d_{22}^{(1)}\cdot\mathbf{q}$ vektort is hozzáadjuk a KM-ben már meglevő $\mathbf{D}_1$ inverz első sorához (a $\mathbf{D}_1$ második sorában csak egy elem, a $d_{22}^{(1)}\neq 0$). Ez azt jelenti, hogy át kell vinnünk a KM-ből a $\mathbf{D}_1$ inverz első sorát a GM-be, ott elvégezni ezen és a $\mathbf{D}_1$ második során is a (3.2) képletben szereplő diádmódosítást, majd a módosított sorokat vissza kell vinni a KM-be.

Ehhez hasonlóan $k=3, 4, \ldots, n$-re elvégezzük a fenti lépéseket. Tegyük fel, hogy $\mathbf{D}_{k-1}$-et már kiszámoltuk és sorfolytonosan elhelyeztük a KM-ben.

A (3.2) képlet számolása a k-adik lépésben a következőt jelenti: beolvassuk az $\mathbf{A}$ k-adik sorának nem 0 elemeit és oszlopindexeit. A (3.2) képletben szereplő összegek elkészítéséhez $j_i<k$-ra beolvassuk a $\mathbf{D}_{k-1}$ inverz j_i-edik sorát, ha $a_{kj_i}\neq 0$, és az összes eddigi részletösszeghez hozzáadjuk az $a_{kj_i}d_{j_im}$ tagot, $m=1, \ldots, n$. Kiszámítjuk a (3.2) képletben szereplő $p=\frac{1}{1+\sum_{i=1}^{s} a_{kj_i}d_{j_ik}^{(k-1)}}$ hányadost és összeállítjuk a

$$\mathbf{q}=\left\{p\cdot\sum_{i=1}^{s} a_{kj_i}d_{j_im}^{(k-1)}\right\},\quad m=1, \ldots, n$$

vektort. Ha $d_{jk}^{(k-1)}\neq 0$ (ez csak $j<k$-ra állhat fenn), akkor a $d_{jk}^{(k-1)}\cdot\mathbf{q}$ vektort hozzáadjuk a $\mathbf{D}_k$ inverz j-edik sorához a KM-ben. Ez úgy történik, hogy átvisszük a $\mathbf{D}_k$ mátrix j-edik sorát GM-be, ott hozzáadjuk a $d_{kj}\mathbf{q}$ vektort, majd visszavisszük a KM-be. Mindez tehát egy KM⇒GM átvitelt (egy visszaléptetést) és egy GM⇒KM átvitelt jelent.

Az inverz számolásához a k-adik lépésben ilyen szervezés mellett annyi KM és GM átvitelre van szükség, ahány darab nem 0 elem van az $\mathbf{a}_k^T$ vektorban és annyi GM⇒KM és KM⇒GM átvitelre ahány nem 0 elem van a $\mathbf{d}_k^{(k-1)}$ vektorban. A (3.2) képletben szereplő szorzások száma is az $\mathbf{a}_k^T$, illetve a $\mathbf{d}_k^{(k-1)}$ nem 0 elemeinek a számával arányos. Ily módon az átvitelek számát (és a szorzások számát is) az invertálandó mátrix struktúrája határozza meg.

4. A módszerek összehasonlítása

Az összehasonlításhoz legyen a mátrixunk rendje $n=1000$. Először a 2. szakaszban ismertetett *Doolittle-féle módszerrel* foglalkozunk. [2] szerint a 2. feltevés teljesülése esetén előre kijelölhető az eliminációban fellépő nem 0 elemek helye. Tegyük fel, hogy ez az egész mátrix 20%-ban lép fel. Legyen először $\mathbf{T}_1=\mathbf{T}_2=1000$. Az $\mathbf{E}_1, \mathbf{E}_2, \ldots, \mathbf{E}_p$ álljon csak egy oszlopból. Ekkor $r=200$, $p=1000$. Ebben az esetben az átvitelek száma

$$\frac{(r+1)(r+2)}{2}+\frac{3}{2}rp-1\sim 320\,300$$

és egy-egy átvitt egység hossza 1000 szám, így összesen $3{,}2\cdot 10^8$ szám átvitelére van szükség.

Az átvitelek száma az átviendő tömbök méretének növelésével erősen csökkenthető.

Legyen most $\mathbf{T}_1=\mathbf{T}_2=10\,000$ és az $\mathbf{E}_1, \mathbf{E}_2, \ldots, \mathbf{E}_p$ álljon 10 oszlopból (ehhez is 10 000 méretű tömb kell). Ekkor $r=20$, $p=100$ és

$$\frac{(r+1)(r+2)}{2}+\frac{3}{2}rp-1 \sim 3230,$$

és egy-egy átvitt egység hossza 10 000 szám, így összesen $3{,}2\cdot 10^7$ szám átvitelére van szükség. Vagyis a második esetben csak tized annyi szám mozgatására volt szükség, mint az első esetben.

A 3. pontban ismertetett módosításos módszer általános esetben igen nagyszámú adatmozgatást igényel, ezért általában ez a módszer nem javasolható.

Tegyük fel, hogy az $\mathbf{A}$ mátrixunkban a főátló alatti nem 0 elemek az $i_1, i_2, \ldots$ $\ldots, i_s$-edik sorban vannak. Tegyük fel, hogy $s=25$, és átlagosan egy sorban 100 db nem 0 elem van. (Ez a nem 0 elemszámra vonatkozó feltevésünk nagyjából megegyezik a fentivel). Használjuk a 3. pontban ismertetett módszert.

A (3.2) képletben kijelölt összegzésekhez összesen körülbelül $25\cdot 100=2500$ átvitelre van szükség. A $\mathbf{d}_{i_p}$, $p=1, \ldots, s$ vektorokban a nem 0 elemek gyakorisága (a ritkasági feltevésünk mellett) átlag 100, ami az inverz képzés folyamán sűrűsödik. Átlagosan soronként 300-zal számolva a (3.2) képlet diádmódosításaihoz összesen

$$2\cdot 25\cdot 300 = 15\,000$$

átvitelre van szükség.

Az átvitt tömbök mérete itt mindig a mátrix egy sora, illetve egy oszlopa (vagyis 1000 szám). Így a fenti feltevésünk mellett az inverz képzéshez $1{,}75\cdot 10^7$ szám átvitelére van szükség.

Megjegyezzük, hogy ha a vizsgált mátrixot a *Doolittle-féle módszer* segítségével invertáltuk volna, sokkal több adat átvitelére lett volna szükségünk, mint a 3. pontban ismertetett módszer esetén. Ugyanis az eliminációs részben még nagyon jól ki tudtuk volna használni, hogy csak kevés sor tartalmaz nem 0 elemeket, de a visszahelyettesítési részben már nem. Ha például $i_s=n$, vagyis az $\mathbf{A}$ utolsó sora tartalmaz nem 0 elemeket, a visszahelyettesítésnél (az $\mathbf{U}(\mathbf{LE})$ szorzásnál) az $\mathbf{U}$ teljes végigolvasására szükség lesz, ami rp átvitelt jelent. Így az átvitelek száma lényegesen nem csökkenthető.

Tételezzünk fel egy olyan számítógépet, amelyben a műveleti sebességek olyanok, hogy egy 100×100-as teljes mátrix invertálása a GM-ben 1 perc időt vesz igénybe, a KM és a GM közti adatátvitel sebessége pedig 10^4 szám/sec. A fent kiszámolt adatátvitelekhez szükséges idők ekkor

$3{,}2\cdot 10^8$ szám átvitele ~ 10 óra,
$3{,}2\cdot 10^7$ szám átvitele ~ 1 óra,
$1{,}75\cdot 10^7$ szám átvitele $\sim 1/2$ óra.

Ha a ritkasági struktúra miatt az inverz számolási ideje 10%-ra csökkenthető, akkor az 1000×1000 ritka mátrix invertálási ideje 100 perc, ami az adatátvitelekre fent számolt idők nagyságrendjébe esik. A feltett gépi adottságok mellett a számolási idők és a KM és GM közti adatátviteli idők jól összemérhetők és azonos nagyságrendbe esnek, ezért nagyon fontos az invertálandó módszer jó megválasztása. Célszerű a módszereket úgy szervezni, hogy a nagytömegű adatmozgatás minimális legyen.

IRODALOM

[1] TINEY, W. F., AND WALKER, J. W., "Direct solution of sparse network equations by optimally ordered triangular factorization", *Proc. IEEE* **55** (1967) 1801—1809.
[2] GERGELY, J., „Numerikus módszerek sparse mátrixokra", *MTA SZTAKI Tanulmány* **26**/1974.
[3] GERGELY, J., „Ritka mátrixok invertálása", *MTA SZTAKI Közlemények* **11** (1973) 51—53.
[4] TEWARSON, R. P., *Sparse Matrices*, (Academic Press, New York and London, 1973).

(Beérkezett: 1978. június 26.)

DR. GERGELY JÓZSEF
MTA SZÁMÍTÁSTECHNIKAI ÉS AUTOMATIZÁLÁSI KUTATÓ INTÉZET
1250 BUDAPEST, ÚRI U. 49.

INVERSION OF SPARSE MATRICES

J. GERGELY

As the inverse of a sparse matrix can be full, in case of great sizes the computer implementation of the inversion strongly depends on the quantity of the transfers between the operational store and the magnetic tape or disk. The paper analyses these problems by comparing several methods.

Alkalmazott Matematikai Lapok 4 (1978) 151—164

OSZTOTT ADATBÁZISOK TERVEZÉSE*

ERICH J. NEUHOLD
NSZK

Az utóbbi egy-két évben a miniszámítógép hálózatok jelentősége a jövőbeni számítógép alkalmazások szempontjából rohamosan növekszik. A miniszámítógép hálózatok megjelenésével lehetővé vált, hogy a nagy, integrált adatbázisok elveit átültessük kisgépes környezetbe. Az ilyen adatbázisok szét vannak osztva a különböző gépek között, és az egyes felhasználóknak nem szükséges tudomással bírniuk arról, hogy adataik a hálózaton belül hol helyezkednek el. Az adatbázis szemantika kell, hogy alapjául szolgáljon minden konzisztens és megbízható adatbázis rendszernek. Miután ismertettük ennek alapelveit, áttekintjük egy osztott adatbázisrendszer architektúráját. A rendszer a stuttgarti egyetemen jelenleg áll fejlesztés alatt.

1. Bevezetés

A kereskedelem és az ipar gyors számítógépesítése egyre bonyolultabb file kezelőrendszerek kifejlesztését vonta maga után. Az az igény, hogy nagymennyiségű adatot hatékonyan, konzisztens és megbízható módon lehessen kezelni, oda vezetett, hogy az adatokat egyetlen rendszerbe integrálták, amelyet a különböző felhasználóknak saját igényeiknek megfelelően kellett kezelniük.

Annak ellenére azonban, hogy a file kezelési lehetőségek beépültek a magasszintű nyelvekbe, mint pl. a COBOL-ba, hamarosan kiderült, hogy

a) a felhasználóknak ismerniök kellett a többi felhasználó számára felépített adatstruktúrákat is; továbbá

b) az adatokat olyan szabályoknak megfelelően kellett kezelniük, amelyek legnagyobb részben a rendelkezésre álló tárolóeszközök fizikai tulajdonságaiból fakadtak.

Ezek a hiányosságok vezettek el az integrált adatbáziskezelő-rendszerekhez, amint az a CODASYL *Feature Analysis* [10] és a CODASYL DBTG *Report* [11] kiadványokban tükröződik. A következő években sok vita folyt a DBTG javaslat hasznosságáról és teljességéről.

Azok a vizsgálatok például, amelyek E. F. CODD *Relációs Modelljéhez* [12] kapcsolódtak, megmutatták, hogy lehetséges és több szempontból kívánatos az adatbázisoknak egy olyan megközelítése, amely az eddiginél jobban a nem-programozó felhasználó igényeihez igazodik (lásd pl. a különböző ACM-SIGMOND/SIGFIDET konferenciák anyagait [28]—[33], valamint az IFIP TC2 adatbázis konferenciáit [24]—[27], ezek közül pedig különösen az *„Adatbázisok leírása"* konferenciát [25]).

* A *II. Magyar Számítástudományi Konferencián* elhangzott előadás. A fordítást a szerző engedélyével LŐCS GYULA készítette.

Ha a felhasználót felmentjük az adatbázis-kezelés fizikai és strukturális tevékenységétől, ezáltal lehetőség nyílik egy sereg fontos, adatbázisokkal kapcsolatos kérdés megoldására:

1. Minthogy minden rendszer folyamatosan módosul, mind az információk maguk, mind az információs struktúra állandóan változik. Mindaddig, amíg a felhasználói tevékenységek (pl. programok) nem függnek közvetlenül a változásoktól, azok nem feltétlenül érintik a felhasználót.
2. Az adatbázis használatának jellemzői hosszabb vagy rövidebb időszakonként változhatnak (pl. nappali és éjszakai műszak). Hatékonysági meggondolásokból célszerű lehet az adatbázis átrendezése. Az ilyen módosítások azonban nem szabad, hogy kihatással legyenek a feladatmegoldásra irányuló felhasználói tevékenységekre.
3. Az egyes felhasználók esetleg csak az adatbázisban tárolt információ egy részét használják. Az egyéni felhasználó szempontjait szem előtt tartó megközelítés szükségtelenné teszi adatbázis olyan részeinek ismeretét, amelyek lényegtelenek az adott feladat szempontjából.

A felhasználó és a rendszer szétválasztásának a fent említett előnyei az adatbáziskezelő-rendszerek célszerű architektúrájának megfogalmazásához vezettek. Az ANSI-SPARC javaslatban [3] egy ilyen architektúrának meglehetősen teljes leírása található, ami jelentősen hozzájárult az adatbáziskezelő-rendszerek terminológiájának egységesítéséhez. Annak a ténynek az elfogadása, hogy ugyanannak az adatbáziskezelő-rendszernek különböző felhasználói igényelhetik mind a nem eljárás orientált (relációs), mind az eljárás orientált (vezérelt, navigációnális) adatkezelési lehetőségeket, az ún. koexisztencia architektúra kialakulását eredményezte, amint azt G. M. NIJSSEN javasolta [20].

Ezek a kutatások azonban főként az adatok strukturálásának és kezelésének problémájára irányultak, akárcsak az a számos kísérlet, amely az adatbáziskezelő-rendszerek formális leírását célozta meg (E. J. NEUHOLD [18], H. BILLER és mások [6], [7]). Hamarosan nyilvánvalóvá vált, hogy az adatbázisok konzisztencia, illetve helyességi kritériumainak leírásához az adatbázisban tárolt adatok *jelentésének (szemantikájának)* lényegesen jobb megértése szükséges. Ebben a tárgyban egy-egy korai munka volt R. ABRIAL [1], és B. LANGEFORS és B. SUNDGREN dolgozata [23], amelyeket később több javaslat követett arra vonatkozóan, amit később adatbázismodellnek (vagy az ANSI-SPARC terminológiával, fogalmi sémának) neveztek el. Ezeket az ajánlásokat részletesen elemezték és összehasonlították KERSHBERG és társszerzői [16], valamint BILLER és társszerzői [8]. Ugyanott részletes irodalomjegyzék is található. A 2. fejezetben rövid bevezetést adunk ebbe a témakörbe, felhasználva a BILLER—NEUHOLD cikkben megalapozott terminológiát [9].

A különböző architekturális elképzelések alapján létrehozott adatbáziskezelő-rendszerek, vagy nagy számítógépekre készültek, mint pl. az M. ASTRAHAN és mások által fejlesztett *System—R* kutatási rendszer [4], amely sokirányú felhasználási igényt fed le, vagy miniszámítógépes rendszerekre, mint pl. az INGRES-rendszer (STONEBRAKER és mások [21]), amelynél a tárolókapacitás korlátai miatt szemmel láthatóan az egyféle felhasználás elve érvényesül.

Az utóbbi években a miniszámítógépek felhasználása lényegesen gyorsabban terjedt, mint a nagy és költséges rendszereké, és a jövőben ez a tendencia várhatólag

tovább fog folytatódni, kiegészülve a miniszámítógép-hálózatok fokozódó terjedésével. A hálózatok alkalmazásának lehetősége enyhíteni fogja azokat a korlátokat, amelyek egyetlen miniszámítógép esetén a műveleti sebességből és a külső tárolókapacitásból következnek, feltéve, hogy sikerül kidolgoznunk a miniszámítógép-hálózatokra alapozott osztott adatbáziskezelő-rendszerek tervezési koncepcióját. Ezen a területen széles körű kutatás folyik, amint azt az osztott adatbázisokkal foglalkozó szimpoziumok növekvő száma ([2], [5]), valamint az egyre szaporodó irodalmi publikációk is mutatják (pl. LEVIN és mások [17], KARP és mások [15], STONEBRAKER és NEUHOLD [22]).

Egy osztott adatbázis környezetében — akárcsak bármilyen más adatbázis architektúra esetén — a felhasználó számára lehetővé kell tenni, hogy a saját, szemantikai orientációjú szemléletével tekinthessen az adatbázisban tárolt adatokra. Nem szabad, hogy foglalkoznia kelljen az információ rendszeren belüli tárolásának és elosztásának módjával. Ennek következtében az osztott adatbázisok architektúrájának kialakításakor a legfontosabb problémák, amelyeket meg kell oldani az adatok tárolási és feldolgozási színhelyeinek célszerű megválasztása, valamint az egyes színhelyek közötti hatékony adatátvitelek megszervezése. Egy, a fenti elvek alapján megtervezett osztott adatbázis architektúráját a jelen cikk 3. fejezetében ismertetjük. A 4. fejezetben az osztott adatbázisokkal kapcsolatos néhány, ma még megoldatlan kérdésre mutatunk rá.

2. Adatbázisok szemantikai modellje

Az ANSI-SPARC jelentés [3] a fogalmi sémát úgy definiálja mint egy intézmény leírását, helyesebben a szervezetnek egy olyan részét, mely azon személyek számára szükséges, akik az adatbázissal kommunikálni kívánnak.

A fogalmi séma ennélfogva

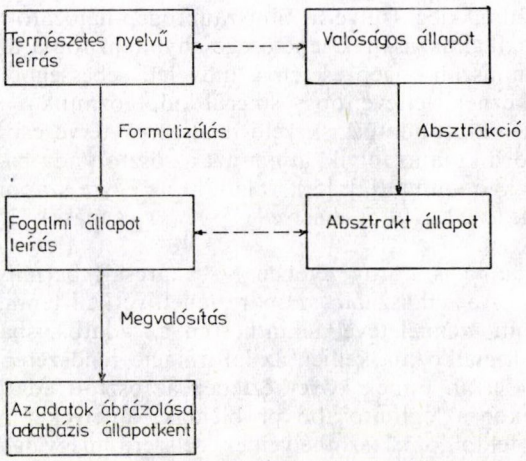
a) az adatbázis felhasználóit a bázisban tárolt adatok egységes értelmezésére kényszeríti, és

b) a megvalósító számára leírást nyújt azokról az adatokról, amelyeknek az adatbázisban tárolható formában kell lenniük.

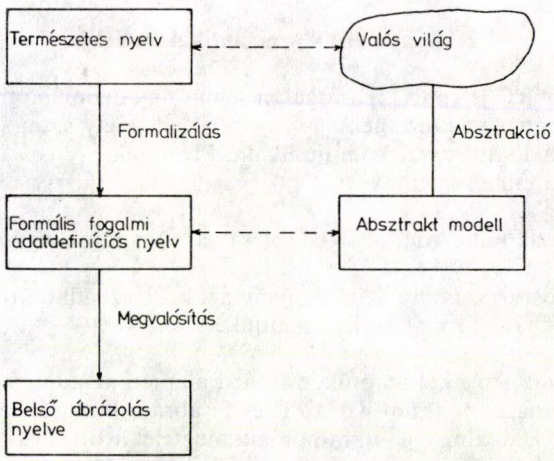
A fogalmi séma fenti két alaptulajdonsága alapján az adatbázisnak *szemantikai jelentést* tulajdoníthatunk, amint azt az 1. és 2. ábrán illusztrált absztrakciós folyamat is mutatja. (Az alkalmazott terminológia megfelel a BILLER—NEUHOLD cikknek [9], és az Olvasó itt találja meg a szabatos definíciókat is.)

Az 1. ábra a valós világ és az absztrakt világ kapcsolatát mutatja, egy adott t időpillanatban. Az ábra bal oldali fele azokat a „nyelveket" szimbolizálja, amelyeket az egyes állapotok ábrázolására általában használni szoktak. Minthogy egy adatbázis nemcsak egy izolált időpillanatban létezik, az egyes állapotok fogalmát ki kell terjesztenünk az *elképzelhető* állapotok halmazának fogalmára, amelyek a valós világot alkotják.

Mielőtt rátérnénk a fenti megközelítés elveinek rövid tárgyalására, felidézünk a „szemantikai" fogalmi sémával (vagy absztrakt modellel) szemben támasztott néhány követelményt, ahogyan azokat az IFIP 2.6 munkacsoportban, és más, adatbázisokkal foglalkozó közösségekben megfogalmazták.



1. ábra. A valóságos állapot
és a hozzá tartozó absztrakt állapot egy t időpillanatban



2. ábra. A valós világ és a hozzá tartozó absztrakt modell

A. Egy absztrakt modellnek

1. könnyen megfogalmazhatónak, és érthetőnek kell lennie, vagyis jól kell tükröznie a valós világot,
2. teljesnek kell lennie,
3. könnyen változtathatónak kell lennie (fejlődés),
4. figyelembe kell vennie a különböző (felhasználói) szempontokat,
5. stabilnak kell lennie.

B. Az absztrakt modellt leíró adatdefiníciós nyelvnek

1. formális nyelvnek kell lennie, amelynek mondataihoz az absztrakt modell, és ezáltal egyszersmind a valós világ fogalmaira történő szemantikai leképezések segítségével jelentést lehet hozzárendelni,
2. teljesnek kell lennie,
3. a felhasználók számára érthetőnek, vagyis „természetesnek" kell lennie,
4. ortogonálisnak kell lennie, vagyis az absztrakt modell különböző fogalmait különböző nyelvi fogalmakkal kell kifejezni, és az absztrakt modell minden egyes fogalmát pontosan egy nyelvi fogalomnak kell leírnia.

A valós világ, ahogyan azt az adatbázisban tárolt információ leírja, esetleg végtelen számú (elképzelhető) állapotból áll, amelyek bármelyike valamely időpillanatban ténylegesen elő is állhat. Ezeket a valóságos állapotokat általában valamilyen természetes nyelven megfogalmazott kijelentő mondatok egy halmazának segítségével írjuk le, amelyek a valós világ állapotára vonatkozó valamilyen *tényt* közölnek, mint pl. „JIMMY CARTER az *Egyesült Államok* elnöke". Strukturális szempontból ezek a kijelentések két csoportba sorolhatók:

a) objektumok tulajdonságai, melyekkel a valós világ bizonyos objektumai vagy bírnak, vagy nem. Pl.: „JIMMY CARTER az *Egyesült Államok* elnöke".

b) objektumok közötti viszonylatok (relációk), amelyek által az objektumok egy valóságos állapotban kombinálódnak egymással, mint pl. „JIMMY CARTER édesanyja *Lillian asszony*".

A valóságos állapot általában igen nagyszámú ilyenfajta tényt tartalmaz. Mielőtt ezeket a tényeket beépítenénk az adatbázisba, egy absztrakciós és osztályozási eljárást kell végrehajtanunk, amelynek során létrejön a valós világ absztrakt modellje, amelyben valamennyi lehetséges állapotot formális és tömör formában írunk le. A BILLER—NEUHOLD által ajánlott osztályozási eljárás [9] az egyes objektumokhoz *entitásokat*, a tulajdonságokhoz *típusokat*, a viszonylatokhoz pedig *relációkat* rendel. A fentieken kívül megad egy formális adatdefiníciós nyelvet, amelynek keretében ezek a kategóriák ábrázolhatók és megadhatók azok a konzisztencia feltételek is, amelyek segítségével a lehetséges absztrakt állapotok halmazát az elképzelhető állapotok absztrakcióinak halmazára szűkítjük le. Ez a nyelv alkalmas a fogalmi sémák definiálására, amint azt az ANSI-SPARC jelentés [3] megköveteli. A fogalmi sémákat azonban ebben az esetben az adatbázis felhasználójának valós világából egy szemantikusan vezérelt absztrakciós eljárás során vonatkoztattuk el, és ezek mindazt a szemantikusan lényeges információt tartalmazzák, amelyet a felhasználónak figyelembe kell vennie az adatbázis megfelelő kezelése szempontjából.

Az adat definíciós nyelv a fogalmi adatbázisban szereplő objektumoknak és maguknak a tényeknek az ábrázolását a nyelvben szereplő ábrázolási lehetőség segítségével adja meg. A fogalmi séma meghatározza, hogy hogyan kell az objektumokat és tényeket ebben a nyelvben egyértelmű módon ábrázolni. Ez lehetővé teszi a felhasználó értelmes támogatását, amikor ő rekonstruálja a „reprezentáció→absztrakt világ fogalom→valós világi tény" láncot, és biztosítja az adatbázisban tárolt adatok helyes értelmezését, valamint a valós világ újabb tényeinek helyes beillesztését.

Az adatbázisok tervezésének fenti, szemantika orientált elveit minden adatbázis kezelő rendszer megvalósításánál szem előtt kell tartani. Bármilyen jellegű legyen is a fizikai megvalósítás (több felhasználós környezet, többfeladatos környezet, többprocesszoros környezet), rendkívül fontos, hogy az adatbázisok szemantika orientált

értelmezésének elve érintetlen maradjon. Csakis ez biztosíthatja, hogy az adatbázis tervezés számos problémáját, mint például a konzisztencia feltételek, biztonsági előírások, hiba utáni helyreállítás, eltérő felhasználói szempontok, hatékonysági meggondolások stb. anélkül oldjuk meg, hogy egyidejűleg tönkretennénk az adatbázisban tárolt adatok helyes értelmezhetőségét.

3. Egy osztott adatbázis architektúrája

Mielőtt az osztott adatbázisokat részleteikben tárgyalnánk, definiáljuk az osztott adatbázis fogalmának a jelen cikkben használt értelmezését.

Egy osztott adatbázis a következő meghatározó tulajdonságokkal rendelkezik:

1. Az adatbázis egy számítógép hálózaton van megvalósítva, ahol a különböző számítógépek mind tárolják, mind feldolgozzák az adatokat.
2. A különböző gépeken jelenlevő adatok kapcsolatban vannak egymással, vagyis minden egyes időpillanatban egyetlen absztrakt állapotnak, illetve a valós világ egyetlen állapotának ábrázolását adják.
3. Az osztott adatbázis felhasználói (elvben) nincsenek tudatában annak, hogy adatmanipulációs feladataik mely helyszín(ek)en oldódnak meg. Ez a szabály egyaránt kiterjed a tárolási és a feldolgozási helyszínekre.

A 3. számú szabályt különböző okokból néha megsértik. Az egyes gépek különböző utasításrendszere vagy eltérő tároló berendezései például kizárhatják annak lehetőségét, hogy bármely tranzakciót bármelyik gépen végre lehessen hajtani. Hatékonysági meggondolások pedig szükségessé tehetik, hogy a felhasználó ismerje adatainak tárolási helyét. Általában azonban minden osztott adatbázisnak ki kell elégítenie a fenti feltételeket.

A fenti alapvető tulajdonságokon túl az osztott adatbázisoknak számos járulékos sajátságuk is van. Ezeket távlati céloknak is tekinthetjük, amelyeket a jövőben megvalósítandó adatbázisok fognak majd kielégíteni, mint például:

a) Az optimális teljesítmény elérése céljából a rendszerhez tartozó gépek terhelésének automatikus (vagyis az adatbázis adminisztrátorának segítsége nélküli) kiegyensúlyozása

— az adattárolás,
— az adatokhoz való hozzáférés és a rajtuk történő manipuláció,
— a feldolgozás

vonatkozásában. Következésképpen az adatbázis adminisztrátora csupán egy hagyományos fogalmi adatbázis sémát határoz meg, tekintet nélkül a hálózatban szereplő különböző számítógépekre.

b) A rendszer automatikusan gondoskodik valamennyi sértetlenségi, biztonsági, megbízhatósági és konzisztencia feltétel kielégítéséről. Magától értetődik, hogy egy ilyen rendszer megvalósítása csak akkor lesz lehetséges, ha ugyanezeket a problémákat egyetlen számítógépre telepített adatbázisok esetére már megoldottuk. Érzésünk szerint ennek elérésétől ma még messze vagyunk, de a 2. pontban leírt adatbázis szemantika végül is sikerre fog vezetni.

c) A nem programozó és nem adatbázis adminisztrátor szakember képes kell, hogy legyen arra, hogy az adatbázist döntés előkészítő (vagy probléma megoldó) rendszerként használja. Az adatbázisokban tárolt információ tömege általában olyan hatalmas lesz, hogy egy alkalmi felhasználó, vagyis olyasvalaki, aki a rendszert, vagy annak egy részét nem „teljes munkaidőben" használja, nem tudja majd nélkülözni a rendszer aktív támogatását feladata megoldásában. A rendszernek információt kell szolgáltatnia a benne tárolt adatokról, ezek értelmezéséről mind az absztrakt, mind a valós világ fogalmaival leírva (beleértve a speciális külső séma terminológiáját is), a rendelkezésre álló adatfeldolgozási algoritmusokról, valamint az új adat manipulációs tevékenységek létrehozásakor figyelembe veendő szabályokról.

Az osztott adatbázisok lehetséges alkalmazásai szempontjából a közeljövőben elsősorban azok a területek jöhetnek szóba, ahol a miniszámítógép-hálózatok különösen rohamosan terjednek, így:

— Bankalkalmazások, ahol a különböző fiókok egy minigép-hálózatba vannak bekapcsolva. Jelenleg ezek a hálózatok többnyire tartalmaznak egy nagy központi gépet, ahol az adatbázist tárolják. Véleményünk szerint ez a helyzet rohamosan meg fog változni, mihelyt az osztott adatbázisok kereskedelmi forgalomban kaphatók lesznek.

— Szupermarket- és áruházi hálózatok, ahol a különböző üzleteknek saját minigépük van a helyi adminisztráció céljaira, de kapcsolatban vannak egy nagy központi adatbázissal is, a rendelések és eladások nyilvántartása, személyzeti nyilvántartás, stb. céljából. Az ilyen központi szervezetek összes veszélyeikkel együtt (megbízhatóság hiánya, az adatok integritásának veszélyeztetettsége) ismét csak el fognak tűnni, mihelyt az osztott adatbázisok rendelkezésre állnak.

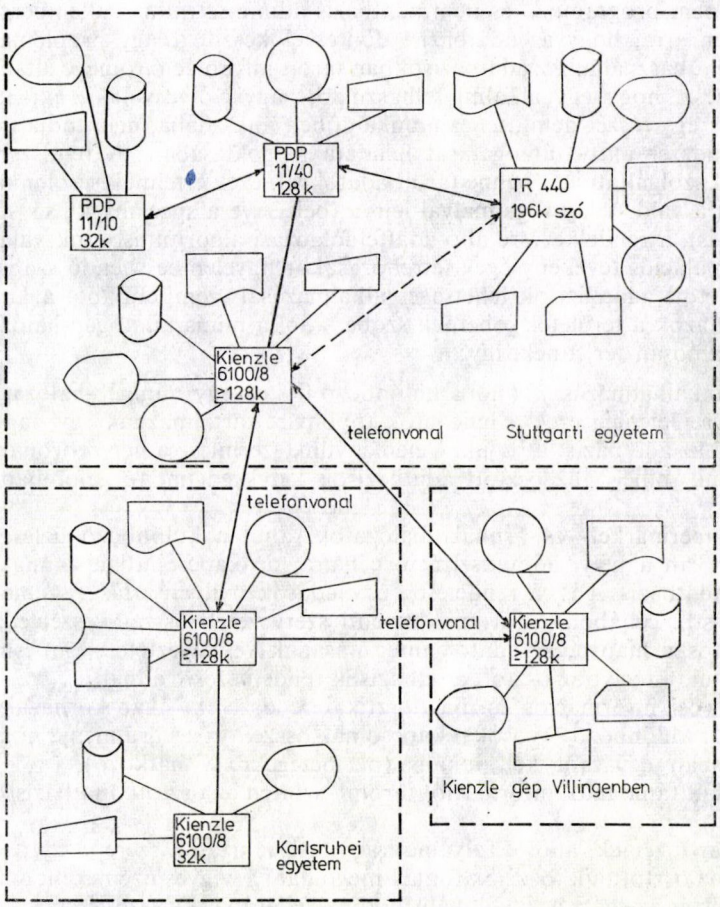
— Orvosi információs és diagnosztikai rendszerek, akár kórházakat szolgálnak ki, akár különböző orvosokat kapcsolnak össze. Az a körülmény, hogy ebben az alkalmazásban egyaránt kell lokális (pl. betegekre vonatkozó), és globális (pl. diagnosztikát támogató) információt tárolni, szintén az osztott adatbázisok irányába mutat.

— Ipari üzemek, ahol a folyamatirányítási hálózatok szerepe egyre növekszik. E téren is az osztott adatbázisok fogják megoldani az egyes üzemkomponensek helyi vezérlésének és az egész üzem globális optimalizálásának és vezérlésének kettős problémáját.

A fenti definíciók alapján először annak az osztott adatbázisnak a hardware konfigurációját ismertetjük, amely a *stuttgarti egyetemen* és a hozzá csatlakozó telephelyeken jelenleg áll fejlesztés alatt [19].

Az osztott adatbázis hardware kiépítése

Az adatbázis céljára létesítendő számítógép-hálózat több miniszámítógépből és egy nagyszámítógépből, a TR 440-ből áll. A gépeket jelenleg meglehetősen lassú (9600 baud-os távhívásos vagy állandó) telefonvonalak kötik össze, de a kommunikációs protokollok kiválasztása lehetővé teszi, hogy a jövőben magasabb átviteli sebességekre térjünk át. A hardware elrendezést a 3. ábra mutatja. A legkisebb miniszámítógépek grafikus megjelenítő egységeikkel, a rendszer interaktív problémamegoldási feladatait fogják támogatni, mint intelligens terminálok.



3. ábra. A hardware elrendezés

Az osztott adatbázis-kezelő rendszer

Az alábbiakban ismertetendő osztott adatbázis-kezelő rendszer architektúrájának megtervezésénél többféle irányelvet is követnünk kellett. Ezek nagyrészt az osztott adatbázisok azon tulajdonságainak következményei, amelyeket a jelen fejezet elején ismertettünk.

A legfontosabb tervezési elvek a következők voltak:

— A rendszernek problémamegoldó funkciói vannak → párbeszédes üzemmód.

— A felhasználónak az adatbázis csak azon részeit kell ismernie, amelyeknek köze van az ő saját feladatához → eltérő adatbázis-szemléletek.

— A rendszernek egyaránt lehetővé kell tennie a

a) procedurális adatkezelést → befogadó nyelv

b) nemprocedurális adatkezelést→igen magasszintű, halmazorientált nyelv.

— A felhasználó nincs tudatában annak, hogy az ő adatai hol helyezkednek el a rendszerben, és hogy hol kerülnek feldolgozásra→automatikus adat-, és feldolgozás elosztási technikák.

— A felhasználó nem kíván a rendszer hatékonysági problémáival foglalkozni→automatikus tárolás és hozzáférés optimalizálási technikák.

— A rendszer biztosítsa a fogalmi sémának tömör és természetes ábrázolását→relációs adatformátumok, relációs, igen magasszintű nyelv.

— Valamennyi (mind a rendszerre, mind a felhasználóra vonatkozó) adat relációs formában tárolódjék→az implementáció nyelve egy befogadó nyelv.

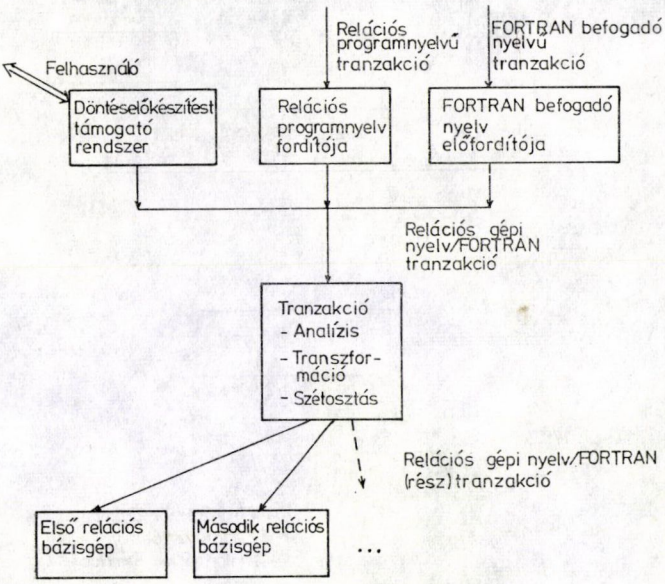
— A rendszer könnyen átvihető legyen más adatfeldolgozó rendszerekre→magasszintű implementációs nyelv (a befogadó nyelv FORTRAN IV).

— Az operációs rendszerekben és a rendelkezésre álló kommunikációs lehetőségekben rejlő lehetőségek a lehető legnagyobb mértékben ki legyenek használva→a kommunikációs protokollokban felhasználjuk a meglevő file kezelési és terminál szimulációs eszközöket.

— A hálózatban szereplő valamennyi számítógép (beleértve az intelligens terminálokat is) ugyanazt az osztott adatbázis-architektúrát támogatja→virtuális gépek elvének alkalmazása, relációs és FORTRAN IV interface-szel. (Az egyes gépek kapacitása ennek az elvnek az alkalmazási lehetőségét esetleg korlátozhatja. Ebben az esetben egy szigorú tartalmazási elvnek kell érvényesülnie.)

A fenti elveknek megfelelően a rendszer három fő komponensből áll:

1. Programkonstrukciós komponensek.
2. Analízis és transzformációs komponens.
3. Végrehajtási komponens.



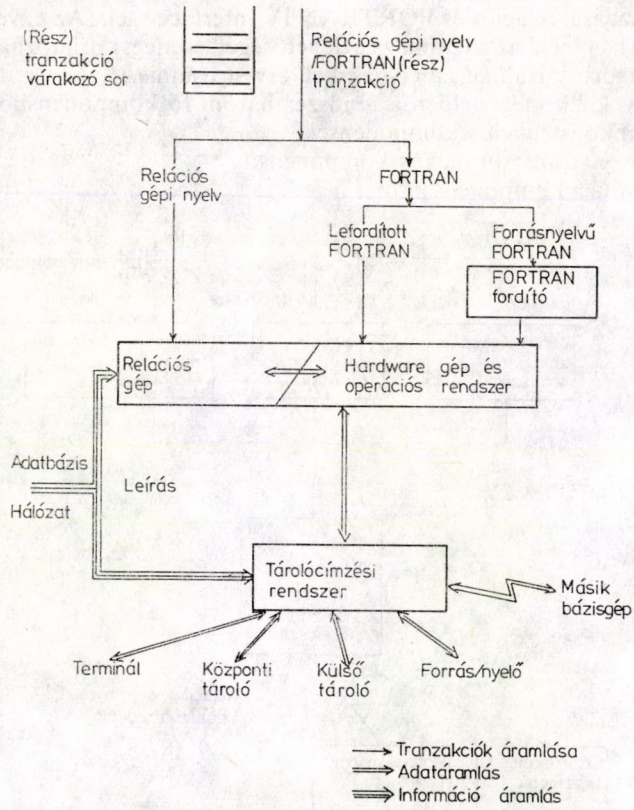
4. ábra. A tranzakciók útja a rendszerben

A fenti komponensek működését a 4. ábra illusztrálja, és egyben mutatja azt is, hogy hogyan dolgoz fel az osztott adatbázis egy programot (illetőleg felhasználói *tranzakciót*).

A felhasználó akár interaktív problémamegoldási technikával, akár valamely hagyományosabb programozási eljárással előállítja azt a tranzakciót, amelyet a géppel végre kíván hajtani. A relációs igen magas szintű nyelven írt tranzakciót egy fordítóprogram fordítja le a relációs gépi nyelvre, míg a FORTRAN tranzakciókat egy előfordító fordítja le standard FORTRAN nyelvre.

A tranzakciókat ezután ellenőrzésnek vetjük alá, és a hatékonyabb végrehajtás érdekében esetleg széttördeljük. A rész-tranzakciókat szétosztjuk az egyes relációs bázisgépek között, amelyek a hálózat azon számítógépének felelnek meg, ahol az adott rész-tranzakció végrehajtódik.

Az 5. ábra mutatja be a végrehajtás folyamatát, ahogyan az a relációs bázisgépeken történik. Az egyes relációs bázisgépre küldött (rész)tranzakciók először egy várakozó sorba kerülnek, ahonnan a rendszer akkor veszi elő, amikor a megfelelő gép hozzáférhetővé válik.

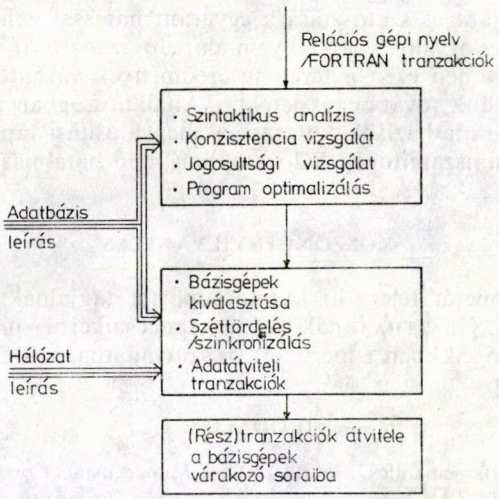


5. ábra. A tranzakció végrehajtása a relációs bázisgépen.

A FORTRAN tranzakciók esetleg előre lefordítottak is lehetnek. Ekkor a programkönyvtárból kell őket behívni, egyébként előzetes fordítást igényelnek. Megjegyezzük, hogy mindig annak a gépnek a FORTRAN fordítóját használjuk, amelyen a tranzakció végrehajtása történik. Ugyanis inhomogén hálózattal állunk szemben, amelyben minden egyes gépnek különbözhet a gépi nyelve.

A FORTRAN programokat, valamint a relációs gépnek az operációs rendszerrel kapcsolatos igényeit a hardware/operációs rendszer rész elégíti ki, ill. hajtja végre. A relációs gép rész ezzel szemben az adatbázis manipulációk végrehajtásával foglalkozik. Hatékonysági meggondolásokból a relációs gépben egy optimalizációs fázis is helyet kapott, amely a hozzáférési információ felhasználásával kijelölheti a tranzakció végrehajtásának leghatékonyabb módját. Valahányszor a rendszeren belül valamely adatra van szükség, a tároló címzési rendszer meghatározza az adat helyét a lokális tárolón, vagy a többi gép tároló címzési rendszerének segítségével, az adat aktuális tárolási helyén.

Mielőtt a (rész)tranzakció bekerülne a különböző relációs gépek várakozó soraiba, a rendszer tranzakció analizáló és transzformáló komponense, főleg helyességi és hatékonysági szempontokból megvizsgálja. A különböző tevékenységeket a 6. ábra mutatja be. A szintaktikus analízis is tartalmaz olyan mélységű konzisztencia- és jogosultsági vizsgálatokat, amelyek még nem igénylik a tranzakció konkrét végrehajtását.



6. ábra. A tranzakció analizáló és transzformálási komponens

Optimalizálásról itt az igen magas szintű nyelvek értelmében beszélünk, mint pl. a közbülső eredményhalmazoknak, bonyolult ciklusoknak, elem behelyettesítéseknek kiküszöbölése stb.

Az optimalizálás után a tranzakciókat az adatmanipulációs funkcióknak, valamint az adatoknak a hálózatban való szétosztása szempontjából vizsgáljuk meg. Amennyiben ez lehetséges, a tranzakciókat olyan részekre bontjuk fel, amelyeket kü-

lönböző bázisgépeken lehet végrehajtani. A megfelelő végrehajtási sorrend biztosítása céljából szinkronizációs mechanizmusokról, a hatékonyság érdekében pedig alkalmas átviteli műveletekről (pl. egy reláció résznek valamely másik gépre való átvitele) kell gondoskodnunk.

4. Összefoglalás és következtetések

Az előző pontban felvázolt architektúrából kiindulva, még sok további kutatást kell elvégeznünk. Ezek olyan lényeges kérdésekre terjednek ki, mint pl.

— a vezérlő információ nyilvántartása
— a hiba utáni helyreállítás elvei
— az adatok szétosztását érintő paraméterek vizsgálata
— a működést javító adat redundancia
— hozzáférési utak optimalizálása
— párhuzamosság és szinkronizáció.

A fenti problémák némelyikének megoldására már születtek javaslatok az irodalomban (pl. STONEBRAKER—NEUHOLD [22]), másokat behatóan kutatnak. Fontosnak tartjuk megjegyezni, hogy a lehetséges megoldások legtöbbje az adatbázisok szemantikai értelmezésének jegyében született. A hibajavításra, a paralelizmusra, az adatok redundanciájára és szétosztására egyaránt hatással van az adatbázisnak az absztrakt modellen és a valós világ tényein alapuló szabatosan definiált információ tartalma. A közeljövőben ezen a téren új eredmények várhatók, e cikk keretében azonban nem mehetünk további részletekbe. Általánosságban azonban elmondhatjuk, hogy az osztott adatbázisok kutatási és megvalósítási tapasztalatai már eddig is megmutatták a miniszámítógép hálózatokban rejlő hatalmas lehetőségeket.

KÖSZÖNETNYILVÁNÍTÁS

A szerző köszönetét fejezi ki kutatócsoportja tagjainak, elsősorban DR. H. BILLERnek és DR. W. GLATTHAARnak hasznos tanácsaikért és munkájukért, amellyel hozzájárultak a jelen cikkben ismertetett osztott adatbázis architektúra kifejlesztéséhez.

IRODALOM

[1] ABRIAL, I. R., "Data semantics", in: Data Base Management, Corsika, ed. Klimbie, North Holland, Amsterdam, 1974.
[2] *AFCET-IP Journées: Bases de données reparties,* Paris, 1977.
[3] ANSI: ANSI/X3/SPARC Study Group on Data Base Management Systems, Interim Report, ANSI 75-02-08. In: *ACM SIGMOD Newsletter FDT* 7 2 1975.
[4] ASTRAHAN, M. M. ET AL.; "System R: A relational approach to data base management", *IBM Research Division,* Report RJ 1738, 1976.
[5] *Berkeley Workshop on Distributed Data Management and Computer Networks,* Berkeley, 1976.
[6] BILLER, H. AND GLATTHAAR, W., On the Semantics of Data Bases: The Semantics of Data Definition Languages, in: *Lecture Notes in Computer Sciences,* Vol. **34** (Springer, Heidelberg, 1975).
[7] BILLER, H., GLATTHAAR, W. AND NEUHOLD, E. J., "On the semantics of data bases: The semantics of data manipulation languages", in: Modelling in Data Base Management Systems, Freudenstadt, ed. G. M. Nijssen, North Holland, Amsterdam, 1976.

[8] BILLER, H. AND NEUHOLD, E. J., "Concepts for the conceptual schema", in: Modelling in Data Base Management Systems, Nice, ed. G. M. Nijssen, North Holland, Amsterdam, 1977.
[9] BILLER, H. AND NEUHOLD, E. J., "Semantics of data bases: The semantics of data models", *Information Systems*, 1977.
[10] *CODASYL: Feature Analysis of Generalized Data Base Management Systems*, ACM, New York, 1971.
[11] *CODASYL: Data Base Task Group Report*, ACM, New York, 1971.
[12] CODD, E. F., "A relational model for large shared data banks", *Communications of the ACM*, **13** (1970).
[13] GLATTHAAR, W. UND PETER, G., „Eine Dialogmaschine für den Betrieb eines Terminals", *Informatikfachberichte* **3** (1976).
[14] *IFIP TC* **2**: *Data Base Description*, Wepion, 1975, North Holland, Amsterdam, 1975.
[15] KARP, P., "TELENET", in: Berkeley Workshop on Distributed Data Management and Computer Networks, Berkeley, 1976.
[16] KERSHBERG, L., KLUG, A. AND TSICHRITZIS, D., "A taxonomy of data models", in: Systems for Large Data Bases, ed. P. C. Lockemann and E. J. Neuhold, North Holland, Amsterdam, 1977.
[17] LEVIN, K. AND MORGAN, H., "Optimizing distributed data bases — A framework of research", in: Proceedings of the Nat. Comp. Conference, AFIPS Press, 1975.
[18] NEUHOLD, E. J., "A formal hierarchical and relational view", Institut für Angewandte Informatik, Bericht Nr. **10**, Universität Karlsruhe, 1973.
[19] NEUHOLD, E. J., "Project proposal for a distributed data base system", Institut für Informatik, Universität of Stuttgart, 1977.
[20] NIJSSEN, G. M., "A gross architecture for the next generation data base management systems", in: Modelling in Data Base Management Systems, Freudenstadt, ed. G. M. Nijssen, North Holland, Amsterdam, 1976.
[21] STONEBRAKER, M. ET AL., "The design and implementation of INGRES", *ACM Transactions on Data Base Systems*, 1976.
[22] STONEBRAKER, M. AND NEUHOLD, E. J., "A distributed data base version of INGRES", Electrical Research Laboratory, University of California, Berkeley, Memo No. ERL-M612, 1976.
[23] SUNDGREN, B., "An infological approach to data bases", Ph. D. Thesis, Urval Nr. 7, Statistika Centralbyran, Stockholm, 1973.

Konferencia anyagok

IFIP TC 2 Adatbázis munkaülések (North Holland, Amsterdam)

[24] KLIMBIE (ed.), *Data Base Management*, Corsika, 1974.
[25] DOUQUÉ, B. C. M. AND NIJSSEN, G. M. (ed.), *Data Base Description*, Wepion, 1975.
[26] NIJSSEN, G. M. (ed.), *Modelling in Data Base Management Systems*, Freudenstadt, 1976.
[27] NIJSSEN, G. M. (ed), *Modelling in Data Base Management Systems*, Nice, 1977.

SIGMOD/SIGFIDET munkaülések (ACM, New York)

[28] CODD, E. F. AND DEAN, A. L. (ed.), *Data Description Access and Control*, San Diego, 1971.
[29] DEAN, A. L. (ed.), *Data Description Access and Control*, Denver, 1972.
[30] RUSTIN, R. (ed.), *Data Description Access and Control*, Ann Arbor, 1974.
[31] KING, W. F. (ed.), *ACM-SIGMOD International Conference on Management of Data*, San Jose, 1975.
[32] POTHNIE, J. B. (ed.), *ACM-SIGMOD International Conference on Management of Data*, Washington D. C., 1976.
[33] *Data Abstraction, Definition and Structure*, Salt Lake City, 1976.

(Beérkezett: 1978. július 29.)

ERICH J. NEUHOLD
INSTITUT FÜR INFORMATIK UNIVERSITÄT STUTTGART
AZENBERGSTRASSE 12, STUTTGART 1, D—7000

THE DESIGN OF DISTRIBUTED DATA BASES

E. J. NEUHOLD

In the last one or two years the importance of networks of minicomputers for the future development of computer applications has increased drastically. Through the development of minicomputer networks it now becomes feasible to introduce the concepts of large integrated data bases into the minicomputer environment. These data bases are distributed over the different computers and the various users do not have to know about the positioning of their data in the network. After introducing the concepts of data base semantics, which have to form the basis for all consistent and reliable data base systems, we introduce the architecture of a distributed data base system as it is currently under development at the *University of Stuttgart*.

A kiadásért felel az Akadémiai Kiadó igazgatója
Műszaki szerkesztő: Marton Andor
A kézirat a nyomdába érkezett: 1979. jan. 5. Terjedelem; 14,35 (A/5) iv
79-84 — Szegedi Nyomda — Felelős vezető: Dobó József

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 a felelős szerkesztő címére kell beküldeni:

Prékopa András, felelős szerkesztő, MTA SZTAKI
1502 Budapest XI., Kende u. 13—17.

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 Ungelichungen«, *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: 42,— Ft

INDEX: 26498
ISSN 0133—3399

TARTALOMJEGYZÉK

INDEX

Alkalmazott matematikai lapok

1978/3-4

A MAGYAR TUDOMÁNYOS AKADÉMIA
MATEMATIKAI ÉS FIZIKAI TUDOMÁNYOK
OSZTÁLYÁNAK KÖZLEMÉNYEI

4.

KÖTET

AKADÉMIAI KIADÓ, BUDAPEST

A MAGYAR TUDOMÁNYOS AKADÉMIA
MATEMATIKAI ÉS FIZIKAI TUDOMÁNYOK OSZTÁLYÁNAK
ALKALMAZOTT MATEMATIKAI LAPJA

IV. 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.

Kéziratok a következő címre küldendők:

Prékopa András, felelős szerkesztő
1502 Budapest XI., Kende u. 13—17.

Ugyanerre a címre küldendő minden szerkesztőségi levelezés.

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 84 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 4 (1978) 165—191

AZ OPTIMALIZÁLÁSELMÉLET KIALAKULÁSÁNAK TÖRTÉNETÉRŐL

PRÉKOPA ANDRÁS

Budapest

LAGRANGE 1788-ban publikálta a függvények egyenlőséges feltételek melletti szélsőértékeinek meghatározására vonatkozó multiplikátoros módszerét, *Mécanique Analytique* című híres könyvében. Több mint 150 év telt el ezután, míg KARUSH, JOHN, KUHN és TUCKER, egyenlőtlenséges feltételek melletti szélsőértékekre vonatkozó munkái megjelentek. Ebben a dolgozatban felhívjuk a figyelmet arra, hogy fontos, idevágó eredmények, analitikus mechanikai felfedezések gyanánt, már a múlt században megszülettek. Megmutatjuk, hogy az ún. *Fourier-féle mechanikai elv* duális alakja, melyet COURNOT írt fel és FARKAS GYULA bizonyított be először, lényegében azonos a nemlineáris programozásban ismert, optimalitási (szükséges) feltételekkel. Megemlítjük még GAUSS, OSZTROGRADSZKIJ és HAMEL idevágó munkásságát, továbbá kitérünk a lineáris egyenlőtlenségekre vonatkozó *Farkas-féle tétel*, valamint néhány ehhez közelálló tétel bizonyos vonatkozásainak a tárgyalására.

1. Bevezetés

FARKAS GYULÁnak a *Crelle Journalban* 1901-ben publikált [23] híres dolgozata a matematikai programozási szakirodalomban egyike lett a leggyakrabban említett dolgozatoknak KUHN és TUCKER 1951-ben publikált *„Nonlinear Programming"* című [54] munkájának megjelenése óta. Ebben az utóbbi dolgozatban a lineáris egyenlőtlenségek alaptétele, *Farkas tétele,* a nemlineáris programozási feladatra vonatkozó szükséges optimalitási feltételek levezetésében nyert felhasználást. A kapott eredmények elősegítették a nemlineáris programozás rohamos fejlődését. Az optimalitásra vonatkozólag FRITZ JOHN kapott hasonló, de gyengébb szükséges feltételeket 1948-ban [51]. Dolgozata régóta ismert már szakmai körökben, ám mindössze néhány éve vált általánosan ismertté KARUSH 1939-ben közölt [52] munkája, melyben megtaláljuk azt az optimalitási tételt, amely KUHN és TUCKER már említett, 1951-ben megjelent dolgozatának fő eredménye.

Dolgozatunkban felhívjuk a figyelmet néhány további munkára, melyek még a múlt században, illetve azelőtt jelentek meg és megmutatjuk, hogy már akkor megszülettek az egyenlőtlenséges feltételek melletti nemlineáris optimalizálás, optimalitásra vonatkozó szükséges feltételei az analitikus mechanika keretein belül. E munkák közül a legfontosabbak szerzői FOURIER, COURNOT, FARKAS, továbbá GAUSS, OSZTROGRADSZKIJ és HAMEL.

Az optimalizáláselmélet korai történetének a felkutatásában sokat segít, ha egy pillantást vetünk FARKAS GYULA [23] dolgozatának első két mondatára:

„Die naturgemässe und zugleich systematische Behandlung der analytischen Mechanik muss das zuerst von FOURIER *und dann später von* GAUSS *formulierte Ungleichheitsprincip der virtuellen Verschiebungen zur Grundlage haben.*

Die Möglichkeit einer solchen Behandlung erfordert aber gewisse Kenntnisse über die homogenen linearen ganzen Ungleichungen, welche bisher so zu sagen, gänzlich gefehlt haben."[1]

A fentiekből látjuk, hogy FARKAS jól meghatározott cél érdekében fejlesztette ki a lineáris egyenlőtlenségek elméletét. Az elméleti fizika professzora volt a *Kolozsvári Egyetemen.* Feltételezhetjük tehát, hogy a lineáris egyenlőtlenségekre vonatkozó eredményeit maga is alkalmazta a mechanikai egyensúly problémájára. Valóban, a sajnos már itthon sem eléggé ismert, egykori akadémiai folyóiratokban meg is találjuk a [13], [14] dolgozatokat, melyekben a szerző a *Magyar Tudományos Akadémián,* 1894. december 17-én, a *Fourier-féle mechanikai elv* alkalmazásairól tartott előadásának anyagát publikálta. A *Crelle Journalban* megjelent [23] legismertebb cikke lényegében összefoglalása a lineáris egyenlőtlenségekre vonatkozó, korábban már közölt eredményeinek, ebben azonban nem tesz említést a *Fourier-féle elvre* vonatkozó alkalmazás mikéntjéről. Emiatt munkásságának ez a vonatkozása nem vált nemzetközileg ismertté. Ennek oka az is, hogy az analitikus mechanikában nyert eredmény optimalizáláselméleti interpretálása akkor nem történt meg, márpedig úgy tűnik, hogy ilyen irányú jelentősége fontosabb, mint a mechanikai.

Az említett interpretációt dolgozatunkban elvégezzük, ám egyben a *Fourier-elvre* vonatkozó, FARKAS előtti legfontosabb eredmények ismertetésére is kitérünk. Mindenekelőtt röviden áttekintjük azokat a mechanikai elveket, amelyek témánk szempontjából a legfontosabbak.

2. A mechanikai egyensúly elvei

A virtuális munka elvét BERNOULLI JÁNOS mondta ki 1717-ben. Ez VARIGNON egy könyvében jelent meg először ugyanebben az évben. Az elv megfogalmazását LAGRANGEtól idézzük [56/I, 21—22. oldal]:

„Le principe des vitesses virtuelles peut être rendu très général de cette manière:

Si un système quelconque de tant de corps ou points que l'on veut, tirés chacun par des puissances quelconques, est en équilibre, et qu'on donne à ce système un petit mouvement quelconque, en vertu duquel chaque point parcoure un espace infiniment petit qui exprimera sa vitesse virtuelle, la somme des puissances, multipliées chacune par l'espace que le point où elle est appliquée parcourt suivant la direction de cette même puissance, sera toujours égale à zéro, en regardant comme positifs les petits espaces parcourus dans le sens des puissances, et comme négatifs les espaces parcourus dans un sens opposé.

JEAN BERNOULLI *est le premier, que je sache, qui ait aperçu cette grande généralité du principe des vitesses virtuelles, et son utilité pour résoudre les problèmes de Statique. C'est ce qu'on voit dans une de ses Lettres à* VARIGNON, *datée de 1717, que ce dernier a plaeée à la tête de la Section neuvième de sa Nouvelle Mécanique, Section employée*

[1] Az analitikus mechanika természetes és szisztematikus tárgyalásának alapját az előbb FOURIER, és később GAUSS által megfogalmazott egyenlőtlenségi elvnek kell alkotnia.

Egy ilyen tárgyalás lehetősége azonban megkövetel bizonyos ismereteket a homogén, lineáris, egész egyenlőtlenségekkel kapcsolatban, melyek eddig úgyszólván teljesen hiányoztak.

tout entière à montrer par différentes applications la vérité et l'usage du principe dont il s'agit."[2]

A *Bernoulli-elvet* LAGRANGE a mechanika axiómájának tekintette. Idézett könyvének 27. oldalán a következőket olvassuk:

„C'est dans cette loi que consiste ce qu'on appelle communément le principe des vitesses virtuelles, principe reconnu depuis longtemps pour le principe fondamental de l'équilibre, ainsi que nous l'avons montré dans la Section précédente, et qu'on peut, par conséquent, regarder comme une espèce d'axiome de Mécanique."[3]

Ha a mechanikai rendszerre ható erők konzervatívak, vagyis komponenseiket egyetlen skalár függvény parciális deriváltjaiként származtathatjuk, akkor a COURTIVRON által kimondott elv írja le az egyensúlyi állapotot. Ezzel kapcsolatban LAGRANGE a következőket írja [56/I., 70. oldal]:

„...ce qui donne cet autre principe de Statique, que, de toutes les situations que prend successivement le système, celle où il a la plus grande ou la plus petite force vive est aussi celle où il le faudrait placer d'abord pour qu'il restât en équilibre. [VOIR COURTIVRON, *les Mémories de l'Académie des Sciences de* 1748 *et* 1749.]"[4]

LAGRANGE elégséges feltételt adott arra vonatkozólag, hogy a potenciál minimális értéket vesz fel. BERTRAND (aki LAGRANGE összegyűjtött műveit jegyzetekkel látta el) megjegyezte, hogy a feltétel bizonyítása nem teljes és hogy DIRICHLET később ugyanerre korrekt bizonyítást adott [10], [11].

FOURIER 1798-ban mondta ki a róla elnevezett mechanikai elvet, *„Mémoire sur la Statique"* című munkájában. Az elv az egyenlőtlenséges kényszerfeltételek esetére vonatkozik. FARKAS úgy vélte [13, 458. oldal], hogy az elv kimondása nem a legfontosabb mondanivalója FOURIER említett dolgozatának. Erre enged következtetni a dolgozat alcíme is: *„Contenant la démonstration du principe des vitesses virtuelles."* Ám ezt a túlságosan általános „bizonyítást", melyet a szerző a fizikai rendszerek széles körére vonatkoztat és nem tesz különbséget a fizikai valóság és (mai szakkifejezéssel élve) annak matematikai modellje között, a mechanika művelői

[2] „A virtuális sebességek elve nagyon általánosan a következő alakra hozható:
Ha valamely, tetszésünk szerint testekből, vagy pontokból álló rendszer, melyre valamilyen erők hatnak, egyensúlyban van és e rendszernek valamilyen kis mozgást adunk, melynek alapján mindegyik pont befut egy végtelenül kis szakaszt, mely virtuális sebességét fogja kifejezni, összeadva az erők mindegyikének szorzatát azzal a szakasszal, amelyet a támadás helyén levő pont az erő hatására befut, ez mindig zéróval lesz egyenlő, ha az erő irányával megegyezően befutott szakaszt pozitívnak, az ellenkező irányban befutott szakaszt negatívnak tekintjük.
BERNOULLI JÁNOS volt az első, tudomásom szerint, aki a virtuális sebességek elvének ezt az általánosságát és a statika problémái megoldásában való hasznosságát észrevette. Ez az, amit látunk az egyik, 1714-ben kelt levélben a VARIGNONhoz írt Levelek közül, amelyet ez utóbbi *Nouvelle Mécanique* című könyve kilencedik Fejezetének elején helyezett el és az egész Fejezetet a szóban forgó elv igazsága és felhasználása különböző alkalmazásokon keresztül való megmutatásának szentelte."

[3] „Ebben áll a közönségesen virtuális sebességek elvének nevezett elv, melyet hosszú ideje az egyensúly alapelvének ismernek el, amint azt az előző Fejezetben mi is megmutattuk és amelyet következésképpen, a Mechanika egyfajta axiómájának tekinthetünk."

[4] „...ez a Statikának azt a másik elvét szolgáltatja, miszerint az összes helyzetek közül, melyet a rendszer egymás után felvesz, az, amelyben a legnagyobb, vagy a legkisebb eleven erővel bír, ugyanaz, mint amelybe azt előbb helyezni kellene, hogy egyensúlyban maradjon. [Lásd COURTIVRON, *Mémoires de l'Académie des Sciences* 1748 és 1749.]"

nem fogadták el (lásd pl. a [77] cikket). Emiatt, FARKAS véleményével ellentétben, mégiscsak az elv kimondását tekintjük FOURIER fent említett dolgozata legfőbb mondanivalójának. FOURIER [34] dolgozata 488. oldaláról idézzük a következőket:

„Comme il arrive souvent que les points du système s'appuient seulement sur les obstacles fixes, sans y être attachés, il est évident qu'il y a des déplacements possibles qui ne satisfont pas aux équations de condition: on voit encore que, par ces déplacements, le moment des résultantes est nécessairement positif, puisque la direction de ces forces doit être perpendiculaire aux surfaces resistantes. Ainsi la somme des moments des forces appliquées est positive pour tous les déplacements de cette espèce; mais il est impossible que l'on dérange un corps dur, en équilibre, de sorte que le moment total des forces appliquées soit négatif. Au reste, si l'on considère les résistances comme des forces, ce qui fournit, comme on le sait, le moyen d'estimer ces résistances, le corps peut être regardé comme libre, et la somme des moments est nulle pour tous les déplacements possibles."[5]

Egy mechanikai rendszerre ható $P, Q, R, \ldots$ erők momentumán (pillanatnyi munkáján) az alábbi skaláris szorzatösszeget értjük:

$$P\delta p + Q\delta q + R\delta r + \ldots, \tag{2.1}$$

ahol $\delta p, \delta q, \delta r, \ldots$ az elmozdulás variációi. A rendszer állapotát (a fázistérben) meghatározó pontból kiindulva a variációk nem sértik meg a feltételeket. A *Bernoulli-elv* kimondja, hogy a (2.1) összeg 0-val egyenlő, a *Fourier-elv* pedig kimondja, hogy (2.1) nempozitív, ha a rendszer stabilis egyensúlyi állapotban van. FOURIER az elmozdulások helyett „fluxiókkal" dolgozott, melyek az elmozdulások -1-szeresei, amint az idézett dolgozatának egyéb részeiből kitűnik. Ez magyarázza azt, hogy az elv kimondásakor FOURIER az erők momentumának nemnegativitását kívánta meg.

Ha potenciál létezik és ezt V jelöli, vagyis fennállnak az alábbi egyenlőségek:

$$P = -\frac{\partial V}{\partial p}, \quad Q = -\frac{\partial V}{\partial q}, \quad R = -\frac{\partial V}{\partial r}, \ldots,$$

ahol a jobb oldalakon álló deriváltak a komponensenként vett deriváltak tömör jelölései, akkor az a követelmény, hogy a (2.1) összeg nempozitív, a következő alakot ölti:

$$\frac{\partial V}{\partial p}\delta p + \frac{\partial V}{\partial q}\delta q + \frac{\partial V}{\partial r}\delta r + \ldots \leqq 0. \tag{2.2}$$

A bal oldalon $\delta p, \delta q, \delta r, \ldots$ helyébe a közelítőleg egyenlő $dp, dq, dr, \ldots$ differenciálokat helyettesítve, a V skalár függvény teljes differenciálját kapjuk. A (2.2) egyen-

[5] „Minthogy gyakran megtörténik, hogy a rendszer pontjait a rögzített kényszerek csak megtartják, anélkül, hogy azok ott rögzítve lennének, világos hogy vannak olyan lehetséges elmozdulások, melyek nem tesznek eleget egyenlőséges feltételeknek: azt is látjuk, hogy az eredőknek ezekkel az elmozdulásokkal vett momentuma szükségképpen pozitív, minthogy ezeknek az erőknek az iránya az ellenállási felületekre merőleges kell, hogy legyen. Így a ható erők momentumainak összege pozitív minden ilyen fajta elmozdulás esetén; és nem lehetséges egy szilárd testet egyensúlyi helyzetéből oly módon elmozdítani, hogy a ható erők teljes momentuma negatív legyen. Végül, ha az ellenállásokat erőknek tekintjük, miáltal, mint tudjuk, eszközt kapunk az ellenállások becslésére, a test szabadnak tekinthető és a momentumok összege nulla minden lehetséges elmozdulás esetén."

lőtlenségre, illetve az általánosabb elv érvényességére vonatkozólag FOURIER munkájában — a mai értelemben vett — korrekt bizonyítás nem található. Másfelől, ma már tudjuk azt, hogy alkalmas regularitási feltétel *(constraint qualification)* nélkül (2.2) érvényessége nem is vezethető le a V függvény minimumának szükséges feltételeként. FOURIER megjegyezte, hogy ha az általa kimondott egyenlőtlenségi elv a (2.2) feltételre redukálódik, akkor az egyensúlyi állapotot a V függvény minimalizálása (!) révén határozhatjuk meg.

GAUSS 1829-ben újból kimondta az egyenlőtlenségi elvet, nem ismervén FOURIER munkáját. Az erre vonatkozó szövegrészt egy lábjegyzettel is ellátta. Mind a kettőt alább idézzük [38, 234. oldal]:

„Nach dem Princip der virtuellen Gesehwindigkeiten erfordert dies Gleichgewicht, dass die Summe der Producte aus je drei Factoren, nemlich jeder der Massen m, m', m'', usw., den Linien $cb, c'b', c''b''$ usw., und irgend welchen auf letztere resp. projicirten, vermöge der Bedingungen des Systems möglichen Bewegungen jener Punkte, immer $=0$ sei, wie man es gewöhnlich ausspricht, oder richtiger, dass jene Summe niemals positiv werden könne."[6]

Lábjegyzet:

„Der gewöhnliche Ausdruck setzt stillschweigend solche Bedingungen voraus, dass die jeder möglichen Bewegung entgegengesetzte gleichfalls möglich sei, wie z. B., dass ein Punkt auf einer bestimmen Fläche zu beliben genöthigt, dass die Entfernung zweier Punkte von einander unveränderlich sei u. dergl. Allein dies ist eine unnöthige und der Natur nicht immer angemessene Beschränkung. Die Oberfläche eines undurchdringlichen Körpers zwingt einen auf ihr befindlichen materiellen Punkt nicht, auf ihr zu beliben, sondern verwehrt ihm bloss das Austreten auf die Eine Seite; ein gespannter, nicht ausdehnbarer aber biegsamer Faden zwischen zwei Punkten macht nur die Zunahme, nicht die Abnahme der Entfernung unmöglich usw. Warum wollten wir also das Gesetz der virtuellen Geschwindigkeiten nicht lieber gleich anfangs so ausdrücken, dass es alle Fälle umfasst?"[7]

1834-ben a *Szent-Pétervári Birodalmi Tudományos Akadémia* egyik ülésén tartott előadásában OSZTROGRADSZKIJ is kimondta az egyenlőtlenségi elvet. VOSS megjegyezte [77, 74. oldal], hogy „daher heisst in *Russland* das *Princip von Fourier* auch wohl das von OSTROGRADSKY".[8] Nagyon érdekes olvasni, hogy mit írt OSZTROGRADSZKIJ az egyenlőtlenségi elvről:

[6] "A virtuális sebességek elve szerint ez az egyensúly megkívánja, hogy mindhárom faktor szorzatösszege, melyek az m, m', m'' etc. tömegek, a $cb, c'b', c''b''$ etc. egyenes szakaszok és az egyes pontok olyan lehetséges mozgásai fentiekre eső vetületei, melyek a kényszerfeltételeknek megfelelnek, mindig $=0$ legyen, ahogy ezt szokás szerint kimondjuk, vagy helyesebben, hogy ez az összeg sohasem lehessen pozitív."

[7] „A szokásos kifejezés hallgatólagosan olyan feltételeket tételez fel, hogy minden lehetséges mozgás ellentétje ugyanúgy lehetséges legyen, mint pl., hogy egy pont egy meghatározott felületen maradjon, hogy két pont közötti távolság változatlan legyen és ehhez hasonlók. Mindazonáltal ez egy szükségtelen és a természetnek nem mindig megfelelő korlátozás. Egy áthatolhatatlan test felülete egy rajta található anyagi pontot nem abban korlátoz, hogy rajta maradjon, hanem csak a másik oldalra való áthaladásban gátolja meg; egy két pont között kifeszített, nem nyújtható, de hajlítható fonal a távolságnak csak a növekedését akadályozza meg, a csökkenését nem etc. Miért ne akarnánk tehát a virtuális sebességek elvét inkább már kezdetben oly módon megfogalmazni, hogy az minden esetet átfogjon?"

[8] „Ezért *Oroszországban a Fourier-elvet Osztrogradszkij-elvnek* is nevezik."

„Il est très surprenant de voir que dans la nouvelle édition de la Mécanique analytique, édition publiée à l'époque où l'on connaissait déjà toute l'étendue du principe des vîtesses virtuelles, LAGRANGE *non seulement n'a fait aucun usage de l'observation, que dans l'équilibre des forces le moment total pouvait acquérir une valeur négative, mais qu'il l'a en quelque sorte écartée quand elle s'est présentée comme d'elle même, dans la démonstration qu'il a donnée du principe des vîtesses virtuelles; cependant, faute d'y avoir eu égard, ce grand géomètre a incomplètement énuméré les déplacemens possibles dans la plupart des questions de la première partie de la Mécanique analytique, et il est facile de reconnaître que les déplacemens qu'il a négligé de considérer, ne sont empêchés par aucune condition, en sorte que, toutes les équations qu'il a établies pour l'équilibre étant satisfaites, l'équilibre cependant pourrait n'avoir pas lieu.*

Nous nous proposons, dans ce mémoire, d'exposer l'analyse relative à l'emploi du principe des vîtesses virtuelles considéré dans toute sa généralité et de compléter la solution de plusieurs questions traitées dans la première partie de la Mécanique analytique."[9]

LAGRANGE multiplikátoros módszerét a *Mécanique Analytique* című könyve első kötetében (77—79. oldal) publikálta. Érvényességét algebrai módon bizonyította. (Ez a zseniális módszer hosszú ideig nem volt képes a matematikus egyetemi hallgatók érdeklődését felkelteni. Manapság sok előadó már azt az utat követi, hogy előbb bebizonyítja a megfelelő tételt az egyenlőtlenséges feltételek esetére, tehát — ahogyan ez 1951 óta elterjedt — az ún. *Kuhn—Tucker-tételt,* mely geometriailag jól szemléltethető és utána megemlíti, hogy az egyenlőséges feltételek esetére vonatkozó tétel ebből könnyen adódik. A tapasztalat szerint ez a módszer sokkal vonzóbbá teszi a témát a hallgatóság számára.) Később meglátjuk, hogy milyen eredmények születtek a *Fourier-elvvel* kapcsolatban. Most néhány oldalt az ún. *Farkas-tétel* tárgyalásának szentelünk.

3. A lineáris egyenlőtlenségekre vonatkozó Farkas-tétel

FARKAS GYULA 1847-ben született a *Fejér* megyei *Sárosd* községben és *Pestszentlőrincen* halt meg 1930-ban, itt is van eltemetve. Először a *Pesti Egyetem Jogi Karának* hallgatója lett és egyidejűleg zenét is tanult, később azonban úgy döntött, hogy a *Bölcsészettudományi Karon* folytatja tanulmányait, fizikát, kémiát és matematikát tanult. A középiskolai tanári oklevelet 1876-ban szerezte meg. 1887-ben a *Kolozsvári Egyetemen* az elméleti fizika professzora lett és ezt a pozíciót 1915-ig

[9] „Nagyon meglepő azt látni, hogy a *Mécanique analytique* új kiadásában, mely olyan korban jelent meg, amikor a virtuális sebességek elvének teljes kiterjedése már ismert volt, LAGRANGE nemcsak nem használta fel azt az észrevételt, hogy az erők egyensúlyában a teljes momentum negatív értéket is elérhet, hanem azt bizonyos módon kizárta, amint az a virtuális sebességek elvére általa adott bizonyításból kitűnik; eközben, nem véve ezt figyelembe, ez a nagy matematikus hiányosan számlálta össze a lehetséges elmozdulásokat a *Mécanique analytique* első részének legtöbb kérdésében és könnyű felismerni, hogy az általa figyelembe nem vett elmozdulásokat nem akadályozza meg semmilyen feltétel, úgyhogy, ha teljesülnek is az általa az egyensúlyra vonatkozólag felállított egyenletek, közben az egyensúly nem biztos, hogy bekövetkezik.

Ebben a memoárban javasoljuk az analízis alkalmazását a virtuális sebességek elvére, azt teljes általánosságában tekintve és a *Mécanique analytique* első részében tárgyalt különféle kérdések megoldásának teljessé tételét."

tartotta. Részt vett az egyetem vezetésében is, a dékáni és a rektori tisztség betöltése révén. A *Magyar Tudományos Akadémia* 1896-ban levelező, 1914-ben pedig rendes tagjává választotta. Egyre erősödő szembaja arra késztette, hogy 1915-ben visszavonuljon az egyetemi tanárságtól, ekkor költözött *Pestszentlőrincre.* Fiatal emberként több tanulmányt írt a numerikus analízis, a differenciálegyenletek, az elliptikus függvények és a zeneelmélet köréből. Legfontosabb tudományos eredményei a *Fourier-féle mechanikai elvvel,* a lineáris egyenlőtlenségekkel és a termodinamika különböző problémáival kapcsolatosak. Emlékére a *Bolyai János Matematikai Társulat* 1974-ben *Farkas Gyula-díjat* alapított, melyet évente nyernek el fiatal matematikusok az alkalmazott matematika terén elért tudományos eredményükért. FARKAS GYULA életére és munkásságára vonatkozó további információkat az olvasó a [32, 33, 66, 67, 74] művekből szerezhet.

1901-ben megjelent [23] dolgozatát elsősorban a homogén lineáris egyenlőtlenségekre vonatkozó tétele miatt idézik. E tételt alább megfogalmazzuk. A továbbiakban általában a vektorokat oszlopos felírásban képzeljük el, vagyis minden vektor egy egyetlen oszlopból alkotott mátrix. A mátrixok transzponáltját vesszővel jelöljük.

Legyenek $\mathbf{g}_1, \ldots, \mathbf{g}_m, \mathbf{g}$ m-komponensű vektorok és tekintsük az alábbi lineáris egyenlőtlenségeket

$$\mathbf{g}_i'\mathbf{x} \geqq 0, \quad i = 1, \ldots, M, \tag{3.1}$$

$$\mathbf{g}'\mathbf{x} \geqq 0. \tag{3.2}$$

Ha a (3.2) egyenlőtlenség fennáll minden olyan $\mathbf{x}$ esetére, melyre a (3.1) egyenlőtlenségek fennállnak, akkor azt mondjuk, hogy (3.2) következménye a (3.1) egyenlőtlenségeknek. *Farkas tétele* a következő:

3.1. TÉTEL. A (3.2) egyenlőtlenség akkor és csak akkor következménye a (3.1) egyenlőtlenségeknek, ha léteznek olyan nemnegatív $\lambda_1, \ldots, \lambda_M$ számok, hogy fennáll a

$$\mathbf{g} = \lambda_1\mathbf{g}_1 + \ldots + \lambda_M\mathbf{g}_M \tag{3.3}$$

egyenlőség.

FARKAS ezt a tételt először 1894-ben publikálta [13]. A dolgozat német változata 1895-ben jelent meg [14]. Az ezekben közölt bizonyítás azonban nem teljes. Ugyancsak hiányos a második bizonyítása, melyet 1896-ban közölt a magyar nyelvű [15] dolgozatban. Ugyanez német nyelven is megjelent 1899-ben [16]. Az 1898-ban magyarul és az 1899-ben németül írt cikkekben [17, 18] szereplő bizonyítás korrekt. Ez azonos azzal a bizonyítással, mely azután a *Crelle Journalban* 1901-ben megjelent cikkében is szerepel,

Az első bizonyítások hézagos voltát akkoriban már felismerték. Erre utal FARKAS GYULÁNAK RÉTHY MÓRhoz írt levele (1903. május 21.):

„Kedves Barátom!

Az 1896-ban Math. és Phys. Lapokban közölt bizonyításom is hézagos. Azonban az Értesítőben közölt és vele teljesen egyezően a Crelleben közölt bizonyításom teljes és minden részében egészen korrekt. Ezt úgy tudom, mint azt, hogy $2\times2=4\ldots$. *A Math. és Phys. Lapok 1898. évi kötetében közölt dolgozatom egy rendszer paramé-*

teres megoldását szolgáltatja, amiben a multiplikátoros tételre támaszkodik. Ez Crelleben a VI. pontban van. Róla is egészen bizonyos vagyok benne, hogy teljes és korrekt, valamint mindenről, amit a Crelleben közöltem. Ez alkalommal tudatom Veled, hogy amikor Crelleben dolgozatom megjelent, levelet kaptam MINKOWSKItól *(aki most göttingeni professzor), hogy az általam tárgyalt kérdések nagy részét ő is elintézte „Geometrie der Zahlen 1896" munkájában. Ebben (Erste Lieferung) a 40., 41., 42., 43. lapon a paraméteres megoldás egy módját adja; a 44. és 45. lapon a multiplikátoros tételekhez ugyancsak geometriai fogalmakat használ, persze többdimenziós térben.*

A legmelegebben üdvözöl igaz barátod

Farkas Gyula"

Az első bizonyítás kijavítása csak a mai matematikai programozási tudás alapján lehetséges. Nincs tudomásunk arról, hogy a hiányosság mibenlétét akkoriban felfedezték volna. FARKAS GYULA valószínűleg nem tekintette ezt komolynak. Az 1918-ban megjelent [29] dolgozatában ugyanis azt írta, hogy tételének hat bizonyítása ismeretes. Utalt MINKOWSKI [61] és HAAR [42] munkáira, a fent említett három bizonyításra és egyre, mely egyetemi jegyzetében jelent meg. Utolsó, lineáris egyenlőtlenségekre vonatkozó [31] dolgozata 1926-ban jelent meg, melyben szintén bizonyítani véli tételét. Ebben az öregkori dolgozatban valójában a tételt csak szemlélteti, de nem bizonyítja. A jó bizonyítása a teljes indukció elvén alapul. MINKOWSKI és HAAR (az utóbbi az általánosabb, inhomogén egyenlőtlenségekre vonatkozó tételt bizonyította) bizonyításai ettől különböznek. Lássuk, milyen hibát követett el FARKAS GYULA a tétel első „bizonyításában". A [14] dolgozatból fogunk idézni.

Először megmutatja, miszerint az általánosság megszorítása nélkül feltehetjük, hogy a lineárisan független (gradiensű) lineáris egyenlőtlenségek száma ugyanannyi, mint a változók száma. Ezután a következmény-egyenlőtlenség együttható vektorát előállítja a többi egyenlőtlenség együttható-vektorainak lineáris kombinációjaként. Ha most, az egyszerűség kedvéért feltételezzük, hogy az M számú egyenlőtlenség közül az első n egyenlőtlenség lineárisan független, akkor a lineáris előállításában ezek szorzóit a többi szorzó segítségével kifejezhetjük az alábbi módon:

$$\begin{aligned} \lambda_1 &= I_0 + I_1\lambda_{n+1} + I_2\lambda_{n+2} + \dots \\ \lambda_2 &= K_0 + K_1\lambda_{n+1} + K_2\lambda_{n+2} + \dots \\ &\vdots \\ \lambda_n &= \dots \end{aligned} \tag{3.4}$$

Az ezután következőket a részletesebb [14] német nyelvű dolgozatból idézzük. FARKAS azt írja, hogy [14, 270. oldal]:

„Nur dann ist es unmöglich, dass alle λ-Grössen zugleich nichtnegative Werte erhalten können, wenn in (8) [(3.4)] *wenigstens eine rechte Seite- oder eine Summe von nicht-negativen vielfachen der rechten Seiten die doppelte Beschaffenheit aufweist, dass ihr erstes Glied Negativ ist, und die Coëfficienten ihrer Glieder entweder negativ sind, oder höchstens verschwinden (da durch Eliminationen positiver Glieder mittels entsprechender negativen, aus einem mit positiven λ durchaus unerfüllbarem Teile von* (5) [(3.2)] *nach und nach aequivalente Teil-Systeme erhalten werden können). In diesem Falle hat aber* (4) [(3.1)] *schon Auflösungen, durch welche* (5) [(3.2)] *nicht*

befriedigt wird. Sei es nämlich, dass diese doppelte Eigenschaft der rechten Seite der ersten Gleichung in (8) [(3.4)] *zukommt, dass also*

$$I_0 < 0; \quad I_1 \leqq 0, \quad I_2 \leqq 0, \ldots"\ ^{10}$$

FARKAS e helyen anticipálta a *Lemke-féle ún. duál-módszert* [58]. Ez azonban nem feltétlenül véges, amint azt HOFFMAN [50] és BEALE [2] megmutatta. A duál módszer lexikografikus változatának [58] (mely a *Charnes-féle perturbációs gondolatokon* alapul) az alkalmazásával FARKAS bizonyítása kijavítható. A bizonyítás így sem lesz bonyolultabb, ha a lexikografikus duál módszerre a [71] könyvben nyújtott tárgyalás igen egyszerű voltát figyelembe vesszük. Röviden összefoglaljuk a FARKAS-tételre ily módon származó bizonyítást.

Farkas-tételének bizonyítása. Először megjegyezzük, hogy ha a (3.1), (3.2) egyenlőtlenségekben alkalmazzuk az $\mathbf{y}=\mathbf{Bx}$ lineáris transzformációt, ahol $\mathbf{B}$ nemszinguláris négyzetes mátrix, akkor az új (3.2) egyenlőtlenség az új (3.1) egyenlőtlenségek következménye marad. Továbbá, ha bebizonyítjuk az új egyenlőtlenségekre *Farkas tételét,* akkor ugyanezekkel a $\lambda_1, \ldots, \lambda_M$ szorzókkal a (3.3) egyenlőség is érvényes lesz.

Először feltételezzük, hogy (3.1)-ben a lineárisan független egyenlőtlenségek száma ugyanannyi, mint amennyi a változók száma. Tekintsük az alábbi lineáris programozási feladatot:

$$\text{minimalizálandó} \quad (0\lambda_1+\ldots+0\lambda_M)$$

$$\text{(3.5)} \qquad \text{feltéve, hogy} \quad \mathbf{g}_1\lambda_1+\ldots+\mathbf{g}_M\lambda_M=\mathbf{g},$$

$$\lambda_1 \geqq 0, \ldots \lambda_M \geqq 0.$$

Tetszőleges bázisból kiindulva (mindegyik bázis duál-megengedett) és alkalmazva a lexikografikus duál módszert, végül eljutunk a (3.5) feladat egyenlőséges feltételeinek egy olyan (3.4) típusú alakjához, melyben vagy fennáll az, hogy $I_0 \geqq 0, K_0 \geqq 0, \ldots$, vagy van olyan sor, hogy a jobb oldalon a konstans tag negatív, míg a változók valamennyi együtthatója nemnegatív. A második eset nem fordulhat elő. Ugyanis a lexikografikus duál módszer garantálja, hogy a (3.4)-beli egyenlőség alapján megalkotható n-méretű oszlopvektorok a $\mathbf{g}_1, \ldots, \mathbf{g}_m, \mathbf{g}$ vektorokból egy (közös) nemszin-

[10] „Csak akkor lehetetlen, hogy valamennyi λ-érték egyidejűleg nemnegatív értéket vehessen fel, ha (8)-ban [(3.4)-ben] legalább egy jobb oldal, vagy a jobb oldalak nemnegatív számokkal vett szorzat-összege azzal a kettős tulajdonsággal bír, hogy első tagja negatív és a többi tagok együtthatói vagy negatívak, vagy legfeljebb eltűnnek (minthogy a pozitív tagnak a megfelelő negatív tagok segítségével történő eliminációjával az (5)-nek [(3.2)-nek] pozitív λ számokkal semmiképpen ki nem elégíthető részéből egymás után ekvivalens részrendszerek kaphatók). Ebben az esetben azonban (4)-nek [(3.1)-nek] vannak olyan megoldásai, melyek (5)-nek [(3.2)-nek] nem tesznek eleget. Legyen ugyanis (8)-ban [(3.4)-ben] az első egyenlet olyan, mely ezzel a kettős tulajdonsággal bír, tehát

$$I_0 < 0; \quad I_1 \leqq 0; \quad I_2 \leqq 0, \ldots"$$

guláris mátrixú lineáris transzformációval nyerhetők, tehát a fent mondottak az $y_1, \ldots, y_n$ változóban felírt

$$
\begin{aligned}
y_1 &\geqq 0 \\
y_2 &\geqq 0 \\
&\ddots \\
y_n &\geqq 0 \\
-I_1 y_1 - K_1 y_2 - \ldots &\geqq 0 \\
-I_2 y_1 - K_2 y_2 - \ldots &\geqq 0 \\
&\vdots
\end{aligned} \tag{3.6}
$$

lineáris egyenlőtlenségeknek következménye az alábbi:

$$I_0 y_1 + K_0 y_2 + \ldots \geqq 0. \tag{3.7}$$

Ha mármost mondjuk $I_0 < 0$, $I_1 \leqq 0$, $I_2 \leqq 0, \ldots$, akkor az $y_1 = 1$, $y_2 = \ldots = y_n = 0$ komponensekből alkotott vektor eleget tesz a (3.6) egyenlőtlenségeknek, de nem tesz eleget a (3.7) egyenlőtlenségnek, tehát (3.7) nem következménye (3.6)-nak. Minthogy ez nem áll fenn, ilyenformán beláttuk, hogy valóban csak az $I_0 \geqq 0, K_0 \geqq 0, \ldots$ eset lehetséges. Ezzel *Farkas tételét* a vizsgált speciális esetre bebizonyítottuk.

Ha a (3.1) egyenlőtlenségek között a lineárisan függetlenek száma $h (< M)$ és egyszerűség kedvéért feltételezzük, hogy $\mathbf{g}_1, \ldots, \mathbf{g}_h$ lineárisan függetlenek, akkor választunk olyan $\mathbf{d}_1, \ldots, \mathbf{d}_{n-h}$ vektorokat, hogy a $\mathbf{B} = (\mathbf{g}_1, \ldots, \mathbf{g}_h, \mathbf{d}_1, \ldots, \mathbf{d}_{n-h})$ mátrix nemszinguláris legyen, majd alkalmazzuk az $\mathbf{y} = \mathbf{Bx}$ transzformációt a (3.1), (3.2) egyenlőtlenségekre. Ekkor az új (3.1) egyenlőtlenségekben csak az $y_1, \ldots, y_h$ változók szerepelnek és minthogy ezeknek következménye az új (3.2) egyenlőtlenség, az $y_{h+1}, \ldots, y_n$ változóknak az utóbbiban is 0-val egyenlő együtthatói vannak. Most, elfelejtve az $y_{h+1}, \ldots, y_n$ változókat, a fenti bizonyítást alkalmazzuk a csak $y_1, \ldots, y_h$ változókat tartalmazó egyenlőtlenségekre. Ennek végén visszavesszük az $y_{h+1}, \ldots, y_n$ változókat 0 együtthatókkal és utána visszaállítjuk az eredeti egyenlőtlenségeket az $\mathbf{y} = \mathbf{Bx}$ transzformáció révén. A nyert $\lambda_1, \ldots, \lambda_M$ számokkal — mint említettük — fennáll a (3.3) egyenlőség. Ezzel *Farkas tételét* bebizonyítottuk. Egyben bizonyítottuk még az alábbi tételt is.

3.2. TÉTEL. Ha a (3.2) egyenlőtlenség nem következménye a (3.1) egyenlőtlenségnek, akkor létezik olyan, nemszinguláris $\mathbf{B}$ mátrixszal vett $\mathbf{y} = \mathbf{Bx}$ lineáris transzformáció, hogy az új egyenlőtlenségekben valamely y_i változó (3.2)-beli együtthatója negatív, míg (3.1)-beli együtthatói mind nemnegatívak.

4. Az egyensúly szükséges feltétele

Tegyük fel, hogy a vizsgált mechanikai rendszer állapotai leírhatók az R^n tér alábbi feltételeknek eleget tevő $\mathbf{x}$ vektoraival:

$$g_i(\mathbf{x}) \geqq 0, \quad i = 1, \ldots, m. \tag{4.1}$$

Jelöljék $X_1, \ldots, X_n$ a rendszerre ható erők komponenseit és tegyük fel, hogy az

$\mathbf{x}^* \in R^n$ pontban a rendszer egyensúlyban van. Ekkor, a *Fourier-elv* szerint fennáll az alábbi egyenlőtlenség:

$$X_1\delta x_1 + X_2\delta x_2 + \ldots + X_n\delta x_n \leqq 0, \tag{4.2}$$

ahol $\delta x_1, \ldots, \delta x_n$ az $x_1, \ldots, x_n$ koordináták variációi, más szóval olyan kis megváltozásai, hogy az

$$x_1^* + \delta x_1, \ldots, x_n^* + \delta x_n \tag{4.3}$$

komponensekből alkotott vektor eleget tesz a (4.1) feltételeknek. A továbbiakban a tárgyalást kissé elnagyoljuk, ám nem nehéz megadni azokat a nem nagyon megszorító feltételeket, amelyek mellett állításaink szabatosakká tehetők.

Nevezzük az $\mathbf{x}^*$ pontban inaktívnak azokat a (4.1)-beli feltételeket, amelyek $\mathbf{x}^*$ esetében határozott egyenlőtlenséggel teljesülnek. Ezek a feltételek az $\mathbf{x}^*$ ponthoz képest bekövetkező kis megváltozásokat nem akadályozzák meg. Ezért a kis megváltozásokra vonatkozó feltételeket az $\mathbf{x}^*$ pontban aktív (vagyis egyenlőséggel teljesülő) feltételekre támaszkodva kell megfogalmaznunk. Tegyük fel, hogy (4.1)-ben az első M feltétel aktív, a többi nem és állapodjunk meg abban, hogy a dx_i szimbólumot alkalmazzuk δx_i helyett. Ekkor a *Fourier-elv* azt kívánja meg, hogy teljesüljön az

$$X_1 dx_1 + \ldots + X_n dx_n \leqq 0 \tag{4.4}$$

egyenlőtlenség, ahol $dx_1, \ldots, dx_n$ eleget tesznek az alábbi egyenlőtlenségeknek:

$$\frac{\partial g_i(\mathbf{x}^*)}{\partial x_1} dx_1 + \ldots + \frac{\partial g_i(\mathbf{x}^*)}{\partial x_n} dx_n \geqq 0, \quad i = 1, \ldots, M. \tag{4.5}$$

Ezután már elfeledkezhetünk a $dx_1, \ldots, dx_n$ mennyiségek nagyságrendjéről. Eddig olyan kicsinynek kellett lenniük, hogy az $\mathbf{x}^*$ pontban inaktív feltételek érvényben maradjanak. Most már azonban csak a (4.5) egyenlőtlenségeket fogjuk előírni, melyek, ha fennállnak adott $dx_1, \ldots, dx_n$ esetén, akkor fennállnak $tdx_1 \ldots, tdx_n$ esetén is, ahol t tetszőleges nemnegatív szám.

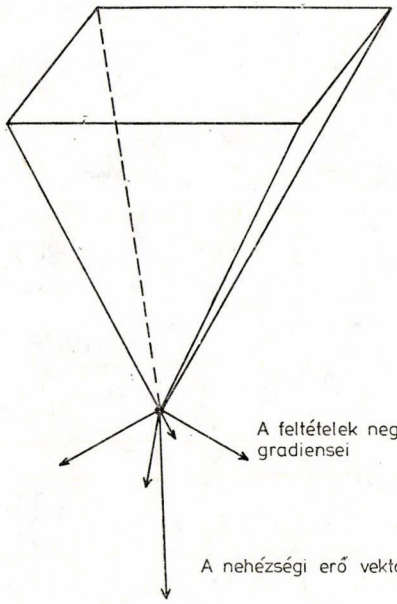
Jelölje $\mathbf{X}$ az $X_1, \ldots, X_n$ komponensekből alkotott vektort. Ha ezt a vektort komponenseivel adjuk meg, akkor (kivételesen) az utóbbiakat egy sorban helyezzük el. Abban is megállapodunk — a legtöbb tankönyvben alkalmazott konvenciót követve —, hogy a gradiens sor-vektor lesz.

Ha $\mathbf{x}^*$ egyensúlyi pont, akkor a *Fourier-elv* alkalmazása révén azt találjuk, hogy a (4.4) lineáris egyenlőtlenség következménye a (4.5) lineáris egyenlőtlenségeknek. Következésképpen, *Farkas tétele* szerint léteznek olyan $\lambda_1, \ldots, \lambda_M$ nemnegatív számok, hogy fennáll az

$$\mathbf{X} + \lambda_1 \nabla g_1(\mathbf{x}^*) + \ldots + \lambda_M \nabla g_M(\mathbf{x}^*) = \mathbf{0}'. \tag{4.6}$$

A (4.6) egyenlet érvényességét egy speciális esetben az 1. ábra szemlélteti. Ha az erők konzervatív rendszert alkotnak, más szóval, létezik olyan $V(\mathbf{x})$ függvény, hogy

$$X_i = -\frac{\partial V}{\partial x_i}, \quad i = 1, \ldots, n, \tag{4.7}$$



1. ábra. *A kúp aljában levő tömegpont egyensúlyi helyzetben van. Az ábrán látható, hogy a nehézségi erő vektora benne van a feltételek negatív gradiensei által generált kúpban*

akkor a (4.4) egyenlőtlenség az alábbi alakot ölti

$$(4.8) \qquad \frac{\partial V}{dx_1} dx_1 + \ldots + \frac{\partial V}{\partial x_n} dx_n \geqq 0.$$

A bal oldalon a V függvény teljes differenciálja áll. Ha létezik potenciál, akkor a *Fourier-elv* helyett kiindulhatunk a speciálisabb *Courtivron-elvből,* amely szerint egyensúly esetén a potenciálnak lokális minimuma van. Innen pedig a (4.8) egyenlőségre könnyen következtethetünk (patologikus esetektől eltekintve).

Feltesszük a kérdést: ki volt az, vagy kik voltak azok, akik először következtettek a (4.6) egyenletre az egyenlőtlenségekre vonatkozó egyensúlyi elv, tehát a *Fourier-elv* alapján? E kérdés megválaszolásához célszerű áttanulmányozni azokat a dolgozatokat, amelyekre FARKAS GYULA a *Crelle Journalban* 1901-ben közölt cikkében hivatkozik, továbbá célszerű a század elején publikált *Enzyklopädie der Mathematischen Wissenschaften* nagy alapossággal megírt mű analitikus mechanikával foglalkozó cikkeit is tanulmányozni. A mechanikáról írott négy kötetben két olyan cikk található, melyekben a szerzők egyenlőtlenséges feltételek melletti statikai, illetve dinamikai problémákat említenek. E két cikk szerzője VOSS [77] és STÄCKEL [72]. FARKAS és e két szerző körülbelül harminc könyvet és dolgozatot említ. Ezek közül néhányat nem sikerült megszereznünk, ám a leglényegesebb referenciákat FARKAS, VOSS és STÄCKEL minden bizonnyal külön kiemeli és ilyenformán a kérdéskör történetéről valószínűleg jó képet kapunk. VOSS a következőket írja [77, 74. old.]:

„Auf diesen von LAGRANGE *nicht berücksichtigten Fall hat unabhängig von* FOURIER *erst wieder* GAUSS *dann* OSTROGRADSKY *hingewiesen ... daher heisst in Russland das Princip von Fourier auch wohl das von Ostrogradsky. In Frakreich ist Fourier's Princip nicht so unbeachtet geblieben;* A. A. COURNOT *entwickelt schon 1827 die Gleichungen von Ostrogradsky;"*[11]

OSZTROGRADSZKIJ [68] dolgozatát FARKAS is említi. STÄCKEL megemlíti még [69] munkáját is.

VOSS nem említi FARKAST, ám STÄCKEL hivatkozik a [14, 16] dolgozatokra. STÄCKELnek megbocsátható, hogy nem ismeri fel FARKAS dolgozatainak jelentőségét, hiszen cikkének címe: *„Elementare Dynamik",* márpedig a cikk keletkezése előtt

[11] „Erre a LAGRANGE által nem vizsgált esetre először, FOURIERtől függetlenül, GAUSS utalt, később pedig OSZTROGRADSZKIJ ... ezért *Oroszországban* a *Fourier-elvet Osztrogradszkij-elvnek is* nevezik. Franciaországban a *Fourier-elv* nem kerülte el a figyelmet; A. A. COURNOT már 1827-ben megfogalmazta az *Osztrogradszkij-egyenleteket.*"

FARKAS főként csak statikával foglalkozott. STÄCKEL igen nagyra értékelte MAYER munkáit:

„Statt holonomer oder nicht holonomer Bedingungsgleichungen können auch Bedingungsungleichheiten auftreten; auch können Bedingungen plötzlich fortfallen oder plötzlich durch andere ersetzt werden. Nachdem bereits früher solche Fälle behandelt worden waren, hat, durch E. STUDY *angeregt,* A. MAYER, *vom Gausschen Principe des kleinsten Zwanges ausgehend, dargelegt, wie man hier die Bewegung des Punktes in allen Fällen bestimmen kann".*[12]

MAYER említett kiváló dolgozatai [59, 60] dinamikai problémákkal foglalkoznak, tehát témánkhoz csak kevéssé kapcsolódnak. Mégis, nagyon tanulságos e dolgozatok olvasása, melyekben bőségesen találunk jó ötleteket, példákat és a *Lagrange-szorzók* nemnegativitására vonatkozó bizonyításokat speciális esetekben. Érdekes bizonyítást adott ZERMELO sokkal általánosabb esetre vonatkozólag. Saját tételének birtokában FARKAS a dinamikai problémákra vonatkozólag is levonta az előbbieknél általánosabb következtetéseit [24] dolgozatában. Az egyenlőséges feltételek mellett dinamikai problémákat tárgyaló dolgozatok körében megemlíthetjük még GIBBS [39] dolgozatát.

FOURIER, COURNOT és FARKAS tűnnek a statikai egyensúlyra vonatkozó szükséges feltételek kialakítóinak. COURNOT (és később OSZTROGRADSZKIJ) felírta a (4.6) egyenletet sejtés formájában, ahogyan ma mondanánk. FARKAS bebizonyította a (4.6) egyenlet érvényességét, miközben a bizonyítás első felét illetően FOURIER munkájára, bizonyítására hagyatkozott. Pontosabban, elfogadta, hogy — konzervatív rendszerekre szorítkozva — egyensúly esetén a (4.8) egyenlőtlenség fennáll. Ma már tudjuk, hogy a (4.8) egyenlőtlenség nem következik abból, hogy a V függvénynek lokális minimuma van a (4.1) feltételek mellett, még akkor sem, ha a g_i, $i=1, \ldots, M$ és a V függvények tetszőleges rendű parciális deriváltjai léteznek. Szükség van még egy regularitási feltételre *(constraint qualification)*. Ám a múlt század mechanikai szerzői számára az analízisnek ez a mélysége szükségtelennek tűnt (a mechanikával foglalkozók közül sokak számára bizonyára még ma is az). Ebben a tekintetben tehát a (4.6) egyenlet FARKAS által adott levezetése elnagyolt.

Meg kell még említenünk, hogy COURNOT nagy mértékben támaszkodott POINSOT [70] munkájára.

Mind FOURIER, mind pedig FARKAS igen fontos eredményekkel gazdagították az optimalizálás elméletét. Röviden összefoglaljuk, melyek az idevágó legfontosabb eredményeik.

FOURIER — felvetette a lineáris programozási problémát [36/II., 325—380. old.], [40], (1824);
— megfogalmazta a mechanikai egyensúly egyenlőtlenséges elvét [34], (1798);
— kezdeményezte a homogén lineáris egyenlőtlenségek parametrikus megoldását [35], (1826).

[12] „Homogén és inhomogén feltételi egyenletek helyett egyenlőtlenséges feltételek is előfordulhatnak; egyes feltételek átmenetileg el is tűnhetnek, vagy helyt adhatnak más feltételeknek. Miután ilyen esetekkel korábban már foglalkoztak, a *Gauss-féle legkisebb kényszer elvből* kiindulva és E. STUDY által felbátorítva, A. MAYER megmutatta, hogyan kell az összes esetekben a pont mozgását meghatározni."

FARKAS — bebizonyította a homogén lineáris egyenlőtlenségek alaptételét, először [17]-ben, (1898);
— bizonyítást adott a *Fourier-féle mechanikai elv* duális alakjára [13, 14], (1894);
— elegáns parametrikus előállítást adott a homogén lineáris egyenlőtlenségek megoldásai számára; először a [19] dolgozatban, (1898).

5. Cournot munkája a mechanikai egyensúllyal kapcsolatban

ANTOINE — AUGUSTIN COURNOT, híres francia matematikus, közgazdász és filozófus, 1801-ben született és 1877-ben halt meg. Középiskolai tanulmányai elvégzése után *Párizsba* ment és az *École Normale Supérieur* diákja lett. Az *Écolt* egy év múlva hirtelen feloszlatták és ő titkára lett GOUVION SAINT CYR tábornoknak, aki egykor NAPOLEON szolgálatában állt. Közben folytatta egyetemi tanulmányait. Az 1823-tól 1833-ig terjedő tíz évben legfőbb tevékenysége katonai memoárok összeállítása volt, megírta — többek között — alkalmazójának száz oldalas életrajzát. Ám talált időt arra is, hogy közben doktoráljon. A *docteur ès sciences* fokozatot 1829-ben szerezte meg, eredményeit az [5, 6] dolgozatokban publikálta. Ebben az időben, 1827-ben jelent meg az a munkája [4], amelyből idézni fogunk. Disszertációjában a [4] dolgozat eredményeit dinamikai problémákra alkalmazza. Az 1826-ban megjelent [3] dolgozata egyenlőtlenségekről szól, ám a *„Farkas-tétel"* ebben semmilyen formában nem szerepel. A FÉRUSSAC folyóiratában megjelent [3, 4] dolgozatokat csak monogramjával látta el (*A.C.*), emiatt ezeket néhány szerző az azonos monogramú CAUCHYnak tulajdonította. Későbbi pályafutása során sok segítséget kapott mentorától és barátjától, POISSONtól. Először a matematika professzora lett *Lyons*-ban 1834-ben. Egy év múlva a *Grenoble* környéki iskolák főigazgatója, továbbá tanfelügyelője lett, ez utóbbi pozícióban AMPÈRE utódaként, aki akkor halt meg. Később, 1838-ban utazó tanfelügyelő lett párizsi székhellyel. Ezt követően, 1854-ben ismét főigazgató lett *Dijonban*, 1862-ben nyugalomba vonult és *Párizsba* költözött. Leghíresebb könyvét, melynek címe: *Recherche dans les Principes Mathématiques de la Théorie des Richesses*, még *Grenoble*-ban írta. E munkája révén COURNOT a matematikai közgazdaságtan korszakalkotó klasszisa lett. Ezt azonban csak a huszadik század elején fedezték fel, közgazdasági műveit akkoriban érdektelenség vette körül. COURNOT maga írja később egy halála évében megjelent műve előszavában, hogy ő az egyetlen francia közgazdász, akit a többiek nem idéztek. A *Théorie des Richesses* ebben a században több kiadásban és több nyelven megjelent. A [7] kiadásban a műhöz WALRAS, PARETO és BERTRAND írt megjegyzéseket, továbbá LUTFALLA közölt COURNOT életrajzot, publikációs jegyzékkel. A [8] angol nyelvű kiadáshoz IRVING FISCHER írt megjegyzéseket. COURNOT szinte egész életében szembajban szenvedett, már e könyve írásakor is igen gyengén látott. A könyv eredetijében emiatt sok elírás van. Adminisztratív munkáját el tudta látni, de senkit nem tudott találni, aki matematikai szöveget felolvasott volna neki. 1873-ban azt írta WALRASnak, hogy harminc évvel ezelőtt a matematikát fel kellett adnia. Élete végén felcsillant az akadémiai tagság lehetősége, ám mielőtt ezt elnyerhette volna, 1877-ben meghalt.

COURNOT — számunkra most legfontosabb — [4] dolgozata nem eléggé ismert, a [7] könyvben közölt publikációs jegyzékben sem szerepel. A dolgozatban FOURIER munkájára nem hivatkozott. Valószínűleg nem tudta, hogy FOURIER 1798-ban már

megfogalmazta a mechanikai egyensúlyra vonatkozó egyenlőtlenséges elvet. COURNOT ezt újból felfedezte, ám megadta az egyensúly szükséges feltételét is. Megjegyezzük, hogy a mechanikában nyert eredmények közgazdasági alkalmazásáról nincs szó *Théorie des Richesses* című könyvében. Rövid dolgozatából az alábbiakat idézzük:

„...il arrive souvent que les liaisons du système ne peuvent être exprimés par des équations; il résulte de ces liaisons des conditions d'équilibre que l'on a toujours regardées comme ne pouvant se déduire du principe des vitesses virtuelles et étant en dehors de ce principe (Mécanique de M. POISSON, *tom. I, p. 241). Notre but est ici de faire voir que, lorsque les liaisons du système peuvent s'exprimer algébriquement par le moyen des signes d'inégalité, le principe des vitesses virtuelles, convenablement modifié, s'y applique encore, et fournit, par une méthode uniforme, les conditions de l'équilibre."*[13]

Ugyanennek a dolgozatnak a 167. és a 168. oldalain a következőket olvassuk:

„Admettons qu'un système soit soumis à un certain nombre de liaisons semblables, exprimées par les inégalités

$$I > 0, \quad I' > 0, \quad \text{etc.} \quad J < 0, \quad J' < 0, \quad \text{etc.}$$

les signes > et < étant toujours supposés ne pas exclure le cas d'égalité.

On ne peut pas, en général, différentier, une inégalité comme une équation, ni en déduire une relation entre les incrémens des variables; mais lorsqu'il s'agit de la recherche des conditions d'équilibre, deux cas seulement peuvent se présenter. Ou le situation du système est telle que l'on pourrait supprimer les liaisons, sans changer son état; comme il arrive, si les fils qui joignent certains de ses points ne sont pas tendus, ou si les points auxquels certaines surfaces opposent des obstacle impénétrables, ne sont point contigus à ces surfaces: dans ce cas l'équilibre doit avoir lieu, indépendamment des conditions de liaison, et il est superflu d'en tenir compte; ou bien le système est placé de manière à ce que les liaisons produisent leur effet et puissent diminuer le nombre des conditions nécessaires pour l'équilibre: alors on a pour les valeurs actuelles des coordonnées $I=0$, $\mathfrak{I}=0$, etc., et si l'on fait varier ces valeurs, leurs incrémens ne seront pas entièrement arbitraires; ils devront, en vertu des conditions primitives, satisfaire aux relations:

(a) $$\delta I > 0, \quad \delta I' > 0, \quad \text{etc.} \quad \delta J < 0, \quad \delta J' < 0, \quad \text{etc.}$$

D'ailleurs, quand le système est soumis à des liaisons de l'espèce de celles que nous considérons ici, il est clair qu'on pourrait les supprimer, pourvu qu'on lui appliquât certaines forces, capables d'en tenir lieu; ainsi la résistance d'une surface peut être remplacée par l'application d'une force normale à cette surface et dirigée dans le sens convenable, celle d'un fil tendu, par l'application d'une force dirigée dans le sens de la tension du fil, etc. Soient, $F, F', \ldots$ les forces directement appliquées au système, suivant les directions $f, f', \ldots$ et $P, P', \ldots$ les forces auxiliaires qui tiendraient

[13] „...gyakran megtörténik, hogy a rendszerre vonatkozó kényszerek nem fejezhetők ki egyenlőségekkel; e kényszerekből olyan egyensúlyi feltételek következnek, melyeket mindig a virtuális sebességek elvéből le nem vezethető, sőt, eme elven kívülieknek tekintettek (*Poisson Mechanikája*, I. kötet, 241. old.). A mi célunk itt az, hogy megmutassuk, hogy ha a rendszer kényszerei algebrailag egyenlőtlenségi jelek segítségével fejezhetők ki, a virtuális sebességek elve itt is érvényes, a megfelelő módosítással, és egységes módszerrel szolgáltatja az egyensúly feltételeit."

lieu de l'existence des obstacles, et seraient dirigées suivant les lignes $p, p' \ldots$; *on aura, en vertu du principe des vitesses virtuelles, l'équation fondamentale*

$$F\delta f + F'\delta f' + \text{ etc. } + P\delta p + P'\delta p' + \text{ etc. } = 0.$$

Or il est aisé de voir que tous les mouvemens virtuels, pour lesquels P *et* δp, P' *et* $\delta p'$, *etc., seraient de signes contraires, tendraient à surmonter les obstacles que les surfaces, les fils, etc., opposent aux differens points du système, et par conséquent sont incompatibles avec les liaisons du système, exprimées par les inégalités (a). Il n'y a de compatibles avec ces liaisons que les mouvemens pour lesquels* P *et* δp, p' *et* $\delta p'$ *sont de même signe, et pour lesquels par conséquent les quantités* $P\delta p$, $P'\delta p'$, *etc. sont essentiellement positives.*

Donc, si l'on fait abstraction des forces auxiliaires P, P', *etc., et des termes qu'elles introduisent dans l'équation des vitesses virtuelles, on aura, par la nature des liaisons auxquelles le système est soumis, l'inégalité*

$$F\delta f + F'\delta f' + \text{ etc. } < 0, \tag{b}$$

laquelle devra subsister, ainsi que les inégalités (a), pour tous les mouvemens virtuels, compatibles avec l'état du système, et de là se déduisent les conditions de l'équilibre, ..."[14]

[14] „Engedjük meg, hogy egy rendszer legyen alávetve bizonyos számú hasonló kényszernek, melyeket az

$$I > 0, \quad I' > 0, \quad \text{stb.} \quad J < 0, \quad J' > 0, \text{ stb.}$$

egyenlőtlenségek fejeznek ki, ahol a $>$ és a $<$ jelek feltételezés szerint nem zárják ki az egyenlőségeket.

Egy egyenlőtlenséget általában nem lehet úgy differenciálni, mint egy egyenlőséget, úgyszintén nem lehet belőle a változók növekményére vonatkozó relációt levezetni; de ha az egyensúly feltételeinek a kutatásáról van szó, csak két eset fordulhat elő. Vagy olyan a rendszer helyzete, hogy a kényszereket el lehetne távolítani, anélkül, hogy állapotát megváltoztatnánk; amint az előfordul, ha azok bizonyos pontjait összekötő fonalak nem feszesek, vagy ha azok a pontok, amelyek számára bizonyos felületek áthatolhatatlan akadályul szolgálnak, egyáltalában nem érintkeznek ezekkel a felületekkel: ebben az esetben az egyensúlynak be kell következnie, a kényszerfeltételektől függetlenül, és szükségtelen azokat számításba venni; vagy pedig a rendszer helyzete olyan, hogy a kényszerek kifejthetik hatásukat és csökkenteni tudják az egyensúly szükséges feltételeinek a számát: ekkor a koordináták aktuális értékeire $I=0$, $J=0$ stb.; és ha ezeket az értékeket variáljuk, megváltozásaik nem lesznek egészen tetszőlegesek; az eredeti feltételek alapján eleget kell, hogy tegyenek a

$$\delta I > 0, \quad \delta I' > 0, \quad \text{stb.} \quad \delta J < 0, \quad \delta J' < 0, \quad \text{stb.} \tag{a}$$

relációknak.

Egyébként, ha a rendszer alá van vetve olyan fajta kényszereknek, amilyeneket mi most itt tekintetbe veszünk, világos, hogy ezeket eltávolíthanánk, feltéve, hogy alkalmaznánk olyan erőket, melyek ugyanazt a hatást fejtik ki; ily módon egy felületnek az ellenállása helyettesíthető egy olyan erő alkalmazásával, mely a felületre merőleges és megfelelő irányítású, egy kifeszített fonalé, a fonal feszültségének az irányába mutató erő alkalmazásával, stb. Legyenek $F, F', \ldots$, a rendszerre közvetlenül ható erők, melyek iránya $f, f', \ldots$ és $P, P', \ldots$ segéderők, melyek az előforduló ellenállásokat helyettesítik és a $p, p' \ldots$ egyeneseknek megfelelő irányúak; a virtuális sebességek elve alapján a következő alapegyenletet kapjuk:

$$F\delta f + F'\delta f' + \text{ stb. } + P\delta p + P'\delta p' + \text{ stb. } = 0.$$

Mármost könnyen látható, hogy az összes virtuális elmozdulások, melyekre P és δp, P' és $\delta p'$ stb., ellenkező előjelűek volnának, azoknak az ellenállásoknak a legyőzésére törekednének, amelyeket a felületek, a fonalak, stb. jelentenek a rendszer különböző pontjai számára, és következésképpen összeférhetetlenek a rendszernek az (a) egyenlőtlenségek által kifejezett kényszereivel.

COURNOT nem írta fel a (4.6) egyenletet teljes általánosságában, hanem azt csak példákon keresztül illusztrálta. Ezek taglalására nem térünk ki.

A fenti egyenletre COURNOT által adott bizonyítás nem egyéb, mint a szemléletünkre való rövid utalás. Amikor az xy-síkkal párhuzamos síkokon elhelyezkedő tömegpontokból álló mechanikai rendszerrel foglalkozik, POINSOT [70] könyvére hivatkozik, amely először 1803-ban jelent meg. Erre a speciális esetre vonatkozólag ugyanis már POINSOT is felfedezte az egyensúly feltételét. Tegyük fel, egyszerűség kedvéért, hogy egy testet egy, az xy-síkkal párhuzamos síkon az $A, B, C, D, \ldots$ pontjai tartják. A [70] könyv 125. oldalán ezzel kapcsolatban a következőket olvassuk:

„...toutes les forces appliquées au système doivent se réduire à une seule, perpendiculaire au plan fixe, et dont la valeur ne soit pas positive.

En second lieu, je dis que sa direction doit rencontrer ce plan dans l'intérieur du polygone formé par les points d'appui $A, B, C, D, \ldots$"[15]

6. Osztrogradszkij munkája a mechanikai egyensúllyal kapcsolatban

MIHAIL VASZILJEVICS OSZTROGRADSZKIJ (1801—1862) fontos eredményekkel gazdagította (többek között) a mechanikát. *Párizsban* tanult, hallgatta FOURIER, POISSON, CAUCHY és más híres francia matematikus előadását. *Oroszországba, Szent-Pétervárra* 1828-ban tért vissza. A *Szent-Pétervári Akadémia* 1830-ban levelező, majd 1832-ben rendes tagjává választotta.

A mechanikai egyensúly problémájával kapcsolatban két dolgozatát kell megemlítenünk. FARKAS csak a [68] dolgozatra hivatkozik (melyet 1834-ben mutatott be a *Szent-Pétervári Akadémián*), ám a [69] dolgozat elméletének javított változatát tartalmazza. A korábbi dolgozatban négy rendszerre vonatkozólag fejti ki az alkalmazás lehetőségét. E négy rendszer a következő: a.) egy pont, mely valamely felület által határolt térben mozoghat csak; b) a libegő sokszög; c) a rugalmas huzal; d) az összenyomhatatlan folyadék. További, dinamikai jellegű problémákat is említ.

A [69] dolgozatban a (mai szóhasználattal élve) *Farkas-tételt* nyilvánvaló algebrai tényként említi és erre támaszkodva levezeti az egyensúlyra vonatkozó (4.6) egyenletet. A [69] dolgozat 589. és 590. oldaláról idézünk:

Csak azok a mozgások összeférhetők ezekkel a kényszerekkel, amelyekre P és δp, P' és $\delta p'$ megegyező előjelűek és amelyekre következésképpen a $P\delta p, P'\delta p'$ stb. mennyiségek lényegében pozitívak.

Ha tehát elvonatkoztatunk a P, P' stb., erőktől és azoktól a tagoktól, amelyeket ezek bevezetnek a virtuális sebességek elvének egyenletébe, azoknak a kényszereknek a természete révén, amelyeknek a rendszer alá van vetve, a következő egyenlőtlenséget kapjuk:

$$F\delta f + F'\delta f' + \text{ stb. } < 0, \tag{b}$$

melynek fenn kell állnia, úgyszintén az (a) egyenlőtlenségeknek, az összes olyan virtuális elmozdulásokra, amelyek a rendszer állapotával összeférhetők és innen adódnak az egyensúly feltételei...".

[15] „...a rendszerre ható összes erők egyetlen olyan erőre kell, hogy redukálódjanak, mely a rögzített síkra merőleges, és melynek értéke nempozitív.

Második helyen, azt mondom, hogy ennek az iránya az $A, B, C, D, \ldots$ tartó pontok által meghatározott poligonon belül kell, hogy találkozzék a síkkal."

„Supposons que les quantités $\delta s, \delta s', \delta s'', \delta s''', \ldots$ appartiennent non seulement à ceux des déplacemens du système, dont les forces perdues sont capables, mais encore, à tous les autres déplacemens tant possibles que non, ou plutôt considérons $\delta s, \delta s', \delta s'', \delta s''', \ldots$ comme tout-à-fait arbitraires. Nous devons exprimer que les forces perdues $R, R', R'', R''', \ldots$ sont incapables de produire aucun déplacemens des systèmes satisfaisant aux conditions

$$(15) \qquad \begin{cases} \delta L > 0 \\ \delta L_1 > 0 \\ \delta L_2 > 0 \\ \delta L_3 > 0 \\ \ldots\ldots\ldots \end{cases}$$

le signe $>$ n'exclut point celui de l'égalité.

Or on sait qu'un système des forces est capable de tout déplacement qui fournit, pour son moment total, une valeur positive et aucun de ceux qui correspondent aux valeurs négatives ou zéro du moment total. Ainsi, pour que les forces perdues soient incapables de produire aucun des déplacemens satisfaisans aux conditions (15), *il faut que leur moment soit négatif ou zéro pour ces déplacemens, c'est-à-dire il faut que la fonction*

$$R\delta s \cos\psi + R'\delta s' \cos\psi' + R''\delta s'' \cos\psi'' + R'''\delta s''' \cos\psi''' + \ldots$$

dans laquelle $\psi, \psi', \psi'', \psi''', \ldots$ désignent respectivement les angles $\widehat{R\delta s}, \widehat{R'\delta s'}$ $\widehat{R''\delta s''}, \widehat{R'''\delta s'''}, \ldots$ et qui par conséquent représente le moment des forees $R, R', R'', R''', \ldots$ soit negative ou zéro toutes les fois que $\delta s, \delta s', \delta s'', \delta s''', \ldots$ remplissent les conditions (15).

La solution de la question qui consiste à rendre la fonction

$$R\delta s \cos\psi + R'\delta s' \cos\psi' + R''\delta s'' \cos\psi'' + R'''\delta s''' \cos\psi''' + \ldots$$

négative ou zéro toutes les fois que les fonctions de même nature, $\delta L, \delta L_1, \delta L_2, \delta L_3, \ldots$ sont positives ou zéro appartient à l'algèbre la plus élémentaire. Il est nécessaire, et il suffit que $R\delta s \cos\psi + R'\delta s' \cos\psi' + R''\delta s'' \cos\psi'' + R'''\delta s''' \cos\psi''' + \ldots$ puisse se réduire à une fonction linéaire de $\delta L, \delta L_1, \delta L_2, \delta L_3, \ldots$ avec des coëfficiens négatifs. Ainsi il n'y a qu'à faire quels que soient $\delta s, \delta' s, \delta'' s, \delta''' s, \ldots,$

$$R\delta s \cos\psi + R'\delta s' \cos\psi' + R''\delta s'' \cos\psi'' + R'''\delta s''' \cos\psi''' + \ldots$$
$$= \lambda\delta L + \lambda_1\delta L_1 + \lambda_2\delta L_2 + \lambda_3\delta L_3 + \ldots$$

et à y ajouter la condition que les λ sont tous négatifs. Ou bien, si l'on veut éviter de considérer les λ comme négatifs, on peut faire

$$R\delta s \cos\psi + R'\delta s' \cos\psi' + R''\delta s'' \cos\psi'' + R'''\delta s''' \cos\psi''' + \ldots$$
$$= -(\lambda\delta L + \lambda_1\delta L_1 + \lambda_2\delta L_2 + \lambda_3\delta L_3 + \ldots)$$

alors tous les λ seront positifs. Il est évident, par la dernière équation, comme par celle qui la précède, que le moment $R\delta s \cos\psi + R'\delta s' \cos\psi' + R''\delta s'' \cos\psi'' +$

$+R'''\delta s'''\cos\psi'''+\ldots$ *sera négatif ou zéró toutes les fois que les fonctions* $\delta L, \delta L_1, \delta L_2, \delta L_3, \ldots$ *seront positives ou zéro.*

En transportant tous les termes d'un même côté, l'équation de l'équilibre des forces perdues deviendra

$$R\delta s\cos\psi+R'\delta s'\cos\psi'+R''\delta s''\cos\psi''+R'''\delta s'''\cos\psi'''+\ldots$$
$$+\lambda\delta L+\lambda_1\delta L_1+\lambda_2\delta L_2+\lambda_3\delta L_3\ldots=0 \tag{16}$$

elle doit avoir lieu quelles que soient $\delta s, \delta s', \delta s'', \delta s''', \ldots$ *tant en grandeur que pour la direction. Mais il ne faut pas oublier d'ajouter à l'équation* (16) *les inégalités*

$$\begin{cases}\lambda > 0\\ \lambda_1 > 0\\ \lambda_2 > 0\\ \lambda_3 > 0\\ \ldots\end{cases} \tag{17}$$

[16] „Tegyük fel, hogy a $\delta s, \delta s', \delta s'', \delta s''', \ldots$ mennyiségek a rendszernek nemcsak azokhoz az elmozdulásaihoz tartoznak, amelyeknek megfelelő elveszett erői hatni képesek, hanem az összes elmozdulásokhoz, akár lehetségesek, akár nem, vagy tekintsük inkább a $\delta s, \delta s', \delta s'', \delta s''', \ldots$ mennyiségeket egészen tetszőlegeseknek. Ki kell fejeznünk azt, hogy az $R, R', R'', R''', \ldots$ elveszett erők nem képesek semmilyen elmozdulást produkálni az alábbi feltételeknek eleget tevő rendszerek számára:

$$\begin{cases}\delta L > 0\\ \delta L_1 > 0\\ \delta L_2 > 0\\ \delta L_3 > 0\\ \ldots\ldots\end{cases} \tag{15}$$

ahol a $>$ jel nem zárja ki az egyenlőséget.

Márpedig tudjuk, hogy az erők egy rendszere képes minden olyan elmozdulást létrehozni, mely a teljes momentum számára pozitív értéket eredményez és semmi olyat, mely a teljes momentum negatív vagy zéró értékeihez tartozik. Ennélfogva, ahhoz, hogy az elveszett erők ne produkálhassanak semmilyen, a (15) feltételeknek eleget tevő elmozdulást, szükséges, hogy ezeknek az elmozdulásoknak az esetében momentumuk negatív, vagy zéró legyen, illetve szükséges, hogy az

$$R\delta s\cos\psi+R'\delta s'\cos\psi'+R''\delta s''\cos\psi''+R'''\delta s'''\cos\psi'''+\ldots$$

függvény, amelyben $\psi, \psi', \psi'', \psi''', \ldots$ az $\widehat{R\delta s}, \widehat{R'\delta s'}, \widehat{R''\delta s''}, \widehat{R'''\delta s'''}$ szögeket jelentik és amely következésképpen az $R, R', R'', R''', \ldots$ erők momentumát képviseli, minden olyan esetben, amikor $\delta s, \delta s', \delta s'', \delta s''', \ldots$ teljesítik a (15) feltételeket, negatív vagy zéró legyen.

Annak a kérdésnek a megoldása, hogy az

$$R\delta s\cos\psi+R'\delta s'\cos\psi'+R''\delta s''\cos\psi''+R'''\delta s'''\cos\psi'''$$

függvényt negatívvá, vagy zéróvá tegyük, valahányszor az azonos természetű $\delta L, \delta L_1, \delta L_2, \delta L_3, \ldots$ függvények pozitívak, vagy zérók, a legelemibb algebrához tartozik. Szükséges és elegendő, hogy $R\delta s\cos\psi+R'\delta s'\cos\psi'+R''\delta s''\cos\psi''+R'''\delta s'''\cos\psi'''$ a $\delta L, \delta L_1, \delta L_2, \delta L_3, \ldots$ lineáris függvénye legyen negatív együtthatókkal. Ennélfogva, akármilyenek legyenek is $\delta s, \delta s', \delta s'', \delta s''', \ldots$, csak azt kell felírnunk, hogy

$$R\delta s\cos\psi+R'\delta s'\cos\psi'+R''\delta s''\cos\psi''+R'''\delta s'''\cos\psi'''=\lambda\delta L+\lambda_1\delta L_1+\lambda_2\delta L_2+\lambda_3\delta L_3+\ldots$$

OSZTROGRADSZKIJ két hibát követett el. Az egyik abban áll, hogy a *Farkas-tétel* állítását egyszerűen deklarálta, nem bizonyította. (OSZTROGRADSZKIJ tekintélye oly nagy volt, hogy több szerző később sem firtatta a bizonyítás szükségességét [59, 60, 65, 77].) A második hibát MAYER közlése szerint [59, 225. old.] STUDY fedezte fel. A [69] dolgozat 583. oldalán OSZTROGRADSZKIJ arra következtet, hogy bizonyos feltételi egyenlőtlenségek az egyensúly esetén egyenlőséggel teljesülnek, vagyis a bennük szereplő függvények zéróval egyenlők. Ez lehetővé tette volna az egyensúly egyenleteiben szereplő szorzók meghatározását. Azt gondolta, hogy ha a feltétel nélküli egyensúly keresés feladatában bizonyos feltételek egyensúly esetén negatívvá válnak, akkor a feltételes egyensúlykeresés esetében ezek szükségképpen zéróval egyenlők. HAMEL adott erre ellenpéldát [47, 41—42. oldal].

7. Farkas munkája a mechanikai egyensúllyal kapcsolatban

A [13, 14] dolgozatok lényegében azonosak, itt most a [13] magyar változatból idézünk. A dolgozat bevezető részében a következőket olvassuk:

„Ebben áll a Fourier-féle elv, a melyet egyenlőtlenségi elvnek is fogok nevezni, a közönségeset pedig egyenlőségnek. Emezt, mint speciálisat magában foglalja, úgy hogy mikor a kényszert csupa egyenlőségek fejezik ki, emerre redukálódik. Ugyanis egyenlőségi kényszerben a virtuális elmozdulások minden érték-rendszerének az ellentétese is érvényes lévén, ekkor a virtuális momentum csupán úgy nem lehet pozitív, ha $=0$.

Az egyenlőtlenségi elv éppúgy, mint az egyenlőségi, független a különböző pontszerkezetek sajátosságaitól s éppen azért tekintendő GALILEI *az egyenlőségi elv feltalálójának, mert előtte csak egy bizonyos szerkezetre, s e szerkezetnek mintájába könnyen beilleszthetőkre nézve fogamzott meg, nevezetesen* ARISTOTELES*nél, s miután*

és hozzátennünk azt a feltételt, hogy a λ értékek mind negatívak legyenek. Vagy, ha el akarjuk kerülni, hogy a λ értékek negatívak legyenek, felírhatjuk azt, hogy

$$R\delta s\cos\psi + R'\delta s'\cos\psi' + R''\delta s''\cos\psi'' + R'''\delta s'''\cos\psi''' + \ldots = -(\lambda\delta L + \lambda_1\delta L_1 + \lambda_2\delta L_2 + \lambda_3\delta L_3 + \ldots),$$

akkor mindegyik λ pozitív lesz. Az utolsó és úgyszintén az azt megelőző egyenlet alapján evidens, hogy az $R\delta s\cos\psi + R'\delta s'\cos\psi' + R''\delta s''\cos\psi'' + R'''\delta s'''\cos\psi''' + \ldots$ momentum negatív, vagy zéró lesz, valahányszor a $\delta L, \delta L_1, \delta L_2, \delta L_3, \ldots$ függvények pozitívak, vagy zérók lesznek.

Mindegyik tagot ugyanarra az oldalra hozva, az elveszett erők egyensúlyának egyenlete a következő lesz:

$$R\delta s\cos\psi + R'\delta s'\cos\psi' + R''\delta s''\cos\psi'' + R'''\delta s'''\cos\psi''' + \ldots$$

$$+\lambda\delta L + \lambda_1\delta L_1 + \lambda_2\delta L_2 + \lambda_3\delta L_3 \ldots = 0 \tag{16}$$

ennek fenn kell állnia, akármilyenek legyenek $\delta s, \delta s', \delta s'', \delta s''', \ldots$ mind nagyságban, mind irányukban. De nem kell elfelejteni a (16) egyenlethez hozzávenni a

$$\begin{cases} \lambda > 0 \\ \lambda_1 > 0 \\ \lambda_2 > 0 \\ \lambda_3 > 0 \\ \ldots\ldots \end{cases} \tag{17}$$

egyenlőtlenségeket."

rengeteg ideig feledésben volt, UBALDInál, *t. i. az emeltyű-szerkezetre nézve. Az elv használatba vételéhez azonban természetesen szükség van már az adott pont-kapcsolatok sajátosságainak tekintetbe vételére, és pedig kinematikájuknak legalább oly mérvű ismeretére és számon-tartására, hogy a kényszer kifejezések felállíthatók legyenek. Ezen a követelményen túl* LAGRANGE *óta az egyenlőségi elv alkalmazása már csak a tiszta analysis dolga s minden egyes esetben kész eljárásokkal megszerkeszthető egyenletek megoldásának problémája.*

Az egyenlőtlenségi elv nem jutott ennyire. Feltalálója FOURIER *csak más elvekből való levezetését mutatta egy értekezésben, amelynek nem is az elv felállítása volt a főczélja.* GAUSS-*ig szünetelt az elv s* GAUSS *csak újra kimondta, még pedig abban az igen rövid közleményében, a melylyel a legkisebb kényszer elvét állította fel. Néhány évvel később még* OSTROGRADSKY *orosz mathematikus foglalkozott vele; ő hozzáfogott az alkalmazásba vételéhez, de nem fogta fel abban az egész általánosságban, a mely neki tulajdonítható, mert a kényszer-viszonyoknak bizonyos faját tartotta csak szem előtt, azt, a melyben a kényszer-kifejezések száma nem nagyobb, mint a virtuális elmozdulási componensek száma; így aránylag nagyon szűk korlátok közt kellett maradnia. Úgy látszik, senki más nem kisérlette meg az elv alkalmazásba vételét, sőt az a vélemény is látszik uralkodni, hogy nem használatképes az elv. Ma jóformán éppen csak némi hire van még."*

„Az itt előadandók főczélja kimutatni, hogy bizonyos módosítással a Lagrange-féle multiplikátoros módszer Fourier elvét is megilleti."

A dolgozat első részében a lineáris egyenlőtlenségekre vonatkozó tételével foglalkozik. Ez a bizonyítás azonban, amint azt már említettük, nem teljes. A második részben levezeti az egyensúly szükséges feltételét:

„Ily módon az esetleg még fennmaradó kényszer-egyenlőségek ezekbe mentek legyen át:

$$\Sigma F\delta q = 0, \quad \Sigma G\delta q = 0, \ldots \tag{11}$$

a kényszer-egyenlőtlenségek pedig a következőkbe;

$$\Sigma S\delta q \geqq 0, \quad \Sigma T\delta q \geqq 0, \ldots; \tag{12}$$

s végre az elvi egyenlőtlenség ezt az alakot öltse:

$$\Sigma Q\delta q \leqq 0 \quad \text{vagyis} \quad -\Sigma Q\delta q \geqq 0, \tag{13}$$

a mely a (3)′*-ból, vagy a* (3)″*-ból származott legyen a szerint, a mint aequilibriumról vagy mozgásról van szó.*

A (11) *alatti egyenleteket is egyenlőtlenségi alakokkal fejezzük ki, úgy, hogy kényszer-kifejezéseiknek a rendszere így jelentkezzék:*

$$\begin{aligned} \Sigma F\delta q \geqq 0, &\quad \Sigma G\delta q \geqq 0, \ldots, \\ -\Sigma F\delta q \geqq 0, &\quad -\Sigma G\delta q \geqq 0, \ldots, \\ \Sigma S\delta q \geqq 0, &\quad \Sigma T\delta q \geqq 0, \ldots. \end{aligned} \tag{14}$$

A Fourier-féle elv azt kívánja, hogy a (13) *mindazokkal a* δq *értékrendszerekkel teljesüljön, a melyekkel* (14) *teljesül. Ez pedig, mint láttuk, csak akkor történik, ha léteznek olyan pozitív értékű multiplikátorok, a melyekkel a* $-Q$ *coëfficiensek mint*

az $F, G, \ldots, -F, -G, \ldots$ *és* $S, T, \ldots$ *coëfficiensek lineáris függvényei kifejezhetők. Jelöljék ezeket a pozitív multiplikátorokat* $\varphi', \psi', \ldots, \varphi'', \psi'', \ldots$ *és* $\lambda, \mu, \ldots$ *. Abban áll a követelés, hogy legyen*

$$-Q = (\varphi' - \varphi'')F + (\psi' - \psi'')G + \ldots + \lambda S + \mu T + \ldots .$$

De $\varphi' - \varphi'', \psi' - \psi'', \ldots$ *negatívok is lehetnek. Így: a Fourier-féle elvnek azokkal a* Q *értékekkel van elég téve, a melyek kifejezései*

$$Q + \varphi F + \psi G + \ldots + \lambda S + \mu T + \ldots = 0,$$

a hol $\varphi, \psi, \ldots$ *és* $\lambda, \mu, \ldots$ *minden* Q *kifejezésében ugyanazok;* $\varphi, \psi, \ldots$ *egészen tetszőlegesek,* $\lambda, \mu, \ldots$ *pedig tetszőleges nemnegatívok. Viszont a* (15) *alatti rendszerből* (11) *és* (12) *alapján könnyen felismerhető módon* (13) *következik."*

A dolgozat egy további részében az alkalmazás két fő típusát említi meg, ezekben az érintkező szilárd testekre és a nem szilárd testekre nyerünk egyensúlyi egyenleteket FARKAS módszerével.

8. A regularitási feltételről

A regularitási feltétel (KUHN és TUCKER terminológiája szerint *„constraint qualification"*) mind a nemlineáris optimalizálás-elmélet, mind pedig a mechanika számára alapvető feltétel.

Századunk elején a matematikai és a fizikai tudományok axiómatikus megalapozása igen fontos tudományos tevékenység volt. HILBERT híres [49] dolgozatában sürgeti a matematikusokat, hogy nyújtsanak axiomatikus megalapozást két „fizikai diszciplina" számára, melyek a valószínűségelmélet és a mechanika. HAMEL 1909-ben megjelent [46] dolgozata igen komoly kísérlet a klasszikus mechanika axiomatikus megalapozására. Sajnos, a *Fourier-féle egyenlőtlenségi elvet* nem vette be rendszerébe. Az 1927-ben megjelent [47] dolgozatban az axiómarendszer egy fejlettebb változatát közli és ebben már helyet kap a *Fourier-elv.* A 17. oldalon találjuk meg Axiom II 2k elnevezéssel. A potenciálos esettel foglalkozva a 33. oldalon a *„Das Energieprinzip"* című részben az Axiom II 5cß jelű axióma rögzíti, hogy a $\delta U \geqq 0$ reláció az egyensúly szükséges feltétele, ahol U a potenciál. E két axióma között Hamelnak konzisztenciát kellett felállítani és ezt csak egy regularitási feltétel segítségével lehetett megtenni. Meg is találjuk ezt a 33. oldalon Axiom II 5cγ gyanánt. Ez előírja, hogy minden tömegponthoz tartozzék egy u skaláris függvény oly módon, hogy teljesüljön a

$$\delta U = \Sigma\, dm \nabla u \delta r \tag{8.1}$$

egyenlőség, ahol dm a tömegpont tömegét, r pedig az állapotát jelenti. Ha (8.1) helyett azt kívánjuk meg, hogy

$$\delta U = \nabla u \delta r, \tag{8.2}$$

ahol r az egész rendszer állapotát jelöli, akkor lényegében a *Karush—Kuhn—Tucker-féle regularitási feltételhez* jutunk. A (8.2) feltétel maga után vonja a (8.1) feltételt,

ugyanis az u függvényeket az U függvényből egyszerűen származtathatjuk oly módon, hogy U-ban minden változót rögzítünk, kivéve azokat, amelyek egy tömegponthoz tartoznak és ezt minden tömegpont esetében megtesszük.

Sajnos, HAMEL nem ismerte FARKAS tételét. Még az 1949-ben publikált *Theoretische Mechanik* című könyvében sem említi a tételt, noha a 69—70, 517—518. oldalakat a *Fourier-féle elv* tárgyalásának szenteli. A regularitási feltétellel kapcsolatos gondolatmenetei heurisztikusak, egzakt matematikai tárgyalást erre vonatkozólag más munkájában sem találunk.

9. További, lineáris egyenlőtlenségekkel kapcsolatos megjegyzések

FARKAS GYULA a *Fourier-elvvel* és a lineáris egyenlőtlenségekkel kapcsolatos legfontosabb eredményeit a [13, 14, 19, 23] dolgozatokban foglalta össze.

A lineáris egyenlőtlenségek megoldásainak parametrikus alakban való előállítását FOURIER [35] kezdeményezte. MINKOWSKI e feladatot az extremális irányok segítségével oldotta meg, ennek fogalma is tőle származik [61]. A FARKAS által adott reprezentációban [19] azoknak a sugaraknak a meghatározása, amelyek nemnegatív súlyokkal vett lineáris kombinációjaként valamennyi megoldás előállítható, igen egyszerű. Ha a (4.5) és a (4.6) relációkban $dx_1, \ldots, dx_n$ helyett $x_1, \ldots, x_n$ szimbólumokat írunk, majd az összes megoldást parametrikus alakban előállítjuk és ezt (4.4)-be helyettesítjük, akkor az egyensúlyi problémát bizonyos esetekben ily módon megoldhatjuk, különösen akkor, ha az aktív feltételek és a változok száma nem nagy. Ezt a módszert javasolta FARKAS [19]. A *Farkas-féle parametrikus reprezentációnak* a lineáris programozásra való felhasználása talán UZAWA [75] dolgozatában a legszerencsésebb. Előbb egy egyenlőtlenséget egyenlőséggé egészít ki, ha szükséges, eliminálja a többi egyenlőtlenségből a jobb oldali számokat, előállítja parametrikus alakban az így nyert homogén egyenlőtlenségek megoldásait, behelyettesíti a kapott kifejezést a megmaradt egyenlőségbe és a célfüggvénybe, ahonnan az optimum értéke és az optimális megoldás egyszerűen meghatározható.

A [21, 22] magyar és német nyelvű dolgozatok tartalma azonos. Ezekben FARKAS a lineáris egyenlőtlenségekkel kapcsolatos tételére további mechanikai alkalmazást adott. Bebizonyította, hogy a mechanikai rendszerekre ható erők felbonthatók lökésszerű erők és olyanok összegére, melyek eleget tesznek az elmozdulásokra vonatkozó feltételek negatívjainak. (VOSS megjegyezte [77, 75. oldal]), hogy az erők ilyenfajta megkülönböztetése PAINLEVÉ gondolata.) A dolgozat matematikai eredményeit [23] is tartalmazza tömör formában. A [20] könyv, melyet magyarul írt, korábbi dolgozataihoz és a [23] dolgozathoz képest új eredményt nem tartalmaz.

Az 1901 és 1917 közötti időben FARKAS nem publikált a lineáris egyenlőtlenségekről. Miután azonban HAAR 1917-ben általánosította FARKAS tételét (az erről szóló cikke 1918-ban jelent meg), megszülettek a [28, 29, 31] dolgozatok. A [25, 26, 27, 30] dolgozatok csak részben kapcsolatosak a lineáris egyenlőtlenségekkel; [30] és [25] tartalmilag nem különbözik.

HAAR három dolgozatot írt a lineáris egyenlőtlenségekről, ezek [42, 43, 44]. Az első kettő ugyanannak a dolgozatnak a magyar és a német változata. Ezekben FARKAS tételére a következő általánosítást adta (a 3. szakasz jelöléseit fogjuk alkalmazni):

9.1. TÉTEL. Ha a

$$\mathbf{g}'\mathbf{x} \geqq b \tag{9.1}$$

egyenlőtlenség következménye a

$$\mathbf{g}_i'\mathbf{x} \geqq b_i, \quad i = 1, \ldots, M \tag{9.2}$$

egyenlőtlenségeknek, vagyis minden, (9.2)-nek eleget tevő $\mathbf{x}$ (9.1)-nek is eleget tesz, akkor léteznek olyan $\lambda, \lambda_1, \ldots, \lambda_M$ nemnegatív számok hogy minden $\mathbf{x} \in R^n$ esetén fennáll az alábbi egyenlőség:

$$\mathbf{g}'\mathbf{x} - b = \sum_{i=1}^{M} \lambda_i(\mathbf{g}_i'\mathbf{x} - b_i) + \lambda. \tag{9.3}$$

Ez ugyanaz az állítás, mint amelyre KUHN és TUCKER hivatkozik NEUMANN JÁNOS [62] kéziratához írt egyik megjegyzésében. NEUMANN e kéziratban közel állt a lineáris programozás dualitás tételének bizonyításához, de szövegéből az tűnik ki, hogy HAAR tételét nem ismerte és hogy (az adott helyen hibásan) FARKAS tételére hivatkozik; nevet azonban nem említ.

HAAR tételét mai szóhasználattal élve oly módon is megfogalmazhatjuk, hogy: ha egy lineáris programozási feladatnak van megengedett megoldása és véges optimuma, akkor a duális feladatnak van megengedett megoldása. Lényegében ehhez az állításhoz az orosz származású varsói professzor, G. VORONOÏ is eljutott [76, 134. kötet 215—216. old.]. Ha azonban FARKASnak a mechanikai egyensúlyra vonatkozó munkáját alaposan szemügyre vesszük, akkor láthatjuk, hogy a *Fourier-elv* FARKAS által bizonyított duális alakja a lineáris potenciál és a lineáris mellékfeltételek esetére tulajdonképpen HAAR tételét szolgáltatja.

Amint HAAR is említi [42, 44], a lineáris egyenlőtlenségek elmélete FARKAS és MINKOWSKI műve. MINKOWSKI híres [61] könyve ezeket az eredményeket (mindkét kiadásban) a 39—45. oldalakon tartalmazza. Véges sok homogén lineáris egyenlőtlenséget vizsgált két szempontból: megadni a megoldások parametrikus előállítását és felfedezni a fölösleges egyenlőtlenségeket, melyek tehát a megoldáshalmaz megváltozása nélkül elhagyhatók. Ez az utóbbi probléma vezette el őt egy, a *Farkas-tételhez* nagyon közel álló tétel bizonyításához. Teljesség kedvéért reprodukáljuk MINKOWSKI tételének pontos megfogalmazását.

Feltesszük, hogy (3.1)-ben a lineárisan független egyenlőtlenségek száma megegyezik n-nel, az $\mathbf{x}$ változó vektor komponenseinek a számával. Amint MINKOWSKI megjegyezte, e feltétel nem szorítja meg az általánosságot. Felteszzük, hogy van olyan $\mathbf{x}$ vektor, melyre (3.1) minden egyenlőtlensége a határozottan nagyobb jellel teljesül. Az ilyen $\mathbf{x}$ vektort valódi megoldásnak nevezzük. Egy $\mathbf{x}$ valódi megoldást a (3.1) kúp extremális sugarának nevezünk, ha nem állítható elő két olyan vektor összegeként, melyek egyike sem 0, nem is konstansszorosa egyik a másiknak, és melyek elemei a (3.1) feltételekkel meghatározott kúpnak.

9.2. TÉTEL. A $\mathbf{g}_i'\mathbf{x} \geqq 0$, $i = 1, \ldots, M$ lineáris formák közül azok, amelyek egyenlőséggel teljesülnek $n-1$ lineárisan független extremális megoldás esetén, azzal a tulajdonsággal bírnak, hogy mind lényegesek és minden további egyenlőtlenség ezek nemnegatív súlyokkal vett lineáris kombinációjaként kifejezhető.

A lineáris egyenlőtlenségekkel kapcsolatos korai dolgozatok túlnyomó többségének adatait megtalálhatjuk a [9] dolgozat irodalomjegyzékében.

IRODALOM

[1] ARROW, K. J., HURWICZ, L. and UZAWA, K. (szerk.), *Studies in Linear and Nonlinear Programming* (Stanford University Press, Stanford, California, 1958).

[2] BEALE, E. M. L., „Cycling in the dual simplex algorithm", *Naval Research Logistics Quarterly* **2** (1955) 269—276.

[3] COURNOT, A., „Sur le calcul des conditions d'inégalité, annoncé par M. Fourier", *Bulletin des Sciences Mathématiques* (Première *Section du Bulletin Universel des Sciences et de l'Industrie, publié sous la dir. de Férussac*) **6** (1826) 1—8.

[4] COURNOT, A., „Extension du principe des vitesses virtuelles au cas où les conditions de liaison du système sont exprimées par des inégalités", *Bulletin des Sciences Mathématiques* (Première *Section du Bulletin Universel des Sciences et de l'Industrie publié sous la dir. de Férussac*) **8** (1827) 165—170.

[5] COURNOT, A. A., „Mémoire sur le mouvement d'un corps rigide, soutenu par un plan fixe", *Journal für die Reine und Angewandte Mathematik* **5** (1830) 133—162, 223—249.

[6] COURNOT, A. A., „Du mouvement d'un corps sur un plan fixe, quand on a égard à la résistance du frottement", *Journal für die Reine und Angewandte Mathematik* **8** (1832) 1—12.

[7] COURNOT, A., *Recherches sur les Principes Mathématiques de la Théorie des Richesses* (Nouvelle Édition, Paris, Marcel Rivière, 1938).

[8] COURNOT, A., *Researches into the Mathematical Principles of the Theory of Wealth* (R. D. Irwin, Homewood, Ill., 1963).

[9] DINES, L. L. and MCCOY, N. H., „On linear inequalities", *Proceedings and Transactions of The Royal Society of Canada Third Series Section III. Mathematical Physical and Chemical Sciences* **27** (1933) 37—70.

[10] DIRICHLET, L., „Sur un moyen général de vérifier l'expression du potentiel relatif à une masse quelconque, homogène ou hétérogène" *Journal für die Reine und Angewandte Mathematik* **32** (1846) 80—84.

[11] DIRICHLET, L., „Über die Stabilität des Gleichgewichts", *Journal für die Reine Und Angewandte Mathemetik* **32** (1846) 85—88.

[12] *Encyklopädie der Mathematischen Wissenschaften* (Leipzig, Teubner, 1901—1908) Vierter Band, Mechanik, Erster Teilband.

[13] FARKAS, GY., „A Fourier-féle mechanikai elv alkalmazásai", *Mathematikai és Természettudományi értesítő* **12** (1894) 457—472.

[14] FARKAS, J., „Über die Anwendungen des mechanischen Princips von Fourier", *Mathematische und Naturwissenschaftliche Berichte aus Ungarn* **12** (1896) 263—281.

[15] FARKAS, GY., „A Fourier-féle mechanikai elv alkalmazásainak algebrai alapjáról", *Mathematikai és Physikai Lapok* **5** (1896) 49—54.

[16] FARKAS. J., „Die algebraischen Grundlagen der Anwendungen des Fourierschen Principes in der Mechanik", *Mathematische und Naturwissenschatliche Berichte aus Ungarn* **15** (1899) 25—40.

[17] FARKAS, GY., „A Fourier-féle mechanikai elv alkalmazásának algebrai alapja", *Mathematikai és Természettudományi Értesítő* **16** (1898) 361—364.

[18] FARKAS, J., „Die algebraische Grundlage der Anwendungen des mechanischen Princips von Fourier", *Mathematische und Naturwissenschaftliche Bereichte aus Ungarn* **16** (1899) 154—157.

[19] FARKAS, GY., „Paraméteres módszer Fourier mechanikai elvéhez", *Mathematikai és Physikai Lapok* **7** (1898) 63—71.

[20] FARKAS, GY., *Vektortan és az egyszerű inaequatiok tana* (Kolozsvár, 1900).

[21] FARKAS, GY., „Általános mechanikai elvek az aether számára", *Mathematikai és Természettudományi Értesítő* **19** (1901) 99—122.

[22] FARKAS, J., „Allgemeine Principen für die Mechanik des Aethers", *Archives Néerlandaises, Série II. Vol.* **4** (*Livre jubiliaire dédié* à H. A. LORENTZ).

[23] FARKAS, J., „Theorie der einfachen Ungleichungen", *Journal für die Reine und Angewandte Mathematik* **124** (1901) 1—27.

[24] FARKAS, J., „Beiträge zu den Grundlagen der analytischen Mechanik", *Journal für die Reine und Angewandte Mathematik* **131** (1906) 165—201.

[25] FARKAS, GY., „Biztos egyensúly potenciál nélkül", *Mathematikai és Természettudományi Értesítő* **32** (1915) 339—354.

[26] FARKAS, GY., „Nemvonalas egyenlőtlenségek vonalassá tétele", *Mathematikai és Természettudományi Értesítő* **35** (1917) 41—50.

[27] FARKAS, GY., „Multiplicatoros módszer négyzetes alakokhoz", *Mathematikai és Természettudományi Értesítő* **35** (1917) 51—53.
[28] FARKAS, GY., „Egyenlőtlenségek alkalmazásának új módjai", *Mathematikai és Természettudományi Értesítő* **36** (1918) 297—308.
[29] FARKAS, GY., „A lineáris egyenlőtlenségek következményei", *Mathematikai és Természettudományi Értesítő* **36** (1918) 397—408.
[30] FARKAS, J., „Stabiles Gleichgewicht ohne Potential", *Mathematische und Naturwissenschaftliche Berichte aus Ungarn* **32** (1922) 43—50.
[31] FARKAS, GY., „Alapvetés az egyszerű egyenlőtlenségek vektorelméletéhez", *Matematikai és Természettudományi Értesítő* **43** (1926) 1—3.
[32] FÉNYES, I., „Megjegyzések és kiegészítések a mechanika elveinek *Farkas Gyula-féle tárgyalásmódjához*", *Fizikai Szemle* **4** (1945) 99—102.
[33] FILEP, L., „*Farkas Gyula* élete és munkássága", egyetemi doktori értekezés, Debrecen, 1977.
[34] FOURIER, J., „Mémoire sur le statique", *Journal de l'Ecole Polytechnique* **5** (1798), [36] *Vol. II.* 478—520.
[35] FOURIER, J., „Solution d'une question particulière du calcul des inégalités", *Nouveau Bulletin des Sciences par la Société Philomatique de Paris* (1826), [36] Vol. II. 317—319.
[36] FOURIER, J., *Oeuvres* (Gauthier—Villars, Paris, 1888).
[37] GALE, D., KUHN, H. W. and TUCKER, A. W., „Linear programming and the theory of games", [53] 317—329.
[38] GAUSS, C. F., „Über ein neues allgemeines Grundgesetz der Mechanik", *Journal für die Reine und Angewandte Mathematik* **4** (1829) 232—235.
[39] GIBBS, J. W., „On the fundamental formulae of dynamics", *American Journal of Mathematics* **2** (1879) 49—64.
[40] GRATTAN-GUINNESS, I., „*Joseph Fourier*'s anticipation of linear programming", *Operational Research Quarterly* **21** (1796) 361—364.
[41] Григорян, А., Т., *Очерки Истории Механики в России* (Акад. Наук СССР., Москва,1961).
[42] HAAR, A., „A lineáris egyenlőtlenségekről", *Mathematikai és Természettudományi Értesítő* **36** (1918) 279—296.
[43] HAAR, A., „*Die Minkowskische Geometrie* und die Annäherung an stetige Funktionen", *Mathematische Annalen* **78** (1918) 294—311.
[44] HAAR, A., „Über linearen Ungleichungen", *Acta Scientiarium Mathematicarum* **2** (1924) 1—14.
[45] HAAR, A., *Gesammelte Arbeiten* (Akadémiai Kiadó, Budapest, 1959).
[46] HAMEL, G., „Über die Grundlagen der Mechanik", *Mathematische Annalen* **66** (1909) 350—397.
[47] HAMEL, G., „Die Axiome der Mechanik", [48] 1—42.
[48] *Handbuch der Physik* (*H. Geiger* és *K. Scheel szerk.*) *Band V., Grundlagen der Mechanik, Mechanik der Punkte und Starren Körper* (Springer, Berlin, 1927).
[49] HILBERT, D., „Mathematische Probleme (*Vortrag gehalten auf dem internationalen Mathematiker-Kongress zu Paris* 1900)", *Nachrichten von der Königl. Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse* (1900) 253—297.
[50] HOFFMAN, A. J., „Cycling in the simplex algorithm", *National Bureau of Standards Report, No.* 2974, 1953.
[51] JOHN, F., „Extremum problems with inequalities as subsidiary conditions", [73] 187—204.
[52] KARUSH, W., Minima of Functions of Several Variables with Inequalities as Side Conditions, Master's Thesis (Department of Mathematics, University of Chicago, 1939).
[53] KOOPMANS, T. C. (szerk.). *Activity Analysis of Production and Allocation* (Wiley, New York, 1951).
[54] KUHN, H. W. and TUCKER, A. W., *Nonlinear Programing*, [63] 481—492.
[55] KUHN, H. W., „Nonlinear Programming: a historical view", [64] 1—26.
[56] LAGRANGE, J. L., *Mécanique Analytique I—II* (Paris, 1788).
[57] LAGRANGE, J. L., *Ouvres* (Gauthier—Villars, Paris 1888).
[58] LEMKE, C. E., „The dual method for solving the linear programming problem", *Naval Research Logistics Quarterly* **1** (1954) 36—47.
[59] MAYER, A., „Über die Aufstellung der Differentialgleichungen der Bewegung für reibungslose Punktsysteme, die Bedingungsungleichungen unterworfen sind", *Berichte über die Verhandlungen der Königlich Sächsischen Gesellschaft der Wissenschaften zu Leipzig, Mathematisch-Physische Classe* **50** (1898) 224—244.
[60] MAYER, A., „Zur Regulierung der Stosse in reibungslosen Punktsystemen, die dem Zwange von Bedingungsungleichungen unterliegen", *Berichte über die Verhandlungen der Königlich*

Sächsischen Gesellschaft der Wissenschaften zu Leipzig, Mathematisch-Physische Classe **50** (1898) 245—264.
[61] MINKOWSKI, H., *Geometrie der Zahlen* (Teubner, Leipzig und Berlin, első kiadás 1896, második kiadás 1910).
[62] VON NEUMANN, J., „Discussion of a maximum principle," *Collected Works, Vol.* VI., 89—95.
[63] NEYMAN, J., (szerk.) *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability* (University of California Press, Berkeley, California, 1950).
[64] *Nonlinear Programming, SIAM-AMS Proceedings Vol.* IX. (American Mathematical Society, Providence, Rhode Island, 1976).
[65] NORDHEIM, L., „Die Prinzipe der Dynamik", [48] 43—90.
[66] ORTVAY, R., „*Farkas Gyula* tudományos működése", *Matematikai és Fizikai Lapok* **34** (1927) 5—25.
[67] ORTVAY, R., „*Farkas Gyula* emlékezete", az *MTA Tagjai Felett Tartott Emlék-Beszédek* **21** (1933) 15 füzet.
[68] OSTROGRADSKY, M., „Considérations générales sur les momens des forces", *Mémoires de l'Académie Impériale des Sciences de Saint-Pétersbourg, Sixième Série* **1** (1838) 129—150.
[69] OSTROGRADSKY, M., „Mémoire sur les déplacemens instantanés des systèmes assujettis a des conditions variables", *Mémoires de l'Académie Impériale des Sciences de Saint-Pétersbourg, Sixième Série* **1** (1838) 565—600.
[70] POINSOT, L., *Éléments de Statique* (Douzième Édition, Paris, Gauthier—Villars, 1877, első kiadás 1803).
[71] PRÉKOPA, A., *Lineáris Programozás I.* (Bolyai János Matematikai Társulat, Budapest, 1968).
[72] STÄCKEL, P., „Elementare Dynamik der Punktsysteme und starren Körper", [12] 435—691.
[73] *Studies and Essays, presented to Courant on his 60th birthday* (Interscience, New York, 1948).
[74] SZÉNÁSSY, B., *A Magyarországi Matematika Története* (Akadémiai Kiadó, Budapest, 1970).
[75] UZAWA, H., „An elementary method for linear programming", [1] 179—188.
[76] VORONOÏ, G., „Nouvelles applications des parameters continus à la théorie des formes quadratiques", *Journal für die Reine und Angewandte Mathematik, Premier Mémoire:* **133** (1908) 97—178, *Deuxième Mémoire:* **134** (1908) 198—287, **136** (1909) 67—178.
[77] VOSS, A., „Die Prinzipien der rationellen Mechanik" [12] 3—121.
[78] ZERMELO, E., „Über die Bewegung eines Punktsystemes bei Bedingungsungleichungen" *Nachrichten von der königl. Gesellschaft der Wissenschaften zu Göttingen. Mathematisch-Physikalische Klasse.* (1898) 306—310.

(Beérkezett: 1979. március 2.)

PRÉKOPA ANDRÁS
BME GÉPÉSZMÉRNÖKI KAR MATEMATIKA TANSZÉK
1521 BUDAPEST, STOCZEK U. H ÉP. IV. EM.

ON THE DEVELOPMENT OF OPTIMIZATION THEORY

A. PRÉKOPA

The method of LAGRANGE for finding extrema of functions subject to equality constraints was published in 1788 in the famous book *Mécanique Analytique.* The works of KARUSH, JOHN, KUHN and TUCKER concerning optimization subject to inequality constraints appeared more than 150 years after that. The purpose of this paper is to call attention to important papers, published as contributions to mechanics, containing fundamental ideas concerning optimization theory. The most important works in this respect were done primarily by FOURIER, COURNOT, FARKAS and further by GAUSS, OSTROGRADSKY and HAMEL.

Alkalmazott Matematikai Lapok 4 (1978) 193—196

A SZÁLLÍTÁSI PROBLÉMÁRA ALKALMAZOTT SZIMPLEX ALGORITMUS CIKLIZÁLÁSÁNAK A LEHETŐSÉGÉRŐL

KÉRI GERZSON
Budapest

GASSNER még 1964-ben példákkal bizonyította a hozzárendelési és a szállítási probléma ciklizálásának a lehetőségét a szimplex módszerrel történő megoldási kísérlet során, ezeket a példákat azonban nem siettek átvenni a szállítási problémát is tárgyaló matematikai programozási kézikönyvek. Ezért elég általános az a nézet, mely szerint a szállítási probléma ciklizálásának a lehetősége vagy lehetetlensége elméletileg még ma is eldöntetlen. A kérdésnek az alaposabb tisztázását kívánjuk elősegíteni GASSNER egy példájának a felelevenítésével. E példát úgy módosítjuk, hogy a ciklizálásra vonatkozó állítás helyessége nagyon gyorsan, mindössze két iterációs lépés elvégzésével belátható legyen.

1. A szimplex módszer ciklizálásának a lehetőségéről általában, illetve a szállítási problémára szorítkozva

A szimplex módszer ciklizálásának azt a jelenséget nevezik, amikor megengedett bázisoknak ugyanaz a sorozata ciklikusan ismétlődik az eljárás során. Ha a kiválasztási stratégiát és a szimplex módszer szabad paramétereinek az értékét menet közben nem változtatjuk meg, és ezenkívül a kerekítési hibák lehetőségét sem engedjük meg, akkor a ciklizálást úgy is jellemezhetjük, hogy egy korábban már előfordult bázis (szállítási probléma esetén báziscellarendszer) később újra fellép. Ilyenkor az ismétlődő bázishoz, valamint a sorrendben közbe eső bázisok mindegyikéhez ugyanaz a bázismegoldás tartozik. Ha viszont nem lép fel ciklizálás, akkor a szimplex módszer eljárása garantáltan véges sok lépés után befejeződik.

Mint ismeretes, a szimplex módszer szabad paramétereket is tartalmaz, mivel például a bázisba belépő vagy a bázisból kilépő vektor kiválasztására gyakran több lehetőség áll rendelkezésre. A szimplex módszer klasszikus tárgyalásmódjában a megoldás menetének a többértelműségét tudomásul véve, a szimplex módszer bázisváltásának a szabályait egyszerűen úgy fogalmazhatjuk meg, hogy egy báziscsere legális, ha a bázisba belépő vektorhoz negatív árnyékár tartozik (maximum feladat esetén), a kilépő vektor választása pedig garantálja az új bázis megengedettségét. Nevezhetjük az e szabálynak megfelelő algoritmust „leglazább" szimplex algoritmusnak is, a bázisba belépő és az onnan kilépő vektorra vonatkozó kiválasztási szabály valamilyen módon történő egyértelmű előírása által kapott algoritmusokat pedig „szoros" szimplex algoritmusoknak. E két típus között helyezkedik el például a lexikografikus szimplex módszer, amely garantálja a ciklizálás lehetetlenségét, és amelynél csak a bázisból kilépő vektor meghatározása egyértelmű, a bázisba belépő vektoré azonban nem. A lexikografikus szimplex módszernél egyszerűbben végre-

hajtható és a ciklizálás ellen ugyanolyan jó védelmet biztosító egyéb eljárásokat ismertet a [3] cikk.

A továbbiakban szimplex módszer alatt a „leglazább" szimplex algoritmust értjük és ennek megfelelően értelmezzük a szimplex módszer szabályait.

A szimplex módszer ciklizálásának elvi lehetőségét még az ötvenes években HOFFMAN [7] és BEALE [2] egy-egy példával bizonyította. BEALE példája, a primál szimplex módszerre való átfogalmazásban megtalálható e dolgozat irodalomjegyzékében szereplő hivatkozások közül az [1], [6], [10] és [11] könyvekben.

E helyen szeretnék röviden kitérni arra is, hogy a szimplex módszernek az általában vett lineáris programozási problémák esetén történő ciklizálását illetően már kezd módosulni az a korábbi általános nézet, mely szerint a ciklizálásra csak mesterségesen kiagyalt példák ismeretesek, hogy a ciklizálás gyakorlati problémák esetén soha (vagy szinte soha) nem fordul elő, és hogy a ciklizálás kizárására szolgáló technikák alkalmazása mindig fölösleges a ciklizálás rendkívüli ritkasága miatt. KOTIAH és STEINBERG a [8] cikkben beszámolnak egy olyan esetről, hogy a lineáris programozás egy gyakorlati alkalmazása során ciklizálást észleltek. Ugyanők a [9] cikkükben is hivatkoznak erre a felfedezésre, majd hozzáteszik, hogy szóbeli közlés formájában másoktól is hallottak hasonló esetekről. A [9] cikkben a szerzők részletesen kifejtik, hogy mi a véleményük a ciklizálás helyes megítélését illetően. E vélemény egy mondatba sűrített lényege a következő: Ciklizálás esetenként előfordul a gyakorlatban és ez nem véletlen a degeneráció rendkívül gyakori volta miatt.

GASSNER a [4] cikkben megmutatta, hogy a szimplex módszer a szállítási (sőt a hozzárendelési) problémánál is ciklizálhat. Ennek ellenére a kézikönyvek egy része — az irodalomjegyzékben szereplő hivatkozások közül az [5], [6], [10] és [11] könyvek — eldöntetlen kérdésként említi a ciklizálás elméleti lehetőségét a szállítási probléma esetére. Ez az elméleti jellegű bizonytalanság adott esetben helytelenül befolyásolhat annak a gyakorlati kérdésnek az eldöntésében, hogy egy, a szállítási probléma megoldására szolgáló gépi programba — a futtatandó konkrét feladatoktól, a gép konfigurációjától, software adottságoktól stb. függően — beépítsünk-e olyan eljárást, amely bizonyítottan garantálja a ciklizálás lehetetlenségét.

A szállítási problémára alkalmazott szimplex módszer, mint ismeretes, rendelkezik azzal a speciális tulajdonsággal is, hogy egész értékű bemenő adatok esetén a módszer csak egész értékű aritmetikát használ. E tulajdonság miatt a kerekítési hibáknak a ciklizálás ellen működő jótékony hatása sem tud érvényre jutni a szállítási probléma esetén.

2. Egy numerikus példa a szállítási probléma ciklizálására

A továbbiakban bemutatásra kerülő példa lényegében azonos GASSNER egyik példájával. A költségmátrix elemei GASSNERnél paraméteres alakban szerepeltek, itt azonban a paraméterek értékét rögzíteni fogjuk olyan értékekkel, hogy ezáltal a példa nemcsak áttekinthetőbbé, hanem szimmetrikussá, és így könnyen megjegyezhetővé válik.

Tekintsük tehát a következő — az 1. ábrán megadott mátrixú — hozzárendelési problémát,

1	0	0	0
0	0	1	1
0	1	0	1
0	1	1	0

1. ábra

és ezt fogalmazzuk meg a lineáris programozási, illetve a szállítási probléma szokásos megadási módjában. Az utóbbihoz mindjárt 3, egymásból a szimplex módszer szabályainak megfelelően adódó szállítási táblát is megadunk (2. ábra). A táblázatokban a bázisváltozók értékét bekarikázzuk.

Lineáris programozási probléma formájában:

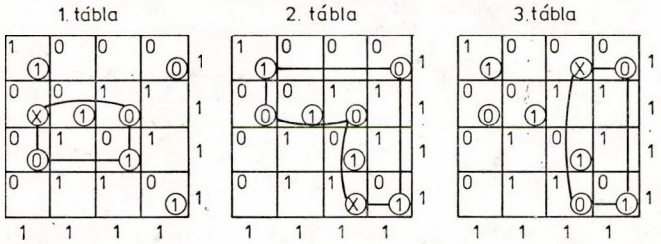
$$\sum_{j=1}^{4} x_{ij} = 1, \quad (i = 1, 2, 3, 4)$$

$$\sum_{i=1}^{4} x_{ij} = 1, \quad (j = 1, 2, 3, 4)$$

$$x_{ij} \geqq 0, \quad (i, j = 1, 2, 3, 4)$$

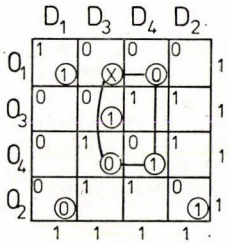
$$\min x_{11} + x_{23} + x_{24} + x_{32} + x_{34} + x_{42} + x_{43}$$

A szállítási táblák:



2. ábra

A karikába írt x-szel jelölt, bázisba bevonandó cellához tartozó $z_{ij} - c_{ij} = 1$ mindhárom tábla esetében. Ha most meg tudjuk mutatni, hogy a 3. tábla a sorok és oszlopok sorrendjétől, s ezenkívül a sorok és oszlopok szerepének esetleges felcserélésétől (azaz a főátlóra való tükrözéstől) eltekintve, teljesen azonos az 1. táblával, akkor ezzel egyúttal azt is belátjuk, hogy a szimplex módszer szabályainak megfelelő báziscserék — az 1. táblától a 3. táblához vezető báziscserékhez hasonló módon — vég nélkül ismételhetők, tehát a szimplex módszer ciklizál az így meghatározott útvonalon. Ez valóban így van, mert az 1. táblát előbb a főátlóra tükrözve (3. ábra), majd a 2., 3., illetve 4. sort a 3., 4., illetve 2. sorba átrakva, és egyúttal az oszlopokat is ugyanígy permutálva, a 2. ábrában levő 3. táblával teljesen azonos táblát kapunk.



3. ábra

Végül megjegyezzük, hogy a fenti gondolatmenet nem bizonyítja a ciklizálás előfordulásának a lehetőségét a báziscsere bármely szokásos egyértelmű szabályozásának az alkalmazása esetén és tudomásom szerint még nem ismeretes kidolgozott numerikus példa a szállítási problémára alkalmazott „szoros" szimplex algoritmusok ciklizálására.

IRODALOM

[1] BAZARAA, M. S. and JARVIS, J. J., *Linear Programming and Network Flows* (John Wiley and Sons, New York, 1977).
[2] BEALE, E. M. L., „Cycling in the dual simplex algorithm", *Naval Research Logistics Quarterly* **2** (1955) 269—276.
[3] BLAND, R. G., „New finite pivoting rules for the simplex method", *Mathematics of Operations Research* **2** (1977) 103—107.
[4] GASSNER, B. J., „Cycling in the transportation problem", *Naval Research Logistics Quarterly* **11** (1964) 43—58.
[5] GEGESI KISS, P., KOCSIS, A., RACSKÓ, P., SCHNEIDER, E. és SOMOGYI, K., *Operációkutatási módszerek* (SZÁMOK, Budapest, 1977).
[6] HADLEY, G., *Linear Programming* (Addison—Wesley, Reading, 1962).
[7] HOFFMAN, A. J., „Cycling in the simplex algorithm", *National Bureau of Standards Report*, No. 2974, Dec. 1953.
[8] KOTIAH, T. C. T. and STEINBERG, D. I., „Occurrances of cycling and other phenomena arising in a class of linear programming models", *Communications of the ACM* **20** (1977) 107—112.
[9] KOTIAH, T. C. T. and STEINBERG, D. I., „On the possibility of cycling with the simplex method", *Operations Research* **26** (1978) 374—375.
[10] PRÉKOPA, A., *Lineáris programozás I.* (Bolyai Társulat, Budapest, 1968).
[11] ZIONTS, S., *Linear and Integer Programming* (Prentice-Hall, Englewood Cliffs, 1974).

(Beérkezett: 1978. augusztus 10.)

KÉRI GERZSON
MTA SZÁMÍTÁSTECHNIKAI ÉS AUTOMATIZÁLÁSI KUTATÓ INTÉZET
1250 BUDAPEST, I., URI U. 49.

ON THE POSSIBILITY OF CYCLING IN THE SIMPLEX ALGORITHM APPLIED TO THE TRANSPORTATION PROBLEM

G. KÉRI

GASSNER proved in 1964 through some examples that cycling in the simplex method is possible, even if the problem to be solved is a transportation or assignment problem. However, these examples did not found their way into the handbooks of mathematical programming, therefore it is a rather general opinion that the possibility or impossibility of cycling in the transportation problem is still undecided. In order to make clearer this point, an example of GASSNER is recalled, with some modifications and with a very short proof of its validity. For the proof only two iterational steps are performed.

Alkalmazott Matematikai Lapok 4 (1978) 197—228

VÉLETLEN KERESŐ ELJÁRÁSOK KONVERGENCIÁJÁNAK ÉS NUMERIKUS HATÉKONYSÁGÁNAK VIZSGÁLATA

PINTÉR JÁNOS
Budapest

A dolgozatban egy sztochasztikus optimalizálási módszer vizsgálatával foglalkozunk. Az 1. pont a véletlen kereső eljárások alapfogalmait és a módszerrel foglalkozó néhány korábbi munka összefoglaló ismertetését tartalmazza. A 2. pontban véletlen kereső eljárások általános konvergenciatulajdonságait részletezzük. Ezen belül először — a [28] dolgozat eredménye alapján — folytonosan differenciálható függvényekkel képzett matematikai programozási problémák esetén vizsgáljuk a módszer konvergenciáját. Ezzel kapcsolatban megmutatjuk, hogy a hatékony irányok módszere sztochasztikus gradiensekkel felírt iránykereső feladat esetén is alkalmazható. Ezt követően konvergenciabizonyítást adunk folytonos, feltétel nélküli feladatra is. A 3. pont egy — a konvergenciavizsgálatokban szereplő elméleti eljárások alapján konstruált — véletlen kereső algoritmust ismertet. Végül a 4. pontban a módszer gépi realizációjának tapasztalatait részletezzük és hasonlítjuk össze korábbi determinisztikus és sztochasztikus algoritmusokra vonatkozó numerikus tapasztalatokkal.

1. Bevezetés

A dolgozatban egy sztochasztikus függvényminimalizáló eljárástípus vizsgálatával foglalkozunk. A sztochasztikus jelző arra utal, hogy a szóban forgó módszerek lépései nem teljesen determináltak, hanem a véletlentől is függenek. Az eljárás gradiensmentes, kereső lépései az újabb vizsgált pontra való áttérésből és a pontbeli függvényérték meghatározásából állnak.

Elsőként BROOKS [4] és RASZTRIGIN [18] foglalkozott sztochasztikus függvényminimalizáló eljárásokkal. RASZTRIGIN későbbi vizsgálatai [30, 31] szerint a véletlen kereső módszerek bizonyos esetekben kevesebb függvényérték kiszámítását igénylik adott pontosságú megoldás meghatározásához, mint a gradiens módszer numerikus differenciáláson alapuló változatai. Alkalmazásuk előnyös lehet olyan esetekben is, amikor a célfüggvény gradiensét az optimalizálás során nem használhatjuk fel (pl. mert a célfüggvény alakja nem, vagy csak közelítőleg ismert, vagy differenciálhatósága nem biztosított). Ez a helyzet igen gyakori lehet például sztochasztikus programozási problémák esetében, ahol a determinisztikus alapmodell bizonyos paramétereinek véletlen jellegét figyelembe véve a feladat matematikai és számítástechnikai szempontból egyaránt bonyolultabbá válik.

A későbbiekben SCHUMER és STEIGLITZ [21] az n-dimenziós tér egységgömbjének felületén egyenletes eloszlású véletlen irányoknak megfelelő lépéseket feltételezve vizsgálta a lépéshossz szerepét a $\sum_{i=1}^{n} x_i^2$ kvadratikus célfüggvény esetén. Erre az elméleti esetre meghatározták azt az optimális lépéshosszt, amellyel az adott közelítő megoldásvektor ismeretében a függvényérték relatív javulásának várható értéke maximális. Az elméleti modell alapján kialakított adaptív lépéshosszú algoritmus

hatékonyságát a $\sum_{i=1}^{n} x_i^2$ és $\sum_{i=1}^{n} c_i x_i^2$ (ahol c_i, $i=1, \ldots, n$ a [0,1; 1] intervallumon egyenletes eloszlású független valószínűségi változóknak az adott feladat megoldása során rögzített realizációi) célfüggvényekre vizsgálták, megállapítva az egységgömb felületéről induló kereső eljárásnak az optimum adott pontosságú meghatározásához szükséges átlagos lépésszámát. A módszert WHITE és DAY [26] továbbfejlesztette és hatékonyságát a fenti célfüggvények esetén a FLETCHER—POWELL-módszer [8] effektivitásával hasonlította össze. SCHRACK és BOROWSKY [19] szintén vizsgálta az adaptív lépéshosszú kereső módszert és azt további véletlen optimalizáló eljárásokkal vetette össze. Véletlen kereső eljárásoknak és determinisztikus módszereknek nagyméretű négyzetösszeg-minimalizálási feladatokon való összehasonlításával foglalkozott SMITH [22] is.

UGRAY [25] az adaptív lépéshosszt a sikertelen lépést követő ellenkező irányú kereséssel kombinálja (l. még SCHRACK és CHOIT [20]). A módszer effektivitását a fenti célfüggvényekre vonatkozóan NELDER és MEAD flexibilis poliéder módszerével [12], a NEWTON—RAPHSON-módszerrel (l. pl. [23]) és a FLETCHER—POWELL-algoritmussal, továbbá SCHUMER—STEIGLITZ és WHITE—DAY eredményeivel hasonlítja össze.

A gradiens irány sztochasztikus becslése felhasználásával történő optimalizálás módszerét először JERMOLJEV [6] és NYIKOLAJEV [13, 14, 15] vizsgálta. Ezt a gondolatot módosította BARTOLOMEI [3], majd ARCHETTI [1], aki a súlyozott véletlen gradiens módszer effektivitását vizsgálta a $\sum_{i=1}^{n} x_i^2$, $\sum_{i=1}^{n} c_i x_i^2$, $\sum_{i=1}^{n} c_i x_i^4$ célfüggvények esetén, és azt a korábbi adaptív lépéshosszú algoritmusokénál jobbnak találta. ($0, 1 \leqq c_i \leqq 1$, $i=1, \ldots n$ független, egyenletes eloszlású valószínűségi változók rögzített realizációi).

A fent említett numerikus tapasztalatokat összefoglalva megállapítható, hogy a véletlen kereső módszerek strukturális egyszerűségük mellett általában véve hatékonyan alkalmazhatók a felsorolt tesztfeladatok közelítő megoldására. Ugyanakkor megjegyezzük, hogy az említett munkák egyike sem foglalkozik véletlen kereső módszerekre vonatkozó, általános érvényű konvergenciatulajdonságok vizsgálatával. Emellett a tesztfeladatok köre lényegileg az igen egyszerű, kvadratikus típusú célfüggvényekre korlátozódik, és nem található a célfüggvény értékét torzító véletlen zajhatásokra vonatkozó szisztematikus vizsgálat sem.

BOGOMOLOV és KARMANOV [28] bizonyították be egy elméleti véletlen kereső algoritmus konvergenciáját folytonosan differenciálható függvényekkel képezett matematikai programozási feladat esetére. A bizonyítás gondolatmenetét ismertetjük és ennek kapcsán megmutatjuk, hogy a hatékony irányok módszere [27] sztochasztikus gradiensekkel felírt iránykereső feladat esetében is alkalmazható. Ezt követően konvergencia-bizonyítást adunk folytonos, feltétel nélküli feladatra vonatkozóan is.

A konvergencia-vizsgálatok után egy olyan algoritmust ismertetünk, amelynek kialakítása során elsősorban RASZTRIGIN és ARCHETTI munkáit vettük figyelembe. Az algoritmus a korábbi dolgozatunkban [16] alkalmazott heurisztikus véletlen kereső eljárás továbbfejlesztett változata.

A dolgozat utolsó pontjában a módszer gépi realizációjának tapasztalatait részletezzük, ennek során a fentebb ismertetett egyszerű célfüggvényeken túlmenően számos más tesztfeladatra vonatkozó eredményt is bemutatunk és ezeket egyéb sztochasztikus és determinisztikus optimalizáló algoritmusok eredményeivel vetjük össze.

2. Véletlen kereső eljárások konvergenciájának vizsgálata

Konvergencia folytonosan differenciálható célfüggvény és feltételi függvények esetén

BOGOMOLOV és KARMANOV [28] dolgozatukban a

$$\min_{\mathbf{x}\in X} f(\mathbf{x}) \tag{2.1}$$

$$X = \{\mathbf{x}: \mathbf{x}\in R^n, f_i(\mathbf{x}) \geqq 0,\ i = 1,\ldots, m\}$$

matematikai programozási feladatnak — ahol az f, f_i $i=1, \ldots, m$ függvények folytonosan differenciálhatók valamely, az X halmazt tartalmazó H nyílt, konvex halmazon — egy elméleti véletlen kereső módszerrel történő megoldása konvergenciáját igazolják. A továbbiakban a vizsgálatainkat egyszerűsítő $\min_{\mathbf{x}\in X} f(\mathbf{x})\geqq 0$ feltevéssel élünk (ez alulról korlátos célfüggvény konstanssal való növelése útján mindig elérhető), és használjuk az $f_0(\mathbf{x})=f(\mathbf{x})$ jelölést. Feltesszük továbbá, hogy $f_0(\mathbf{x})$ legfeljebb véges számú lokális minimumhellyel rendelkezik. Az algoritmus tetszőleges $\mathbf{x}_0\in X$ pontból indul. A k-adik lépésben ismert $\mathbf{x}_k$ megoldáshoz olyan $\mathbf{d}_k\in R^n$ egységvektort és $\alpha_k\geqq 0$ lépéshosszt keresünk, amelyekkel teljesül

$$\begin{aligned} &\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \mathbf{d}_k \\ &[\mathbf{x}_k, \mathbf{x}_{k+1}] \subset X \\ &f_0(\mathbf{x}_{k+1}) \leqq f_0(\mathbf{x}_k) \end{aligned} \tag{2.2}$$

Itt $\mathbf{d}_k$-t azon lehetséges $\mathbf{d}$ vektorok közül választjuk ki, amelyekhez létezik $\bar{\alpha}>0$ úgy, hogy $\alpha\in[0, \bar{\alpha}]$ esetén $\mathbf{x}_k-\alpha\mathbf{d}\in X$.

Legyen $\bar{\alpha}_k=\sup \bar{\alpha}$ rögzített $\mathbf{x}_k$ és $\mathbf{d}_k$ esetén. Egy lehetséges irány kiválasztása után az α_k lépéshossz meghatározása úgy történik, hogy teljesüljön

$$f_0(\mathbf{x}_k-\alpha_k\mathbf{d}_k) \leqq \varepsilon_k f_0(\mathbf{x}_k)+(1-\varepsilon_k) \inf_{0\leqq\alpha\leqq\bar{\alpha}_k} f_0(\mathbf{x}_k-\alpha\mathbf{d}_k). \tag{2.3}$$

Itt minden k-ra $0<\varepsilon_k\leqq\varepsilon<1$, ε pedig k-tól függetlenül adott érték.

(Megjegyezzük, hogy a (2.3) feltétel az $[\mathbf{x}_k, \mathbf{x}_k-\bar{\alpha}_k\mathbf{d}_k]$ szakaszon felvett célfüggvényértékek ismeretlen infinumát is tartalmazza, ezért ebben a formában közvetlenül nem használható a gyakorlatban.)

Ha a (2.3) egyenlőtlenség nem teljesül, akkor újabb $\mathbf{d}$ iránnyal próbálkozunk.

A (2.1) problémával kapcsolatban tekintsük rögzített $\mathbf{x}$ esetén a ZOUTENDIJK [27] által vizsgált következő iránykereső eljárást:

Keresendő olyan $\mathbf{d}\in R^n$ vektor és σ valós szám, amely megoldja az alábbi feladatot:

$$\begin{aligned} &\max \sigma \\ &(\nabla f_i(\mathbf{x}), \mathbf{d})+\sigma \leqq 0, \quad i\in I(\mathbf{x}, v) = \{i:\ 1\leqq i\leqq m,\ 0\leqq f_i(\mathbf{x})\leqq v\}, \quad v\geqq 0 \\ &-(\nabla f_0(\mathbf{x}), \mathbf{d})+\sigma \leqq 0 \\ &\|\mathbf{d}\| \leqq 1 \end{aligned} \tag{2.4}$$

(Itt és a továbbiakban $\|\mathbf{v}\|$ a $\mathbf{v}$ vektor euklideszi normáját jelöli.)

Jelölje a (2.4) feladat optimális megoldását $\bar{\mathbf{d}}(\mathbf{x}, v)=\bar{\mathbf{d}}$ és $\bar{\sigma}(\mathbf{x}, v)=\bar{\sigma}$. Egyszerűen igazolható a

2.1. LEMMA: Ha $\mathbf{x}$ a (2.1) probléma lokális minimuma, akkor $v>0$ esetén (2.4) optimális $(\bar{\mathbf{d}}, \bar{\sigma})$ megoldásában $\bar{\sigma}=0$ ([27]).

Bizonyítás: A lemma igazolására először is megjegyezzük, hogy $\mathbf{d}=\mathbf{0}\in R^n$, $\sigma=0$ (2.4) megengedett megoldása $\mathbf{x}\in X$ esetén. Ezután $\bar{\sigma}$ nempozitív voltának bizonyítása indirekt úton történik.

Tegyük fel, hogy $\mathbf{x}$ lokális minimumhely, és (2.4)-nek létezik olyan $(\mathbf{d}, \sigma)$ megengedett megoldása, amelyben $\sigma>0$.

Ekkor elég kis pozitív α-val és alkalmas $0\leqq\theta_i\leqq 1$, $i\in\{0\}\cup I(\mathbf{x}, v)$ számokkal a gradiensvektorok folytonossága miatt teljesül

$$
\begin{aligned}
& f_i(\mathbf{x}-\alpha\mathbf{d}) \geqq 0, \; i \notin I(\mathbf{x}, v), \\
& f_i(\mathbf{x}-\alpha\mathbf{d}) = f_i(\mathbf{x})-\alpha\left(\nabla f_i(\mathbf{x}-\theta_i\alpha\mathbf{d}), \mathbf{d}\right) \geqq f_i(\mathbf{x}), \quad i\in I(\mathbf{x}, v), \\
& f_0(\mathbf{x}-\alpha\mathbf{d}) = f_0(\mathbf{x})-\alpha\left(\nabla f_0(\mathbf{x}-\theta_0\alpha\mathbf{d}), \mathbf{d}\right) < f_0(\mathbf{x}),
\end{aligned} \tag{2.5}
$$

tehát $\mathbf{x}-\alpha\mathbf{d}$ (2.1) megengedett, $\mathbf{x}$-nél jobb megoldása, ami ellentmond $\mathbf{x}$ lokális minimumhely voltának.

Az alábbiakban a 2.1. lemma felhasználásával megmutatjuk, hogy a (2.4) feladatban szereplő gradienseket sztochasztikus gradienssel becsülve kapott iránykereső feladat is alkalmas hatékony irány kiválasztására: ennek igazolása BOGOMOLOV és KARMANOV konvergencia-bizonyításától független.

Legyen $\tilde{\mathbf{d}}_j$, $j=1, \ldots, n$ az R^n-beli egységgömb felületén egyenletes eloszlású valószínűségi változó n számú független realizációja (vagyis n számú véletlen irány). Ezekből 1 valószínűséggel képezhető $\mathbf{d}_j$, $j=1, \ldots, n$ sztochasztikus ortonormált vektorrendszer *(Gram-Schmidt ortogonalizációs eljárással)*.

Mivel adott $\mathbf{d}_j$, $j=1, \ldots, n$ ortonormált bázis esetén tetszőleges folytonosan differenciálható f függvényre igaz

$$
\nabla f(\mathbf{x}) = \sum_{j=1}^{n} \left(\nabla f(\mathbf{x}), \mathbf{d}_j\right)\mathbf{d}_j, \tag{2.6}
$$

tehát a

$$
\mathbf{g} = \mathbf{g}(\mathbf{x},\alpha) = \sum_{j=1}^{n} \frac{f(\mathbf{x}+\alpha\mathbf{d}_j)-f(\mathbf{x})}{\alpha}\,\mathbf{d}_j \tag{2.7}
$$

jelöléssel fennáll az alábbi határérték-reláció:

$$
\lim_{\alpha\to 0} \mathbf{g}(\mathbf{x},\alpha) = \sum_{j=1}^{n} \left(\nabla f(\mathbf{x}), \mathbf{d}_j\right)\mathbf{d}_j = \nabla f(\mathbf{x}). \tag{2.8}
$$

Tekintsük ezért a következő (a $\mathbf{d}$, σ változókat tartalmazó) feladatot a $\mathbf{g}_i=\mathbf{g}_i(\mathbf{x}, \alpha)$, $i=0, 1, \ldots, m$ vektorokkal, ahol $\mathbf{g}_i$ a $\nabla f_i(\mathbf{x})$ vektor becslése, $\alpha>0$ paraméter,

$\mathbf{x}$ pedig rögzített, a (2.1) feladat iteratív megoldása során talált legjobb vektor:

$$\begin{gathered} \max \sigma \\ (\mathbf{g}_i, \mathbf{d}) + \sigma \leqq 0, \quad i \in I(\mathbf{x}, v), \quad v > 0 \\ -(\mathbf{g}_0, \mathbf{d}) + \sigma \leqq 0, \\ \|\mathbf{d}\| \leqq 1. \end{gathered} \tag{2.9}$$

A (2.9) feladatra vonatkozóan fennáll a

2.2. LEMMA: Tegyük fel, hogy az f_i, $i=0, 1, \ldots, m$ függvények folytonos differenciálhatósága mellett a következő feltételek is teljesülnek:

1. A (2.9) feladat célfüggvénye és a feltételi függvények a $\mathbf{d}, \sigma, \alpha$ paraméterek kétszer folytonosan differenciálható függvényei a (2.4) feladat $(\bar{\mathbf{d}}, \bar{\sigma})$ megoldásának egy környezetében.
2. A (2.9) feladat $\alpha=0$ melletti $(\bar{\mathbf{d}}, \bar{\sigma})$ optimális megoldáspontjában az aktív feltételi függvények gradiensvektorai lineárisan függetlenek.
3. A $(\bar{\mathbf{d}}, \bar{\sigma})$ pontban teljesülnek a szigorú komplementaritási feltételek, valamint a másodrendű elégséges feltétel (l. [7]).

Ekkor, ha $\mathbf{x}$ a (2.1) feladat lokális minimuma, akkor az elég kis pozitív α-val konstruált (2.9) feladat bármely $(\mathbf{d}, \sigma)$ megengedett megoldásában $\sigma<\tau$ tetszőleges $\tau>0$ mellett.

Másrészt, ha $\mathbf{x}$ nem lokális minimumhely, amelyhez létezik olyan $-\mathbf{d}$ hatékony irány és σ pozitív szám, hogy fennáll

$$\begin{gathered} -(\nabla f_0(\mathbf{x}), \mathbf{d}) + \sigma \leqq 0, \\ (\nabla f_i(\mathbf{x}), \mathbf{d}) + \sigma \leqq 0, \quad i \in I(\mathbf{x}, 0) \end{gathered} \tag{2.10}$$

akkor elég kis pozitív α mellett (2.9) $(\bar{\mathbf{d}}(\alpha), \bar{\sigma}(\alpha))$ optimális megoldásában $\bar{\sigma}(\alpha) > \frac{1}{2}\sigma$.

Bizonyítás: A lemma 1—3. feltételeinek fennállása esetén alkalmazható ARMACOST és FIACCO [2] általános nemlineáris programozási feladat érzékenységi vizsgálatára vonatkozó eredménye, amely szerint elég kis α-ra a (2.9) feladat $(\bar{\mathbf{d}}(\alpha), \bar{\sigma}(\alpha))$ optimális megoldásainak sorozata $\alpha \to 0$ esetén folytonosan differenciálható görbén tart a $(\bar{\mathbf{d}}, \bar{\sigma}) = (\bar{\mathbf{d}}(0), \bar{\sigma}(0))$ megoldáshoz.

Indirekt módon okoskodva, legyen $\mathbf{x}$ (2.1) lokális minimumhelye, továbbá $\tau>0$ tetszőleges érték, amelyhez létezik (2.9)-nek $(\mathbf{d}(\alpha), \sigma(\alpha))$ $\sigma(\alpha) \geqq \tau$ megoldása minden pozitív α mellett.

Mivel $i \in I(\mathbf{x}, v) \cup \{0\}$ esetén $\lim_{\alpha \to 0} \mathbf{g}_i(\mathbf{x}, \alpha) = \nabla f_i(\mathbf{x})$, [2] idézett eredménye szerint az optimális $(\bar{\mathbf{d}}(\alpha), \bar{\sigma}(\alpha))$ megoldás $\alpha \to 0$ esetén tart $(\bar{\mathbf{d}}, \bar{\sigma})$-hoz, ahol $\bar{\sigma} \geqq \tau > 0$. Így a 2.1. lemma értelmében ellentmondásra jutottunk.

A lemma második állítása hasonlóan igazolható.

Megjegyzések

1. Ahhoz, hogy a 2.2. lemma 1—2. feltételei teljesüljenek, elegendő a $\mathbf{g}_i(\mathbf{x}, \alpha)$ függvények α szerinti kétszeri folytonos differenciálhatóságát, továbbá a $\nabla f_i(\mathbf{x})$, $i \in I = I(\mathbf{x}, 0) \cup \{0\}$ vektorok lineáris függetlenségét megkövetelni. Emellett a másodrendű elégséges feltétel teljesülése könnyen igazolható. Írjuk fel ugyanis a (2.4) feladat *Lagrange-függvényét* (a $\|\mathbf{d}\| \leqq 1$ feltételt $\|\mathbf{d}\|^2 \leqq 1$-gyel helyettesítve):

$$L(\mathbf{d}, \sigma, u_i, u) = \sigma - \sum_{i \in I(\mathbf{x}, v)} u_i\big((\nabla f_i, \mathbf{d}) + \sigma\big) - u_0\big(-(\nabla f_0, \mathbf{d}) + \sigma\big) - u\big((\mathbf{d}, \mathbf{d}) - 1\big)$$
$$u_i \geqq 0, \quad i \in I, \quad u \geqq 0. \tag{2.11}$$

Az optimalitás elsőrendű feltételei a következők:

$$u_0 \nabla f_0 - \sum_{i \in I^*} u_i \nabla f_i - 2u\mathbf{d} = 0, \quad I^* = I(\mathbf{x}, 0)$$
$$1 - \sum_{i \in I^* \cup \{0\}} u_i = 0,$$
$$u_i\big((\nabla f_i, \mathbf{d}) + \sigma\big) = 0, \quad i \in I(\mathbf{x}, v), \tag{2.12}$$
$$u_0\big((-\nabla f_0, \mathbf{d}) + \sigma\big) = 0,$$
$$u\big((\mathbf{d}, \mathbf{d}) - 1\big) = 0,$$
$$u_i \geqq 0, \quad i \in I, \quad u \geqq 0.$$

A ∇f_i, $i \in I^* \cup \{0\}$ vektorok lineáris függetlensége és (2.12) második egyenlősége miatt $u > 0$ (tehát a kvadratikus feltételben teljesül a szigorú komplementaritás). Másrészt tetszőleges $\mathbf{y} \in R^{n+1}$, $\mathbf{y} \neq \mathbf{0}$, és az $\mathbf{y}^T(\nabla f_i, 1) = 0$, $i \in I(\mathbf{x}, v)$, $\mathbf{y}^T(-\nabla f_0, 1) = 0$ és $\mathbf{y}^T(2\mathbf{d}, 0) = 0$ skalárszorzat-feltételek szerinti vektor esetén nyilvánvalóan fennáll

$$\left(\nabla^2_{d,\sigma} L(\mathbf{d}, \sigma, u_i, u)\mathbf{y}, \mathbf{y}\right) = -2u \sum_{i=1}^{n} y_i^2 < 0, \tag{2.13}$$

ami éppen a másodrendű elégséges feltétel.

2. A 2.2. lemma szerint a $\mathbf{g}_0(\mathbf{x}, \alpha)$, $\mathbf{g}_i(\mathbf{x}, \alpha)$, $i \in I(\mathbf{x}, v)$ vektorokkal képezett (2.9) iránykeresési probléma is alkalmas hatékony irány kiválasztására, megfelelően kis α esetén. Ez a tény a (2.1) matematikai programozási feladatnak gradiensbecsléseket felhasználó módszerekkel való megoldási lehetőségére utal.

3. A (2.9) feladattal kapcsolatban végül megjegyezzük, hogy az egyszerű transzformációval kvadratikus programozási problémává alakítható át (l. pl. [29], 123—127.). Bevezetve a $\xi = -\sigma$, $\mathbf{h}_0 = -\mathbf{g}_0$, $\mathbf{h}_i = \mathbf{g}_i$, $i = 1, \ldots, m_0$ jelöléseket, (2.9) a $(\mathbf{d}, \xi) \in R^{n+1}$ vektorváltozóra vonatkozó

$$\min \xi$$
$$(\mathbf{h}_i, \mathbf{d}) \leqq \xi, \quad i = 0, 1, \ldots, m_0, \tag{2.14}$$
$$\|\mathbf{d}\| \leqq 1$$

feladatként írható fel. (A jelölések egyszerűsítése kedvéért itt feltesszük, hogy a $\mathbf{h}_i$, $i = 0, 1, \ldots, m_0$ vektorok a $\mathbf{h}_i$, $i \in I(\mathbf{x}, v) \cup \{0\}$ vektorok maximális lineárisan

független alrendszerét képezik: ez az i indexeket alkalmasan felcserélve mindig elérhető.)

Mivel **d**-re vonatkozóan a normáló feltételen kívül csak a $\mathbf{h}_i$, $i=0, 1, \ldots, m_0$ vektorokkal képezett skalárszorzat-feltételek szerepelnek, (vagyis **d**-nek a $\mathbf{h}_i$, $i=0, 1, \ldots, m_0$ vektorok által kifeszített altér R^n-beli kiegészítő alterére vonatkozó vetülete a normafeltétel kivételével irreleváns) célszerű a **d** vektort

$$\mathbf{d} = \sum_{j=0}^{m_0} \xi_j \mathbf{h}_j \tag{2.15}$$

alakban keresni.

Bevezetve a $h_{ij}=(\mathbf{h}_i, \mathbf{h}_j)$ jelölést, (2.14)-ből adódik a fenti ξ_j, $j=0, 1, \ldots, m_0$ és $\xi_{m_0+1}=\xi$ változókra vonatkozó alábbi probléma:

$$\min \xi_{m_0+1}$$

$$\sum_{j=0}^{m_0} h_{ij}\xi_j \leqq \xi_{m_0+1}, \quad i = 0, 1, \ldots, m_0 \tag{2.16}$$

$$\sum_{i,j=0}^{m_0} h_{ij}\xi_i\xi_j \leqq 1.$$

A (2.16) feladat *Lagrange-függvénye:*

$$F(\xi_j, \zeta_i, \ i, j = 0, 1, \ldots, m_0+1) = \tag{2.17}$$

$$= \xi_{m_0+1} + \sum_{i=0}^{m_0} \zeta_i \left(\sum_{j=0}^{m_0} h_{ij}\xi_j - \xi_{m_0+1} \right) + \zeta_{m_0+1} \left(\sum_{i,j=0}^{m_0} h_{ij}\xi_i\xi_j - 1 \right).$$

A *Lagrange-függvény* nyeregpontját a

$$\max_{\substack{\zeta_i \geqq 0 \\ i=0,1,\ldots,m_0+1}} \min_{\substack{\xi_j \\ j=0,1,\ldots,m_0+1}} F(\xi_j, \zeta_i) \tag{2.18}$$

feladat megoldása szolgáltatja. Felírva a

$$\frac{\partial F}{\partial \xi_j} = 0, \quad j = 0, 1, \ldots, m_0+1 \tag{2.19}$$

relációkat, adódik

$$\sum_{i=0}^{m_0} h_{ij}(\zeta_i + 2\zeta_{m_0+1}\xi_i) = 0, \quad j = 0, 1, \ldots, m_0,$$

$$\sum_{i=0}^{m_0} \zeta_i = 1. \tag{2.20}$$

A (2.20) relációkat együttesen figyelembe véve és felhasználva azt, hogy a $\mathbf{H}=(h_{ij})$ mátrix nem-szinguláris, adódik $\zeta_{m_0+1} \neq 0$, továbbá

$$\xi_i = -\frac{\zeta_i}{2\zeta_{m_0+1}}, \quad i = 0, 1, \ldots, m_0. \tag{2.21}$$

Így (2.18) a

$$-\min_{\substack{\zeta_i \geqq 0 \\ i=0,1,\ldots,m_0+1}} \left(\frac{1}{4\zeta_{m_0+1}} \sum_{i,j=0}^{m_0} h_{ij}\zeta_i\zeta_j + \zeta_{m_0+1} \right) \tag{2.22}$$

alakban írható fel. Itt a pozitív ζ_{m_0+1} szerint differenciálva adódik

$$\zeta_{m_0+1} = \frac{\left(\sum_{i,j=0}^{m_0} h_{ij}\zeta_i\zeta_j \right)^{\frac{1}{2}}}{2}, \tag{2.23}$$

tehát a következő, a (2.9) feladattal ekvivalens kvadratikus feladatot nyerjük a ζ_i változókra vonatkozóan:

$$\begin{aligned} &\min \sum_{i,j=0}^{m_0} h_{ij}\zeta_i\zeta_j \\ &\sum_{i=0}^{m_0} \zeta_i = 1 \\ &\zeta_i \geqq 0, \quad i = 0, 1, \ldots, m_0. \end{aligned} \tag{2.24}$$

A (2.24) problémát megoldva, a ζ vektor ismeretében a (2.9) feladat megoldását

$$\mathbf{d} = -\sum_{j=0}^{m_0} \frac{\zeta_j}{2\zeta_{m_0+1}} \mathbf{h}_j \tag{2.25}$$

szolgáltatja (a megfelelő σ érték $\mathbf{d}$-nek a (2.9) feltételeibe való behelyettesítésével egyszerűen meghatározható).

A (2.1) matematikai programozási problémára, illetve az azzal kapcsolatos (2.4) feladatra visszatérve definiáljuk az

$$X^* = \{\mathbf{x}^* : \mathbf{x}^* \in X,\ f_0(\mathbf{x}^*) \leqq f_0(\mathbf{x}_0),\ \bar{\sigma}(\mathbf{x}^*, 0) = 0\} \tag{2.26}$$

halmazt, amely a (2.4) feladatnak az algoritmus kiindulópontjánál kisebb függvényértékű stacionárius pontjait tartalmazza. Legyen továbbá

$$\varrho(\mathbf{x}, X^*) = \inf_{\mathbf{x}^* \in X^*} \|\mathbf{x} - \mathbf{x}^*\|$$

tetszőleges $\mathbf{x}$ pontnak az X^* halmaztól mért távolsága.

BOGOMOLOV és KARMANOV igazolták, hogy a (2.2)—(2.3) típusú véletlen kereső séma 1 valószínűséggel az X^* halmaz valamelyik pontjához konvergál, azaz

$$P\left(\lim_{k \to \infty} \varrho(\mathbf{x}_k, X^*) = 0\right) = 1. \tag{2.27}$$

Megjegyezzük, hogy az X^* halmaz a 2.1. lemma értelmében a (2.1) feladat összes lokális minimumhelyét tartalmazza.

A bizonyításhoz a következő feltételek teljesülte szükséges:

$$X_0 = \{\mathbf{x} : \mathbf{x} \in X,\ f_0(\mathbf{x}) \leqq f_0(\mathbf{x}_0)\} \quad \text{korlátos, zárt halmaz.} \tag{2.28}$$

Létezik olyan $L>0$ konstans, hogy f_i, $i=0, 1, \ldots, m$ és $[\mathbf{x}, \mathbf{y}] \subset X$ esetén fennáll

$$\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\| \leqq L\|\mathbf{x}-\mathbf{y}\|. \tag{2.29}$$

Az alábbiakban a (2.27) reláció igazolásának gondolatmenetét ismertetjük; az itt nem részletezett lemmák bizonyítását az idézett [28] dolgozat tartalmazza.

Tegyük fel, hogy az $\mathbf{x}_k$ pontban létezik a (2.4) feladatnak olyan $\sigma=\sigma_k$ és $\mathbf{d}_k \in R^n$ megoldása, ahol $\sigma>0$. Ekkor fennáll az alábbi két lemma:

2.3. LEMMA: a $-\mathbf{d}_k$ irányú lépések $\bar{\alpha}_k$ maximális lehetséges hosszára teljesül

$$\bar{\alpha}_k \geqq \min\left\{\frac{v}{M}, \frac{\sigma}{L}\right\}, \text{ ahol } M = \max_{i=1,\ldots,m} \max_{\mathbf{x}\in X_0} \|\nabla f_i(\mathbf{x})\|, \tag{2.30}$$

$v>0$ pedig a (2.4) feladatban szereplő paraméter.

Ez azt jelenti, hogy $v>0$ esetén pozitív hosszúságú megengedett lépés tehető $-\mathbf{d}_k$ irányban. A következő lemma azt mutatja meg, hogy az $\mathbf{x}_k$-ról $\mathbf{x}_{k+1}$-re való áttéréskor a célfüggvényérték szigorúan csökken.

2.4. LEMMA:

$$f_0(\mathbf{x}_k) - f_0(\mathbf{x}_{k+1}) \geqq \frac{1}{2}(1-\varepsilon_k)\sigma \min\left\{\bar{\alpha}_k, \frac{\sigma}{L}\right\} \geqq \tag{2.31}$$

$$\geqq \frac{1}{2}(1-\varepsilon)\sigma \min\left\{\frac{v}{M}, \frac{\sigma}{L}\right\} > 0.$$

Itt ε_k és ε a (2.3)-ban definiált értékek.

A 2.3. és 2.4. lemmák bizonyítása az f_i, $i=0, 1, \ldots, m$ függvények folytonosan differenciálhatóságából következik. A következő két lemma igazolása szintén ezen alapszik. Ezek a stacionárius pontok meghatározására definiált feladatokhoz konstruálható pozitív v paraméterű (2.4) segédfeladat létezésére és ennek megoldására vonatkoznak. Az előző két lemma értelmében viszont a segédfeladatból adódik tényleges függvényérték-csökkenést eredményező $\mathbf{x}_{k+1}$ vektor.

2.5. LEMMA: Bármely $\bar{\mathbf{x}} \in X$-hez található $\bar{v}=\bar{v}(\bar{\mathbf{x}})>0$ és $\delta=\delta(\bar{\mathbf{x}})>0$, amelyekkel $v\in[0, \bar{v}]$, $\mathbf{x}\in U_\delta(\bar{\mathbf{x}})=\{\mathbf{x}: \mathbf{x}\in X, \|\mathbf{x}-\bar{\mathbf{x}}\|<\delta\}$ esetén

$$I(\mathbf{x}, v) \subset I(\bar{\mathbf{x}}, 0). \tag{2.32}$$

2.6. LEMMA: Jelölje $(\bar{\mathbf{d}}, \bar{\sigma})$ a (2.4) iránykereső feladat optimális megoldását $\mathbf{x}=\bar{\mathbf{x}}$, $I=I(\bar{\mathbf{x}}, 0)$ mellett, amelyben $\bar{\sigma}>0$.

Ekkor létezik $\bar{v}>0$, $\delta>0$ és $a\in(0, 1)$, hogy tetszőleges

$$\|\mathbf{d}\| = 1, \quad (\mathbf{d}, \bar{\mathbf{d}}) \geqq 1-a, \quad v\in[0, \bar{v}], \quad \mathbf{x}\in U_\delta(\bar{\mathbf{x}}) \tag{2.33}$$

esetén a (2.4) iránykereső feladatnak $\mathbf{d}$ és $\sigma=\frac{1}{4}\bar{\sigma}$ megengedett megoldása.

Megjegyezzük, hogy a 2.2. lemma feltételei mellett a 2.6. lemma megfelelője igaz marad pl. $\sigma=\frac{1}{8}\bar{\sigma}$ esetén akkor is, ha a (2.4)-et közelítő (2.9) feladatot vizsgál-

juk. Ehhez a gradiensek folytonossága miatt csak α-t kell megfelelően kicsire választanunk (l. a 2.2. lemma bizonyítását).

Jelölje $A_k = A_k(\bar{\mathbf{d}}, a)$ azt az eseményt, hogy az $\mathbf{x}=\mathbf{x}_k$-hoz definiált (2.4) feladat $\bar{\mathbf{d}}$ megoldásához közeli $\mathbf{d}_k (\|\mathbf{d}_k\| = 1)$ véletlen irányt választunk ki a k-adik lépésben, azaz $\mathbf{d}_k$-ra teljesül $(\bar{\mathbf{d}}, \mathbf{d}_k) \geqq 1 - a$ ($a \in (0, 1)$ rögzített érték).

Ha a szokásos módon $\bar{A}_k$ az A_k esemény komplementerét, $P(B|C)$ pedig B-nek a C feltétel melletti valószínűségét jelöli, akkor alkalmas $p = p(\bar{\mathbf{d}}, a) > 0$ konstanssal megköveteljük a

$$P(A_k | \bar{A}_{k-1}, \bar{A}_{k-2}, \ldots, \bar{A}_0) \geqq p, \quad k = 1, 2, 3, \ldots \tag{2.34}$$

relációk fennállását.

A (2.2)—(2.3) algoritmus pl. akkor teljesíti a (2.34) feltételt, ha a $\mathbf{d}$ irány kiválasztása minden lépésben a korábbi lépésektől független, az n-dimenziós egységgömb felületén egyenletes eloszlású véletlen vektor előállításával történik. Ez esetben ugyanis

$$P(A_k | \bar{A}_{k-1}, \bar{A}_{k-2}, \ldots, \bar{A}_0) = P(A_k) > 0,$$

mivel $a > 0$, és így a $(\bar{\mathbf{d}}, \mathbf{d}_k) \geqq 1 - a$ feltétel $\bar{\mathbf{d}}$ pozitív valószínűségű kúpkörnyezetébe eső véletlen vektor kiválasztását írja elő. Természetesen más, a (2.34) tulajdonsággal rendelkező véletlen kereső eljárások is definiálhatók, amelyekkel a konvergencia sebessége ehhez az egyszerű irányválasztáson alapuló algoritmushoz képest várhatóan növekszik.

Elemi valószínűségszámítási meggondolásokkal igazolható a

2.7. LEMMA: A (2.34) feltétel fennállása esetén $P\left(\bigcup_{k=0}^{\infty} A_k\right) = 1$, tehát az A_k események valamelyike 1 valószínűséggel bekövetkezik.

A 2.1—2.7. lemmák segítségével a (2.2)—(2.3) elméleti véletlen kereső eljárás konvergenciája bizonyítható. Pontosabban, fennáll a következő

2.1. TÉTEL: Ha a $k = 0, 1, 2, \ldots$ indexek tetszőleges $\{k_i\}$ részsorozatára teljesül $P(A_{k_i} | \bar{A}_{k_i - 1}, \bar{A}_{k_i - 2}, \ldots, \bar{A}_{k_0}) \geqq p > 0$, akkor igaz a (2.27) reláció, vagyis az $\mathbf{x}_k$ sorozat bármely határpontja a (2.1) feladat lokális minimumhelyeit tartalmazó X^* halmazba esik.

Bizonyítás: Nyilvánvaló, hogy a (2.34)-gyel kapcsolatos korábbi megjegyzésünk a lépésenkénti független irányok felhasználása miatt igaz marad tetszőleges $\{k_i\}$ index-részsorozat esetén is.

Tegyük fel indirekt módon, hogy (2.27) nem teljesül, tehát az $\{\mathbf{x}_k\}$ sorozat valamely $\{\mathbf{x}_{k_i}\}$ részsorozatának $\bar{\mathbf{x}}$ határpontjában a (2.4) iránykereső feladat megoldása $I(\bar{\mathbf{x}}, 0)$ mellett olyan $(\bar{\mathbf{d}}, \bar{\sigma})$ pár, amelyben $\bar{\sigma} > 0$.

A 2.5. és 2.6. lemmából következik, hogy ekkor létezik olyan $\bar{\nu} > 0$, $\delta > 0$ és $a \in (0, 1)$, hogy $\nu \in [0, \bar{\nu}]$, $\mathbf{x} \in U_\delta(\bar{\mathbf{x}})$ és $\|\mathbf{d}\| = 1$, $(\mathbf{d}, \bar{\mathbf{d}}) \geqq 1 - a$ esetén a (2.4) feladatnak $\mathbf{d}$ és $\sigma = \frac{\bar{\sigma}}{4} > 0$ megengedett megoldása lesz. A 2.3. és 2.4. lemmát felhasználva, a (2.31) reláció értelmében létezik tehát olyan $c > 0$ konstans, amelyre igaz $\mathbf{x}_{k_i} \in U_\delta(\bar{\mathbf{x}})$ esetén, tehát $i \geqq i_1$-től kezdve

$$f_0(\mathbf{x}_{k_i}) - f_0(\mathbf{x}_{k_i + 1}) \geqq c, \tag{2.35}$$

valahányszor egy $A_{k_i}(\bar{\mathbf{d}}, a)$ esemény bekövetkezik. Az A_{k_i}-k valamelyike a 2.7. lemma értelmében 1 valószínűséggel bekövetkezik, tehát (2.35) is 1 valószínűséggel igaz valamely i-re, $\{\mathbf{x}_{k_i}\}$, $i \geqq i_1$ tetszőleges részsorozata esetén.

Vegyük figyelembe másrészt, hogy az $\mathbf{x}_{k_i}$ vektorok és természetesen az $f_0(\mathbf{x}_{k_i})$ függvényértékek sorozata is konvergens. Létezik tehát olyan i_0 index, hogy $i \geqq i_0$ esetén fennáll

$$f_0(\mathbf{x}_{k_i}) - f_0(\mathbf{x}_{k_{i+1}}) < c, \tag{2.36}$$

ami a (2.35) 1 valószínűségű relációnak ellentmond.

Megjegyezzük, hogy a $\bar{\mathbf{d}}$ irányhoz közeli $\mathbf{d}_k$ irány kiválasztása akkor is lehetséges, ha $\mathbf{d}_k$-t az n-dimenziós egységgömbön egyenletes eloszlású $\mathbf{v}$ vektor, valamint a $\mathbf{w}$ egységnyi „emlékezetvektor" felhasználásával a

$$\mathbf{d}_k = \frac{\mathbf{v} + \varkappa_k \mathbf{w}}{\|\mathbf{v} + \varkappa_k \mathbf{w}\|}, \quad 0 \leqq \varkappa_k \leqq \varkappa < 1 \tag{2.37}$$

kifejezéssel állítjuk elő (l. [28]).

A korábbi lépések tapasztalatainak valamilyen figyelembevétele gyakran előnyösnek bizonyul (pl. a [30] munka tartalmaz erre vonatkozó számítási tapasztalatokat).

Konvergencia folytonos, feltétel nélküli feladat esetén

A fenti gondolatmenet során lényegesen kihasználtuk az f_i, $i=0, \ldots, m$ függvények gradienseinek folytonosságát. Ebben az esetben a (2.2)—(2.3) elméleti véletlen kereső algoritmus alkalmazható a (2.1) matematikai programozási feladat megoldására.

A továbbiakban véletlen kereső eljárások konvergenciáját arra az esetre vonatkozóan vizsgáljuk, ha az f feltétel nélküli célfüggvényről csak folytonosság tételezhető fel. Ilyen feladatra vezet pl. a folytonos f_i, $i=0, 1, \ldots, m$ függvényekkel konstruált (2.1) matematikai programozási problémából képezhető alábbi feltétel nélküli, büntetőtagokkal kiegészített célfüggvényű feladatok sorozata (a módszert általánosan [7] tárgyalja):

$$\min_{\mathbf{x} \in R^n} Q^{(j)}(\mathbf{x}, q^{(j)}) =$$

$$= \min_{\mathbf{x} \in R^n} \left\{ f_0(\mathbf{x}) + \sum_{i=1}^{m} q_i^{(j)} [\min(0, f_i(\mathbf{x}))]^2 \right\}, \tag{2.38}$$

ahol $0 < q_i^{(j)} \leqq q_i^{(j+1)}$, $j = 0, 1, 2, \ldots$, $\lim_{j \to \infty} q_i^{(j)} = +\infty$ minden $i = 1, \ldots, m$ esetén.

Folytonos $f(\mathbf{x})$ célfüggvényre vonatkozó vizsgálataink során a következő feltételek fennállását írjuk elő:

1. Létezik olyan $t>0$, amelyre az $S=\{\mathbf{x}: f(\mathbf{x}) \leqq t\}$ nívóhalmaz kompakt.
2. Az algoritmus $\mathbf{x}_0$ kiindulópontja eleme S-nek, és a k-adik lépésbeli $\mathbf{x}_k$ pontról

$\mathbf{x}_{k+1}$-re úgy térünk át, hogy az algoritmus által kiválasztott $\mathbf{d}_k$ egységvektor irányú Δ_k hosszúságú szakaszon egzakt keresést hajtunk végre, tehát

$$\mathbf{x}_{k+1} = \arg \min_{\alpha \in [0, \Delta_k]} f(\mathbf{x}_k + \alpha \mathbf{d}_k), \quad k = 0, 1, 2, 3, \ldots . \tag{2.39}$$

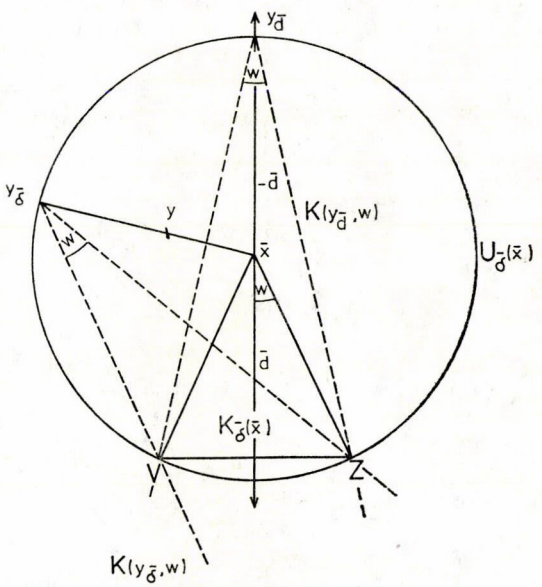
Itt $\Delta_k \geqq \Delta$ minden k-ra, $\Delta > 0$ pedig rögzített konstans.

3. Ha $\bar{\mathbf{x}}$ f-nek nem lokális minimumhelye, akkor léteznek olyan $\bar{\delta} = \bar{\delta}(\bar{\mathbf{x}}) > 0$ és $\bar{a} = \bar{a}(\bar{\mathbf{x}}) \in (0, 1)$ számok, valamint $\bar{\mathbf{d}} = \bar{\mathbf{d}}(\bar{\mathbf{x}}) \in R^n$, $\|\bar{\mathbf{d}}\| = 1$ irány, amelyekkel $0 < \|\mathbf{y} - \bar{\mathbf{x}}\| \leqq \bar{\delta}$ és $\left(\frac{\mathbf{y} - \bar{\mathbf{x}}}{\|\mathbf{y} - \bar{\mathbf{x}}\|}, \bar{\mathbf{d}}\right) \geqq 1 - \bar{a}$ esetén teljesül $f(\mathbf{y}) < f(\bar{\mathbf{x}})$.

2.2. TÉTEL: Az 1—3. feltételek fennállása esetén a véletlen kereső eljárással nyert $\{\mathbf{x}_k\}$ sorozat minden $\{\mathbf{x}_{k_i}\}$ konvergens részsorozata 1 valószínűséggel $f(\mathbf{x})$ egy S-beli lokális minimumhelyéhez konvergál.

Bizonyítás: Ismét indirekt módon okoskodunk. Az $\{\mathbf{x}_k\} \subset S$ sorozatból S kompaktsága miatt kiválasztható valamely, $\bar{\mathbf{x}} \in S$ határponttal bíró $\{\mathbf{x}_{k_i}\}$ konvergens részsorozat. Mivel f folytonos, az $f(\mathbf{x}_{k_i})$ függvényértékek sorozatának határértéke $f(\bar{\mathbf{x}})$. Tegyük fel, hogy $\bar{\mathbf{x}}$ $f(\mathbf{x})$-nek nem lokális minimumhelye. Ekkor a 3. feltétel szerint létezik olyan $\bar{\delta} > 0$, $\bar{a} \in (0, 1)$ és $\bar{\mathbf{d}}$, $\|\bar{\mathbf{d}}\| = 1$, hogy $0 < \|\mathbf{y} - \bar{\mathbf{x}}\| \leqq \bar{\delta}$ és $\left(\frac{\mathbf{y} - \bar{\mathbf{x}}}{\|\mathbf{y} - \bar{\mathbf{x}}\|}, \bar{\mathbf{d}}\right) \geqq 1 - \bar{a}$ esetén $f(\mathbf{y}) < f(\bar{\mathbf{x}})$ teljesül.

Tekintsük az $\bar{\mathbf{x}}$ csúcspontú, $\bar{\mathbf{d}}$ tengelyirányú, $\arccos(1 - \bar{a}) = w$ félnyílásszögű n-dimenziós K forgáskúpot. A 3. feltétel következtében a $K_{\bar{\delta}}(\bar{\mathbf{x}}) = \{\mathbf{y} : \mathbf{y} \in K,\ 0 < \|\mathbf{y} - \bar{\mathbf{x}}\| \leqq \bar{\delta}\}$ halmaz pontjaiban fennáll $f(\mathbf{y}) < f(\bar{\mathbf{x}})$ (l. az 1. ábrát).



1. ábra

Megmutatjuk, hogy a 2. feltétel szerint haladó véletlen kereső algoritmus 1 valószínűséggel vizsgál $K_\delta(\bar{\mathbf{x}})$-ba eső pontot, tehát talál $f(\bar{\mathbf{x}})$-nál kisebb függvényértékű pontot is.

Világos, hogy tetszőlegesen kicsinyre választott δ esetén az $\{\mathbf{x}_{k_i}\}$ sorozat elég nagy indexű tagjai az $U_\delta(\bar{\mathbf{x}})=\{\mathbf{x}: \|\mathbf{x}-\bar{\mathbf{x}}\| \leqq \delta\}$ halmazba esnek. Másrészt $\delta \leqq \frac{\Delta}{2}$ mellett tetszőleges $\mathbf{x}_{k_i} \in U_\delta(\bar{\mathbf{x}})$-ból kiinduló, $\Delta_{k_i} \geqq \Delta$ hosszúságú szakaszon végrehajtott egzakt kereső eljárás találhat az $\bar{\mathbf{x}}$-nál jobb pontot. Ennek elégséges feltétele, hogy az $\mathbf{x}_{k_i}$-ben kiválasztott kereső irány a $K_\delta(\bar{\mathbf{x}})$ halmazon haladjon át. Jelölje A_{k_i} az $\mathbf{x}_{k_i}$ pontban a $K_\delta(\bar{\mathbf{x}})$ halmazt metsző irány véletlen kiválasztásának eseményét. Bebizonyítjuk, hogy az n-dimenziós tér egységgömbjén egyenletes eloszlású, lépésenként függetlenül generált véletlen irányt felhasználó kereső stratégia esetén teljesül $P(\mathbf{x}_{k_i})=P(A_{k_i})=P(A_{k_i}|\bar{A}_{k_{i-1}}, \ldots, \bar{A}_{k_1}) \geqq p>0$, ugyanis az $\mathbf{y} \in U_\delta(\bar{\mathbf{x}})$ pontokra analóg módon definiált $P(\mathbf{y})$ valószínűségek p infinuma pozitív.

Jelölje $\mathbf{y}_d$ az $\bar{\mathbf{x}}$-ból kiinduló, $-\mathbf{d}$ irányú félegyenesnek az $U_\delta(\bar{\mathbf{x}})$ határával való metszéspontját (l. az 1. ábrát). Érvényes a következő

2.8. LEMMA: A $K_\delta(\bar{\mathbf{x}})$ halmazt metsző irány kiválasztásának valószínűsége az $U_\delta(\bar{\mathbf{x}})$ pontjait tekintve az $\mathbf{y}_d$ pontban minimális, továbbá ez a valószínűség pozitív.

Bizonyítás: Legyen $\mathbf{y}$ az $U_\delta(\bar{\mathbf{x}})-K_\delta(\bar{\mathbf{x}})$ halmaz belsejének tetszőleges pontja. (Nyilvánvalóan nem kell foglalkoznunk a $K_\delta(\bar{\mathbf{x}})$ belsejébe eső pontokkal, ahol a lemma állításában szereplő valószínűség 1. Egyszerűség kedvéért nem foglalkozunk $K_\delta(\bar{\mathbf{x}})$ határpontjaival sem, hiszen úgyis „csak" 1 valószínűségű konvergenciát akarunk igazolni, a határpontok pedig nullmértékű n-dimenziós halmazt alkotnak.) Jelölje $\mathbf{y}_\delta$ az $\bar{\mathbf{x}}$-ból kiinduló, $\mathbf{y}-\bar{\mathbf{x}}$ irányú félegyenesnek az $U_\delta(\bar{\mathbf{x}})$ halmaz határával való metszéspontját. Ekkor teljesül $P(\mathbf{y}) \geqq P(\mathbf{y}_\delta)$, hiszen ha $\mathbf{y}$ az $\bar{\mathbf{x}}$ és $\mathbf{y}_\delta$ pontokat összekötő szakasz belső pontja, akkor a $K_\delta(\bar{\mathbf{x}})$ halmaz (K-t meghatározó, $U_\delta(\bar{\mathbf{x}})$ határán végződő) alkotó szakaszai nagyobb szög alatt látszanak $\mathbf{y}$-ból, mint $\mathbf{y}_\delta$-ból, és éppen a fenti alkotókat (először) metsző irányok definiálnak $K_\delta(\bar{\mathbf{x}})$-on áthaladó irányokat. Elég ezért azt igazolni, hogy $P(\mathbf{y}_\delta) \geqq P(\mathbf{y}_d)$.

Tekintsük az $\bar{\mathbf{x}}, \mathbf{y}_\delta$ és $\mathbf{y}_d$ pontok által meghatározott síkot (l. az 1. ábrát). Ez az $U_\delta(\bar{\mathbf{x}})$ környezetből (amely n-dimenziós gömb) egy C kört metsz ki. Jelölje $\mathbf{v}$ és $\mathbf{z}$ a K kúp alkotóinak a C határvonalára eső pontjait. Az elemi geometriából ismert tétel szerint az $\mathbf{y}_\delta$ és $\mathbf{y}_d$ pontokból a $\mathbf{v}$ és $\mathbf{z}$ pontokat összekötő szakasz azonos w szög alatt látszik.

Képezzük azt a $K(\mathbf{y}_\delta, w)$ n-dimenziós kúpot, amelynek csúcsa $\mathbf{y}_\delta$, alkotói pedig a $K_\delta(\bar{\mathbf{x}})$ halmaz alkotóinak $U_\delta(\bar{\mathbf{x}})$ határával való metszéspontjain (tehát a $\frac{\mathbf{v}+\mathbf{z}}{2}$ középpontú, $\|\mathbf{v}-\mathbf{z}\|$ átmérőjű $n-1$ dimenziós G gömb határán) haladnak át. Tekintve, hogy $\frac{\Delta}{2} \geqq \delta$, a $K(\mathbf{y}_\delta, w)$-ből történő irányválasztás esetén az $\mathbf{y}_\delta$-ból kiinduló, legalább Δ hosszúságú szakaszon végrehajtott egzakt keresés nyilván eredményez $K_\delta(\bar{\mathbf{x}})$-ba eső pontot. Vegyük még figyelembe, hogy $\mathbf{y}_\delta$-ból a fenti G gömb összes átmérői legalább w szög alatt látszanak. (Ez abból következik, hogy az $\mathbf{y}_\delta$ pontot a G-t tartalmazó hipersíkra merőlegesen vetítve, a talppontból minden átmérő nagyobb szög alatt látszik, mint a $\mathbf{v}$ és $\mathbf{z}$ által meghatározott átmérő, továbbá a merőleges affinitás a látószögekre vonatkozóan rendezéstartó.) Az $\mathbf{y}_\delta$-ból $K(\mathbf{y}_\delta, w)$-be

mutató irány kiválasztásának valószínűsége tehát legalább akkora, mint az $\mathbf{y}_\delta$ pontból a G gömb felhasználásával analóg módon definiálható $K(\mathbf{y}_\delta, w)$ forgáskúpba eső irányválasztásé. Ez a valószínűség viszont megegyezik az $\mathbf{y}_\delta$-ból $K_\delta(\bar{\mathbf{x}})$-ba mutató irány kiválasztási valószínűségével és $\bar{a}>0$ miatt (l. a tétel 3. feltételét) pozitív. Ezért valóban teljesül minden $\mathbf{y}\in U_\delta(\bar{\mathbf{x}})$ esetén $P(\mathbf{y})\geqq P(\mathbf{y}_\delta)=p$, amivel a lemmát igazoltuk.

A 2.8. lemma értelmében tehát létezik olyan $p>0$ konstans, amellyel $\mathbf{x}_{k_i}\in U_\delta(\bar{\mathbf{x}})$ esetén fennáll a $P(A_{k_i}|\bar{A}_{k_{i-1}}, \ldots, \bar{A}_{k_1})\geqq p$ reláció.

A 2.7. lemma szerint viszont ekkor az A_{k_i} események valamelyike 1 valószínűséggel bekövetkezik, tehát a 2. feltétel szerinti egzakt keresés eredményeképpen 1 valószínűséggel létezik olyan k_i index, amelytől kezdve teljesül

$$f(\mathbf{x}_{k_{i+l}})\leqq f(\mathbf{x}_{k_i})<f(\bar{\mathbf{x}}),\quad l=1, 2, 3, \ldots, \tag{2.40}$$

ami ellentmond $\bar{\mathbf{x}}$ limeszpont voltának.

Megjegyzések

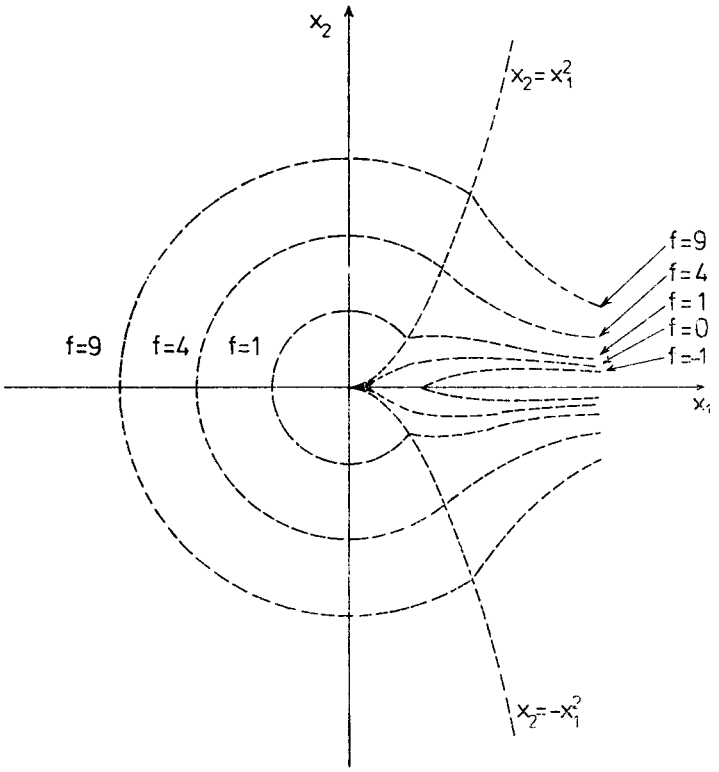
1. Egyszerűen belátható, hogy $(\mathbf{x}_{k_i}\in U_\delta(\bar{\mathbf{x}}), \delta\leqq\frac{\Delta}{2}$ mellett) a $P(A_{k_i}|\bar{A}_{k_{i-1}}, \ldots, \bar{A}_{k_1})\geqq$ $\geqq p>0$ reláció tetszőleges olyan véletlen kereső stratégia esetén teljesül, amely a lépésenkénti irányválasztásnál figyelembe veszi legalább egy, a korábbiaktól független, az n-dimenziós egységgömbön egyenletes eloszlású, véletlen próbalépés eredményét is (esetleges egyéb próbalépések mellett). Ez a tény a 2.8. lemmában szereplő egyszerű véletlen keresésnél rugalmasabb, effektívebb optimalizáló algoritmusok kialakítását teszi lehetővé.

2. A tétel 3. feltétele az f függvény bizonyos értelemben vett regularitását írja elő, nevezetesen azt, hogy f tetszőleges nem lokális minimumpontjában található olyan irány, amelynek valamely kúp környezetében az algoritmus javuló függvényértékeket eredményező pontokon haladhat tovább. Megmutatjuk, hogy ha nem követeljük meg a 2. és 3. feltételben szereplő Δ, ill. $\bar{a}$ pozitivitását (tehát az egzakt keresés lépéshossza alulról nem korlátos és csak csökkenő függvényértékű irány létezését írjuk elő nem lokális minimumpontokban), akkor megadható olyan folytonos függvény, és lépésenként független, az n-dimenziós egységgömbön egyenletes eloszlású véletlen irányokat felhasználó kereső stratégia, amely (alkalmasan választott kezdőpontból) pozitív valószínűséggel nem konvergál a függvény (lokális) minimumához.

Tekintsük ugyanis a következő kétváltozós, folytonos függvényt:

$$f(x_1, x_2)=\begin{cases}\dfrac{x_1^2+x_1^4+x_1}{x_1^2}\,|x_2|-x_1, & \text{ha}\quad x_1>0,\quad |x_2|\leqq x_1^2;\\ x_1^2+x_2^2, & \text{egyébként.}\end{cases} \tag{2.41}$$

(Ez a függvény az $f(x_1, x_2)=x_1^2+x_2^2$ forgásparaboloidból úgy származtatható, hogy azt az $x_2=x_1^2$ és $x_2=-x_1^2 (x_1\geqq 0)$ felületekkel elmetszve, a metszésvonalak $(x_1, x_2, f(x_1, x_2))$ — megfelelően $(x_1, x_1^2, x_1^2+x_1^4)$, illetve $(x_1, -x_1^2, x_1^2+x_1^4)$ — pontjait és az $(x_1, 0, -x_1)$ pontot egyenesszakaszokkal kötjük össze minden $x_1\geqq 0$ esetén. Az f függvény könnyen módosítható úgy, hogy alulról korlátos legyen, és a 2.2. tétel 1. feltétele fennálljon. A függvény néhány szintvonalát és a metszési felületek (x_1, x_2) síkra eső vetületét a 2. ábrán vázoltuk.)



2. ábra

Nyilvánvaló, hogy f-nek (0, 0) nem lokális minimumhelye, ennek ellenére ott nem definiálható olyan irány, amelynek valamely kúp környezetében az algoritmus továbbhaladása 1 valószínűséggel biztosítható volna, tetszőlegesen kicsiny lépéshosszt megengedő kereső stratégia mellett.

Tekintsük a (0, 0) pontból kiinduló véletlen keresést. (Ebben a pontban az egyetlen függvényérték-csökkenést eredményező lokális irányt az x_1-tengely pozitív iránya szolgáltatja.)

Legyen $p_k, k=0, 1, 2, 3, \ldots$ annak valószínűsége, hogy a (0, 0) pontból kiindulva — a k-adik lépésben, adott Δ_k hosszúságú szakaszon végrehajtott egzakt minimalizálás mellett — csökkenő függvényértékű pontot találunk. Definiáljuk úgy a $0<p_k<1$ értékeket, hogy fennálljon a

$$\sum_{k=0}^{\infty} |\ln(1-p_k)| < \infty \tag{2.42}$$

reláció. (Ez pl. az $|\ln(1-p_k)|=q^{k+1}(0<q<1$ konstans) előírással valósítható meg. Nyilvánvaló, hogy a p_k valószínűség és Δ_k, a k-adik kereső lépés hossza kölcsö-

nösen egyértelműen feleltethető meg egymásnak. Megjegyezzük, hogy (2.42)-ből következik $\lim_{k\to\infty} p_k = 0$ és $\lim_{k\to\infty} \Delta_k = 0$.) A (2.42) feltételből viszont következik

$$\prod_{k=0}^{\infty} (1-p_k) > 0 \tag{2.43}$$

tehát a fenti lépéshossz-választás mellett a (0, 0) pontból kiinduló keresés pozitív valószínűséggel nem talál kedvező irányt a (2.41) függvény esetén. Nyilvánvaló, hogy a $\prod_{k=0}^{\infty} (1-p_k)$ valószínűség 1-hez tetszőlegesen közeli értéket is felvehet, elég gyorsan csökkenő Δ_k sorozat esetén.

A 2. pontban elvégzett konvergenciavizsgálatok tetszőleges, igen általános feltételeknek eleget tevő elméleti véletlen kereső algoritmusokra érvényesek. Egy tényleges számítógépi eljárás természetesen egyéb, a konvergencia igazolásához nem szükséges tényezőktől is függ, amelyek azonban a módszer hatékonyságát nagy mértékben befolyásolják: a következő részben leírt algoritmust ilyen szempontok figyelembevételével próbáltuk kialakítani.

3. Az alkalmazott sztochasztikus gradiens algoritmus ismertetése

Az alábbi eljárás az $f(\mathbf{x})$ ($\mathbf{x} \in R^n$) függvény feltétel nélküli lokális minimumának meghatározására irányul. A jelen dolgozatban csak az algoritmus főbb vonásainak leírására szorítkozunk, gépi realizációjának részletesebb ismertetését a [17] tájékoztató cikk tartalmazza. A leírás egyszerűsítése érdekében az eljárás paramétereit nem látjuk el lépésszámláló indexekkel, és továbbra is feltesszük, hogy $f(\mathbf{x}) \geqq 0$, $\mathbf{x} \in R^n$.

Az algoritmus a következő fő részekre bontható:

0. *A kezdeti paraméterek meghatározása*

Itt adjuk meg a megoldás kezdeti $\mathbf{x}$ közelítését, a továbbhaladási irány számításához felhasznált próbalépések $1 \leqq n_d \leqq n$ számát, a kezdeti α lépéshosszt, illetve a lépéshossz-módosítások paramétereit, az új közelítő megoldás elfogadását szabályozó $0 < \varepsilon < 1$ paramétert, valamint a befejezési feltételek paramétereit.

1. *A kereső lépés irányának kiválasztása*

Az aktuális α lépéshossz meghatározása a korábbi lépések tapasztalatait is figyelembe véve történik. Ezután az $\mathbf{x}$ közelítő megoldáspontban a $-\mathbf{d}$ továbblépési irányt α hosszúságú próbalépések alapján határozzuk meg. Ezért először n_d számú, az n-dimenziós tér egységgömbjén független, egyenletes eloszlású véletlen vektorból ortonormált $\mathbf{d}_j$, $j = 1, \ldots, n_d$ irányokat állítunk elő, majd a $-\mathbf{d}_j$ irányú próbalépések alapján kiszámítjuk a

$$\mathbf{d}_s = -\sum_{j=1}^{n_d} \frac{f(\mathbf{x}-\alpha\mathbf{d}_j)-f(\mathbf{x})}{\alpha}\,\mathbf{d}_j \tag{3.1}$$

irányt. Ezzel kapcsolatban megjegyezzük, hogy tetszőleges $\mathbf{d}_j$, $j=1, \ldots, n$ R^n-beli ortonormált bázis és folytonosan differenciálható f függvény esetén f gradiensére teljesül a

$$\nabla f = \sum_{j=1}^{n} (\nabla f, \mathbf{d}_j)\mathbf{d}_j = -\sum_{j=1}^{n} \lim_{\alpha \to 0} \frac{f(\mathbf{x}-\alpha\mathbf{d}_j)-f(\mathbf{x})}{\alpha}\mathbf{d}_j \tag{3.2}$$

reláció, tehát $\mathbf{d}_s$ a gradiensirány sztochasztikus becslését adja, ha $n_d=n$. Így (legalábbis sima függvények esetén) $-\mathbf{d}_s$ várhatóan lokálisan hatékony irány. A $-\mathbf{d}$ továbbhaladási irány végleges megválasztása az

$$f(\mathbf{x}-\alpha\mathbf{d}) = \min\left[f(\mathbf{x}-\alpha\mathbf{d}_j),\ j=1, \ldots, n_d, f\left(\mathbf{x}-\alpha\frac{\mathbf{d}_s}{\|\mathbf{d}_s\|}\right)\right] \tag{3.3}$$

előírás szerint történik, vagyis $-\mathbf{d}$ megegyezik a $-\mathbf{d}_j$, $j=1, \ldots, n_d$ és $-\mathbf{d}_s$ irányú, α hosszúságú próbalépések közül a legkisebb függvényértéket eredményező lépés irányával.

2. *Iránymenti haladás*

Ha az algoritmus 1. pontjában meghatározott $\mathbf{d}$ irány felhasználásával nyert $\mathbf{x}_d=\mathbf{x}-\alpha\mathbf{d}$ pontban teljesül

$$f(\mathbf{x}_d) \leqq \varepsilon f(\mathbf{x}) \tag{3.4}$$

(figyelembe véve $f(\mathbf{x})$ nemnegatív voltát, ez valamivel erősebb feltétel, mint a (2.3) elméleti kritérium), akkor az $\mathbf{x}:=\mathbf{x}_d$ értékadás, és $\alpha:=\alpha\varphi$ ($\varphi>1$ konstans) lépéshossznövelés után a $-\mathbf{d}$ irány mentén továbbhaladunk. A lépéshossznöveléssel egybekötött iránymenti keresést addig végezzük, amíg először be nem következik $f(\mathbf{x}_d)>\varepsilon f(\mathbf{x})$. Ha ezt legalább egy $-\mathbf{d}$ irányú, (3.4) értelmében sikeres lépés megelőzte, akkor az algoritmus 3. pontjára térünk át. Ellenkező esetben (tehát ha az $f(\mathbf{x}-\alpha\mathbf{d}_j)$, $j=1, \ldots, n_d$ és $f\left(\mathbf{x}-\alpha\frac{\mathbf{d}_s}{\|\mathbf{d}_s\|}\right)$ lépések egyike sem sikeres), az $\mathbf{x}_d:=\mathbf{x}+\gamma\alpha\mathbf{d}$ ($0<\gamma<1$ konstans) által definiált pontban határozzuk meg a függvény értékét. Ha ez a lépés sikeres, akkor a (3.4) reláció után leírt módon ($\mathbf{d}$ irányú extrapolációval) haladunk tovább. Ha viszont mind a $-\mathbf{d}$, mind pedig a $\mathbf{d}$ irányú $\mathbf{x}$-ből végrehajtott lépés sikertelen, akkor újabb $\alpha:=\alpha\gamma$ lépéshossz-redukcióval innen térünk át az eljárás 3. pontjára.

3. *Kvadratikus iránymenti approximáció*

A 3. lépés bármelyik típusú elérése esetén rendelkezésre áll három utoljára vizsgált, kollineáris pont, és az ezekben számított függvényértékek. Az iránymenti haladás konstrukciója folytán világos az is, hogy — az említett pontokat rendre $\mathbf{x}_{-1}$, $\mathbf{x}$, $\mathbf{x}_1$-gyel jelölve — mindenesetre teljesül

$$\frac{1}{\varepsilon}f(\mathbf{x}_{-1}) \geqq f(\mathbf{x}) \leqq \frac{1}{\varepsilon}f(\mathbf{x}_1). \tag{3.5}$$

E három pont és a megfelelő függvényértékek alapján a függvénynek a szóban forgó egyenesen elhelyezkedő lokális minimumára vonatkozóan kvadratikus becslést végzünk (vagyis a $g(\mu)=f(\mathbf{x}+\mu\mathbf{d})\approx a\mu^2+b\mu+c$ kvadratikus függvény minimumát — az a, b, c együtthatók kiszámítása után — meghatározzuk). A kvadratikus becslés figyelembevételével kiválasztjuk a megoldás aktuális közelítését, majd visszatérünk az algoritmus 1. pontjára, ha csak a befejezési feltételek valamelyike nem teljesül.

4. *Befejezési feltételek*

Az eljárás a talált legjobb közelítő megoldás elfogadásával véget ér, ha a következő feltételek valamelyike teljesül:

a) A három legjobb közelítő megoldásvektor komponenseinek és a megfelelő függvényértékek páronkénti eltérése az előírt értékeknél nem nagyobb;

b) Az egymást követő, (3.4) értelmében sikertelen kereső lépések száma meghaladja az előírt korlátot;

c) A függvényérték-meghatározások száma meghaladja az előírt korlátot.

A dolgozat 4. részében az ismertetett algoritmus alapján elkészített program teszteredményeit mutatjuk be.

4. Az algoritmus számítógépi realizációjának tapasztalatai

A 3. pontban leírt véletlen kereső eljárást FORTRAN nyelven programoztuk, a futtatásokat az *Egyetemi Számítóközpont* R—10 számítógépén végeztük el.

A tesztfuttatásokat három csoportra osztottuk, az eredményeket és azoknak más algoritmusok eredményeivel való összehasonlítását is eszerint közöljük. Természetes, hogy egy algoritmus használhatósága matematikailag bizonyítható konvergenciáján kívül még számos tényező (pl. gépi program formájában történő megvalósítása, az eljárás paramétereinek megválasztása, a felhasznált számítógép softwarelehetőségei) függvénye, ezért a más körülmények között elért eredményekkel való összehasonlítás távolról sem torzításmentes. Ennek ellenére remélhető, hogy nagyobb számú, különböző jellegű feladatra vonatkozó vizsgálat alapján az algoritmus effektivitásáról mégis eléggé megbízható képet kapunk.

Az eredmények összehasonlítását a célfüggvény optimumának megadott kiindulópontból történő, adott pontosságú meghatározásához szükséges lépésszámok alapján végezzük. Ezt elsősorban az indokolja, hogy ez a mérőszám az esetek döntő többségében rendelkezésre áll. Amint azt a dolgozat bevezető részében megjegyeztük, az algoritmust elsősorban olyan feladatok megoldására kívánjuk alkalmazni, amelyek esetében a szereplő függvények deriváltjait az optimalizálás során nem használhatjuk fel. Ezért eredményeinket elsősorban más gradiensmentes módszerekre vonatkozó eredményekkel hasonlítjuk össze, számos differenciálható tesztfüggvény esetében is.

(i) *Kvadratikus jellegű tesztfüggvények vizsgálata*

Ebben a részben elsősorban a korábbi véletlen kereső eljárásokkal való összehasonlítás eredményeit ismertetjük. A vizsgált célfüggvények:

$$\sum_{i=1}^{n} x_i^2, \quad \sum_{i=1}^{n} c_i x_i^2, \quad \sum_{i=1}^{n} c_i x_i^4. \tag{4.1}$$

A feladatok megoldásának kezdőpontja az n-dimenziós egységgömb felületén megadott egyenletes eloszlásból kiválasztott véletlen pont, $\mathbf{c} \in R_+^n$ pedig először az egységgömb felületének a pozitív ortánsba eső részén definiált egyenletes eloszlásból kiválasztott véletlen pont, amelyre teljesül

$$\min_{i=1,\ldots,n} c_i \geqq 0{,}1 \max_{i=1,\ldots,n} c_i. \tag{4.2}$$

A $\mathbf{c}$ vektort ezután alkalmas konstanssal szorozva úgy normáljuk, hogy fennálljon a megfelelő kezdőpont és feladat esetén

$$\sum_{i=1}^{n} x_i^2 = 1, \quad \sum_{i=1}^{n} c_i x_i^2 = 1, \quad \sum_{i=1}^{n} c_i x_i^4 = 1. \tag{4.3}$$

A fenti célfüggvények minimalizálását $n=5$, 10, 15, 20, 25, 30-ra végeztük el, minden n-re 10—10 független futtatást hajtva végre. Az egyes futásoknál — a korábbi dolgozatok eredményeivel való összehasonlítás céljából — a kilépési kritériumot nem a 3. részben leírt feltételek alapján adtuk meg, hanem 10^{-8}-nál kisebb célfüggvényérték elérését írtuk elő. Az eredményeket az 1—3. táblázatok foglalják össze. A táblázatokban alkalmazott jelölések:

n az $\mathbf{x}$ vektor komponenseinek száma,
M_n, D_n a 10 futtatás lépésszámainak empirikus várható értéke és szórása,
$l_{\min}$, $l_{\max}$ a 10 futtatás során előforduló legkisebb és legnagyobb lépésszám, továbbá

$$q_n = \frac{D_n}{M_n}, \quad r_n = \frac{l_{\max} - l_{\min}}{l_{\max}}, \quad a_n = \frac{M_n}{n}.$$

1. TÁBLÁZAT

A $\sum_{i=1}^{n} x_i^2$ *célfüggvényre vonatkozó futtatások eredményei*

n	M_n	D_n	q_n	$l_{\min}$	$l_{\max}$	r_n	a_n
5	187,0	23,6	0,13	154	247	0,38	37,4
10	394,2	42,2	0,11	313	479	0,35	39,4
15	534,3	31,5	0,06	494	589	0,16	35,6
20	720,3	53,3	0,07	609	795	0,23	36,0
25	855,0	32,9	0,04	831	920	0,10	34,2
30	1028,0	49,2	0,05	951	1102	0,14	34,3

2. TÁBLÁZAT

A $\sum_{i=1}^{n} c_i x_i^2$ *célfüggvényre vonatkozó futtatások eredményei*

n	M_n	D_n	q_n	$l_{\min}$	$l_{\max}$	r_n	a_n
5	246,9	56,3	0,23	177	348	0,49	49,4
10	669,0	151,7	0,23	472	921	0,49	66,9
15	836,5	126,9	0,15	713	1181	0,40	55,8
20	1164,5	193,2	0,17	861	1458	0,41	58,2
25	1323,2	275,9	0,21	937	1843	0,49	52,9
30	1937,0	154,9	0,08	1612	2165	0,26	64,6

3. TÁBLÁZAT

A $\sum_{i=1}^{n} c_i x_i^4$ *célfüggvényre vonatkozó futtatások eredményei*

n	M_n	D_n	q_n	$l_{\min}$	$l_{\max}$	r_n	a_n
5	133,1	33,7	0,25	76	202	0,62	26,6
10	238,8	51,4	0,22	168	338	0,50	23,9
15	337,3	28,3	0,08	294	390	0,25	22,5
20	415,5	28,5	0,07	371	474	0,22	20,8
25	526,5	60,6	0,12	431	647	0,33	21,1
30	614,1	45,6	0,07	550	672	0,18	20,5

Megjegyzések

1. Az 1—3. táblázatok összesen 180 futtatás eredményeit tükrözik. Az optimumhelyet valamennyi futtatás során az előírt pontossággal határoztuk meg, eltérülést nem tapasztaltunk. Ez az algoritmus stabilitását mutatja, legalábbis a fenti egyszerű célfüggvények esetében.

Az eredmények természetesen függenek az alkalmazott paraméterek értékeitől, tapasztalataink szerint azonban ez a függés itt eléggé tág határok között nem alapvető jelentőségű.

2. A $q_n = \frac{D_n}{M_n}$ és $r_n = \frac{l_{\max} - l_{\min}}{l_{\max}}$ hányadosok a nagyobb n értékek felé haladva csökkenő tendenciát mutatnak. Ezt leginkább az 1. táblázat tükrözi, ahol q_n kb. 0,04—0,07 között stabilizálódik. Így — a szimmetriaokokból azonos nehézségű feladatok megoldásához szükséges lépésszám rögzített n-re közelítőleg normális eloszlású voltát feltételezve — a lépésszámra vonatkozóan elég szűk konfidencia-intervallumok becslése válik lehetővé. Ez arra mutat, hogy azonos feladat többszöri, véletlen módszerrel történő megoldásának lépésszámai itt eléggé stabilak.

A 2. és 3. táblázat eredményei a véletlen c_i együtthatók miatt ennél lényegesen nagyobb mértékű ingadozást tükröznek (itt 10—10 különböző feladatot vizsgálunk minden n-re). A nagyobb n értékek felé haladva azonban itt is inkább csökken q_n és r_n, jelezve azt, hogy a több c_i együtthatót tartalmazó, véletlenszerűen generált feladatok „jobban hasonlítanak egymáshoz".

3. Az $a_n = \frac{M_n}{n}$ hányadosok ingadozása a rögzített n-re vonatkozó kísérletek kis száma ellenére sem túl nagy, és nem mutat növekvő tendenciát a nagyobb n értékek felé haladva. Ez azt mutatja, hogy az algoritmus hatékonysága a változók számának növekedtével — kvadratikus jellegű célfüggvények esetén — nem romlik.

A 4. táblázatban az $M_n = mn$ regressziós egyeneseknek a (fenti eredményekből a legkisebb négyzetek módszerével becsült) meredekségére kapott értékeit hasonlítjuk össze korábbi minimalizáló eljárások alkalmazása során nyert értékekkel, [1], [21], [25] és [26] alapján. (A *Nelder—Mead-* és *Newton—Raphson-módszerek* lépésszáma a dimenziónak [25] szerint lineárisnál gyorsabban növekvő függvénye, tehát elég nagy n-re a véletlen módszerek hatékonyabbak.)

A táblázatban FLP a *Fletcher—Powell-módszer* egy realizációját, ASSRS az *adaptív lépéshosszú véletlen kereső eljárást,* ASSRSR ennek sikertelen lépés esetén ellenkező irányú lépés beiktatásával módosított változatát, RWG a *súlyozott véletlen gradiensek módszerét,* STG pedig a dolgozatban leírt *sztochasztikus gradiens algoritmust* jelöli.

4. TÁBLÁZAT

Az $M_n = mn$ regressziós egyenes m paraméterének becsült értékei

	$\sum_{i=1}^{n} x_i^2$	$\sum_{i=1}^{n} c_i x_i^2$	$\sum_{i=1}^{n} c_i x_i^4$
FLP [25]	6,7	91,6	
ASSRS [21]	80,0	84,2	
ASSRS [26]	62,0	81,4	
ASSRS [1]	55,2	76,1	40,1
ASSRSR [25]	70,4	94,4	
RWG [1]	32,2	44,5	26,6
STG	34,9	59,3	21,1

Az eredmények összehasonlítását nehezíti, hogy az egyes szerzők eredményeiket a 4. rész elején említett különböző feltételek mellett érték el. Ezen túlmenően a tesztfeladatok kiinduló megoldásait sem teljesen azonos elv szerint választják ki. Több esetben pl. a $\sum_{i=1}^{n} c_i x_i^2$ célfüggvény c_i együtthatóit a [0, 1; 1] intervallumon egyenletes eloszlású, független valószínűségi változók adják a (4.3) normálás nélkül, továbbá az egyes esetekre végzett tesztfutások száma csak 4—5. Ezek a körülmények az eredmények ingadozását az 1—3. táblázatban közöltekéhez képest várhatóan növelik, erre vonatkozó adatokat azonban egyik korábbi dolgozat sem ad meg. Megjegyezzük még, hogy az RWG módszer erősen kihasználja a célfüggvény kvadratikus jellegét, ugyanis három kollineáris pontból — az egyenes további pontjainak vizsgálata nélkül — végez a célfüggvénynek az egyenesen felvett minimumára vonatkozó becslést. Ez az eljárás a $\sum_{i=1}^{n} x_i^2$ és $\sum_{i=1}^{n} c_i x_i^2$ célfüggvények esetében természetesen pontos, általában azonban már nem (pl. a $\sum_{i=1}^{n} c_i x_i^4$ célfüggvényre vonatkozóan sem). Ezzel

szemben a leírt sztochasztikus gradiens algoritmus a kvadratikus becslést az iránymenti keresés eredményétől függően végzi, ezért általános célfüggvények esetén is alkalmazható. Jól látható ez az RWG és STG módszerek m regressziós paramétereinek összehasonlításából, valamint a további eredményekből is.

Véletlen zajjal módosított kvadratikus tesztfüggvényre vonatkozó eredmények

Mivel a véletlen kereső eljárást sztochasztikus függvényekkel konstruált feladatokra is alkalmazni szeretnénk, illusztrációképpen az egyszerű $\sum_{i=1}^{n} x_i^2$ célfüggvényt zajjal módosítva nyert

$$f(\mathbf{x}, r) = \left(\sum_{i=1}^{n} x_i^2\right)(1+r) \tag{4.4}$$

függvényre vonatkozó, az előzőekben leírtakhoz hasonló statisztikai vizsgálatokat is végeztünk. A (4.4)-ben szereplő r a $[-r_0, r_0]$ intervallumban egyenletes eloszlású, a célfüggvény minden kiszámításánál véletlenszerűen generált érték. Az 5. és 6. táblázatban az $n=5, 10, 15, 20$-ra vonatkozó 10—10 futtatás eredményeit foglaltuk össze (a korábban alkalmazott jelölésekkel), $\pm 5\%$ és $\pm 10\%$ közötti zajszint mellett.

5. TÁBLÁZAT

$A\left(\sum_{i=1}^{n} x_i^2\right)(1+r)$ *célfüggvényre vonatkozó futtatások eredményei,* $r_0 = 0{,}05$

n	M_n	D_n	q_n	$l_{\min}$	$l_{\max}$	r_n	a_n
5	190,9	30,9	0,16	129	247	0,48	38,2
10	373,0	24,8	0,07	332	423	0,22	37,3
15	605,5	24,3	0,04	569	649	0,12	40,4
20	837,7	46,5	0,06	736	905	0,19	41,9

6. TÁBLÁZAT

$A\left(\sum_{i=1}^{n} x_i^2\right)(1+r)$ *célfüggvényre vonatkozó futtatások eredményei,* $r_0 = 0{,}1$

n	M_n	D_n	q_n	$l_{\min}$	$l_{\max}$	r_n	a_n
5	208,7	52,5	0,25	153	300	0,49	41,7
10	436,8	49,5	0,11	372	531	0,30	43,7
15	680,3	51,0	0,08	606	784	0,23	45,4
20	1039,3	80,3	0,08	930	1213	0,23	52,0

Az eredményeket az 1. táblázatban foglaltakkal összevetve látható, hogy itt is lényegileg hasonló tendenciák érvényesülnek. Az 1—3. táblázatok utáni megjegyzések most is általában helytállóak. A feladatok adott pontosságú megoldásához szükséges lépésszámok természetesen általában növekszenek, a növekedés mértéke azonban nem tűnik indokolatlannak. Ezek az eredmények természetesen nem álta-

lánosíthatók közvetlenül más feladatokra, mégis tükrözik azt a tendenciát, hogy a leírt módszer hatékonyságát a szóban forgó nagyságrendű hibák nem befolyásolják döntően.

Amint azt korábban megjegyeztük, az egyszerű (4.1) tipusú függvények esetében az eljárás paraméterei tág határok között választhatók meg, az eredmények lényeges megváltozása nélkül. Példaként említhető, hogy az 1—3. táblázatok eredményeit $n_d=n/3$ számú lokális kereső irány felhasználásával értük el, míg az 5—6. táblázatokban összefoglalt futtatások esetén $n_d=n/2$ volt. A 6. táblázattal való összehasonlítás céljából közöljük a 7. táblázatot, amely ugyanarra a feladatra vonatkozó futások eredményeit foglalja össze $n_d=n$ mellett.

7. TÁBLÁZAT

A $\left(\sum_{i=1}^{n} x_i^2\right)(1+r)$ *célfüggvényre vonatkozó futtatások eredményei,* $r_0 = 0{,}1$

n	M_n	D_n	q_n	l_{min}	l_{max}	r_n	a_n
5	164,0	22,8	0,14	116	195	0,41	32,8
10	421,4	35,5	0,08	359	473	0,24	42,1
15	705,0	42,1	0,06	634	761	0,17	47,0
20	1007,3	54,4	0,05	924	1106	0,17	50,4

Az eredmények átlagai között lényeges különbség általában nem tapasztalható, a nagyobb n_d érték azonban stabilizáló hatású (kisebb q_n, r_n értékek).

(ii) *További feltétel nélküli tesztfüggvények vizsgálata*

Az alábbiakban a sztochasztikus gradiens algoritmus effektivitását más gradiensmentes algoritmusokéval vetjük össze. Az összehasonlítást elsősorban HIMMELBLAU [10] dolgozatában közölt tesztfüggvények és az ezekre vonatkozó számítási eredmények alapján végezzük.

A tesztfeladatok kiinduló megoldása minden függvénynél adott, az algoritmusok befejezési kritériuma pedig egységesen a következő:

$$|f(\mathbf{x}_{k+1})-f(\mathbf{x}_k)| < 10^{-5}, \quad \text{ha} \quad f^* = 0; \quad \text{egyébként}$$

$$\left|\frac{f(\mathbf{x}_{k+1})-f(\mathbf{x}_k)}{f(\mathbf{x}_k)}\right| < 10^{-5};$$

(4.5) $\qquad i = 1, \ldots, n$ esetén

$$|x_{k+1,i}-x_{k,i}| < 10^{-5}, \quad \text{ha} \quad \mathbf{x}^* = \mathbf{0}; \quad \text{egyébként}$$

$$\left|\frac{x_{k+1,i}-x_{k,i}}{x_{k,i}}\right| < 10^{-5}.$$

Itt $\mathbf{x}_k$ és $\mathbf{x}_{k+1}$ az eljárás során nyert egymást követő két megoldásvektor, i a komponenseik indexe, $\mathbf{x}^*, f^*$ pedig az optimális megoldást és optimumot jelöli. A (4.5) feltételt a 3. részben leírt 4a) feltétellel összehasonlítva látható, hogy az ott alkalma-

zott kritérium szigorúbb. Ennek ellenére biztonsági okokból a 4a) feltételt alkalmazzuk, ami lépésszámainkat némileg növeli.

A vizsgált tesztfüggvényeket, $\mathbf{x}_0$ kiinduló megoldásukat, $\mathbf{x}^*$ optimális megoldásukat és f^* optimumukat a következőkben soroljuk fel. (A tesztfüggvények eredeti előfordulási helyét általában [10] irodalomjegyzéke tartalmazza.)

1) $f(\mathbf{x}) = 100(x_1^2 - x_2)^2 + (1 - x_1)^2,$

$\mathbf{x}_0 = (-1,2;\ 1), \quad \mathbf{x}^* = (1;1), \quad f^* = 0.$ *(Rosenbrock)*

2) $f(\mathbf{x}) = \frac{1}{15}(16x_1^2 + 16x_2^2 - 8x_1x_2 - 56x_1 - 256x_2 + 991) + 28{,}2,$

$\mathbf{x}_0 = (3;\ 8), \quad \mathbf{x}^* = (4;\ 9), \quad f^* = 10.$ *(Zangwill)*

3) $f(\mathbf{x}) = 100(x_2 - x_1^3)^2 + (1 - x_1)^2,$

$\mathbf{x}_0 = (-1,2;\ 1), \quad \mathbf{x}^* = (1;\ 1), \quad f^* = 0.$ *(White—Holst)*

4) $f(\mathbf{x}) = x_1^3 + x_2^2 - 3x_1 - 2x_2 + 12,$

$\mathbf{x}_0 = (0;\ 2), \quad \mathbf{x}^* = \begin{cases} (1;1) \\ (-\infty; x_2) \end{cases}, \quad f^* = \begin{cases} 9 \\ -\infty \end{cases}.$ *(Himmelblau)*

5) $f(\mathbf{x}) = [1{,}5 - x_1(1 - x_2)]^2 + [2{,}25 - x_1(1 - x_2^2)]^2 + [2{,}625 - x_1(1 - x_2^3)]^2,$

$\mathbf{x}_0 = (1{,}0;\ 0{,}8), \quad \mathbf{x}^* = (3{,}0;\ 0{,}5), \quad f^* = 0.$ *(Beale)*

6) $f(\mathbf{x}) = x_1^4 + x_2^4 + 2x_1^2x_2^2 - 4x_1 + 3,$

$\mathbf{x}_0 = (0{,}5;\ 2), \quad \mathbf{x}^* = (1;\ 0), \quad f^* = 0.$ *(Engvall)*

7) $f(\mathbf{x}) = \sum_{j=1}^{10} [\exp(-x_1 t_j) - \exp(-x_2 t_j) - \exp(-t_j) + \exp(-10t_j)]^2,$

$t_j = 0{,}1j, \quad j = 1, \ldots, 10,$

(4.6) $\mathbf{x}_0 = (5;0), \quad \mathbf{x}^* = (1;10), \quad f^* = 0.$ *(Box)*

8) $f(\mathbf{x}) = (x_1 - x_2 + x_3)^2 + (-x_1 + x_2 + x_3)^2 + (x_1 + x_2 - x_3)^2,$

$\mathbf{x}_0 = (100;\ -1;\ 2{,}5), \quad \mathbf{x}^* = (0;\ 0;\ 0), \quad f^* = 0.$ *(Zangwill)*

9) $f(\mathbf{x}) = \sum_{j=1}^{5} f_j^2(\mathbf{x})$

$f_1(\mathbf{x}) = x_1^2 + x_2^2 + x_3^2 - 1, \quad f_2(\mathbf{x}) = x_1^2 + x_2^2 + (x_3 - 2)^2 - 1,$

$f_3(\mathbf{x}) = x_1 + x_2 + x_3 - 1, \quad f_4(\mathbf{x}) = x_1 + x_2 - x_3 + 1,$

$f_5(\mathbf{x}) = x_1^3 + 3x_2^2 + (5x_3 - x_1 + 1)^2 - 36,$

$\mathbf{x}_0 = (1;2;0), \quad \mathbf{x}^* = (0;0;1), \quad f^* = 0.$ *(Engvall)*

10) $f(\mathbf{x}) = x_1(2x_1 - x_3 - 1) + x_2(x_2 - 3) + x_3(2x_3 + x_4 + 1) + x_4(x_4 - 1) + 10,$

$\mathbf{x}_0 = (20;\ 20;\ 20;\ 20),\quad \mathbf{x}^* = (0{,}1538;\ 1{,}500;\ -0{,}3845;\ 0{,}6921),$

$f^* = 7{,}13462.$ (*Sargent—Sebastian*, módosított alakban)

11) $f(\mathbf{x}) = (x_1 + 10x_2)^2 + 5(x_3 - x_4)^2 + (x_2 - 2x_3)^4 + 10(x_1 - x_4)^4,$

$\mathbf{x}_0 = (3;\ 1;\ 0;\ -1),\quad \mathbf{x}^* = (0;\ 0;\ 0;\ 0),\quad f^* = 0.$ *(Powell)*

12) $f(\mathbf{x}) = [\exp(x_1) - x_2]^4 + 100(x_2 - x_3)^6 + \operatorname{tg}^4(x_3 - x_4) + x_1^8 + (x_4 - 1)^2,$

$\mathbf{x}_0 = (1;\ 2;\ 2;\ 2),\quad \mathbf{x}^* = (0;\ 1;\ 1;\ 1),\quad f^* = 0.$ *(Cragg—Levy)*

13) $f(\mathbf{x}) = 100(x_2 - x_1^2)^2 + (1 - x_1)^2 + 90(x_4 - x_3^2)^2 + (1 - x_3)^2 +$

$+ 10{,}1(x_2 - 1)^2 + (x_4 - 1)^2 + 19{,}8(x_2 - 1)(x_4 - 1),$

$\mathbf{x}_0 = (-3;\ -1;\ -3;\ -1),\quad \mathbf{x}^* = (1;\ 1;\ 1;\ 1),\quad f^* = 0.$ *(Wood)*

14) $f(\mathbf{x}) = \dfrac{1}{[1 - (x_1 - x_2)^2]^2} + \sin(x_2 + x_3 + \pi) + \exp\left[\left(\dfrac{x_1 + x_3}{x_3} - 2\right)^2\right],$

$\mathbf{x}_0 = (0{,}5;\ 0{,}5;\ 0{,}5),\quad \mathbf{x}^* = \left(\dfrac{\pi}{4}; \dfrac{\pi}{4}; \dfrac{\pi}{4}\right)\ \left(\dfrac{\pi}{4} \approx 0{,}785398\right),$

$f^* = 1.$ (*Schmidt—Vetters*, módosított alakban)

A [10] dolgozat a következő deriváltmentes algoritmusokra vonatkozó eredményeket közli: (az algoritmusok ismertetését pl. [11] tartalmazza):

HJ: *Hooke—Jeeves,*
NM: *Nelder—Mead,*
P: *Powell,*
R: *Rosenbrock,*
S: *Stewart.*

8. TÁBLÁZAT

A (4.6) tesztfüggvényekre vonatkozó futtatások eredményei

CFS	M_{LS}	D_{LS}	Q	LS_{min}	LS_{max}	R
1.	584,8	123,5	0,21	385	761	0,49
2.	53,0	14,5	0,27	35	70	0,50
3.	314,0	101,5	0,32	200	475	0,58
4.	89,6	20,2	0,23	59	111	0,47
5.	184,6	71,3	0,39	114	275	0,59
6.	56,0	5,3	0,10	48	64	0,25
7.	749,6	314,0	0,42	612	1079	0,43
8.	171,2	25,2	0,15	139	202	0,31
9.	820,0	85,5	0,10	734	970	0,24
10.	192,4	36,1	0,19	122	200	0,39
11.	1050,2	237,6	0,23	761	1362	0,44
12.	680,6	178,6	0,26	426	877	0,51
13.	2065,8	981,3	0,48	803	3183	0,75
14.	177,4	38,3	0,22	108	219	0,51

9. TÁBLÁZAT

Feltétel nélküli tesztfeladatok megoldási lépésszámai deriváltmentes algoritmusok alkalmazásával

Algoritmus \ Tesztfüggvény	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.
HJ	*c*	80	651	640	205	64	498	130	81	*c*	77	9283	836	177
NM	*c*	185	359	190	230	210	268	810	561	*c*	1022	563	797	279
P—P	*c*	29	284	24	134	96	161	84	315	*c*	966	3480	276	86
P—G	*c*	218	156	220	396	264	278	502	652	*c*	1783	3103	850	312
R	*c*	62	294	163	218	119	314	297	457	*c*	801	955	1043	158
S—P	*c*	16	256	*a*	*b*	119	177	37	150	*c*	622	1662	715	75
S—G	*c*	84	194	*a*	161	137	406	108	304	*c*	1117	3749	905	127
STG	585	53	314	90	185	56	750	171	820	192	1050	681	2066	177

A táblázatban használt jelölések:
a: a globális minimumhely felé konvergál az eljárás
b: az eljárás eltérült
c: a [10] dolgozatban nem szereplő célfüggvény

A fenti algoritmusoknál alkalmazott egydimenziós kereső módszerek: aranymetszés, illetve egy POWELLtől származó módszer. Ezek esetleges használatát az algoritmus rövidítése utáni —G, ill. —P szimbólum jelzi a továbbiakban.

Az ismertetett tesztfeladatokra 5—5 futtatást végeztünk az STG sztochasztikus gradiens algoritmussal. Az eredményeket a 8. táblázatban foglaljuk össze, a következő jelölésekkel:

CFS a feladat sorszáma,
M_{LS} az adott feladat megoldásához szükséges lépésszámok empirikus várható értéke (5 futtatás alapján);
D_{LS} a lépésszámok empirikus szórása;
LS_{min} az adott feladatra vonatkozó futtatások során adódott legkisebb lépésszám;
LS_{max} az adódott legnagyobb lépésszám;

$$Q = \frac{D_{LS}}{M_{LS}};$$

$$R = \frac{LS_{max} - LS_{min}}{LS_{max}}.$$

A táblázatból látható, hogy az eredmények általában erősebben ingadoznak, mint a kvadratikus jellegű célfüggvényekre vonatkozó hasonló adatok (vö. 1—3. táblázat). Az egyes feladatokra vonatkozó futtatások kis száma miatt a nyert statisztikai jellemzők természetesen csak tájékoztató jellegűek.

Érdekesebb az itt kapott átlagos megoldó lépésszámok összevetése a [10] dolgozatban közölt eredményekkel. Ezt a 9. táblázat tartalmazza.

Az eredmények alapján a következőket jegyezzük meg:

1. A sztochasztikus gradiens algoritmus számára láthatóan nehéznek bizonyulnak a meredeken változó, bonyolultabb nívóhalmaz-struktúrájú 1., 3. és 13. függvények (ezek lényegileg ugyanazt a függvénytípust reprezentálják). Emellett a 7. feladatnál gyakori, hogy a keresés a (∞, ∞) „kedvező irány" felé indul el és nehezen korrigál az optimális megoldás irányába. A 9. négyzetösszeg-feladat megoldása is lassú.

2. A további célfüggvények esetében az eljárás eléggé hatékonynak bizonyul. A 6. feladatra kapott eredmények a többi módszer lépésszámainál jobbak, de a 2., 4., 5., 8., 11., 12., 14. feladatok megoldásának átlagos effektivitása is jól állja az összehasonlítást a deriváltmentes algoritmusokéval.

3. A lépésenként részletesen kiíratott futások azt bizonyítják, hogy a módszer általában elég hamar lokalizálja az optimumhelyet, annak környezetében viszont gyakran lelassul. Érdemesnek tűnik ezért olyan hibrid algoritmusokat alkalmazni, amelyek a kezdeti véletlen keresésről determinisztikus optimalizálásra térnek át. (A módszer lelassulását természetesen az utolsó három megoldás összehasonlítására vonatkozó, az eljárást stabilizáló 4a) feltétel is befolyásolja.)

Véletlen zajjal módosított feltétel nélküli tesztfüggvények vizsgálata

Az (ii) pontban vizsgált célfüggvények közül néhánynak zajjal módosított alakjára vonatkozó futtatásokat is végeztünk. A zaj figyelembevétele az (i)-ben leírtakhoz hasonlóan történt, tehát az $f(\mathbf{x})$ célfüggvény helyett az $f(\mathbf{x})(1+r)$ függvényt tekintettük, ahol r a $[-r_0, r_0]$ intervallumban egyenletes eloszlású véletlen szám.

Illusztrációképpen az 5., 6., 12. és 14. feladatra vonatkozó $r_0=0{,}05$, ill. $r_0=0{,}1$ melletti 5—5 futtatás eredményeit foglaljuk össze a következő táblázatban, a 8. táblázat jelöléseivel.

10. TÁBLÁZAT

Néhány véletlen zajjal módosított tesztfüggvényre vonatkozó eredmények

CFS	r	M_{LS}	D_{LS}	Q	LS_{min}	LS_{max}	R
5.	0,05	212,8	139,9	0,66	116	483	0,76
	0,1	357,6	191,2	0,54	105	609	0,83
6.	0,05	54,8	6,3	0,12	51	62	0,27
	0,1	68,2	8,6	0,13	56	78	0,28
12.	0,05	340,6	77,0	0,23	228	469	0,51
	0,1	420,0	124,7	0,30	284	636	0,55
14.	0,05	639,6	268,6	0,42	407	1129	0,64
	0,1	620,8	101,8	0,16	468	707	0,34

Szembetűnő, hogy a véletlen zaj a megoldó lépésszámok ingadozását általában jelentősen megnöveli. Emellett természetesen általában maguk a lépésszámok is növekednek és a kapott megoldások is többnyire pontatlanabbak: a tényleges (zajmentes) megoldástól való eltérések az 5., 12. és 14. feladatok esetében rendre

10^{-10}—10^{-2}, 10^{-7}—10^{-2}, illetve 10^{-3}—10^{-2} nagyságrendűek. Az egyszerű 6. feladat megoldása véletlen zaj esetén is pontos volt, itt az ingadozás sem növekedett jelentősen. (A felsorolt tendenciákat egyébként sztochasztikus feladatok más módszerekkel való megoldása esetén is érvényesnek véljük.)

Megemlítjük még, hogy HILLSTROM [9] dolgozatában különböző algoritmusok összehasonlítását véletlenszerűen generált pontokból történő tesztfeladat-megoldások lépésszámai alapján végezte el. A standard kiindulópontú optimalizálás mellett mi is vizsgáltunk véletlen pontokból történő megoldásokat. Ezek elsősorban a módszer stabilitását igazolták, a nyert statisztikai jellemzők azonban — a futtatások kis száma miatt — távolról sem megbízhatóak, ezért ismertetésüket mellőzzük.

(iii) *Néhány matematikai programozási tesztfeladat megoldása*

A sztochasztikus gradiens algoritmussal kapcsolatos numerikus vizsgálatainkat az irodalomból ismert néhány matematikai programozási tesztprobléma megoldásával zárjuk. Az itt leírtak kizárólag az algoritmus kvalitatív jellemzésére szolgálnak, a módszernek feltételekkel korlátozott problémákra való elvi alkalmazhatóságát illusztrálják. Ennek oka az, hogy a feladatok megoldására csak egy igen egyszerű, heurisztikus SUMT (*Sequential Unconstrained Minimization Technique* [7], vagyis feltétel nélküli függvényminimalizálások sorozatán alapuló) algoritmust alkalmaztunk, amely nagymértékben tökéletesíthető.

Az eljárás az alábbi lépésekből áll:

0) Legyen $j:=0$ esetén $\mathbf{x}^{(0)}$ tetszőleges n-dimenziós vektor, $q_i^{(0)}=q>0$, $i=1, \ldots, m$, ahol q adott konstans. Definiáljuk a sztochasztikus gradiens algoritmus kiinduló paramétereit.

1) A (2.1) alakú matematikai programozási feladat megoldását a korábban bemutatott módon (vö. (2.38)) az

$$\mathbf{x}^{(j+1)} = \underset{\substack{\mathbf{x}\in R^n \\ f(\mathbf{x})\leqq f(\mathbf{x}^{(j)})}}{\arg\min} \{Q^{(j)}(\mathbf{x}, q^{(j)})\} = \underset{\substack{\mathbf{x}\in R^n \\ f(\mathbf{x})\leqq f(\mathbf{x}^{(j)})}}{\arg\min}$$

$$\{f_0(\mathbf{x}) + \sum_{i=1}^{m} q_i^{(j)} [\min(0, f_i(\mathbf{x}))]^2\}, \quad j = 0, 1, 2, \ldots \tag{4.7}$$

$$(0 < q_i^{(j)} \leqq q_i^{(j+1)}, \quad j = 0, 1, 2, \ldots; \quad q_i^{(j)} \to \infty, \quad j \to \infty, \quad i = 1, \ldots, m)$$

feltétel nélküli problémák sorozatán alapuló iterációval határozzuk meg. Ez az eljárás a [7] munkában részletezett feltételek mellett a (2.1) feladat lokális minimumhelyének meghatározására alkalmas.

A (4.7)-ben szereplő minimumfeladatot a sztochasztikus gradiens módszer segítségével megoldjuk.

2) Az $\mathbf{x}^{(j)}$ és $\mathbf{x}^{(j+1)}$, valamint a megfelelő $q_i^{(j)}, q_i^{(j+1)}, q_i^{(j)} \neq q_i^{(j+1)}$ szorzók ismeretében a feladat megoldására vonatkozó lineáris becslést végzünk. Bevezetve a $h_i^{(k)}=\dfrac{1}{q_i^{(k)}}$, $k=j, j+1$ jelölést, a megoldásvektor $\mathbf{x}_i^{(j+1)}$ becslését az

$$\mathbf{x}_i^{(j+1)} = \mathbf{x}^{(j+1)} + \frac{h_i^{(j+1)}}{h_i^{(j)} - h_i^{(j+1)}} (\mathbf{x}^{(j+1)} - \mathbf{x}^{(j)}) \tag{4.8}$$

$$q_i^{(j)} \neq q_i^{(j+1)}$$

formula szolgáltatja. A következő iteráció kiinduló megoldását az $\mathbf{x}^{(j+1)}$ és $\mathbf{x}_i^{(j+1)}$, $q_i^{(j)} \neq q_i^{(j+1)}$ vektorok közül a minimális függvényértékű megoldást adó vektor szolgáltatja, jelöljük ezt $\mathbf{x}^{(j+1)}$-gyel.

3) Az együtthatók transzformációját a következőképpen végezzük:

$$q_i^{(j+1)} = \begin{cases} q_i^{(j)}, & \text{ha } f_i(\mathbf{x}^{(j+1)}) \geqq 0; \\ [q_i^{(j)}]^{c_i}, & \text{ha } f_i(\mathbf{x}^{(j+1)}) < 0; \end{cases} \tag{4.9}$$

Itt

$$c_i = 1 + \frac{f_i(\mathbf{x}^{(j+1)})}{\sum\limits_I f_i(\mathbf{x}^{(j+1)})},$$

$$I = \{i: f_i(\mathbf{x}^{(j+1)}) < 0\}.$$

Ezután $j:=j+1$ értékadás, valamint a sztochasztikus gradiens algoritmus paramétereinek finomítása után visszatérünk az 1. lépésre, ha csak a befejezési feltételek valamelyike nem teljesül.

4) Az eljárás a talált legjobb megoldás elfogadásával ér véget, ha az alábbi kritériumok valamelyike teljesül:

a) A kapott $\mathbf{x}^{(j+1)}$ megoldás ϱ_1-megengedett, vagyis adott $\varrho_1 > 0$ esetén teljesül $f_i(\mathbf{x}^{(j+1)}) + \varrho_1 \geqq 0$ $i = 1, \ldots, m$; továbbá az $\mathbf{x}^{(j+1)}$ és $\mathbf{x}^{(j)}$ egymást követő két megoldáspont komponensenkénti eltérése a $\varrho_2 > 0$ adott konstansnál nem nagyobb;

b) A feltétel nélküli célfüggvény kiértékeléseinek száma meghaladja az adott korlátot.

A fenti heurisztikus eljárást — a 2. pontban bevezetett f_i, $i = 0, 1, \ldots, m$ jelölésekkel — az alábbiakban felsorolt problémák megoldására alkalmaztuk: (a feladatok eredeti előfordulását az [5] dolgozat irodalomjegyzéke tartalmazza). A vizsgált kiinduló megoldásokat $\mathbf{x}_1^{(0)}$, $\mathbf{x}_2^{(0)}$, az optimális megoldásvektort és optimumot pedig $\mathbf{x}^*$ és f^* jelöli. Megjegyezzük, hogy valamennyi feladat esetében $\mathbf{x}_1^{(0)}$ a megengedett tartomány pontja, $\mathbf{x}_2^{(0)}$ pedig külső pont.

1) $f_0(\mathbf{x}) = \dfrac{105}{9} - 8x_1 - 8x_2 - 4x_3 + 2x_1^2 + 2x_2^2 + x_3^2 + 2x_1x_2 + 2x_1x_3 + 2x_2x_3,$

$f_1(\mathbf{x}) = x_1,$

$f_2(\mathbf{x}) = x_2,$

$f_3(\mathbf{x}) = x_3,$

$f_4(\mathbf{x}) = \dfrac{8}{3} - x_1 - x_2 - 2x_3,$

$\mathbf{x}_1^{(0)} = (0{,}5;\ 0{,}5;\ 0{,}5;), \quad \mathbf{x}_2^{(0)} = (1;\ 1;\ 1);$

$\mathbf{x}^* = \left(\dfrac{4}{3}; \dfrac{4}{3}; 0\right), \quad f^* = 1, \quad$ aktív feltételek: $f_3, f_4.$

(*Beale*, módosított alakban)

(4.10)

2) $f_0(\mathbf{x}) = x_1^2+x_2^2+2x_3^2+x_4^2-5x_1-5x_2-21x_3+7x_4+45,$

$f_1(\mathbf{x}) = -x_1^2-x_2^2-x_3^2-x_4^2-x_1+x_2-x_3+x_4+8,$

$f_2(\mathbf{x}) = -x_1^2-2x_2^2-x_3^2-2x_4^2+x_1+x_4+10,$

$f_3(\mathbf{x}) = -2x_1^2-x_2^2-x_3^2-2x_1+x_2+x_4+5,$

$\mathbf{x}_1^{(0)} = (0;\ 0;\ 0;\ 0),\quad \mathbf{x}_2^{(0)} = (2;\ 2;\ 5;\ 0);$

$\mathbf{x}^* = (0;\ 1;\ 2;\ -1),\quad f^* = 1,\quad$ aktív feltételek: $f_1, f_3.$

(Rosen—Suzuki)

3) $f_0(\mathbf{x}) = (x_1-10)^2+5(x_2-12)^2+x_3^4+3(x_4-11)^2+$

$+10x_5^6+7x_6^2+x_7^4-4x_6x_7-10x_6-8x_7,$

$f_1(\mathbf{x}) = -2x_1^2-3x_2^4-x_3-4x_4^2-5x_5+127,$

$f_2(\mathbf{x}) = -7x_1-3x_2-10x_3^2-x_4+x_5+282,$

$f_3(\mathbf{x}) = -23x_1-x_2^2-6x_6^2+8x_7+196,$

$f_4(\mathbf{x}) = -4x_1^2-x_2^2+3x_1x_2-2x_3^2-5x_6+11x_7,$

$\mathbf{x}_1^{(0)} = (1;2;0;4;0;1;1),\quad \mathbf{x}_2^{(0)} = (3;3;0;5;1;3;0);$

$\mathbf{x}^* = (2{,}3305;\ 1{,}9514;\ -0{,}477\,54;\ 4{,}3657;\ -0{,}624\,49;\ 1{,}0381;\ 1{,}5942),$

$f^* = 680{,}630,\quad$ aktív feltételek: $f_1, f_4.$ *(Wong)*

A felsorolt feladatokra vonatkozóan a q kezdeti büntetőszorzó, valamint a sztochasztikus gradiens algoritmus befejezési feltételében (l. 3. rész 4a) feltétel) szereplő paraméterek különböző megválasztásaival végeztünk futtatásokat. Néhány számítás eredményeit foglalja össze a következő táblázat, az alábbi jelölésekkel:

F a vizsgált feladat sorszáma (pl. 3/1 a 3. tesztprobléma, az x_1^0 kezdeti megoldásból kiindulva);
LS az adott feladat megoldásának lépésszáma;
CFV a büntetőtagokkal kiegészített célfüggvény közelítő értéke a megoldáspontban;
AFE az aktív feltételi függvények közelítő értéke a megoldáspontban (a feladatok leírásánál megadott sorrend szerint).

11. TÁBLÁZAT

A (4.10) matematikai programozási feladatokra vonatkozó számítások eredményei

F	LS	CFV	AFE
1/1	160	1,00003	$-7\cdot10^{-5}$, $8\cdot10^{-4}$
1/2	155	1,00007	$6\cdot10^{-4}$, $3\cdot10^{-3}$
2/1	222	1,00472	$3\cdot10^{-5}$, $-3\cdot10^{-5}$
2/2	499	1,00265	$-5\cdot10^{-3}$, -10^{-2}
3/1	586	680,740	$-2\cdot10^{-2}$, $-7\cdot10^{-3}$
3/2	823	682,573	$-7\cdot10^{-4}$, $-3\cdot10^{-2}$

A (4.10) feladatokra a korábbiakhoz hasonló módon figyelembe vett (a $[-0{,}05, 0{,}05]$ intervallumban egyenletes eloszlású) r véletlen zaj esetére is végeztünk futtatásokat, tehát a célfüggvény értékét lépésenként a $[Q^{(j)}(\mathbf{x}, q^{(j)})]\,(1+r)$ kifejezéssel határoztuk meg. Ezek eredményeit az alábbi táblázat tartalmazza.

(Itt CFV a büntetőtagokkal kiegészített célfüggvény véletlen zajjal torzított értéke, míg a többi jelölés változatlan.)

12. TÁBLÁZAT

A véletlen zajjal módosított (4.10) *matematikai programozási feladatokra vonatkozó számítások eredményei*

F	LS	CFV	AFE
1/1	630	1,06271	$2\cdot10^{-3}$, $7\cdot10^{-2}$
1/2	703	1,00805	$6\cdot10^{-3}$, $1\cdot10^{-1}$
2/1	794	1,00706	$2\cdot10^{-2}$, $-4\cdot10^{-2}$
2/2	685	0,97177	$-3\cdot10^{-2}$, $-4\cdot10^{-2}$
3/1	749	663,140	0,59, 5,0
3/2	846	699,050	1,54, −0,51

Amint azt az eredmények mutatják, feltételekkel korlátozott problémák esetén a leírt algoritmus elsősorban az optimális megoldás közelítő meghatározására alkalmas. Itt is érvényesül az a tendencia, hogy a módszer az optimumhely gyors megközelítése után lelassul, ami hibrid algoritmus alkalmazását tenné indokolttá. Emellett a szekvenciális módszernek a bemutatott heurisztikus SUMT eljárásnál jóval hatékonyabb változatai is alkalmazhatók: különösen érvényes ez a zajjal torzított feladatok megoldási stratégiáira. Ezért a (4.10) problémákra vonatkozóan nem végeztünk statisztikai vizsgálatokat, számításainkkal csak az algoritmus elvi alkalmazhatóságát kívántuk illusztrálni.

A szerző köszönetét fejezi ki HEGEDŰS CSABA lektornak a kézirat gondos átolvasásáért és értékes megjegyzéseiért.

IRODALOM

[1] ARCHETTI, F., „Evaluation of random gradient techniques for unconstrained optimization", *Calcolo* **12** (1975) 83—94.
[2] ARMACOST, R. L. and FIACCO, A. V., „Computational experience in sensitivity analysis for nonlinear programming", *Mathematical Programming* **6** (1974) 301—326.
[3] BARTOLOMEI, P. I., „Investigation of the statistical gradient method and its recent modification in the problem of optimizing a multiparameter system", *Automatic Control* **2** (1971) 32—37.
[4] BROOKS, S. H., „Discussion of random methods for locating surface maxima", *Operations Research* **6** (1958) 244—251.
[5] CHARALAMBOUS, C., „Nonlinear least pth optimization and nonlinear programming", *Mathematical Programming* **12** (1977) 195—225.
[6] ERMOLEV, YU, M., „On the method of generalized stochastic gradients and stochastic quasi Fejer-sequences", *Cybernetics* **5** (1969) 208—220.
[7] FIACCO, A. V. and MC CORMICK, G. P., *Nonlinear Programming: Sequential Unconstrained Minimization Techniques* (John Wiley and Sons, Inc. New York, 1968).
[8] FLETCHER, R. and POWELL, M. J. D., „A rapidly convergent descent method for minimization", *Computer Journal* **6** (1963) 163—168.
[9] HILLSTROM, K. E., „A simulation test approach to the evaluation of nonlinear optimization algorithms", *ACM Transactions on Mathematical Software* **3** (1977) 305—315.
[10] HIMMELBLAU, D. M., „A uniform evaluation of unconstrained optimization techniques", in: *Numerical Methods for Nonlinear Optimization*, Ed.: *F. A. Lootsma*, Academic Press, London, 1972, 69—97.

[11] HIMMELBLAU D. M., *Applied Nonlinear Programming* (McGraw—Hill, New York, 1972).
[12] NELDER, J. A. and MEAD, R., „A simplex method for function minimization", *Computer Journal* **7** (1965) 308—312.
[13] NIKOLAEV, E. G., „The random *m*-gradient method", *Automatic Control* **1** (1969) 26—29.
[14] NIKOLAEV, E. G., „Steepest descent based on the random *m*-gradient method", *Automatic Control* **3** (1970) 39—44.
[15] NIKOLAEV, E. G., „Steepest descent with a random choice of directions", *Automatic Control* **5** (1970) 25—31.
[16] PINTÉR, J., Egy sztochasztikus programozási probléma megoldása a véletlen keresés módszerével, Egyetemi doktori disszertáció, ELTE TTK, Budapest, 1975.
[17] PINTÉR, J., „Függvényminimalizálás sztochasztikus gradiens módszerrel", *Egyetemi Számítóközpont 23. sz. Tájékoztatója*, Budapest, 1978, 161—171.
[18] RASTRIGIN, L. A., „Extremal control by the method of random scanning", *Automation and Remote Control* **21** (1960) 891—896.
[19] SCHRACK, G. and BOROWSKY, N., „An experimental comparison of three random searches", in: *Numerical Methods for Nonlinear Optimization*, Ed. *F. A. Lootsma*, Academic Press, London 1972, 137—147.
[20] SCHRACK, G. and CHOIT, M., „Optimized relative step size random searches", *Mathematical Programming* **10** (1976) 230—244.
[21] SCHUMER, M. A. and STEIGLITZ, K., „Adaptive step size random search", *IEEE Transactions on Automatic Control* **AC—13**/3 (1968) 270—276.
[22] SMITH, D. E., „An empirical investigation of optimum — seeking in the computer simulation situation", *Operations Research* **21** (1973) 475—497.
[23] SIMMONS, D. M., *Nonlinear Programming for Operations Research* (Prentice—Hall Inc., Englewood Cliffs, New Jersey, 1975).
[24] SUTTI, C., „Convergence proof of minimization algorithms for nonconvex functions", *Journal of Optimization Theory and Applications* **23** (1977) 203—210.
[25] UGRAY, Cs., Some random search function minimization techniques, *Egyetemi Számítóközpont 21. sz. Kutatási Beszámolója* Budapest, 1977.
[26] WHITE, L. J. and DAY, R. G., „An evaluation of adaptive step size random search", *IEEE Transactions on Automatic Control* **AC—16**/5 (1971) 475—478.
[27] ZOUTENDIJK, G., *Methods of Feasible Directions* (American Elsevier, New York, 1960).
[28] Богомолов, Н. А.—Карманов, В. Г., «О сходимости метода случайного поиска для отыскания стационарных точек общей задачи нелинейного программирования», *Вестник Московского Университета, Серия ВМК* No. 1. (1977) 20—26.
[29] Юдин, Д. Б., *Математические методы управления в условиях неполной информации* (Изд. Советское радио, Москва, 1974).
[30] Растригин, Л. А., *Случайный поиск в задачах оптимизации многопараметрических систем* (Изд. Зинатне, Рига, 1965).
[31] Растригин, Л. А., *Системы экстремального управления* (Изд. Наука, Москва, 1974).

(Beérkezett: 1978. június 7.)

PINTÉR JÁNOS
EGYETEMI SZÁMÍTÓKÖZPONT, ALKALMAZOTT MATEMATIKAI OSZTÁLY
1093 BUDAPEST, DIMITROV TÉR 8.

ON THE CONVERGENCE AND NUMERICAL EFFECTIVENESS OF RANDOM SEARCH PROCEDURES

J. PINTÉR

In this paper a stochastic optimization method is studied. Section 1. introduces basic concepts of random search procedures and summarizes some related previous works. In Section 2. general convergence properties of random search are investigated. Firstly — based on the results of [28] — we show the convergence of the procedure for mathematical programming problems with continuously differentiable functions. In this context, applicability of the method of feasible directions using stochastic gradients is also shown. Further on, a convergence proof is given for continuous unconstrained problems. A computer implementation of the theoretical algorithms, defined for the convergence proofs, is presented in Section 3. Finally, in Section 4., numerical experience with this method is detailed, comparing our results to those of earlier random and deterministic algorithms.

Alkalmazott Matematikai Lapok 4 (1978) 229—243

AUTÓBUSZOK ERŐÁTVITELI LÁNCÁNAK OPTIMÁLIS MÉRETEZÉSE MECHANIKUS SEBESSÉGVÁLTÓ ESETÉN

RAPCSÁK TAMÁS
Budapest

A dolgozatban a mechanikus sebességváltóval rendelkező autóbuszok erőátviteli láncának optimális méretezését egy nemlineáris programozási feladat megoldására vezetjük vissza. Először a feltételek és a célfüggvény típusát adjuk meg, majd egy négyfokozatú és egy hatfokozatú váltó esetén a konkrét feladatokat ismertetjük. Ez az első esetben egy 12 változót és 58 feltételt tartalmazó, a második esetben egy 16 változót és 82 feltételt tartalmazó problémát jelent. A feladatokat a *Magyar Tudományos Akadémia* CDC 3300-as gépén a büntetőfüggvényes módszerek egy külső pontos változatával oldottuk meg. A számítási idők (fordítással együtt) 1 perc körül voltak.

A modell gyakorlati felhasználása esetén, az erőátviteli lánc automatikus tervezésén túl, nagyszámú számítógépes kísérletet is lehetne végezni a különböző sebességváltók, differenciálművek, motorok tesztelésére.

1. Bevezetés

A jelenlegi gyakorlat szerint az autóbuszok erőátviteli láncának elemeit nem egy gyárban készítik és méretezését nem a lánc komplex működését vizsgálva végzik el. Célunk annak bemutatása, hogy egy erőátviteli lánc összes paramétereit megkaphatjuk a jármű igénybevételeinek ismeretében egy matematikai programozási feladat megoldásaként.

A probléma felvetése HALMOS EMILTŐL származik, megoldása az IKARUS gyárral kötött kutatási-fejlesztési szerződés keretében történt.

A feladat megfogalmazásánál az első nehézséget az jelentette, hogy a jármű igénybevételei, illetve az autóbuszok erőátviteli rendszerével szemben támasztott követelmények nincsenek egyértelműen megállapítva.

Ezek a problémák más vonatkozásban is fennállnak. Az [1] tanulmány rámutat, hogy a nemzetközi irodalomban, üzemeltetési gyakorlatban nem létezik egységes kategóriarendszer az autóbuszok számára, sőt országonként is rendszerint többfélét használnak. Bonyolítja a helyzetet az azonos elnevezések eltérő tartalmú használata, illetve egy adott autóbuszkategória többféle megnevezése.

Ezért vált szükségessé egy egységes kategorizálási rendszer és egy részletesen kidolgozott követelményrendszer elfogadása.

Mi a továbbiakban az [1] tanulmánybeli osztályozást követve, az ottani városközi expressz és turista luxus autóbusz típusok erőátviteli láncának a méretezésével foglalkozunk, mivel ezek mechanikus sebességváltóval üzemelnek. Ezekre a buszokra az a jellemző, hogy a nagyobb távolságra fekvő városok közötti menetrendszerű autóbuszforgalmat bonyolítják le, ritkán állnak meg, nagy a végsebességük. *Magyarországon* a jelenleg érvényben levő sebességkorlátozások miatt a maximális sebesség 80 km/ó. Azonban a nemzetközi tendenciákat és a jelenlegi gyakorlatot

valamint az új autóbuszok piaci versenyképességének biztosítását figyelembe véve 100—120 km/ó végsebességre kell az autóbuszokat méretezni. Ezeknek a típusoknak a legkisebb sebességi fokozatban egy 20%-os, a legnagyobb sebességi fokozatban, végsebességnél pedig egy 3%-os lejtőt kell tudniuk leküzdeni. A sebességváltó méretezésénél a mechanikus direkt váltókat kell elsősorban figyelembe venni.

Ezek azok az általános követelmények, amelyek az [1] tanulmányban megtalálhatók. Érdemesnek látszik az erőátviteli rendszerrel szemben is egy részletes, sebességfokozatokra lebontott követelményrendszer kidolgozása. Mi a feladat megfogalmazásánál a megfelelő sebességfokozatokban az alábbi követelményeket is figyelembe vettük:

a) az autóbusz gyorsítóképessége adott fokozatban, adott útellenállás mellett maximális legyen,

b) az autóbusz sebessége adott fokozatban, adott határok között legyen, a motornak annál a fordulatszámánál, ahol a nyomaték a legnagyobb (vagy ahol a fogyasztás a legkisebb),

c) az autóbusz sebessége adott fokozatban maximális legyen, a motornak annál a fordulatszámánál, ahol a nyomaték a legnagyobb (vagy ahol a fogyasztás a legkisebb),

d) az autóbusz végsebessége az egyes fokozatokban adott határok között legyen.

Az előbbi követelmények alapján fogalmazunk meg egy olyan optimalizálási feladatot, melynek megoldása szolgáltatja az erőátvileti lánc tervezési változóit. Esetünkben ezek a sebességváltó fogaskerekeinek osztókör sugarai, az autóbusz kerekének a gördülő sugara és a differenciálmű áttétele.

A bemutatásra kerülő modellel azt akarjuk igazolni, hogy az erőátviteli lánc járműdinamikai feltételek alapján történő optimális méretezése számítógép felhasználásával automatikussá tehető. Nem törekedtünk arra, hogy egy konkrét tervezési feladatnál szükséges összes feltételt figyelembe vegyük, mivel ezek a feladat struktúrájának megváltoztatása nélkül beépíthetők. A modell tovább finomítható, ha a hátsóhíd áttételét a differenciálmű geometriai méreteivel fejezzük ki. A modellben a motor paramétereit adottnak tekintettük, azonban lehetőség van arra, hogy ezeket is változóként kezeljük. Így az igénybevételek ismeretében a motor optimális paraméterei is meghatározhatók, tehát lehetőség van minimális üzemanyag-fogyasztás mellett kielégíteni a követelményeket.

A feladatot egy négyfokozatú és két hatfokozatú sebességváltó mellett oldottuk meg. Első esetben 12 (+1) változót és 58 feltételt tartalmazó, a második esetben 16 (+1) változót és 82 feltételt tartalmazó nemlineáris programozási feladatot kaptunk.

A feladatokat a *Magyar Tudományos Akadémia* CDC 3300-as gépén, a büntetőfüggvényes (SUMT) módszerek egy külső pontos változatával oldottuk meg. A számítási idők (fordítással együtt) 1 perc körüli értékek voltak, amelyek azt mutatják, hogy a modell jó számítástechnikai tulajdonságokkal rendelkezik. Az itt szereplő nemlineáris programozási feladatnál ez annak a következménye, hogy a változószám alacsony, s a függvények értékei könnyen számíthatók. Az előzőekben javasolt változtatások sem rontják el ezeket a tulajdonságokat, ugyanis a változók száma nem lesz lényegesen nagyobb, és az alkalmazott módszer a feltételek számának a növekedését alig érzékeli.

A modell alapján az erőátviteli lánc automatikus tervezésén túl, nagyszámú számítógépes kísérletet is lehetne végezni (pl. különböző sebességváltók, differenciálművek, motorok tesztelésére).

A feladat matematikai tulajdonságai nem ismeretesek, így nem tudjuk, hogy lokális vagy globális optimumot találunk-e. Ha induló pontnak a jelenleg használatos erőátviteli láncok paramétereit adjuk meg, akkor biztos, hogy a gyakorlatban is jól használható eredményeket kapunk.

A dolgozat a bevezetésen kívül 4 részből áll. Az elsőben a gépjármű mozgásegyenletét vizsgáljuk. A másodikban ennek alapján megadjuk a feltételek és a célfüggvény típusát. A harmadik részben konkréten megadjuk egy négyfokozatú és egy hatfokozatú váltó esetén az erőátviteli lánc méretezésére szolgáló feladatot. Végül az utolsó részben a megoldó algoritmust és a futási eredményeket ismertetjük.

Köszönetemet fejezem ki dr. UDVARY LÁSZLÓnak, az IKARUS gyár fejlesztési osztályvezetőjének értékes tanácsaiért, és a jó együttműködésért, HALMOS EMILnek a munka folyamán nyújtott segítségéért, hasznos észrevételeiért és a dolgozat gondos áttanulmányozásáért.

2. A gépjármű mozgásegyenletének vizsgálata

A feltételek meghatározásánál az alapösszefüggés a gépjármű mozgásegyenlete lesz, amely a tervezési változók közötti kapcsolatot adja meg. A mozgásegyenletet [4] alapján tárgyaljuk. Csak azt az esetet vizsgáljuk, mikor a gépjármű hátsókerék meghajtású.

Nézzük meg először, hogyan alakul a gépjármű mozgása adott motorteljesítmény és adott szerkezeti kialakítás mellett. A mozgás irányába eső erők egyensúlyát felírva, azt kapjuk, hogy:

$$X - P_w - G_x - P_T = 0, \tag{2.1}$$

ahol X jelenti a két tengelyre ható összes vonóerőt, amit a motor szolgáltat, P_w a légellenállás, G_x a mozgás irányába ható súlyerő, P_T a tömeg tehetetlenségi ellenállása.

A (2.1) egyenlet azt mondja, hogy a motor által szolgáltatott hajtóerő minden pillanatban a levegőellenállás, a tehetetlenségi ellenállás, és a súlyerő x irányú komponensének legyőzésére szolgál.

Tudjuk azt, hogy

$$P_w = h_w F v^2, \tag{2.2}$$

ahol h_w ismert arányossági tényező, F a gépjármű homlokkeresztmetszete (m^2), v a gépjármű sebessége (m/sec),

$$G_x = G \sin \alpha, \tag{2.3}$$

ahol G a gépjármű súlya, α az emelkedő hajlásszöge,

$$P_T = m \frac{dv}{dt}, \tag{2.4}$$

ahol m a tehetetlen tömeg, $\frac{dv}{dt}$ a gépjármű gyorsulása,

$$m = \frac{G}{g} + \sum_{k=1}^{4} \frac{Q_k}{r^2}, \tag{2.5}$$

ahol g a gravitációs gyorsulás, r a kerekek gördülő sugara, Q_k, $k=1, \ldots, 4$ a kerekek tehetetlenségi nyomatéka.

Számítsuk ki most az X értéket. Mivel hátsó-kerék hajtás esetén az első kerekek csak vontatott kerekek, így $X_1 = -Y_1 f$.

A hátsó kerekek forgatott vontató kerekek, ahol a nyomatéki egyensúly a következő:

$$M - Y_2 fr - X_2 r = 0, \tag{2.6}$$

azaz

$$X_2 = \frac{1}{r} M - Y_2 f. \tag{2.7}$$

Ezekben a képletekben X_1, X_2 az első, illetve a második tengelyre ható vonóerőt; Y_1, Y_2 az első, illetve a második tengelyre ható tengelynyomásokat; M a motor nyomatékát; f pedig a gördülési ellenállást jelenti.

A fenti összefüggésekből kapjuk, hogy

$$X = X_1 + X_2 = \frac{1}{r} M - (Y_1 + Y_2) f, \tag{2.8}$$

másrészt

$$Y_1 + Y_2 = G \cos \alpha, \tag{2.9}$$

így

$$X = \frac{1}{r} M - Gf \cos \alpha. \tag{2.10}$$

A kapott értékeket helyettesítsük be a (2.1) egyenletbe.

$$\frac{1}{r} M - Gf \cos \alpha - h_w F v^2 - G \sin \alpha - m \frac{dv}{dt} = 0. \tag{2.11}$$

A kerekekre ható M forgatónyomaték nem azonos a motor M_e nyomatékával, egyrészt mert a motor és a kerekek között általában valamilyen áttétel van, másrészt pedig a motor teljesítményéből egy rész még a kerekek előtt elvész. Ez a teljesítményveszteség két részre bontható. Van egy valóságos teljesítményveszteség, ami a mozgó alkatrészek súrlódása következtében hővé alakul át, s van egy látszólagos veszteség, ami a motor saját forgórészeinek felgyorsításához használódik fel. (Ez utóbbi veszteség azért látszólagos, mert lassításkor teljes egészében visszakapjuk.)

A fentieket figyelembe véve a kerekeken levő forgatónyomatékok összegét úgy kapjuk meg, hogy a motor nyomatékából levonjuk a saját forgórészeinek felgyorsításához szükséges tehetetlenségi nyomatékot, majd a maradékot megszorozzuk az áttétellel és a hatásfokkal. Tehát

$$M = \left[M_e - Q_m \frac{dw_m}{dt} \right] k\eta, \tag{2.12}$$

ahol M_e a motor effektív nyomatéka, amit a tengelyről valamilyen adott fordulatnál levehetünk, Q_m a motor forgórészeinek a főtengelyre redukált tehetetlenségi nyomatéka, beleértve a lendítőkerék és a tengelykapcsoló motor felőli részének

tehetetlenségi nyomatékát is, w_m a motor szögsebessége, $k=\frac{M}{M_e}$ a kerék és a motor között levő összes áttétel nyomatékmódosítása (nyomatéki áttétel), η a kerék és a motor között levő mechanizmus összhatásfoka. Ezt behelyettesítve a (2.11) egyensúlyi egyenletbe azt kapjuk, hogy

$$\frac{M_e}{r} k\eta - h_w F v^2 - G(f\cos\alpha+\sin\alpha) - \left(m\frac{dv}{dt}+Q_m\frac{dw_m}{dt}\frac{k\eta}{r}\right)=0. \tag{2.13}$$

Legyen $\psi=f\cos\alpha+\sin\alpha$, amit útellenállási tényezőnek nevezünk. A további vizsgálat céljából a gyorsulást tartalmazó kifejezéseket átalakítjuk a $v=rw_k$ összefüggés felhasználásával, ahol w_k a kerék szögsebességét jelenti. Így azt kapjuk, hogy

$$m\frac{dv}{dt}+Q_m\frac{dw_m}{dt}\frac{k\eta}{r}=\frac{G}{g}\frac{dv}{dt}\left(1+\sum_{k=1}^{4}\frac{Q_k}{r^2}\frac{g}{G}+\frac{Q_m}{r^2}\frac{g}{G}\eta k\frac{dw_m}{dw_k}\right). \tag{2.14}$$

Vezessük be a következő jelöléseket:

$$\delta_1=\sum_{k=1}^{4}\frac{Q_k}{r^2}\frac{g}{G} \quad \text{(a keréktömegekre vonatkozó tehetetlenségi tényező),}$$

$$\delta_2=\frac{Q_m}{r^2}\frac{g}{G}\eta \quad \text{(a motortömegekre vonatkozó tehetetlenségi tényező),} \tag{2.15}$$

$$\delta=1+\delta_1+\delta_2 k\frac{dw_m}{dw_k} \quad \text{(az összes forgó tömegek tehetetlenségi tényezője).}$$

Ezeket felhasználva a gépjármű mozgásegyenlete:

$$\frac{M_e}{r}k\eta - h_w F v^2 - G\psi - \frac{G}{g}\frac{dv}{dt}\delta = 0, \tag{2.16}$$

ahol $P=\frac{M_e}{r}k\eta$ a motortól származó vonóerő, amely a mozgást létrehozza, $P_w=h_w F v^2$ a légellenállás, $G\psi$ az útellenállás, $P_T=\frac{G}{g}\frac{dv}{dt}\delta$ a tömeg tehetetlenségi ellenállása.

A (2.16) egyenlet szerint, adott szerkezeti kialakítás mellett a motortól származó vonóerő teljes egészében a légellenállás, az útellenállás és a tömeg tehetetlenségi ellenállásának a legyőzésére fordítódik. (Ez az egyenlet [4]-ben megtalálható.)

Az út és a gépjárműkerék kapcsolatát vizsgálva, megkaphatjuk a kifejthető erő és az elérhető gyorsulás maximális értékeit is.

3. A nemlineáris programozási feladat feltételei és célfüggvénye

A nemlineáris programozási feladat feltételeinek egy részét a (2.16) egyenletből származtatjuk. Ezért az egyenletet más formába írjuk, úgy hogy a későbbi tervezési változók szerepeljenek benne.

Minthogy

$$v=\frac{2\pi}{60}rn\frac{1}{k}, \tag{3.1}$$

ahol n a vizsgált időpillanatban a motor fordulatszámát jelenti és mechanikus sebességváltók esetén egy-egy fokozaton belül a nyomatéki áttétel konstans, így

$$\frac{dw_m}{dw_k} = k_l, \quad l = 1, \ldots, p, \tag{3.2}$$

ahol p a fokozatok számát jelenti.

A $\delta_3 = 1 + \delta_1$ jelölést és a (3.1) és (3.2) összefüggéseket felhasználva nyerjük a

$$\frac{M_e}{r} k\eta - h_w F \left(\frac{2\pi}{60} rn\right)^2 \frac{1}{k_l^2} - G\psi - \frac{G}{g}\frac{dv}{dt}(\delta_3 + \delta_2 k_l^2) = 0, \quad l = 1, \ldots, p, \tag{3.3}$$

egyenleteket.

Az erőátviteli rendszerben igen fontos szerepet játszanak a nyomatéki áttételek reciprokai, amelyeket kinematikai áttételeknek nevezünk, s i-vel jelölünk. Tehát

$$i = \frac{1}{k}. \tag{3.4}$$

Mechanikus sebességváltó esetén a motor és a kerekek közötti kinematikai áttétel előáll, mint a sebességváltó megfelelő fokozatának i_{s_l}, $l=1, \ldots, p$ és a differenciálmű i_d kinematikai áttételének szorzata, azaz

$$i_l = i_{s_l} i_d, \quad l = 1, \ldots, p. \tag{3.5}$$

Azonban a sebességváltó tetszőleges fokozatában

$$i_{s_l} = \frac{r_a}{r_b}\frac{r_l'}{r_l}, \quad l = 1, \ldots, p, \tag{3.6}$$

ahol r_a, r_l, $l=1, 2, \ldots, p$ jelenti a mechanikus sebességváltó nyelestengelyén és kimenőtengelyén levő, r_b, r_l', $l=1, 2, \ldots, p$ pedig az előtéttengelyen levő fogaskerekek osztókörének sugarát. Ez azt jelenti, hogy minden fokozatban az i_{s_l} helyett 4 új változót helyettesíthetünk be. Ez a valóságban nem négy változó, hiszen a sebességváltó tengelytávolsága állandó. Mi a továbbiakban az alábbi összefüggést használjuk:

$$r_a + r_b = r_l + r_l', \quad l = 1, \ldots, p. \tag{3.7}$$

Ezekből következik, hogy

$$r_l' = r_a + r_b - r_l, \quad l = 1, \ldots, p, \tag{3.8}$$

tehát elég csak az r_a, r_b, r_l, $l=1, \ldots, p$ sugarakat új változóként bevezetni. Így

$$i_{s_l} = \frac{r_a}{r_b} \cdot \frac{r_a + r_b - r_l}{r_l}, \quad l = 1, \ldots, p. \tag{3.9}$$

Az előbbieket felhasználva a mozgásegyenlet az alábbi alakra hozható.

$$\frac{M_e}{r}\eta i_l - h_w F \left(\frac{2\pi}{60} rn\right)^2 i_l^4 - \left(G\psi + \frac{G}{g}\delta_3 \frac{dv}{dt}\right) i_l^2 - \frac{G}{g}\delta_2 \frac{dv}{dt} = 0, \tag{3.10}$$

$$l = 1, \ldots, p,$$

ahol az i_l, $l=1, \ldots, p$ értékeit (3.5), (3.9) szerint számítjuk.

Ezekben az egyenletekben már szerepelnek a tervezési változók, amelyek az erőátviteli láncot meghatározzák. Mi a továbbiakban az $r_a, r_b, r_l, l=1, \ldots, p, i_d, r, (G)$ változókat tekintjük tervezési változóknak. Természetesen az erőátviteli lánc műszakilag pontosabb leírása esetén újabb változókat is bevezethetünk (pl. a motor paraméterei, a differenciálmű paraméterei).

A tervezés alapelve az, hogy megadjuk a feltételi függvények segítségével a működés szempontjából lényeges és kritikus helyzeteket minden fokozatban, s a tervezendő rendszertől megköveteljük, hogy ilyen körülmények között is biztosítani tudja az autó folyamatos haladását. A célfüggvény szerepe abban áll, hogy a feltételeket kielégítő rendszerek közül a számunkra legmegfelelőbbet tudjuk kiválasztani. Az erőátviteli rendszerrel szemben támasztott összes követelmények is szerepelnek vagy a feltételekben vagy a célfüggvényben.

Elképzelésünk szerint, ha a lényeges és kritikus helyzetekben az autó tovább tud haladni, akkor a kevésbé kritikus helyzetekben sem lehet probléma, így elég véges sok időpillanatban ellenőrizni az autó mozgását.

Ehhez szükséges a (3.10) egyenlet. Jó tervezés esetén ugyanis minden lényeges és kritikus helyzetben legalább annyi vonóerőnek kell rendelkezésre állnia, amellyel az autó le tudja győzni a légellenállást, az útellenállást és a tehetetlenségi ellenállást, ami a (3.10) bal oldalán levő kifejezés nemnegativitását jelenti.

Most nézzük meg, milyen típusúak lesznek a feltételek és a célfüggvény. Vezessük be az alábbi jelöléseket.

$$z(i_l, r, G, a_l, n, \psi) = \tag{3.11}$$

$$= \frac{M_e}{r}\eta i_l - h_w F\left(\frac{2\pi}{60} rn\right)^2 i_l^4 - \left(G\psi + \frac{G}{g}\delta_3 \frac{dv}{dt}\right) i_l^2 - \frac{G}{g}\delta_2 \frac{dv}{dt}, \quad l = 1, \ldots, p,$$

ahol $a_l = \frac{dv}{dt}$, $l=1, \ldots, p$ az autó gyorsulása a vizsgált időpontban és az i_l, $l=1, \ldots, p$ értékeit pedig a (3.5), (3.9) szerint számítjuk. Legyenek továbbá K_l^a, K_l^f, $K_l^{a_v}$, $K_l^{f_v}$, $l=1, \ldots, p$ az egyes fokozatokban a sebességre, illetve a végsebességre előírt alsó és felső korlátok; n_M a motornak az a fordulatszáma, ahol a nyomaték a legnagyobb, $n_{\max}$ pedig a motor legnagyobb fordulatszáma.

Ha az első sebességi fokozatban az autóbusznak egy 20 fokos lejtőt kell legyőzni, akkor a

$$z(i_1, r, G, a_1, n_M, f\cos 20\% + \sin 20\%) \geqq 0,$$

$$K_1^a \leqq \frac{2\pi}{60} rn_M i_1 \leqq K_1^f \tag{3.12}$$

egyenlőtlenségek teljesülését követeljük meg.

Ha a második fokozatot gyorsító fokozatnak szeretnénk használni, akkor a

$$z(i_2, r, G, a_2, n_M, f) \geqq 0,$$

$$K_2^a \leqq \frac{2\pi}{60} rn_M i_2 \leqq K_2^f \tag{3.13}$$

feltételeket írjuk elő (esetleg több időpillanatban is) és a gyorsulás értékét (értékeit) beépítjük a maximalizálandó célfüggvénybe.

Biztosítani kell azt is, hogy az előre megadott sebességtartomány minden értékét elérjük. Erre szolgálnak a

$$z(i_l, r, G, 0, n_{\max}, f) \geqq 0,$$

$$K_l^{a_v} \leqq \frac{2\pi}{60} r n_{\max} i_l \leqq K_l^{f_v}, \quad l = 1, \ldots, p, \tag{3.14}$$

egyenlőtlenségek.

A követelmények között van az is, hogy az autóbusz sebessége adott fokozatban legyen maximális, feltéve, hogy a motor olyan fordulatszámmal pörög, ahol a nyomaték a legnagyobb. Ha a

$$z(i_l, r, G, 0, n_M, f) \geqq 0 \tag{3.15}$$

egyenlőtlenséget követeljük meg a megfelelő l indexre és a sebességet beépítjük a célfüggvénybe, akkor elérjük ezt a célt.

Nagyon lényeges, hogy a feladat változóira (a tervezési változókra és a gyorsulásokra) alsó és felső korlátokat adjunk meg. Ezek biztosítják többek között, hogy a feltételeket ne megépíthetetlen erőátviteli lánccal elégítsük ki. (Lásd az 1. táblázatot.)

Az előzőekben láttuk, hogy a célfüggvényben különböző dimenziójú mennyiségek (gyorsulások és sebességek) szerepelnek. Ha a maximalizálandó mennyiségeket megfelelő súlyokkal szorozzuk be, akkor a kapott összegben csupa egyforma dimenziójú mennyiség szerepel. (Az itt szereplő első modell célfüggvényében a 2. és 3. fokozatban az egy másodperc alatt egyenletes gyorsulással megtett út és a 3. és 4. fokozatban az egy másodperc alatt egyenletes sebességgel megtett út szerepel.) Emellett a célfüggvény egyes tagjait és a feltételeket a megoldó algoritmus sajátosságait figyelembe véve is kellett súlyozni. Ez azért vált szükségessé, mert a különböző tagok és feltételek értéke nagyságrendekkel különbözött, s a program a számítás során a kisebbeket figyelmen kívül hagyta.

Megjegyezzük, hogy más követelményrendszer esetén a célfüggvény változhat, tehát a modellben szereplő célfüggvény nincs rögzítve.

4. Az optimalizálási feladatok

Ebben a részben egy négyfokozatú és egy hatfokozatú váltó esetén megadjuk a konkrét követelményeket és a hozzájuk tartozó optimalizálási feladatokat, amelyeket számítógépen megoldottunk.

A—I. Az erőátviteli rendszerrel szemben támasztott követelmények négyfokozatú sebességváltó esetén

1. fokozat:

a) az autóbusz küzdjön le egy 20%-os lejtőt,

b) az autóbusz sebessége adott útellenállás mellett adott határok között legyen, feltéve, hogy a motor olyan fordulatszámmal pörög, ahol a nyomaték a legnagyobb,

c) az autóbusz végsebessége adott határok között legyen.

2. fokozat:

a) az autóbusz gyorsítóképessége adott útellenállás mellett maximális legyen,
b) az autóbusz sebessége adott útellenállás mellett adott határok között legyen, feltéve, hogy a motor olyan fordulatszámmal pörög, ahol a nyomaték a legnagyobb,
c) az autóbusz végsebessége adott határok között legyen.

3. fokozat:

a) az autóbusz gyorsítóképessége adott útellenállás mellett maximális legyen,
b) az autóbusz sebessége adott útellenállás mellett adott határok között legyen, feltéve, hogy a motor olyan fordulatszámmal pörög, ahol a nyomaték a legnagyobb,
c) az autóbusz sebessége legyen maximális, feltéve, hogy a motor olyan fordulatszámmal pörög, ahol a nyomaték a legnagyobb (kb. 70—80 km/óra),
d) az autóbusz végsebessége adott határok között legyen.

4. fokozat:

a) az autóbusz a végsebességénél küzdjön le egy 3%-os lejtőt,
b) az autóbusz sebessége adott útellenállás mellett adott határok között legyen, feltéve, hogy a motor olyan fordulatszámmal pörög, ahol a nyomaték a legnagyobb,
c) az autóbusz sebessége legyen maximális, feltéve, hogy a motor olyan fordulatszámmal pörög, ahol a nyomaték a legnagyobb (kb. 100—120 km/óra),
d) az autóbusz végsebessége adott határok között legyen.

A—II. Az optimalizálási feladat négyfokozatú váltó esetén

$$\max\left[\frac{1}{2}(a_2+a_3)+\frac{2\pi}{60}rn_M(i_3+i_4)\right]$$

$$z(i_l, r, G, a_l, n_M, \psi_l) \geqq 0, \quad l = 1, \ldots, 4,$$

$$z(i_l, r, G, 0, n_{\max}, f) \geqq 0, \quad l = 1, \ldots, 4,$$

(4.1)
$$K_l^a \leqq \frac{2\pi}{60} rn_M i_l \leqq K_l f, \quad l = 1, \ldots, 4,$$

$$K_l^{a_v} \leqq \frac{2\pi}{60} rn_{\max} i_l \leqq K_l^{f_v}, \quad l = 1, \ldots, 4,$$

$$K_1 \leqq r_a, r_b, r_l, r_l' \leqq K_2, \quad l = 1, \ldots, 4,$$

$$0 \leqq a_l \leqq K_3 \quad l = 1, \ldots, 4,$$

$$K_4 \leqq i_d \leqq K_5,$$

$$K_6 \leqq r \leqq K_7,$$

$$K_8 \leqq G \leqq K_9,$$

ahol

$$\psi_1 = f\cos 20\% + \sin 20\%; \quad \psi_2 = f; \quad \psi_3 = f; \quad \psi_4 = f\cos 3\% + \sin 3\%.$$

A négyfokozatú váltó esetén a sebességhatárok az alábbiak voltak. (A dimenzió km/h-ban, és mellette zárójelben m/sec-ban van megadva. Ténylegesen az utóbbi értékekkel számolunk.)

$K_1^a = 5(1{,}38)$, $K_1^f = 30(8{,}33)$, $K_1^{a_v} = 25(6{,}94)$, $K_1^{f_v} = 35(9{,}72)$,

$K_2^a = 30(8{,}33)$, $K_2^f = 60(16{,}66)$, $K_2^{a_v} = 55(15{,}28)$, $K_2^{f_v} = 70(19{,}44)$,

$K_3^a = 60(16{,}66)$, $K_3^f = 90(25)$, $K_3^{a_v} = 85(23{,}61)$, $K_3^{f_v} = 95(27{,}78)$,

$K_4^a = 80(22{,}22)$, $K_4^f = 120(33{,}33)$, $K_4^{a_v} = 110(30{,}55)$, $K_4^{f_v} = 120(33{,}33)$.

A négy- és a hatfokozatú váltók esetén is $K_1 = 5$ cm, $K_2 = 14$ cm, $K_3 = 5$ m/sec^2, $K_4 = 4$, $K_5 = 9$, $K_6 = 0{,}4$, $K_7 = 1$, $K_8 = 12\,000$ kg, $K_9 = 20\,000$ kg volt.

B—I. Az erőátviteli rendszerrel szemben támasztott követelmények hatfokozatú sebességváltó esetén

1. fokozat:
Csak a sebességhatárokban különbözik az előbbi váltó 1. fokozatától.

2. fokozat:
a) az autóbusz küzdjön le egy 10%-os lejtőt, a többi követelmény csak a sebességhatárokban különbözik az előbbi váltó 2. fokozatától.

3. és 4. fokozat:
Csak a sebességhatárokban különbözik az előbbi váltó 2. fokozatától.

5. fokozat:
Csak a sebességhatárokban különbözik az előbbi váltó 3. fokozatától.

6. fokozat:
Csak a sebességhatárokban különbözik az előbbi váltó 4. fokozatától.

Az optimalizálási feladat a (4.1) feladathoz teljesen hasonlóan írható fel. A hatfokozatú váltók esetén a sebességhatárok az alábbiak voltak:

$K_1^a = 5(1{,}38)$, $K_1^f = 15(4{,}17)$, $K_1^{a_v} = 10(2{,}78)$, $K_1^{f_v} = 20(5{,}56)$,

$K_2^a = 15(4{,}17)$, $K_2^f = 25(6{,}94)$, $K_2^{a_v} = 20(5{,}56)$, $K_2^{f_v} = 30(8{,}33)$,

$K_3^a = 25(6{,}94)$, $K_3^f = 35(9{,}72)$, $K_3^{a_v} = 30(8{,}33)$, $K_3^{f_v} = 40(11{,}11)$,

$K_4^a = 35(9{,}72)$, $K_4^f = 55(15{,}28)$, $K_4^{a_v} = 45(12{,}5)$, $K_4^{f_v} = 60(16{,}66)$,

$K_5^a = 55(15{,}28)$, $K_5^f = 90(25)$, $K_5^{a_v} = 80(22{,}22)$, $K_5^{f_v} = 95(26{,}39)$,

$K_6^a = 90(25)$, $K_6^f = 120(33{,}33)$, $K_6^{a_v} = 110(30{,}55)$, $K_6^{f_v} = 130(36{,}11)$.

Az előzőekben láttuk, hogy a négyfokozatú váltó esetén az optimalizálási feladat egy 12 (+1) változót (tervezési változók és gyorsulások) és 58 feltételt tartal-

mazó nemlineáris programozási feladat, amelyben 24 nemlineáris és 34 változókra kiszabott alsó, felső korlát feltételünk van.

A hatfokozatú váltó esetén két feladatot oldottunk meg. Az elsőben az 5., a másodikban pedig a 6. fokozat volt direkt. Mindkét esetben a feladat 16 (+1) változót (tervezési változók és gyorsulások) és 82 feltételt tartalmazott, amelyekből 36 volt nemlineáris.

5. A futási eredmények

A nemlineáris programozási feladatokat a *Magyar Tudományos Akadémia* CDC 3300 gépén a büntetőfüggvényes (SUMT) módszerek [2] $\min(0, X)^2$ függvényre épülő külső pontos algoritmusával oldottuk meg. (Minimum feladattá írtuk át őket.) A külső pontos algoritmusoknak az az egyik előnye, hogy nem megengedett pontból is indulhatunk. A feltétel nélküli minimalizálást a sztochasztikus programozási feladatoknál is sikerrel alkalmazott ([5], [6]) HOOKE és JEEVES [3] módszerével végeztük.

A feladatok megoldásánál 4 feltétel nélküli minimalizálást, s a feltétel nélküli minimalizálásokon belül 5—8 iteratív lépést végeztünk. A büntetőparaméter értéke indulásnál 1 volt, s minden lépésben az ötszörösére növeltük. A számítások során a nehezen teljesülő feltételeket, valamint a célfüggvény értékét 1000-rel, 10 000-rel vagy 100 000-rel szoroztuk. Mivel az algoritmus nem megengedett pontból indult (az indulópontot véletlenszerűen választottuk), így a büntetőtagok értékei olyan nagy számok voltak, hogy domináltak az algoritmus folyamán. Ezt mutatja, hogy az első feltétel nélküli minimalizálás után általában 4—5 nagyságrenddel csökkent a büntetőfüggvény értéke. Ezért a büntetőfüggvény értékeit nem írtuk be a táblázatokba. Az eredeti célfüggvény szerepét mutatja az, hogy mikor a 4. fokozat gyorsulását kihagytuk, akkor annak értéke lényegesen kisebb volt a többinél (1.3. táblázat). A feladatban szereplő állandók értékei a következők voltak:

$$\eta = 0{,}9;\ f = 0{,}01;\ h_w = 0{,}018;\ F = 8\ m^2;\ Q_m = 8\ \text{mkp/sec}^2.$$

A $\sum_{i=1}^{4} \frac{Q_k}{r^2}$ érték helyett a $4m_{k_{\text{red}}}$ (=10) értékkel számoltunk. A motorra vonatkozó paraméterértékeket (bemenő nyomaték, $n_{\max}, n_M$) a táblázatokban közöljük. Ezek az IKARUS autóbuszoknál figyelembe vett motorok paraméterei. A táblázatokban közölt megoldásokban a tervezési változók az alábbi sorrend szerint vannak rendezve:

$$r, (G), r_a, r_b, r_1, r_2, r_3, r_4, i_d, a_1, a_2, a_3, a_4$$

a négyfokozatú váltónál és

$$r, (G), r_a, r_b, r_1, r_2, r_3, r_4, i_d, a_1, a_2, a_3, a_4, r_5, r_6, a_5, a_6$$

a hatfokozatú váltóknál.

Az induló pontok a következők voltak:

0,7; 17000; 5,0; 10,0; 6,0; 7,0; 8,0; 9,0; 5,5; 1,5; 1,0; 1,5; 0,0;

illetve

0,7; 16000; 5,0; 10,0; 6,0; 7,0; 8,0; 9,0; 5,5; 1,0; 1,0; 3,0; 3,0; 6,0; 6,0; 1,5; 1,0.

1. TÁBLÁZAT

4 fokozatú váltó

	A motor par.	Szám. idő (ford. együtt.)	Az optimális megoldások	Sebességek	Végsebességek
1. futás	77 mkp 2200 1400	40 mp	1,17; 11 999,9; 4,99; 8,5; 11,66; 8,9; 7,1; 6,38; 6,1; 0,0; 0,312 0,266; 0,0	—	—
2. futás	90 mkp 2200 1400	47 mp	1,12; 11999,8; 5,041; 8,82; 11,82; 8,85; 6,94; 6,31; 6,057; 0,0; 0,374; 0,312; 0,0	3,5; 11,3; 19,9; 23,9	4,2; 13,8; 24,3; 29,3

Az 1. táblázatban látható, hogy a G értéke a megoldásokban mindig egyenlő az alsó korláttal. Ezért a hatfokozatú váltók esetén a G értékét már adottnak vettük. A négyfokozatú váltó esetén az r_l', $l=1, \ldots, 4$ értékeit nem korlátoztuk se alulról, se felülről. Az eredményekből látszik, hogy így megépíthetetlen váltót kaptunk. Az első és a negyedik fokozatban a gyorsítás értéke nulla. Ez azt jelenti, hogy a motor nem elég erős a megadott követelményekhez. Az optimálisnak tekintett megoldás minden esetben csak majdnem megengedett volt, azaz néhány feltételnél kisebb pontatlanságok adódtak (pl. az r értékei), ezek azonban a tervezési probléma szempontjából elhanyagolhatók voltak.

A későbbiekben is azt tapasztaltuk, hogy ha a motornál a bemenő nyomaték nagyobb volt, akkor a feltételek kisebb hibával teljesültek. Az eredményekből az is leolvasható, hogy az erősebbb motor lényegesen nem befolyásolja az erőátviteli lánc paramétereit, hanem csak a gyorsulások és sebességek értékeit.

A 4 és a 6 fokozatú váltók esetén is az erőátvitel akkor volt a legjobb, mikor a kerék sugara a megengedett maximális értéket felvette.

A 2. és a 3. táblázat eredményei azt mutatják, hogy az optimális megoldásokban az első és a második fokozat között nincs különbség. (A gyakorlatban az autóbuszvezetők is a második fokozatból indulnak.) A negatív és kis gyorsulásértékek azt jelzik, hogy a megadott követelményekhez nem elég erős a motor. Egy másik érdekes tanulság, hogy a direkt váltók esetén igen nagy teher hárul a differenciálműre. Korábban láttuk, hogy a differenciálmű értékének felső határa 9. Ezt a feltételt a 2. táblázatban található 1., 2., 3. futás esetén 1000-rel szoroztuk, mégis a bemenő nyomaték nagyságának növekedésével a differenciálmű áttétele is nőtt. Csak a 4. futás esetén csökkent 9 alá, mikor a feltételt 10 000-rel szoroztuk. Ebből a szempontból sokkal jobb, mikor a 6. fokozat direkt. Ekkor súlyként 1000-t használva is elfogadható eredményeket kaptunk.

A 3. táblázatban szereplő első négy futásnál a célfüggvényben nem szerepelt a 4. fokozat gyorsulása. Látható, hogy ezek az értékek lényegesen kisebbek, (még negatív érték is szerepel), mint az 5., 6., 7. futásnál kapottak.

Az első és második futás között csak annyi különbség volt, hogy a célfüggvényt először 10 000-rel, másodjára pedig 100 000-rel súlyoztuk. (A 6 fokozatú váltóknál a célfüggvényt 100 000-rel szoroztuk.)

2. TÁBLÁZAT

6 fokozatú váltó

	A motor par.	Szám. idő (ford. együtt.)	Az optimális megoldások	Sebességek	Végsebességek
1. futás	80 mkp 2200 1800	51 mp	1,068; 16 000; 5,074; 9,507; 9,578; 9,574; 8,249; 7,265; 9,479; −0,14; −0,027; 0,271; 0,279; 9,281; 4,82; 0,182; −0,206	6,1; 6,1; 8,9; 11,7; 21,8; 23,6	7,6; 7,6; 11,1; 14,6; 27,1; 29,4
2. futás	105 mkp 2200 1800	51,5 mp	1,03; 16 000; 5,234; 9,551; 9,797; 9,785; 8,426; 6,937; 10,018; −0,092; 0,027; 0,357; 0,375; 9,344; 4,886; 0,283; −0,098	6,0; 6,0; 8,9; 13,3; 21,5; 23,9	7,5; 7,5; 11,1; 16,7; 26,9; 29,9
3. futás	138 mkp 2400 1900	46,4 mp	1,016; 16 000; 5,36; 9,406; 9,768; 9,756; 8,523; 7,168; 10,074; −0,056; 0,124; 0,478; 0,512; 9,281; 4,909; 0,426; −0,005	6,2; 6,2; 8,8; 12,8; 21,1; 24,2	7,7; 7,7; 11,0; 16,0; 26,4; 30,2
4. futás	138 mkp 2400 1900	59,2 mp	0,922; 16 000; 5,194; 9,452; 9,643; 9,638; 8,365; 6,957; 8,959; −0,067; 0,131; 0,486; 0,518; 9,563; 4,855; 0,418; −0,002	6,1; 6,2; 8,9; 13,1 21,5; 23,9	7,7; 7,7; 11,1; 16,4; 26,9; 29,9

3. TÁBLÁZAT

6 fokozatú váltó (6. fokozat direkt)

	A motor par.	Szám. idő (ford. együtt.)	Az optimális megoldások	Sebességek	Végsebességek
1. futás	80 mkp 2200 1800	48,6 mp	1,08; 16000; 4,882; 9,45; 9,338; 9,32; 8,26; 6,84; 8,89; −0,16; −0,041; 0,27; −0,297; 5,465; 8,98; 0,201; −0,207	6,6; 6,6; 9,0; 13,4; 19,9; 23,7	8,1; 8,2; 11,2; 16,6; 24,7; 29,5
2. futás	80 mkp 2200 1800	60,6 mp	1,07; 16000; 4,887; 9,449; 9,34; 9,336; 8,273; 6,883; 8,784; −0,166; −0,042; 0,273 0,156; 5,799; 9,156; 0,224; −0,2	6,5; 6,6; 9,0; 13,3; 18,0; 23,7	8,1; 8,2; 11,2; 16,5; 22,4; 29,4
3. futás	105 mkp 2200 1800	58,3 mp	1,084; 16000; 4,828; 9,53; 9,36; 9,34; 8,29; 6,81; 9,54; −0,101; 0,008; 0,357; 0,031; 5,328; 9,21; 0,309; −0,098	6,4; 6,5; 8,8; 13,3; 20,4; 23,8	8,0; 8,1; 11,0; 16,7; 25,5; 29,7
4. futás	138 mkp 2400 1900	52,4 mp	1,04; 16000; 4,91; 9,458; 9,38; 9,365; 8,37; 6,984; 9,14; −0,074; 0,098; 0,484; 0,094; 5,914; 9,0; 0,475; 0,0	6,6; 6,7; 8,9; 13,2; 17,8; 23,9	8,3; 8,3; 11,1; 16,5; 22,3; 29,9
5. futás	80 mkp 2200 1800	70 mp	1,073; 16000; 4,887; 9,45; 9,341; 9,336; 8,273; 7,686; 8,784; −0,166; −0,042; 0,273 0,278; 5,78; 9,156; 0,224; −0,2	6,5; 6,6; 9,0; 10,6; 18,0; 23,7	8,1; 8,2; 11,2; 13,2; 22,4; 29,4
6. futás	105 mkp 2200 1800	53,3 mp	1,084; 16000; 4,828; 9,53; 9,364; 9,340; 8,29; 6,83; 9,54; −0,101; 0,008; 0,357; 0,375; 5,33; 9,22; 0,309; −0,098	6,4; 6,5; 8,8; 13,3; 20,4; 23,8	8,0; 8,1; 11,0; 16,6; 25,5; 29,7
7. futás	138 mkp 2400 1900	58,7 mp	1,044; 16000; 4,92; 9,458; 9,381; 9,365; 8,374; 6,984; 9,139; −0,074; 0,098; 0,484; 0,516; 5,914; 9,0; 0,475; 0,0	6,6; 6,7; 8,9; 13,2; 17,8; 23,9	8,3; 8,3; 11,1; 16,5; 22,3; 29,9

IRODALOM

[1] Autóipari Kutató Intézet Jármű Tagozat „*Autóbuszokkal szemben támasztott követelmények*" tanulmány, Budapest, 1976.
[2] FIACCO, A. V. and MCCORMICK, G. P., *Nonlinear Programming Sequential Unconstrained Minimization Techniques* (Wiley and Sons, New York, 1968).
[3] KOVALIK, J. and OSBORNE, M. R., *Methods for Unconstrained Optimization Problems* (Elsevier, New York, 1968).
[4] LÉVAI, Z., *Gépjárműtechnika* (Tankönyvkiadó, Budapest, 1972).
[5] PRÉKOPA, A., RAPCSÁK, T. és ZSUFFA, I., „Egy új módszer sorbakapcsolt tározórendszerek tervezésére", *Alkalmazott Matematikai Lapok* **2** (1976).
[6] RAPCSÁK, T., „A SUMT módszer alkalmazása logaritmikusan konkáv feltételi függvényeket tartalmazó nemlineáris programozási feladat megoldására," *MTA SZTAKI Közlemények* **19** (1978) 17—28.
[7] RAPCSÁK, T., „Az autóbuszok erőátviteli rendszerének optimális számítógépes tervezése", Tanulmány, Budapest, 1977.

(Beérkezett: 1978. július 13.)

RAPCSÁK TAMÁS
MTA SZÁMÍTÁSTECHNIKAI ÉS AUTOMATIZÁLÁSI KUTATÓ INTÉZET
1250 BUDAPEST, ÚRI U. 49.

THE OPTIMAL POWER TRANSMISSION OF MECHANICAL SPEED GEAR

T. RAPCSÁK

In this paper the optimal power transmission of mechanical speed gear is designed by solving a nonlinear programming problem. First the contraints and the objective function are constructed. Next problems are given in case of speed-gears of four and six stages. To solve these problems on a CDC 3300, an exterior penalty method is used. In the final section some runs are tabulated.

Alkalmazott Matematikai Lapok 4 (1978) 245—270

CSOMAGKAPCSOLT SZÁMÍTÓGÉPHÁLÓZATOK TERVEZÉSEKOR FELMERÜLŐ OPTIMALIZÁLÁSI FELADATOK

SZ. TURCHÁNYI PIROSKA
Budapest

A dolgozat célja, hogy részletesen bemutasson két optimalizálási feladatot, melyek megoldása számítógéphálózatok tervezésekor alapvető jelentőségű és fontosságú. Feltételezzük, hogy az olvasó számára a számítógéphálózatok problémaköre nem ismeretlen, de az érthetőség megkönnyítésére az első fejezetben összefoglaljuk az osztott számítógéphálózat felépítését és működési elveit, a második fejezetben pedig vázoljuk az átlagos késési idő meghatározását. A továbbiakban a nemlineáris célfüggvényű *multicommodity* folyamfeladatok feltétel nélküli minimalizálására alkalmas *Flow Deviation módszert* ismertetjük. Megmutatjuk, hogyan oldható meg az optimális útképzési feladat és ennek kapacitástervezéssel bővített változata FD módszerrel.

1. Számítógéphálózatokról röviden

A számítástechnika mind szélesebb körű alkalmazása és a számítógépek jobb kihasználtságának érdekében olyan lehetőségek megteremtésére van szükség, melyek viszonylag egyszerű, a felhasználó számára kényelmes hozzáférést biztosítanak egy vagy akár több számítógéphez.

Első lépésként kifejlesztették a *time-sharing* rendszert: az egy számítógép köré épülő terminálhálózatot.

Következett a számítógépekből álló hálózat kifejlesztése. Ebben több, programfeldolgozást végző ún. erőforrás számítógép áll egymással kapcsolatban oly módon, hogy egy felhasználó igénye szerinti bármelyikhez hozzáférhessen.

Tekintsük át röviden a számítógéphálózatokkal — SZGH — szemben támasztott követelményeket:

— egy SZGH legyen gyors (az adatátviteli idő legyen rövid),
— bármely két számítógép (ill. ezek termináljai) között lehessen kapcsolatot létesíteni,
— a SZGH legyen megbízható (automatikus hibafeltárás és hibajavítással),
— az összekötő vonalak legyenek nagy kapacitásúak,
— az erőforrások kihasználtsága minél jobb legyen,
— természetesen minél alacsonyabb költséggel üzemeljen a hálózat.

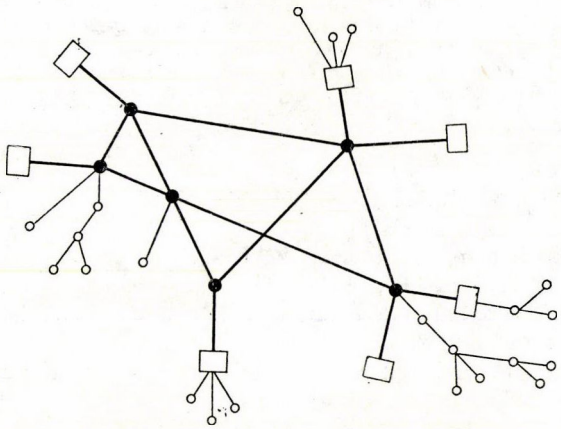
Az SZGH fejlesztésének főbb céljait a hagyományos számítógép funkciók szerint két csoportba oszthatjuk:

A processzálás jellegű alkalmazásokkal kapcsolatos célok:
— nagy teljesítményű, gyors távfuttatás,
— különböző erőforrások (különböző számítógépek software, illetve hardwarejének, különböző programcsomagjainak) elérése,
— azonos típusú erőforrásszámítógépek közötti job-elosztás.

A tárolás jellegű alkalmazásokkal kapcsolatos legfontosabb célok:
— gyors, olcsó adatbázis lekérdezés és módosítás (egészségügyi adatbankok, megrendeléseket nyilvántartó rendszerek stb. esetében),
— igény szerinti erőforrásszámítógép adatbázisainak elérése (pl. egy utazási iroda használhat több helyfoglaló rendszert).

Dolgozatomban osztott számítógéphálózat matematikai modelljét fogom ismertetni.

Egy osztott számítógéphálózat funkcionálisan két részből áll: a programfeldolgozást végző erőforrás számítógépekből és a hozzájuk tartozó helyi és távoli terminálokból, valamint az összeköttetést megvalósító adatátviteli hálózatból *(communication/data network)*, mely utóbbin tágabb értelemben — nemcsak a technikai berendezéseket, hanem minden olyan hardwaret, softwaret és matematikai módszert értünk, melyek az adattovábbítást hivatottak biztosítani.



1. ábra

Fekete korongok és vastag vonalak jelölik az adatátviteli hálózat részeit: a csomópontok ún. kapcsoló számítógépek *(switching-computer)*, az élek nagysebességű — 50—200 kbit/sec — csatornák. Az erőforrás számítógépek — világos négyzetekkel jelölve — a kapcsoló számítógépeken keresztül kötődnek az adatátviteli hálózathoz, szintén nagysebességű csatornákkal. Terminálok — világos körök — is tartozhatnak hozzájuk, de egy terminál közvetlen összeköttetésben is állhat az adatátviteli hálózattal, kis sebességű vonalak segítségével.

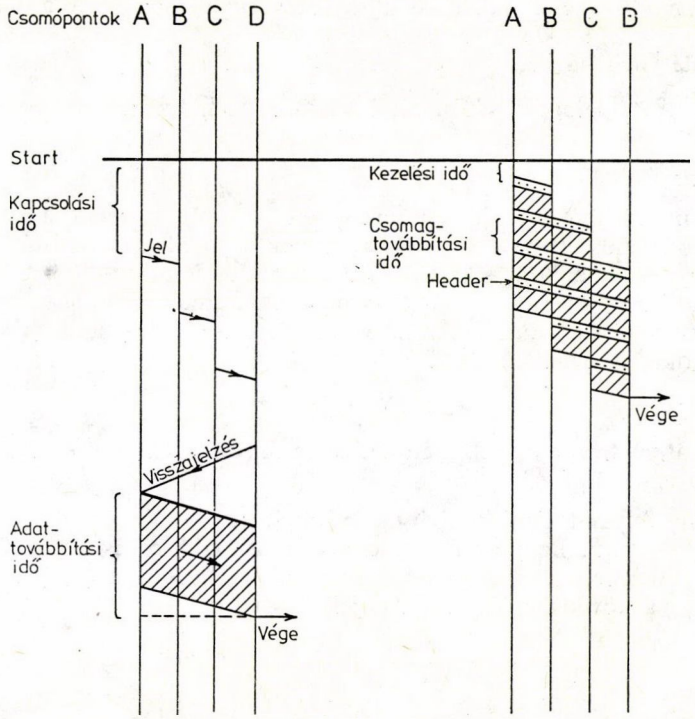
A teljes SZGH egy bizonyos csomópontjában üzenetek jelentkeznek megadott csomópontbeli számítógéphez való eljuttatás céljából. Így egy üzenet szempontjából egy csomópont lehet feladóhely (ez terminál v. számítóközpont), továbbítóhely (kapcsoló számítógép), és rendeltetési hely (számítóközpont).

Az adatátviteli rendszer alapján a hálózat lehet vonalkapcsolt *(circuit/line switching)* vagy csomagkapcsolt *(message/packet-switching)*.

Vonalkapcsolt hálózatban a kapcsoló számítógépek nem rendelkeznek tárolókapacitással, így egy üzenet eljuttatása a feladóhelyről a rendeltetési helyre csak úgy lehetséges, ha a két hely közötti teljes vonal szabad. A feladóhely egy speciális jelet

bocsát ki, amely valamilyen úton eljut a rendeltetési helyre, miközben lefoglalja az utat alkotó összes csatornát. Ezután a rendeltetési helyről visszajelzés érkezik, s megkezdődhet a tényleges üzenettovábbítás. A felhasznált csatornák csak akkor szabadulnak fel, amikor a továbbítás teljesen befejeződött, függetlenül attól, hogy az adott úton az üzenet melyik csatornán tart éppen. Nagymennyiségű adat folyamatos átvitele esetén a vonalkihasználtság jó, egyébként igen alacsony.

A másik fajta adatátviteli rendszerben minden kapcsoló számítógép képes üzenet tárolására, így egy üzenet továbbításához adott időben csak egy csatorna lefoglalása szükséges. Az üzenet először a feladóhelytől a legközelebbi kapcsolópontig halad, s mikor ide teljesen megérkezett, hibaellenőrzés után továbbítják az útvonalában következő kapcsolópont felé. Ha a két kapcsolópont közötti csatorna foglalt, az üzenet „várakozik", sőt „sorbaáll". Ez így folytatódik, míg a rendeltetési helyre nem ér. Az üzenet lehet a teljes adathalmaznak megfelelő karaktersorozat, de lehet ennek csak (fix) hosszúságú része. Így az angol nyelvű irodalomban *message switching,* illetve *packet switching* jelzővel illetik a hálózatot. Magyarul: üzenet, illetve csomagkapcsolt SZGH. Látjuk tehát, hogy itt az üzenetkezelés több adminisztrációt kíván a hálózat minden pontjában, mint a vonalkapcsolásos rendszerben, hiszen pl. az üzeneteket el kell látni a feladó- és rendeltetési helyre vonatkozó információval *(„header"),* ami különösen *packet switching* esetén — növeli az üzenethosszt; ugyancsak *packet switching* esetén a teljes üzenet *(message)* csomagokra



2. ábra. *Üzenettovábbítás*
a) *vonalkapcsolt hálózatban,* b) *csomagkapcsolt (packet switching) hálózatban*

bontása, majd a rekonstruálás is többlet feladat. Óriási előny viszont, hogy a csatornák állandóan továbbításra kész állapotban vannak, egy csatornán egymás után haladhatnak különböző feladó- és rendeltetési helyű üzenetek. E hálózat másik elnevezése, mely működéséről többet árul el: *store and forward* — raktározd és továbbítsd — röviden S/F hálózat.

2. Csomagkapcsolt számítógéphálózat matematikai programozási modellje

Mivel az SZGH-t az üzenetforgalom (útvonal, az üzenet hálózatban töltött ideje) szempontjából fogjuk vizsgálni, elég az adatátviteli hálózatra szorítkoznunk.

Az adatátviteli hálózat matematikailag egy

$G=(N, A)$ véges gráf (szükség esetén az éleknek kettős irányítást adva),

ahol

N: a csúcsok halmaza,
a csúcsok kapcsolószámítógépek, számuk: n,

A: az élek halmaza,
az élek a kapcsolószámítógépeket összekötő nagysebességű duplex csatornák, számuk: b.

Legyen $r_{ij} \geqq 0$ az i-edik csúcsból a j-edik csúcsba továbbítandó üzenetek átlagos mennyisége (bit/sec) $i=1, \ldots, n$ $j=1, \ldots, n$.

Az adatátviteli hálózat átlagos üzenetforgalmát a $\Gamma=((r_{ij}))$ üzenetmátrixszal és a $\gamma=\sum_{i=j}^{n} \sum_{j=1}^{n} r_{ij}$ átlagos összforgalmi értékkel jellemezzük.

Út a gráfban:

Az i-edik csúcsból a j-edik csúcsba vezető π_{ij} útnak egymáshoz csatlakozó, hurokmentes élrendszert nevezünk. Ha az i-edik és j-edik csúcs között több π_{ij} út létezik, az r_{ij} igény eljuttatására több lehetőség van, és az utak közötti különbségtevés az utak hossza szerint történik. Az út hosszát esetenként másként definiáljuk. Igen egyszerű metrika, ha az útban levő élek számát vesszük, de lényegesen bonyolultabb metrikát is definiálhatunk.

Folyam a gráfban:

Az r_{ij} üzenetmennyiségek szétosztását (a feladóhely és a rendeltetési hely közötti utakra) egy

$\Phi = \{f_{kl}^{ij}\}$ folyamfüggvénnyel írjuk le, ahol
f_{kl}^{ij} jelenti, hogy az i-edik csúcsból a j-edik csúcsba továbbítandó r_{ij} üzenetnek mekkora mennyiségét továbbítjuk a (k, l) élen. Tehát Φ a következő feltételekkel jellemezhető:

$$\sum_{k=1}^{n} f_{kl} - \sum_{k=1}^{n} f_{lm} = \begin{cases} -r_{ij}, & \text{ha} \quad l = i \\ +r_{ij}, & \text{ha} \quad l = j \\ 0, & \text{egyébként} \end{cases} \forall_{i,j}, \tag{2.1}$$

$$f_{kl}^{ij} \geqq 0, \quad i, j, k, l = 1, \ldots, n. \tag{2.2}$$

A (2.1), (2.2) feltételeknek eleget tevő Φ folyamfüggvényt *multicommodity folyamnak* (m.c.f.) nevezzük. Egyszerűbb, ha f_{kl}^{ij} helyett f_t^{ij}-vel jelöljük az élenkénti folyamértékeket, $t=1, \ldots, b$.

Legyen

$$\tilde{f}^{ij} = (f_1^{ij}, f_2^{ij}, \ldots, f_b^{ij}) \tag{2.3}$$

az r_{ij} üzenetmennyiséget i-ből j-be juttató (i, j) *elemi folyam.*

Ekkor az f_t^{ij} komponens ennek a t-edik élre jutó részét jelenti. Legyen továbbá

$$\tilde{f} = \sum_{i=1}^{n} \sum_{j=1}^{n} \tilde{f}^{ij} \tag{2.4}$$

a hálózatbeli *teljes folyam.* (Ez nem azonos a Φ m.c. folyammal.)

Egy Γ üzenetmátrixnak megfelelő m.c. folyam (röviden: Γ-folyam) tehát kijelöli minden egyes r_{ij} üzenet útját az i-edik és a j-edik csomópont között.

Φ nemelágazó m.c. folyam: ha az f^{ij} elemi folyam egyetlen π_{ij} úton viszi az r_{ij} üzenetet i-ből j-be minden (i, j) pontpárra.

Φ elágazó m.c. folyam: ha az egyes r_{ij} üzenetmennyiségek számára π_{ij}^k, $k=1, \ldots, K_{ij}$ számú út áll rendelkezésre, és a π_{ij}^k utakon $\alpha_{ij}^k \cdot r_{ij}$ rész halad, ahol $\sum_{k=1}^{K_{ij}} \alpha_{ij}^k = 1$, minden i, j pontpárra. Ekkor $f_t^{ij} = \sum_{k=1}^{K_{ij}} \sum_{t \in \pi_{ij}^k} \alpha_{ij}^k r_{ij}$, $t=1, \ldots, b$.

Φ legrövidebb folyam: ha olyan $\tilde{f}^{ij}$ elemi folyamokból áll, ahol az r_{ij} mennyiségek a legrövidebb π_{ij} utakon „haladnak".

Ugyanezek az elnevezések alkalmazhatók az $\tilde{f}$ teljes (Γ-) folyamra is.

A hálózatbeli matematikai programozási feladatok a Φ m.c. folyam valamely P függvényének optimalizálását jelentik. Csak olyan feladatokkal fogunk foglalkozni, ahol elegendő a $P(\Phi)$ helyett $P(\tilde{f})$ függvényt venni.

$P(\tilde{f})$ általában időfüggvény. Ha már van útképzési algoritmusunk az $\tilde{f}$ Γ-folyam előállítására, megbecsülhető, illetve kiszámolható egy üzenet hálózatban eltöltött idejének várható értéke. Ezt részletezzük a következő fejezetben.

A feladatokban az $\tilde{f}$ Γ-folyamra kapacitásfeltételt is elő kell írnunk:

$$\sum_{i=1}^{n} \sum_{j=1}^{n} f_t^{ij} < C_t, \quad t = 1, \ldots, b,$$

ahol az élekhez rendelt kapacitásértékekben a csatornák sebességét és a csomópontokbeli kapcsolószámítógépek üzenetkezelési sebességét is figyelembe vesszük. Az így értelmezett kapacitásvektorhoz gyakran egy konstrukciós költségfeltétel is járul:

$$\sum_{t=1}^{b} d_t(C_t) \leqq D.$$

Egy csomagkapcsolt számítógéphálózat tervezésekor felmerülő főbb matematikai problémák:

a) az üzenetek *header*jének optimalizálása
(információ elméleti téma, pl. [9]),
b) útképzési algoritmus kidolgozása
(matematikai programozási téma, pl. [2], [3], [5], [6], [10], [12]—[15]),

c) egy üzenet hálózatban eltöltött ideje várható értékének kiszámítása (tömegkiszolgálási téma, pl. [4], [6], [10], [11], [12]),
d) a hálózatban töltött idő várható értékének minimalizálása (matematikai programozási téma, pl. [1], [2], [4], [5], [12]—[15]),
e) a hálózat topológiájának optimális tervezése, pl. [12].

A *b)* és *d)* problémákat általában együtt oldják meg. A feladatokat öt alaptípusba soroljuk: (T minden esetben a hálózatban töltött idő várható értékét jelöli).

(i) *Kapacitástervezési feladat* (CA *problem*).
Adott a $G=(N, A)$ gráf,
Γ üzenetmátrix,
D költségkorlát
és a kapacitásfeltételnek eleget tevő m.c. folyam. Minimalizálandó

T a $\tilde{C}$ kapacitásvektor függvényeként, feltéve, hogy $\sum_{t=1}^{b} d_t(C_t) \leqq D$.

(ii) *Útképzési feladat* (FA *problem*) (prioritás nélküli üzenetek számára).
Adott a $G=(N, A, \tilde{C})$ gráf,
Γ üzenetmátrix.
Minimalizálandó

T az $\tilde{f}$ teljes folyam függvényeként,

feltéve, hogy

$$\sum_{i=1}^{n} \sum_{j=1}^{n} f_t^{ij} \leqq C_t, \quad t = 1, \ldots, b.$$

(iii) Az előbbiek kombinációja:
Kapacitástervezési és útképzési feladat (CFA *problem*).
Adott a $G(N, A)$ gráf,
Γ üzenetmátrix,
D költségkorlát.

Minimalizálandó

T az $\tilde{f}$ és $\tilde{C}$ függvényeként,

feltéve, hogy

$$\sum_{i=1}^{n} \sum_{j=1}^{n} f_t^{ij} \leqq C_t, \quad t = 1, \ldots, b,$$

és

$$\sum_{t=1}^{b} d_t(C_t) \leqq D.$$

(iv) *Topológia és kapacitástervezés, valamint optimális útképzés feladata* (TCFA *problem*).
Adott a D költségkorlát.
Minimalizálandó

T a $\tilde{C}, \tilde{f}$ és a $G=(N, A)$ függvényeként,

feltéve, hogy

$$\sum_{t=1}^{b} d_t(C_t) \leqq D.$$

(v) *Kapacitástervezés és útképzés különböző prioritású üzenetek esetében* („CPFA *problem*").

Megjegyzés:

Ezek a feladatok kevés változót képesek figyelembe venni, megoldásuk mégis igen bonyolult. A bonyolultság nagymértékben függ a költségfüggvénytől: legegyszerűbb eset a lineáris költségfüggvényé, legnehezebb a diszkrété. Az algoritmusok gyakran tartalmaznak heurisztikus lépéseket, főként a (iii)—(v) feladatok esetében.

Az útképzési feladatok két csoportba oszthatók.

Tervezéshez jól használhatók a determinisztikus v. fix technikák, amikor minden (i, j) csomópontpárra előre kijelölik az r_{ij} üzenetek útját. A hálózat üzemeltetéséhez, de már szimulációjához is az adaptív algoritmusok előnyösebbek, melyek csomópontról csomópontra határozzák meg egy üzenet útját.

Multicommodity folyam problémák egzakt megoldhatóságáról kitűnő áttekintést kapunk [1]-ben.

3. Egy üzenet hálózatban töltött idejének kiszámítása

Mint láttuk, csomagkapcsolt számítógéphálózatban az adatátviteli hálózat kapcsolószámítógépei különböző manipulációkat végeznek egy üzenettel. Ennek következtében egy csomópontban általában több üzenet várakozik, míg egynek a feldolgozása, továbbítása megtörténik. Tehát a hálózat egy igen bonyolult tömegkiszolgáló rendszert is jelent.

Vizsgáljunk először egy csomópontot, mint egy csatornás tömegkiszolgáló reodszert.

Legyen $A(x)$ az üzenetek beérkezése között eltelt idő eloszlásfüggvénye,

$B(x)$ a csomópontbeli kiszolgálási idő eloszlásfüggvénye (kiszolgálási idő: egy üzenet feldolgozása és továbbítása),

λ az üzenetek érkezésének gyakorisága [messg/sec],

$\bar{t}$ egy üzenet átlagos kiszolgálási ideje [sec/messg],

$\frac{1}{\mu}$ az átlagos üzenethossz [bit/messg],

$C=\frac{1}{\mu \bar{t}}$ kapacitás: üzenetfeldolgozási, illetve továbbítási sebesség [bit/sec].

Matematikailag rendkívül leegyszerűsíti a rendszer leírását, ha feltesszük, hogy A és/vagy B exponenciális eloszlásúak. Ennek megfelelően több eset lehetséges:

1. eset

(i) Az érkezések között eltelt idők független, egyforma exponenciális eloszlások $(A(x)=1-e^{-\lambda x})$, tehát az érkezések *Poisson-folyamatot* alkotnak.

(ii) A kiszolgálási idők függetlenek, egyforma exponenciális eloszlásúak $(B(x)=1-e^{-\frac{x}{\bar{t}}})$.

Ha a kiszolgálás az érkezés sorrendjében történik, azaz a hálózatban nincs prioritása egy üzenetnek sem, akkor az átlagos kiszolgálási idő várható értéke elemi módon kiszámolható [10], [11]:

$$T=\frac{\bar{t}}{1-\lambda\bar{t}}, \quad \text{vagy} \quad T=\frac{1}{\mu C-\lambda}, \tag{3.1}$$

ahol T az üzenetnek a kiszolgáló csomóponton eltöltött teljes idejét jelenti, amely két részre bontható:

a kiszolgáló foglaltsága miatt esetleg szükséges várakozási időre, valamint a tényleges kiszolgálás idejére. Ennek a felbontásnak megfelelően a T teljes várakozási idő várható értéke így írható:

$$T=\frac{\lambda\bar{t}^2}{1-\lambda\bar{t}}+\frac{1}{\mu C}. \tag{3.2}$$

2. eset

Az érkezések ismét *Poisson-folyamatot* alkotnak, de a kiszolgálási idők tetszőleges eloszlásúak. Ekkor a *Pollaczek—Hincsin-formulából*

$$T=\frac{\lambda s}{2(1-\lambda\bar{t})}+\frac{1}{\mu C}, \tag{3.3}$$

ahol $s=\int x^2\,dP(x)$, a kiszolgálási idő második momentuma.

3. eset

$A(x)$ is, $B(x)$ is tetszőleges eloszlásfüggvények. Ekkor nem ismeretes zárt formula T meghatározására, csupán egy igen jó felső korlát [9]:

$$T\leqq\frac{\lambda(\sigma_1^2+\sigma_2^2)}{2(1-\lambda\bar{t})}+\frac{1}{\mu C}, \tag{3.4}$$

ahol σ_1^2, ill. σ_2^2 az $A(x)$, ill. $B(x)$ eloszlások szórásnégyzetei.

Tekintsük most a teljes hálózatot.

Poisson-érkezési folyamat és exponenciális kiszolgálás esetén, ha az egyes csomópontok, mint tömegkiszolgáló-rendszerek függetlenek, egy üzenetre a hálózatban eltöltött idő kiszámítását a hálózat dekomponálásával — csomópontokra történő lebontásával — végezhetjük.

A csomagkapcsolt hálózat azonban másként működik: a csomópontok egymástól nem függetlenek, s mivel egy üzenet hossza a továbbítás során nem változik, nem változik a csomópontonkénti kiszolgálási idő eloszlása sem.

Ennek ellenére, ha az üzenetek hosszát minden egyes csomópontban valószínűségi változónak tekintjük, (ez sok pontból álló hálózatban, ahol egy csomópontba

több irányból érkeznek és több irányba futnak ki üzenetek, nem is olyan irreális feltételezés) a függetlenség és *Poissonitás* feltételezésével elméleti úton számított késési idő nagyjából megegyezik a szimulációs módszerekkel kapott késési idővel [10], [12]. Ez az ún. *Kleinrock-féle hipotézis*. Ennek alapján kiszámítható az egységnyi üzenet által a hálózatban töltött teljes idő T várható értéke:

Ha $\Gamma=((r_{ij}))$, $i=1, \ldots, n$, $j=1, \ldots, n$ az üzenetmátrix és $\gamma=\sum_{i=1}^{n}\sum_{j=1}^{n} r_{ij}$ a hálózat átlagos összforgalma, akkor a csomópontonkénti lebontás alapján

$$T = \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{r_{ij}T_{ij}}{\gamma}, \tag{3.5}$$

ahol T_{ij} az i-edik csomópontból a j-edikbe küldött üzenet várakozási idejének várható értéke.

Kényelmesebb, ha a várakozási időket nem a csúcsoknál, hanem a csatornákon vesszük figyelembe:

T_t jelöli azt az időt, melyet egy üzenet várakozik, míg a t-edik csatorna egyik végpontjából hibátlanul eljut a másik végpontba. Determinisztikus, illetve determinisztikus alternatív útképzési algoritmus használata esetén a (3.5)-beli T idő várható értéke [12]:

$$T = \sum_{t=1}^{b} \frac{\lambda_t}{\gamma} T_t, \tag{3.6}$$

ahol λ_t a t-edik csatornára jutó átlagos üzenetmennyiség.

Megjegyzések:

1) Érdekességképpen megemlítjük a következő formulát

$$T = \sum_{t=1}^{b} \left(\frac{\lambda_t}{\gamma} T_t^L\right)^{\frac{1}{L}},$$

ahol L alkalmasan választott hatványkitevő, azaz T az élenkénti késések hatványközepe, amely nagy L-ekre érzékenyebb az éleken előforduló legnagyobb késésekre.

A T idő ilyen módon történő számolása a következő esetben hasznos:

Előfordulhat, hogy egy bizonyos csatornán az átlagos késés nagyságrendekkel nagyobb, mint a többin, de ha a forgalom kicsi, ez a jelenség az előbbi módon számított késési időre (3.6) nincs hatással, ennek következtében a felhasználó „rosszul jár" [6]. Hatványközép alkalmazásával ez a negatívum csökkenthető.

2) A T_t idő kiszámítása T_{ij} időéhez hasonló, és többféle pontossággal történhet. Szokás még a (3.6)-beli T időt is tovább finomítani:

$$T = K + \sum_{t=1}^{b} \frac{\lambda_t}{\gamma}(T_t + P_t + K),$$

ahol K a csomóponti processzálási idő,

P_t a t-edik csatornán való tényleges áthaladás ideje, amely azonban csak ezer km-es nagyságrendű földrajzi távolságoknál veendő figyelembe.

4. Két optimalizálási feladat részletes megfogalmazása

Célunk, hogy a 2. fejezetben ismertetett öt alapfeladat közül az útképzési és az útképzés-kapacitástervezési feladatra mutassunk egy igen hatásos megoldási módszert.

Tekintsük a $G=(N, A)$ n csomópontú, b élű hálózatot, a $\Gamma=((r_{ij}))$ $i=1, \ldots, n$ $j=1, \ldots, n$ üzenetmátrixszal és a $\tilde{C}=(C_1, C_2, \ldots, C_b)$ kapacitásvektorral.

Feltesszük, hogy az i-edik csomópontba r_{ij} paraméterű *Poisson-folyamat* szerint érkeznek üzenetek a j-edik csomópont felé történő továbbítás céljából (tehát az i-ből j-be tartó üzenetek mennyiségének várható értéke r_{ij}) és a t-edik élen (csatornán) a kiszolgálási idő eloszlása független C_t paraméterű exponenciális eloszlás $\left(\text{tehát a kiszolgálási idő várható értéke } \frac{1}{C_t}\right)$.

Ekkor a *Kleinrock-féle hipotézis* alapján a t-edik élre $f_t=\sum_{i=1}^{n}\sum_{j=1}^{n} f_t^{ij}$ paraméterű *Poisson-folyamat* szerint érkeznek az üzenetek (lásd (2.1)—(2.3)), így a t-edik élen a várakozási idő várható értéke

$$T_t=\begin{cases}\dfrac{1}{C_t-f_t}, & \text{ha } f_t < C_t,\\ \infty, & \text{különben,}\end{cases} \tag{4.1}$$

tehát ugyancsak a *Kleinrock-féle hipotézis* alapján a teljes hálózatban

$$T=\sum_{t=1}^{b}\frac{f_t}{\gamma}T_t=\frac{1}{\gamma}\sum_{t=1}^{b}\frac{f_t}{C_t-f_t}, \tag{4.2}$$

hacsak $f_t < C_t$, $t=1, \ldots, b$.

Pontosabb meghatározásra most nincs szükségünk.

Most részletesen felírjuk azt a két feladatot, mellyel foglalkozni fogunk.

Jelölje F a (2.1)—(2.4)-ben definiált folyamok halmazát.

Optimális útképzés

Adott $G=(N, A)$ hálózat (n csúcs, b él),
$\Gamma=((r_{ij}))$ üzenetmátrix, $i=1, \ldots, n$, $j=1, \ldots, n$,
$\tilde{C}=(C_1, \ldots, C_b)$ élenkénti kapacitás.

$$\left\{\begin{array}{l}\text{Minimalizálandó}\\ \qquad T(\tilde{f})=\dfrac{1}{\gamma}\displaystyle\sum_{t=1}^{b}\frac{f_t}{C_t-f_t},\\ \text{feltéve, hogy}\\ \qquad \tilde{f}\in F\\ \qquad \tilde{f}\leqq \tilde{C}, \quad \text{azaz } f_t\leqq C_t, \quad t=1, \ldots, b.\end{array}\right. \tag{4.3}$$

Optimális útképzés és csatornakapacitás-tervezés

Adott $G = (N, A)$ hálózat (n csúcs, b él),
$\Gamma = ((r_{ij}))$ üzenetmátrix,
$d(C) = \sum_{t=1}^{b} d_t(C_t)$ költség kapacitás-függvény,
D = költségkorlát.

(4.4) Minimalizálandó

$$T(\bar{C}, \bar{f}) = \frac{1}{\gamma} \sum_{t=1}^{b} \frac{f_t}{C_t - f_t},$$

feltéve, hogy

$$\bar{f} \in F,$$

$$f_t \leqq C_t, \quad t = 1, \ldots, b,$$

$$d(\bar{C}) \leqq D.$$

A (4.4) feladat megoldását $d(\bar{C}) = \sum_{t=1}^{b} d_t C_t$ lineáris költségkapacitás-függvény esetén tárgyaljuk, ezért először nézzük meg, hogyan redukálódik ekkor a probléma.

A feladat célfüggvényét először rögzített $\bar{f}$ mellett, $\bar{C}$-ben minimalizáljuk, a *Lagrange-multiplikátoros módszerrel* [7]. Ennek eredményeképpen

$$C_t = f_t + \frac{d - \sum_{t=1}^{b} f_t d_t}{d_t} \cdot \frac{\sqrt{f_t d_t}}{\sum_{j=1}^{b} f_j d_j},$$

melyet a célfüggvénybe visszahelyettesítve

$$T(\bar{C}, \bar{f}) = T(\bar{f}) = \frac{\left(\sum_{t=1}^{b} f_t d_t\right)}{\gamma\left(D - \sum_{t=1}^{b} d_t f_t\right)}.$$

Könnyen látható, hogy a (4.4) feladatbeli $f_t \leqq C_t$, $t = 1, \ldots b$ és $\sum_{t=1}^{b} d_t C_t \leqq D$ feltételek helyettesíthetők a

$$D - \sum_{t=1}^{b} d_t f_t \geqq 0 \quad \text{feltétellel.}$$

Így a következő feladathoz jutunk:

Optimális útképzés és csatornakapacitás-tervezés lineáris költségkapacitás-függvény esetén

Adott $G = (N, A)$ hálózat (n csúcs, b él),
$\Gamma = ((r_{ij}))$ üzenetmátrix,
$d(\tilde{C}) = \sum_{t=1}^{b} d_t C_t$ költségkapacitás-függvény,
D = költségkorlát.

$$(4.5)\quad \begin{cases} \text{Minimalizálandó} & T = \dfrac{\left(\sum_{t=1}^{b} \sqrt{f_t d_t}\right)^2}{\gamma\left(D - \sum_{i=1}^{b} d_t f_t\right)} \\ \text{feltéve, hogy} & \tilde{f} \in F \\ & D - \sum_{t=1}^{b} d_t f_t \geqq 0. \end{cases}$$

Megjegyzések:

1) A megengedett megoldások halmaza a (4.3) és a (4.5) feladatok mindegyikében konvex, zárt és korlátos.

$$F_1 = F \cap \{\tilde{f} | \tilde{f} \leqq \tilde{C}\},$$

$$F_2 = F \cap \{\tilde{f} | D - \sum_{t=1}^{b} d_t f_t \geqq 0\}.$$

Belátható [8], hogy az F, F_1 és F_2 halmazok extremális pontjai legrövidebb folyamok (l. 2. fejezet).

2) Mind a (4.3), mind a (4.5) feladat esetében, ha egy megengedett megoldás ($\tilde{f} \in F_1$, ill. $\tilde{f} \in F_2$) a feltételhalmaz határához közeledik, akkor a célfüggvény értéke végtelenhez tart, azaz a (4.3) feladat célfüggvénye az $\tilde{f} \leqq \tilde{C}$ feltételt, a (4.5) feladat célfüggvénye a $D - \sum_{t=1}^{b} d_t f_t \geqq 0$ feltételt büntetőfüggvényként magában foglalja. Tehát bármelyik feladatban, ha valamilyen módon találtunk egy induló megengedett megoldást, a célfüggvény feltételnélküli minimalizálására térhetünk át.

3) Mivel a megengedett megoldások halmaza konvex, zárt és korlátos, és mivel (4.3) célfüggvénye konvex, (4.5) célfüggvénye kvázikonkáv (l. 6.1. fejezet), azért mindkét feladatra igaz, hogy ha van megengedett megoldás, akkor van optimális megoldás is. Megengedett megoldás keresése csak a (4.3) feladat esetében jelent problémát.

5. A „Flow Deviation" — FD-módszer

Legyen $F=\{\tilde{f}|\tilde{f}$ m.c. folyam (2.1)—(2.4) szerint$\}$ és P folytonosan differenciálható függvény az F halmazon.

5.1. DEFINÍCIÓ: $\tilde{f}\in F$ stacionárius pontja a P függvénynek, ha bármely kicsi $\delta\tilde{f}$ megváltoztatás esetén mindig teljesül $P(\tilde{f}+\delta\tilde{f})\geqq P(\tilde{f})$, feltéve, hogy $\tilde{f}+\delta f\in F$.

Mivel P lokális és globális minimumhelyei stacionárius pontok, szükséges és elégséges feltételt keresünk ahhoz, hogy $\tilde{f}\in F$ stacionárius pontja legyen P-nek.

5.2. DEFINÍCIÓ: *„Flow Deviation"*. Legyen $\tilde{f}\in F$ rögzített, $\tilde{v}\in F$ tetszőleges pont, és $0\leqq\lambda\leqq 1$.

Tekintsük a következő konvex kombinációt:

$$\tilde{f}' = \tilde{f}+\lambda(\tilde{v}-\tilde{f}), \quad (\tilde{f}'\in F, \text{ hiszen } F \text{ konvex}).$$

Legyen $0<\hat{\lambda}<\varepsilon$, ahol $\varepsilon\ll 1$. Ekkor a *Lagrange-középérték tétel* alapján

$$P(\tilde{f}')-P(\tilde{f}) = \hat{\lambda}\nabla P(\tilde{g})(\tilde{v}-\tilde{f}),$$

ahol

$$\tilde{g} = \tilde{f}+\theta\lambda(\tilde{v}-\tilde{f})\in F, \quad 0\leqq\theta\leqq 1,$$

és

$$\nabla P(\tilde{g}) = \left(\frac{\partial P}{\partial f_1}, \ldots, \frac{\partial P}{\partial f_b}\right)\Bigg|_{\tilde{f}=\tilde{g}},$$

vagyis a P függvény gradiense a $\tilde{g}$ pontban.

De mivel $0<\hat{\lambda}<\varepsilon$, és P folytonosan differenciálható:

$$P(\tilde{f}')-P(\tilde{f}) \approx \hat{\lambda}\nabla\cdot P(\tilde{f})(\tilde{v}-\tilde{f}) = \lambda\sum_{t=1}^{b}\frac{\partial P}{\partial f_t}(v_t-f_t).$$

Jelöljük $\left.\frac{\partial P}{\partial f_t}\right|_{\tilde{f}} = l_t$ $(t=1, \ldots, b)$-vel a P függvény $\tilde{f}$ pontbeli parciális deriváltjait.

Mivel $\hat{\lambda}>0$, az 5.1. definíció szerint $\tilde{f}$ a P függvénynek akkor és csak akkor stacionárius pontja, ha

$$\sum_{t=1}^{b} l_t(v_t-f_t)\geqq 0, \quad \text{bármely } \tilde{v}\in F \text{ esetén.} \tag{5.1}$$

Emlékezve $\tilde{f}$-nek a (2.1)—(2.4)-ben adott definíciójára, $\tilde{f}$ elég kicsi megváltoztatását elérhetjük úgy is, hogy csak egy (tetszőleges rögzített) f^{ij} elemi folyammennyiséget változtatunk meg, azaz a P függvény változtatását csak $\tilde{f}$ koordinátánkénti megváltoztatása esetén vizsgáljuk, vagyis ha $\tilde{v}$ a következőképpen áll elő $\tilde{f}$-ből:

$$\tilde{v}^{ij} = \begin{cases} \tilde{f}^{ij}, & i\neq i_0, \quad j\neq j_0 \\ \tilde{v}^{i_0 j_0}, & (i_0, j_0) \text{ tetszőleges, rögzített } (i,j) \text{ pár} \end{cases}$$

akkor (5.1) fennállása esetén igaznak kell lennie az

$$\sum_{t=1}^{b} l_t(v_t^{ij}-f_t^{ij})\geqq 0 \tag{5.2}$$

egyenlőtlenségnek, mégpedig bármely $\tilde{v}\in F$ folyamra.

Megfordítva: (5.2) fennállása esetén (5.1) triviálisan következik, tehát az (5.1) és (5.2) feltétel ekvivalens.

E két feltételből következnek a

$$\min_{v \in F} \sum_{t=1}^{b} l_t v_t \geqq \sum_{t=1}^{b} l_t f_t, \tag{5.3}$$

illetőleg

$$\min_{v^{ij} \in F^{ij}} \sum_{t=1}^{b} l_t v_t^{ij} \geqq \sum_{t=1}^{b} l_t f_t^{ij} \quad i = 1, \ldots, n, \quad j = 1, \ldots, n \tag{5.4}$$

egyenlőtlenségek, melyek $\tilde{f} \in F$, ill. $\tilde{f}^{ij} \in F^{ij}$ miatt egyenlőséggel kell, hogy teljesüljenek.

Miért lényeges a két feltétel ekvivalenciája? A $\min_{v^{ij} \in F^{ij}} \sum_{t=1}^{b} l_t v_t^{ij}$ meghatározása igen egyszerű feladat:

Egy rögzített (i, j) párra a G hálózatbeli „minimális költségű f^{ij} folyam", avagy más szemlélettel az „$\{l_t,\ t=1, \ldots, b\}$ metrikában legrövidebb f^{ij} folyam" meghatározását jelenti.

Vagyis az (5.3) feltétel bal oldalát minimalizáló *multicommodity folyam* (5.4) alapján könnyen előállítható.

Az FD-módszer alapgondolata a következő:

Ha az f folyamot egyforma abszolút értékű vektorok irányában változtatjuk, a *Lagrange-középértéktétel* alapján úgy tudjuk a P függvényt a legjobban csökkenteni, ha a változtatás iránya a negatív gradiens irányával egyezik meg.

Mivel az F halmazon belül kell haladnunk, nem választhatjuk a negatív gradiens irányát.

Láttuk, hogy ha az (5.4) vagy (5.3) feltétel egyenlőséggel teljesül, a P függvény stacionárius pontjában vagyunk.

Ha a feltétel bal oldalát minimalizáló $\tilde{v}$ pontra nem teljesül az egyenlőség, akkor

$$\sum_{t=1}^{b} l_t v_t^{ij} < \sum_{t=1}^{b} l_t f_t^{ij}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, n, \tag{5.5}$$

illetve

$$\sum_{t=1}^{b} l_t v_t < \sum_{t=1}^{b} l_t f_t, \tag{5.6}$$

tehát $(\tilde{v} - \tilde{f})$ irányban haladva P értéke csökken, és az F halmazon belül ez a legnagyobb csökkenés iránya.

5.3. DEFINÍCIÓ:

$$\mathrm{FD}(\tilde{v}, \lambda)\colon\ F \to F \text{ operátor}$$

$$\mathrm{FD}(\tilde{v}, \lambda) \odot f = \tilde{f} + \lambda(\tilde{v} - \tilde{f}) = \tilde{f}',$$

ahol $\tilde{v}$ az $\{l_t\}$ metrikában legrövidebb m.c. folyam és λ $(0 \leqq \lambda \leqq 1)$ az az érték, melyre $P[(1-\lambda)\tilde{f} + \lambda\tilde{v}]$ minimális λ-ban.

Az FD-módszer főbb lépései

I. fázis

Meghatározunk egy $\tilde{f}^0 \in F$ induló megoldást. (Ennek módja a konkrét feladattól függ.)

II. fázis

1. Legyen $n=0$.
2. $\tilde{f}^{n+1} = \mathrm{FD}(\tilde{v}^n, \lambda^n) \odot \tilde{f}^n$,

   ahol $\tilde{v}^n$ az $\left\{ l_t^n = \dfrac{\partial P}{\partial f_t}\Big|_{\tilde{f}=\tilde{f}^n},\ t=1, \ldots, b \right\}$ metrikában legrövidebb m.c. folyam

$$\lambda^n: \min_{0 \leq \lambda \leq 1} P[(1-\lambda)\tilde{f}^n + \lambda \tilde{v}^n)$$

3. Ha $P(\tilde{f}^n) - P(\tilde{f}^{n+1}) < \varepsilon$, $\left(\text{vagy } \sum_{t=1}^{b} l_t(f_t^n - v_t^n) < \varepsilon' \right)$,

ahol $\varepsilon > 0$, ill. $\varepsilon' > 0$ tetszőleges rögzített hibakorlát, *akkor készen vagyunk*, ellenkező esetben $n=n+1$, és a 2. lépéstől folytatjuk az eljárást.

A módszer konvergens:

5.1. TÉTEL ([5]).

Ha a P függvény az F halmazon alulról korlátos, kétszer differenciálható

$$\frac{\partial P}{\partial f_k} \geqq 0 \quad \text{és} \quad \frac{\partial^2 P}{\partial f_j \partial f_k} \text{ felülről korlátos,} \quad j, k = 1, \ldots, b$$

továbbá P nem degenerált, azaz bármely $\tilde{f}_1 \neq \tilde{f}_2$ *stac.* pont esetén $P(\tilde{f}_1) \neq P(\tilde{f}_2)$, akkor az FD-módszer alkalmazható:

P értéke minden FD-iterációban csökken (nem nő)
a módszer stacionárius ponthoz vezet.

A P függvény első parciálisainak nem-negativitása annak a metrikának a nem-negativitását jelenti, mely szerinti legrövidebb folyamot egy FD-iteráción belül meg kell határoznunk. A nem-negativitási feltétel zárja ki a negatív ciklusok lehetőségét, s ez a feltétel a számítógéphálózat modelljeiben szereplő célfüggvényekre, mint látni fogjuk, teljesül.

Számunkra elég az 5.1. tételnek az a speciális esete, amikor P konvex függvény, és az egyszerűség kedvéért a bizonyítást is csak erre az esetre végezzük el.

6. Az FD-módszer konvergenciája

6.1. TÉTEL.

Ha P szigorúan konvex függvény az F halmazon, kétszer differenciálható és $\dfrac{\partial^2 P}{\partial f_j \partial f_k}$ korlátosak $j, k=1, \ldots, b$, akkor az FD-módszer P globális minimumhelyéhez konvergál.

Bizonyítás:

Feltéve, hogy találtunk $\tilde{f}^0$ induló megengedett megoldást, legyen $\{\tilde{f}^n\}$ az FD módszer által generált sorozat. Mivel

$$P(\tilde{f}^n) - P(\tilde{f}^{n+1}) \geqq 0,$$

azért a $\{P(\tilde{f}^n)\}$ sorozat monoton nem növekvő, a konvexitás miatt alulról korlátos is, tehát

$$\exists \lim_{n\to\infty} P(\tilde{f}^n) = P(\tilde{f}^*), \quad \tilde{f}^* \in F.$$

ÁLLÍTÁS: $\tilde{f}^*$ globális minimumhelye a P függvénynek, azaz $P(\tilde{f}^*) \leqq P(\tilde{f})$ bármely $\tilde{f} \in F$.

Bizonyítás:

Tegyük fel, hogy a $\{P(\tilde{f}^n)\}$ sorozat nem tart a P globális minimumához, azaz

$$\exists t > 0, \quad \text{hogy} \quad P(\tilde{f}^n) \geqq P(\tilde{f}^{\min}) + t \quad \text{bármely } n\text{-re.}$$

Ez szemléletesen azt jelenti, hogy a P függvényen nem jutunk lejjebb, mint a $P(\tilde{f}^{\min})$ feletti t magasságban levő sík — röviden „t-sík" — által a P-ből kimetszett görbe. Ez a görbe korlátos és konvex, hiszen $P(\tilde{f})$ szigorúan konvex. Tetszőleges rögzített $\tilde{f} \in F$ pontra, amelyre $P(\tilde{f})$ még a t-síkon, vagy a t-sík felett van, legyen

$$P(\lambda) = P(\mathrm{FD} \odot f) = P[\tilde{f} + \lambda(\tilde{v} - \tilde{f})].$$

Ekkor ismét a *Lagrange-középértéktételből* $\exists 0 < \xi < \lambda$,

$$P(\lambda) = P(0) + \lambda \left.\frac{dP}{d\lambda}\right|_{\lambda=0} + \frac{1}{2}\lambda^2 \left.\frac{d^2P}{d\lambda^2}\right|_{\lambda=\xi},$$

ahol $\left.\dfrac{d^2P}{d\lambda^2}\right|_{\lambda=\xi} \leqq M < \infty$ a feltevések miatt,

és

$$\left.\frac{dP}{d\lambda}\right|_{\lambda=0} = \sum_{k=1}^{b} \frac{\partial P}{\partial f_k}(v_k - f_k).$$

Ha $\tilde{f} \in F$ olyan, hogy $P(\tilde{f})$ a t-síkon van, és h jelöli $\tilde{f}$ és $\tilde{f}^{\min}$ távolságát, akkor a konvexitásból könnyen belátható, hogy

$$0 < \frac{t}{h_0} \leqq \frac{t}{h} \leqq \left.\frac{dP}{d\lambda}\right|_{\lambda=0},$$

ahol

$$h_0 = \max h < \infty.$$

Tehát

$$\left.\frac{dP}{d\lambda}\right|_{\lambda=0} \leqq -\frac{t}{h_0} = -\Delta < 0,$$

vagyis

$$\sum_{k=1}^{b} \frac{\partial P}{\partial f_k}(v_k - f_k) < 0$$

minden $\tilde{f} \in F$-re, melyre $P(\tilde{f}) = P(\tilde{f}^{\min}) + t$ és következésképpen azokra is, melyekre $P(\tilde{f}) \geqq P(\tilde{f}^{\min}) + t$. Így

$$P(\lambda) - P(0) \leqq -\Delta\lambda + \frac{1}{2}\lambda^2 M = \frac{M}{2}\left(\lambda - \frac{\Delta}{M}\right)^2 - \frac{\Delta}{2M^2}$$

és a becslés egyenletes olyan $\tilde{f} \in F$ pontokban, melyekre $P(\tilde{f}) \geqq P(\tilde{f}^{\min}) + t$.

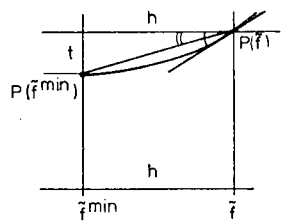
Tudjuk, hogy a λ lépéshossz nagysága $\min_{0\leqq\lambda\leqq 1} P(\lambda)$ meghatározásával történt, a fenti becslésben szereplő majoráló függvény minimum helye pedig $\lambda=\frac{\Delta}{M}\geqq 0$, és mivel Δ mindig választható úgy, hogy $\lambda=\frac{\Delta}{M}\leqq 1$ is teljesüljön, igaz lesz

$$P(\lambda)-P(0)\leqq -\frac{\Delta^2}{2M}=-\varepsilon<0.$$

Ez utóbbi megállapítás azt jelenti, hogy $P(\tilde{f})$ értéke minden egyes FD iterációban legalább ε-nal csökken, tehát

$$P(\tilde{f}^*)-P(\tilde{f}^n)=\sum_{l=n}^{\infty}(P(\tilde{f}^{l+1})-P(\tilde{f}^l))=-\infty,$$

bármely n-re, ami ellentmondás $\lim_{n\to\infty} P(\tilde{f}^n)=P(\tilde{f}^*)$-gal.



3. ábra

Megjegyzések:

1. Ha P nem szigorúan konvex függvény az F halmazon, akkor olyan (heurisztikus) módszert kell találnunk a globális minimum meghatározásához, mely nem vizsgálja meg az összes lokális minimumhelyét a P függvénynek.

2. Igen egyszerű az FD módszer alkalmazása szigorúan kvázikonkáv P függvényre.

6.1. DEFINÍCIÓ:

a) A P függvény kvázikonkáv az $\hat{f}\in F$ pontban, ha minden olyan $\tilde{f}\in F$-re, melyre $P(\tilde{f})\geqq P(\hat{f})$, fennáll:

$P(\hat{f})\leqq P[\hat{f}+\lambda(\tilde{f}-\hat{f})]$, ahol $0\leqq\lambda\leqq 1$

és $\hat{f}+\lambda(\tilde{f}-\hat{f})\in F$, ha F konvex, egyébként az $F\cap\{\hat{f}+\lambda(\tilde{f}-\hat{f})\}$ halmazra vonatkozik a feltétel.

b) A P függvény kvázikonkáv az F halmazon, ha minden $\tilde{f}\in F$ pontban kvázikonkáv.

c) Ha $\tilde{f}\neq\hat{f}$, $\lambda>0$ és $P(\hat{f})<P[\hat{f}+\lambda(\tilde{f}-\hat{f})]$,

akkor a függvényt szigorúan kvázikonkávnak nevezzük.

Szigorúan kvázikonkáv függvény lokális minimumhelyei az F halmazon az F extremális pontjai, azaz „legrövidebb" m.c. folyamok, ennek következtében az FD iterációkban a lépéshossz; $\lambda=1$ (vagyis mindig F extremális pontjain — legrövidebb m.c. folyamokon haladunk).

7. Az optimális útképzési feladat megoldása FD-módszerrel

A könnyeb áttekinthetőség kedvéért újra felírjuk a (4.3) feladatot:

Adott $G = (N, A)$ n csúcsú, b élű gráf,

$\Gamma = ((r_{ij}))$ $i = 1, \ldots, n$, $j = 1, \ldots, n$ üzenetmátrix,

$\tilde{C} = (C_1, \ldots, C_b)$ élenkénti kapacitás.

Minimalizálandó

$$T(\tilde{f}) = \frac{1}{\gamma} \sum_{t=1}^{b} \frac{f_t}{C_t - f_t}$$

feltéve, hogy

$\tilde{f} \in F$ és

$\tilde{f} \leqq \tilde{C}$, azaz $f_t \leqq C_t$, $t = 1, \ldots, b$.

Láttuk a 4. pontban, hogy a $T(\tilde{f})$ célfüggvény az $\tilde{f} \leqq \tilde{C}$ feltételt büntetőfüggvényként magában foglalja, ezért ha találtunk egy $\tilde{f} \leqq \tilde{C}$ megengedett megoldást, akkor ebből indítva az FD-iterációkat, már nem kell ezt a feltételt is szem előtt tartani.

$T(\tilde{f})$ konvex függvény, és a feltételhalmaz konvex halmaz, ezért ha a feladatnak van $\tilde{f}$ megengedett megoldása, akkor van $\tilde{f}^*$ optimális is, melyre viszont $f_t^* < C_t$, $t = 1, \ldots, b$, azaz $\exists \varepsilon > 0$, melyre

$$\tilde{f}^* \in F_{1,\varepsilon} = \{\tilde{f} \mid \tilde{f} \in F \quad \text{és} \quad f_t \leqq C_t - \varepsilon, \quad t = 1, \ldots, b\}.$$

Ezt azt jelenti, hogy elég kis ε esetén elég az $F_{1,\varepsilon}$ halmazon alkalmaznunk az FD-módszert. Ezen viszont a $T(\tilde{f})$ függvény kielégíti az (5.10)-beli konvergenciafeltételeket:

$T(\tilde{f})$ kétszer differenciálható, $\tilde{f} \in F_{1,\varepsilon}$ alulról korlátos függvény (konvex),

$$\frac{\partial T}{\partial f_t} = \frac{1}{\gamma} \frac{C_t}{(C_t - f_t)^2} \geqq 0, \quad t = 1, \ldots, b, \quad \tilde{f} \in F_{1,\varepsilon},$$

$$\frac{\partial^2 T}{\partial f_j \partial f_k} = \begin{cases} 0, & \text{ha} \quad j \neq k, \\ \dfrac{1}{\gamma} \dfrac{2C_k}{(C_k - f_k)^3} < \infty & \text{ha} \quad j = k \quad \text{és} \quad \tilde{f} \in F_{1,\varepsilon}, \quad j, k = 1, \ldots, b. \end{cases}$$

Algoritmus induló megoldás meghatározására:

Induló megoldásnak a Γ igénymátrixnak megfelelő olyan $\tilde{f}$ m.c. folyamot (röviden: Γ-folyamot) kell keresnünk, amely a kapacitásfeltételeket is kielégíti.

Olyan Γ-folyamot, amelytől nem kívánjuk meg a kapacitásfeltételek teljesülését, könnyű találnunk. Az alábbiakban vázolt módszer olyan iteráció, mely Γ-folyamok olyan $\tilde{f}^0, \tilde{f}^1, \ldots$ sorozatát határozza meg, melyek durván szólva, „egyre közelebb visznek" a kapacitásfeltételek teljesüléséhez.

1. Kiindulásként tekintsük az $\tilde{f}\equiv 0$ pontban számolt legrövidebb folyamot, azaz az

$$\left\{l_t = \left.\frac{\partial T}{\partial f_t}\right|_{f_t=0} = \frac{1}{\gamma}\frac{1}{C_t},\ t = 1, \ldots, b\right\}$$

metrikának megfelelő legrövidebb $\tilde{f}^0$ Γ-folyamot. Legyen $n=0$.

2. Ha már ismert az $\tilde{f}^n$ Γ-folyam, legyen

$$\sigma^n = \max_t \frac{f_t^n}{C_t}.$$

3. Az iteráció befejeződik, ha $\sigma^n < 1$.

4. Ha $\sigma^n \geqq 1$, akkor legyen

$$N_{n+1} = [1+\varepsilon(1-\sigma^n)]/\sigma^n,$$

ahol $0<\varepsilon<1$ alkalmasan rögzített konstans, és legyen

$$\tilde{f}^{n+1} = \frac{1}{N_{n+1}} \mathrm{FD} \odot (N_{n+1}\tilde{f}),$$

ahol az FD-operátorhoz a $\tilde{v}$ folyamot és λ lépéshosszt a szokott módon határozzuk meg.

Mivel $N_{n+1}\cdot\sigma^n < 1$, az $N_{n+1}\cdot\tilde{f}^n$ folyam, mely $(N_{n+1}\cdot\Gamma)$-folyam, kielégíti a kapacitásfeltételeket. Heurisztikusan belátható, hogy ha egy, a kapacitásfeltételeket kielégítő folyamra alkalmazzuk az FD-operátort, akkor az eredményül kapott folyam általában az előzőnél kevésbé használja ki a kapacitásokat, ami annak mellékeredménye, hogy az új folyam a $T(\tilde{f})$ célfüggvény értékét csökkenti. Ezért formálisan általában az várható, hogy $\sigma^{n+1}<\sigma^n$. Az új $\tilde{f}^{n+1}$ folyam pedig nyilván Γ-megengedett.

5. Az iteráció befejeződik, ha $\sigma^n<1$, azaz találtunk egy megengedett induló megoldást, illetve, ha megfelelően rögzített θ és δ pozitív tűrési állandók mellett egyszerre teljesül, hogy

$$\left|\sum_{t=1}^{b} l_t(v_t - N_n f_t^n)\right| < \theta,$$

és

$$|\sigma^n - \sigma^{n+1}| < \delta,$$

ez esetben az adott θ, δ tűrés mellett a feladatnak *nincs megengedett megoldása.* Egyébként a 2. lépéstől folytatjuk az eljárást.

Ha meghatároztunk egy induló megengedett megoldást, az (5. fejezetben ismertetett) FD-módszer II. fázisa következik, mely ennél a feladatnál elvezet az optimális megoldáshoz.

[5] szerint a még 21 csomópontból és 26 csatornából álló (az algoritmus során a csatornák számának kétszeresét kell venni, az irányítatlanság miatt) ARPA hálózatra az FD-módszerrel $r_{ij}=r=1{,}187$ kbit/sec igénymátrix esetén $T=0{,}2406$ sec volt a minimális idő. A FORTRAN nyelvű program IBM 360/91 számítógépen 30 sec alatt futott le.

8. Az FD-módszer gyorsítása

Az elágazó, illetve nem-elágazó multicommodity folyam fogalmát a 2. fejezetben már bevezettük.

Adódhat olyan feladat, melyben feltételként szerepel a megoldás nem-elágazó tulajdonsága. Más feladatokban egy nem-elágazó folyam igen jó közelítése lehet az optimális megoldásnak, s nyilván jelentősen gyorsul az FD-algoritmus, ha a feladatot eleve nem-elágazó folyamokra korlátozzuk.

Jelölje F^{nb} a nem-elágazó m.c. folyamok halmazát:

$$F^{nb} = \{\tilde{f} | \tilde{f} \in F, \tilde{f} \text{ nem-elágazó}\}.$$

A nem-elágazási feltételt diszkrét változók bevezetésével írhatjuk fel: minden i, j pontpár közötti f^{ij} elemi folyamok közül ki kell zárni azokat, melyek a π_{ij}^k utak kombinációira oszlanak el. Így azonban már tízes nagyságrendű csomópontból álló hálózatban is olyan nagy számú diszkrét feltételhez jutunk, melyek kezelése diszkrét programozási módszerekkel igen nehézkes. Folytonos módszerek pedig az F^{nb} diszkrét halmazon nyilván nem alkalmazhatók.

Bevezetjük a „nagy és kiegyensúlyozott" hálózat fogalmát és megmutatjuk, hogy egy ilyen hálózatban az FD-módszer speciális egyszerűsítésével haladhatunk kizárólag nem-elágazó folyamokon a nem-elágazó optimális megoldás felé, és ezzel az optimumot igen jól tudjuk közelíteni.

Tekintsük tehát a $G=(N, A)$ n csomópontú, b élű hálózatot a $\Gamma=((r_{ij}))$ $i=1, \ldots, n$, $j=1, \ldots, n$ üzenetmátrixnak megfelelő m.c. folyamokkal.

Jelölés: p_{ij} az (i, j) csomópontok között egy tetszőleges út hossza,
q_{ij} az (i, j) csomópontok közötti legrövidebb út hossza.

A nagy és kiegyensúlyozott hálózat jellemzése érdekében néhány mérőszámot definiálunk. Legyen

$$r = \frac{1}{(n-1)n} \sum_{i=1}^{n} \sum_{j=1}^{n} r_{ij}$$

a hálózatbeli átlagos üzenetmennyiség és

$$m = \max_{i,j} \frac{r_{ij}}{r}.$$

Nyilván $r \geqq \min_{i,j} r_{ij}$ és $m \geqq 1$. Tetszőleges $\tilde{f} \in F$ esetén, ha p_{ij} az f^{ij} elemi folyamhoz tartozó úthossz, akkor az átlagos $\bar{p}$ úthossz

$$\bar{p} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} r_{ij} p_{ij}}{n(n-1)r} \quad \text{és}$$

a legrövidebb folyamhoz tartozó átlagos úthossz ekkor

$$\bar{q} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} r_{ij} q_{ij}}{n(n-1)r}.$$

Vizsgáljuk meg, mit mondhatunk a hálózat élein a teljes folyam és az (i,j) elemi folyam arányáról, azaz vizsgáljuk az

$$\frac{f_t}{f_t^{ij}}, \quad t=1,\ldots,b \quad \text{értékeket,}$$

illetőleg ezek $\frac{1}{b}\sum_{t=1}^{b}\frac{f_t}{f_t^{ij}} = \overline{\left(\frac{f_t}{f_t^{ij}}\right)}$ átlagát (minden (i,j) pontpárra).

Mivel $f_t^{ij} \leqq mr$,

$$\overline{\left(\frac{f_t}{f_t^{ij}}\right)} \geqq \frac{1}{bmr}\sum_{t=1}^{b} f_t = \frac{1}{bmr}\sum_{i=1}^{n}\sum_{j=1}^{n} r_{ij}p_{ij} = \frac{r(n-1)n}{bmr}\bar{p} \geqq \frac{n(n-1)}{bm}\bar{q}.$$

Jelölje:

$$\frac{n(n-1)}{bm}\bar{q} = \frac{1}{\eta}.$$

8.1. DEFINÍCIÓ: A $G=(N,A)$ n csomópontú, b élű hálózatot a $\Gamma=((r_{ij}))$ üzenetmátrixnak megfelelő m.c. folyamokkal nagy és kiegyensúlyozott hálózatnak nevezzük, ha $\eta \ll 1$. Ilyen hálózatban ugyanis $\overline{\left(\frac{f_t}{f_t^{ij}}\right)} \geqq \frac{1}{\eta}$ miatt átlagosan egy-egy élen az f^{ij} elemi folyam elhanyagolható a teljes f folyamhoz képest.

9. FD-módszer nagy és kiegyensúlyozott hálózatban

Legyen most $f \in F^{nb}$,

π_{ij} az f^{ij}-hez tartozó út,

$\hat{\pi}_{ij}$ az $\{l_t\}$ metrikában legrövidebb (i,j) út,

vagyis ha célfüggvényünk P, az $l_t = \left\{\left.\frac{dP}{df_t}\right|_f, \; t=1,\ldots,b\right\}$ értékekkel metrizált hálózatban bármely két (i,j) pontpár között megkeresve egy legrövidebb $\hat{\pi}_{ij}$ utat, a legrövidebb $\tilde{v}$ folyam ($\tilde{v} \in F^{nb}$) nem más, mint a $\hat{\pi}_{ij}$ utakra terhelt r_{ij}, $\forall(i,j)$ párra.

Legyen most az FD-operátor olyan, hogy a π_{ij} útról λr_{ij} mennyiséget tesz át a $\hat{\pi}_{ij}$ útra ($\forall(i,j)$ párra), és vizsgáljuk a

$$P(\lambda) = P[f(1-\lambda)+\tilde{v}\lambda], \quad 0 \leqq \lambda \leqq 1$$

függvényt.

$$P(\lambda) = P(0)+\lambda\sum_{t=1}^{b} l_t(v_t-f_t)+\sigma[\lambda(\tilde{v}-f)], \tag{9.1}$$

ahol $\sigma[\lambda(\tilde{v}-f)]$ tartalmazza a (9.1)-beli *Taylor-sorfejtés* szerinti kifejtésben a magasabb rendű tagokat.

Ha a hálózat nagy és kiegyensúlyozott, az $f_t^{ij} \ll f_t$, $t=1,\ldots,b$. Mivel f és $\tilde{v} \in F^{nb}$, a (v_t-f_t) különbség infinitezimális.

Az 5. fejezetben láttuk, hogy $\sum_{t=1}^{b} l_t(v_t - f_t) \leqq 0$. Ha $\sum_{t=1}^{b} l_t(v_t - f_t)$ elegendően negatív ahhoz, hogy $\sigma[\lambda(\tilde{v} - \tilde{f})]$ (9.1)-ben elhanyagolható legyen, akkor

$$\min_{0 \leqq \lambda \leqq 1} P(\lambda) = P(1),$$

azaz az FD-operátorral $\tilde{f}$-ről teljesen áttérünk $\tilde{v}$-re:

$$\mathrm{FD} \odot \tilde{f} = \tilde{v},$$

s erről tudjuk, hogy nem-elágazó folyam: $\tilde{v} \in F^{nb}$. Ha viszont $\sum_{t=1}^{b} l_t(v_t - f_t) \approx 0$, akkor $\lambda_{\min} < 1$, azaz a $\mathrm{FD} \odot \tilde{f} = \tilde{f}(1 - \lambda_{\min}) + \tilde{v} \cdot \lambda_{\min}$ folyam már elágazó. Ugyanakkor $\sum_{t=1}^{b} l_t(v_t - f_t) \approx 0$ azt jelenti, hogy az optimális megoldás közelében vagyunk (l. 5. fejezet!), vagyis ekkor az $\tilde{f}$ nem-elágazó folyam előre rögzíthető hibakorlát szerinti közelítése az optimális, de elágazó m.c. folyamnak.

A „Nem-elágazó FD-módszer" főbb lépései

I. fázis: A konkrét feladattól függő módon meghatározunk egy megengedett (induló) megoldást.

II. fázis: Indulás az I. fázis végén kapott $\tilde{f}^0$ m.c. folyammal.

Az iteráció n-edik lépése:

1. Meghatározzuk az $\left\{ l_t = \left.\frac{\partial T}{\partial f_t}\right|_{\tilde{f}=\tilde{f}^n},\ t = 1, \ldots, b \right\}$ metrikában legrövidebb $\hat{\pi}_{ij}$, $i = 1, \ldots, n$, $j = 1, \ldots, n$ utakat.
2. $\tilde{g} = \tilde{f}^n$ és minden (i, j) párra elvégezzük a következőket:
   2a) $\tilde{g}$ folyamot úgy változtatjuk, hogy a $\tilde{g}^{ij}$ elemi folyamot a $\hat{\pi}_{ij}$ útra tesszük át.
   2b) Ha a kapott folyam megengedett, és egyúttal a célfüggvény értéke is csökken, ezzel a folyammal folytatjuk az eljárást 2a) lépéstől a következő (i, j) párra. Ha nem-megengedett folyamhoz jutottunk, vagy a célfüggvény nem csökken, az előző folyammal folytatjuk az eljárást (2a) lépéstől a következő (i, j) párra).
3. Ha már minden (i, j) párra sor került, két eset lehetséges:
   $\tilde{g} \equiv \tilde{f}^n$ változatlan maradt,
   ekkor $\tilde{f}^n$ tovább nem javítható nem-elágazó megoldás, az eljárás véget ért.
   $\tilde{g} \not\equiv \tilde{f}^n$, tehát találtunk nem-elágazó $\tilde{f}^n$-nél jobb megoldást, ezzel indítjuk az $(n+1)$-edik iterációt.

Módszerünk véges, hiszen a nem-elágazó folyamok száma véges. A csak nem-elágazó folyamokon haladás kb. hetedrészre csökkenti a (közelítő) optimum megkeresésének idejét [5]:

A 7. fejezet végén említett példában (21 csomópontú, $26 \cdot 2 = 52$ élű hálózatban)

$$\eta = \frac{25}{21 \cdot 20} \frac{1}{\bar{q}} < \frac{52}{21 \cdot 20} = 0{,}12 \ll 1.$$

A nem-elágazó FD módszer már 4 mp alatt igen jól megközelítette a „pontos" FD módszerrel kapott optimumot:
$T=0{,}2438$ volt az eredmény, az előző $T=0{,}2406$ értékkel szemben.

10. Az optimális útképzési és csatornakapacitás-tervezési feladat megoldása lineáris költségfüggvény esetén

Írjuk fel újra a (4.5) feladatot:
Adott G $=(N, A)$ n csúcsú, b élű hálózat,
Γ $=((r_{ij}))$, $i=1, \ldots, n$, $j=1, \ldots, n$ üzenetmátrix,
$d(\bar{C})=\sum_{t=1}^{b} d_t(C_t)$ költségfüggvény,
D költségkorlát.

Minimalizálandó

$$T(\bar{f}) = \frac{\left(\sum_{t=1}^{b} \sqrt{f_t d_t}\right)^2}{\gamma\left(D - \sum_{t=1}^{b} d_t f_t\right)}$$

feltéve, hogy

$$\bar{f} \in F$$

$$D - \sum_{t=1}^{b} d_t f_t \geqq 0.$$

A $T(\bar{f})$ függvény büntetőfüggvényként magában foglalja a $D-\sum_{t=1}^{b} d_t f_t \geqq 0$ feltételt. Hasonlóan a 7. fejezetben a (4.3) feladatra vonatkozó meggondoláshoz, a $T(\bar{f})$ függvényt elég az

$$F_{2,\varepsilon} = \{\bar{f} | \bar{f} \in F \quad \text{és} \quad \sum_{t=1}^{b} d_t f_t \leqq D-\varepsilon\}$$

halmazon vizsgálni, ahol $\varepsilon>0$ alkalmasan választott konstans.
Ekkor viszont $T(\bar{f})$ alulról korlátos és kétszer differenciálható függvény $\bar{f} \in F_{2,\varepsilon}$,

$$\frac{\partial T}{\partial f_t} = \frac{1}{\gamma}\left(\frac{\sum_{j=1}^{b} \sqrt{f_j d_j}}{D - \sum_{j=1}^{b} f_j d_j}\right)\sqrt{\frac{d_t}{f_t}} + \frac{1}{\gamma}\left(\frac{\sum_{j=1}^{b} \sqrt{f_j d_j}}{D - \sum_{j=1}^{b} f_j d_j}\right)^2 d_t \geqq 0.$$

Kiszámolva a második parciálisokat, adódik ezek végessége is, tehát az $F_{2,\varepsilon}$ halmazon az FD módszer alkalmazható.

Megjegyzés.

$$\lim_{f_t \to 0} \frac{\partial T}{\partial f_t} = \infty \quad \text{bármely } t\text{-re,}$$

ami azt jelenti, hogy ha egy FD-iterációban valamely t-re $f_t=0$, azaz valamelyik élen a folyam zérussá redukálódik, ennek az élnek a hossza az $\{l_t\}$ metrikában végtelen lesz, ezért a további FD-iterációkban sem nőhet zérus fölé az f_t folyam. A $T(\tilde{f})$ célfüggvénynek ez a tulajdonsága jól használható, ha a hálózat topológiájának megtervezése a feladatunk: Adott csomópontok mellett a teljes gráfból indulunk és az FD-módszerrel állapítjuk meg, mely éleket hagyunk el a gráfból [5], [10].

A $T(\tilde{f})$ függvény mostani alakját lineáris költség-kapacitás függvénnyel kaptuk. Bizonyítható [5], hogy ekkor, de tetszőleges konkáv $d_t(C_t)$ függvények esetén is $T(\tilde{f})$ kvázikonkáv, hacsak

$$\tilde{f}\in F_2=\{\tilde{f}\,|\,\tilde{f}\in F,\ \sum_{t=1}^{b} f_t d_t \leqq D\}.$$

Ezért az FD-módszer igen leegyszerűsödik (l. 6. fejezet 2. megjegyzés): $\lambda=1$ lesz az egyes iterációkban, vagyis mindig F_2 extremális pontjain: legrövidebb folyamokon haladhatunk, s ez egyúttal egy nem-elágazó FD-módszert is jelent. Két problémánk van még:

1) Induló megengedett megoldást kell keresnünk.
2) A módszer véges ugyan, de csak lokális minimumhoz vezet — mit tegyünk?

A feladatot meglehetősen heurisztikus módszerekkel oldjuk meg:

I. fázis:

Induló megoldás keresése:

a) Az élekhez tetszőleges, de egyenlő hosszt rendelünk.
b) Az így kapott metrikában meghatározzuk a legrövidebb $\tilde{f}\in F^{nb}$ folyamot.
c) Ha $\tilde{f}$ kielégíti a $D-\sum_{t=1}^{b} f_t d_t>0$ feltételt is, akkor indulhat a II. fázis. Ellenkező esetben az $\tilde{f}$ folyamot ejtjük és másik metrikával folytatjuk az eljárást az a) lépéstől.

II. fázis

Lokális optimum meghatározása:

Az I. fázisból kapott megoldással indulunk. Az n-edik iteráció főbb lépései:

1) $\tilde{f}^{n+1}=\mathrm{FD}\odot\tilde{f}^{n}$

ahol $\tilde{f}^{n+1}$ az $\left\{l_t^n=\left.\frac{\partial T}{\partial f_t}\right|_{\tilde{f}=\tilde{f}^n},\ t=1,\ldots,b\right\}$ metrikában meghatározott legrövidebb folyam.

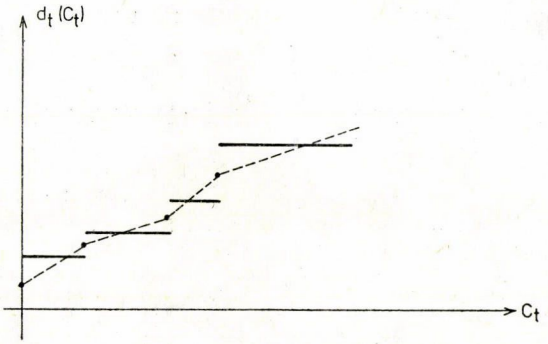
2) Ha $T(\tilde{f}^{n+1})<T(\tilde{f}^{n})$ következik az $(n+1)$-edik iteráció. Ellenkező esetben lokális minimumhoz értünk, így a II. fázis véget ért.

III. fázis — szuboptimális megoldás.

A kapott lokális minimum függ az I. fázis által adott induló megoldástól. Ezért az I. és II. fázist egymás után elég sokszor végrehajtjuk, és a véletlen módon generált induló megoldásokból kapott lokális minimumok minimumát fogadjuk el ún. szuboptimális megoldásnak.

Az ARPA hálózatban felmerült feladatban a költségfüggvény diszkrét (lépcsős) függvény volt, melyet szakaszonként lineáris függvénnyel közelítettek.

A kapott lokális optimumot ezért még hozzá kellett igazítani az eredeti költségfüggvényhez. Ezt úgy végezték, hogy a lokálisan optimális megoldásban kapott élenkénti C_t kapacitásértékeket a lépcsősfüggvény ugráshelyeinek megfelelően



4. ábra

megnövelték (a költség két szomszédos ugráshely közötti intervallumban nem változik!) és egy újabb $\tilde{f}$ megoldást határoztak meg FD-algoritmussal.

Természetesen nem állíthatjuk, hogy ez utóbbi megoldás (lokálisan) optimális, de nem is definiálhatjuk pontosan, mit értsünk lépcsős függvény esetén lokális optimumon. [5]-ben a szerzők szerint eljárásuk nem rosszabb, mint más, hasonló esetben alkalmazott heurisztikus technikák.

IRODALOM

[1] ASSAD, A. A., „Multicommodity network flows — A survey", *Networks* **8** (1978) 37—91.
[2] CANTOR, G. and GERLA, M., „Optimal routing in a packetswitched computer network", *IEEE Trans. on Comp. C*—**23** (1974) 1062—1068.
[3] FRANK, and CHOU, W., „Routing in computer networks", *Networks* **1** (1971) 99—112.
[4] FRANK, H., KAHN, R. E. and KLEINROCK, L., „Computer communication network design — Experience with theory and practice", *Networks* **2** (1972) 135—166.
[5] FRATTA, L., GERLA, M. and KLEINROCK, L., „The flow deviation method: An approach to store- and forward communication network design", *Networks* **3** (1973) 97—133.
[6] FRATTA, L. and MONTANARI, U. *Analytical Techniques for Computer Networks /Analysis and Design*, (Computer Architectures and Networks, North Holland Publ. Company, 1974) 155—185.
[7] LUENBERGER, D. G., *Introduction to Linear and Nonlinear Programming* (Addison—Wesley Publ. Company, 1973).
[8] HU, T. C., *Integer Programming and Network Flows* (Addison—Wesley Publ. Company, 1969).
[9] GALLEGER, R. G., „Basic limits on protocol information in data communication networks", manuscript.
[10] KLEINROCK, L., „Analytic and simulation methods in computer network design", *AFIPS Conference Proceedings SJCC* (1970) 569—579.
[11] KLEINROCK, L., *Queing Systems, Volume* 1: *Theory* (Wiley, 1975).
[12] KLEINROCK, L., *Queing Systems, Volume* 2: *Computer Applications* (Wiley, 1976).
[13] MARUYAMA, K. and TANG, D. T., „Discrete link capacity an priority assignments in communication networks", *IBM J. Res. Develop.* **21** (1977) 254—263.

[14] MARUYAMA, K., FRATTA, L. and TANG, D. T., „Heuristic design algorithm for computer communication networks with different classes of packets" *IBM J. Res. Develop.* **23** (1977) 360—369.
[15] SEGALL, A., „The modeling of adaptive routing in data — communication networks", *IEEE Trans. on Comm.* **25** (1977) 85—94.

(Beérkezett: 1979. január 5.)

SZ. TURCHÁNYI PIROSKA
MTA SZÁMÍTÁSTECHNIKAI ÉS AUTOMATIZÁLÁSI KUTATÓ INTÉZET
1111 BUDAPEST, XI. KENDE U. 13—17.

ON THE OPTIMIZATION PROBLEMS IN PACKET-SWITCHING NETWORKS

P. SZ. TURCHÁNYI

The analysis of stochastic flow in store-and-forward networks and as a result the efficient design of a computer communication network are extremely complex tasks. The problem of the flow assignment and the capacity and flow assignment are considered:

the multicommodity FA problem requires that we minimize the nonlinear function T with respect to the flows on the arcs in such a way that the external flow requirements at the nodes are satisfied; this, of course, is under the assumption of a given capacity assignment;

in the CFA problem we require an optimum choice of channel capacities and a flow configuration, thus, we are no longer able to give globally optimal solutions, but rather we describe procedures that find local minima for T.

The method to be described is due to FRATTA, GERLA and KLEINROCK [5], called FD method, it gives an exact solution to the FA problem and results in an alternate routing flow. The large and balanced network case is also investigated, we see that we have a faster and rather good algorithm of a fixed routing type.

Since CFA algorithm eliminates certain channels as it iterates to a local minimum, we see that it has possibilities in the topological design of computer networks.

Alkalmazott Matematikai Lapok 4 (1978) 271—325

SZABÁLYOS JOB-FOLYAM PÁROK DOMINÁNS ÜTEMEZÉSE

TANKÓ JÓZSEF
Budapest

Egy különleges nem-véges determinisztikus ütemezési problémával foglalkozunk. A feladat két, rendszeresen megújuló igény-sorozat (job-folyam) kiszolgálása egyetlen processzorral és a processzor idejének minél jobb kihasználása.

A dolgozat három fejezetet tartalmaz. Az első fejezetben általában tekintjük át az ütemezési problémák témakörét irodalmi hivatkozások mellett. Bevezetjük a szabályos job-folyamot mint új modellt és megemlítjük annak néhány alkalmazási lehetőségét. A második fejezetben az ütemezési problémáink tárgyalásához szükséges előkészítő ismereteket foglaljuk össze. Elsősorban a lánctört-fejtés és az ún. koincidencia feladatok megoldásának kérdésével foglalkozunk, amelyek eredményeire az ütemezési probléma későbbi tárgyalásához szükség lesz. A két fejezet erősen rövidített változatal a szerző egy tanulmányának [92]. A harmadik fejezetben a szabályos job-folyam párok ütemezésével kapcsolatos fogalmakat definiáljuk, osztályozzuk a lehetséges ütemezéseket (ütemterveket), ütemezési stratégiákat és dominancia tételek segítségével szűkítjük azt a halmazt, amelyben az optimális ütemezést keresni kell.

1. Bevezető

Az utóbbi pár évtizedben megnövekedett az ütemezési problémák jelentősége. Ennek oka az automatizálás és a számítástechnika fejlődése és alkalmazásának növekedése. Az automatizálás és a számítástechnika alkalmazása megkívánja erőforrások és igények rendszeres tervszerű időbeli összerendelését, azaz az erőforrások ütemezését. A rendszertervezésnek az ütemezési kérdések nagyon gyakran központi témái. Ugyanakkor az automatika és a számítógép eszközt is biztosít bonyolultabb és hatékonyabb ütemezések kivitelezésére.

1.1. Az ütemezési problémák irodalmi áttekintése

Az ütemezési problémák jelentőségének növekedése tükröződik a témával foglalkozó publikációk számának rohamos növekedésében is, de elsősorban a különféle modellek számának és kapcsolódásainak gyarapodásában. Egyre több cikk, sőt könyv önálló témája ütemezési probléma vizsgálata. Egy önálló diszciplina, az ütemezés elmélete (scheduling theory) kialakulásáról beszélhetünk. A korábbi ipari alkalmazások [79] mellett elsősorban az operációs rendszerek elméletében [9, 15, 25, 27, 51], multiprogramozási, többprocesszoros és időosztásos rendszerekben és számítógép-hálózatokban kerülnek előtérbe ütemezési problémák [1, 32, 62, 75, 97, 99]. Kifejezetten az ütemezési problémák a témája MUTH—THOMPSON [79], COMWAY—MAXWELL—MILLER [28], BAKER [8] könyvének és egy COFFMAN szerkesztette könyvnek [24]. Ezenkívül számos konferencia és könyv fontos résztémája az ütemezés

[35, 53, 83]. Az ütemezési téma alkalmazásként matematikai módszerekre természetszerűen szerepel számos operációkutatási, kombinatorikai (gráfelméleti) és egyéb matematikai tárgyú (pl. sztochasztikus folyamatok) könyvben is [2, 13, 26, 33, 84]. A témával foglalkozó cikkek listája pedig csak reprezentatív lehet [17, 20, 21, 23, 37, 38, 40, 43, 46, 49, 54, 57, 73, 87].

A legnagyobb irodalma talán a véges determinisztikus esetnek van, amelynél véges számú task igényel kiszolgálást processzortól egyéb erőforrás biztosítása mellett, vagy anélkül. Egy, vagy több, de véges és az egyidejű igényeknél feltétlenül kevesebb számú processzor idejét és az egyéb erőforrásokat kell ütemezni úgy, hogy a task-okat, előírt sorrendjüket meg nem sértve, valamilyen adott kritérium szerint, lehetőleg optimálisan kielégítsék. Az ilyen jellegű problémák meglehetősen változatosak és a legteljesebb áttekintés talán COFFMAN [24] könyvében található. E problémák általában kombinatorikus feladatokra, illetve módszerekhez vezetnek, amelyek számításigénye gyakorlatilag hamar elviselhetetlenül megnövekszik. Ezért ezeknél a problémáknál, illetve a megoldó algoritmusokkal kapcsolatban a komplexitás fogalma döntő kérdéssé válik. E kérdésnek, valamint az ehhez kapcsolódó közelítő megoldási módszereknek a kutatása ugyancsak az utóbbi évtizedekben fokozódott és számos munka témája [14, 29, 45, 59, 76, 89, 94]. Utalunk még a [92] tanulmány 1.3. pontjában mondottakra is. Bizonyos problémákat számítási szempontból kezelhetetlennek tekintenek és csak közelítő módszereket fejlesztenek. Ilyen a problémák ún. **NP**-teljes (**NP**-*complete*) osztálya, amelyek áttekintése KARP klasszikus cikkében [59], AHO—HOPCROFT—ULLMAN [4] könyvében és COFFMAN könyvének [24] ULLMAN írta fejezetében [95] található többek között. Az ütemezési problémák komplexitásával foglalkozó cikkek [16—18, 36, 39, 96] eredményeit nagyrészt áttekintik. A közelítő módszereket is számos könyv [4, 8, 24, 94] és cikk tárgyalja [7, 21, 44, 49, 74, 87, 98]. Az ütemezési módszerek és az ütemezések sok modellnél igen érzékenyek a modell paramétereinek változásaira és meglepő anomáliákat mutatnak [47, 48].

A véges diszkrét determinisztikus modelleken kívül természetesen egyéb modellekre is szükség van ütemezési problémák leírásához, tanulmányozásához és megoldásához. Az ütemezési elmélettel (is) foglalkozó könyvek [8, 24, 25, 28, 35, 83] legtöbbje foglalkozik sztochasztikus modellekkel is, amelyek közül a tömegkiszolgálási modellek [3, 10, 19, 23, 26, 40, 55, 62, 63, 70, 72, 73, 75] és diffúziós közelítések [6, 41, 42, 66, 67] a leggyakoribbak.

Az általunk vizsgált ütemezési problémához talán a job-shop típusú és az időkritikus ütemezési problémák állnak legközelebb (l. [92]-ben 1.4. és 1.5. pontok), bár vizsgálati módszereik egyáltalán nem alkalmazhatók a mi esetünkre. A job-shop típusú probléma lényegében a legegyszerűbb esetekben is kezelhetetlen [39], ezért csak közelítő módszerek ismeretesek megoldására. Speciális esete a job-shop problémának a flow-shop probléma, amelynek két processzor esetére a pontos, több processzor esetére csak közelítő megoldásai ismeretesek. A kétprocesszoros esetre JOHNSON klasszikus algoritmusa ismeretes [12, 57] az egyéb flow-shop és job-shop problémák vizsgálatával kapcsolatban az összefoglaló munkákon kívül számos cikkre is utalhatunk [7, 12, 20, 21, 37, 38, 43, 44, 49, 50, 52, 54, 56, 57, 64, 74, 77, 87].

Az időkritikus ütemezési problémák erősen különbözőek lehetnek és velük kapcsolatban jóval kevesebb eredmény ismeretes, mint a véges determinisztikus, vagy akár a job-shop típusú problémákkal kapcsolatban. Pedig sok gyakorlati problémánál ilyen jellegű az ütemezési igény és az egyéb modellek durva megközelítést ered-

ményezhetnek. Az ilyen modellekben részben a ciklikusan megújuló igények, részben a kiszolgálás késésének korlátozása a jellemző [24, 46, 69, 71, 78, 85, 86, 88, 98]. Bizonyos folyamatok nemcsak időbeni ütemezést, hanem egyéb erőforrást, esetleg térbeli elhelyezést *(allocation)* is igényelnek: memória, háttértárak a számítógépnél, virtuális kapacitások kezelése [11, 30, 31, 80]. Így az ütemezés problémája kapcsolódhat programozástechnikai és programoptimalizálási kérdésekhez is [5, 34, 90].

Sztochasztikus esetben a szabályos job-folyam párok analogonját ARATÓ [6] és TOMKÓ [93] vizsgálták és igen érdekes eredményeket kaptak, elsősorban prioritásos ütemezésekkel kapcsolatban. Természetesen determinisztikus esetben teljesen más módszerekre van szükség, mint sztochasztikus esetben. Ezeket tárgyaljuk a második fejezettől.

1.2. Ütemezési problémák komponensei

Az ütemezési problémák a valóságban annyira különbözőek, hogy a legkülönfélébb modellekkel lehet csak jellemezni azokat. Ennek megfelelően a megfelelő vizsgálati és megoldási módszerek is igen különbözők. Nem beszélhetünk tehát valamiféle egységes ütemezési elméletről. Az ütemezési elmélet *(scheduling theory)* csupán gyűjtőfogalma a modelleknek, módszereknek és eredményeknek. Az alkalmazott modellek alapján is még sokféle szempontból osztályozhatnánk az ütemezési problémákat. Az ütemezési problémákban a legfontosabb közös vonás talán az, hogy egyaránt rendelkeznek bizonyos alapvető komponensekkel. Az ütemezési probléma a valóságban mindig valamilyen rendszer részeként, annak működésében merül fel, ezért a megoldás a rendszertervezés része. Egyes komponensek önkényesen tekinthetők az ütemezési probléma részének, vagy a tágabb rendszer-modell részének is. Az ütemezési modellben általában adott egy erőforráskészlet, adott az igény oldal, adottak bizonyos technológiai és működési feltételek és végül bizonyos szubjektív hatékonysági szempontok és célkitűzések (célfüggvény), amelyeket a megoldásnál figyelembe kell venni.

Az erőforrás-rendszer tartalmazhat processzorokat és egyéb erőforrásokat. A processzorok idejét és az egyéb erőforrások kapacitását kell az igényekhez rendelni az ütemezés folyamán úgy, hogy konfliktus ne lépjen fel. Speciális erőforrástípus az osztott processzor. Az erőforrások oszthatóságára és rendelkezésre állására különféle feltételek (kvantum, valószínűség stb.) lehetnek előírva. A processzoroknak sebessége is lehet értelmezve.

Az igényrendszer a lehető legkülönbözőbb lehet. A leggyakoribb a taskkészlet. A task jellemzője, hogy egyetlen meghatározott típusú processzortól kíván adott, véges idejű kiszolgálást. Ezenkívül a kiszolgálás idejére egyéb erőforrásokat is igényelhet. A task-ok száma lehet véges, vagy megszámlálhatóan végtelen. Határeset a nagyon rövid kiszolgálási idejű task-sorozatok folytonos paraméterű folyamattal történő modellezése (pl. diffúziós közelítés [6, 41, 42, 66, 67]). A task-ok időigénye lehet diszkrét (pl. egész), vagy folytonos értékű. A task-ok között általában előidejűségi relációk *(precedence constraints)* lehetnek előírva, amelyeket irányított gráf segítségével adhatunk meg. A relációk speciális esetben a task-okat részhalmazokba rendezhetik, amelyeken belül lineáris (esetleg fa típusú) rendezés van. Ezeket nevezik általában job-oknak. A task-oknak és a job-oknak lehet érkezési (rendelkezésre állási) és határideje *(deadline)*, sőt sok esetben csupán az azonnali

kiszolgálás megengedett (időkritikus problémák). A task-ok igénye lehet determinisztikus és sztochasztikus (valószínűségi változó). Lehet ciklikusan visszatérő azonos (eloszlású) igény is.

A feltételrendszer magában foglalja a már említett feltételeket az erőforrásrendszer és az igényrendszer működésére, ill. kiszolgálhatóságára vonatkozóan, de ezenkívül egyéb technológiai feltételeket is tartalmazhat. A legfontosabb feltételek az erőforrások kapacitásai, a task-ok előidejűségi viszonyai, a határidők és a task-ok kiszolgálásának megszakíthatóságára vonatkozó feltételek. Az utóbbi feltételek határozzák meg, hogy a task-ok egybefüggően, megszakítás nélkül *(non-preemptive)*, vagy megszakítással *(preemption)* ütemezhetők-e és megszakítás esetén újra kell-e kezdeni *(preempt-repeat)*, vagy folytatni lehet a kiszolgálást *(preempt-resume)*. A technológiai feltételrendszerhez sorolhatjuk még a bizonyos költségek árán megsérthető feltételeket és a költségeket meghatározó adatokat is.

A szubjektív hatékonysági szempontok és célkitűzések szintén igen változatosak lehetnek és részben meghatározottak a rendszer többi komponense által. Az ütemezés célja általában nemcsak a konfliktusmentes kiszolgálás, hanem a lehető „leggazdaságosabb" kiszolgálás. A gazdaságossági szempont sokszor több, részben ellentétes szempont egyensúlya, vagy egy (ritkán több) célfüggvény szélső értékéhez való közelsége. A leggyakrabban használt optimalizálási szempontok megtalálhatók az ütemezéssel foglalkozó könyvekben [8, 24]. Véges esetben a teljes igényegyüttes kiszolgálásának időtartama, nem véges esetben a processzor(ok) idejének kihasználási foka (hatékonyság) a leggyakoribb célfüggvény. Az ütemezés módjával kapcsolatban további megszorítások alkalmazhatók, amelyek például önkényesen szűkítik a lehetséges (vizsgált) ütemezési stratégiák körét (pl. lista szerinti ütemezésekre, vagy prioritásos ütemezésekre stb.).

1.3. Szabályos job-folyam párok ütemezési problémája

A szabályos job-folyam párok ütemezésének speciális problémája a következő. Adva van egy $\mathscr{P}=(P_A, P_{B1}, P_{B2})$ processzor-hármas és egy $Q=(Q^{(1)}, Q^{(2)})$ job-folyam pár. A $Q^{(i)}$, $i=1, 2$, job-folyam $C_{ij}=(A_{ij}, B_{ij})$ task-párok $j=1, 2, \ldots$, megszámlálhatóan végtelen sorozata, amelynél az A_{ij} task-ok a P_A processzoron $\tau_i^A \geqq 0$, és a B_{ij} task-ok a P_{Bi} processzoron $\tau_i^B \geqq 0$ idejű kiszolgálást igényelnek. A $\tau_i^X \geqq 0$, $X=A, B$, igények adott állandók és j-től nem függnek. Mivel a $Q^{(1)}$ és a $Q^{(2)}$ job-folyamok egyaránt igénylik a P_A processzort, igényeik ütközhetnek. A konfliktust az ütemezéssel oldjuk fel, amelynél a processzorok idejét úgy ütemezzük a Q igényeinek kielégítésére, hogy egyidejűleg minden processzor legfeljebb egy task-hoz legyen hozzárendelve, sőt egyéb kiegészítő feltételek is teljesüljenek. Ilyen feltétel az, hogy a B_{ij} task kiszolgálása nem kezdhető meg, amíg az A_{ij} kiszolgálása be nem fejeződött, és az $A_{i,j+1}$ task nem kezdhető meg, amíg a B_{ij} be nincs fejezve. A task-ok kiszolgálási sorrendje tehát szigorúan $A_{i1}, B_{i1}, A_{i2}, B_{i2}, \ldots$. Járulékos feltétel lehet az, hogy a task-ok csak összefüggően, megszakítás nélkül szolgálhatók ki. A P_{Bi} processzorokon ütemezési konfliktus nem léphet fel. A P_A processzor idejének a kihasználása a Q job-folyam pár τ_i^X, $i=1, 2$, $X=A, B$, igényeinek nagysága mellett az ütemezés módjától függ és célkitűzésünk ennek maximalizálása. Egy Q job-folyam pár ütemtervét az ún. *Gantt-diagramon* szemlél-

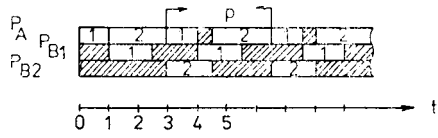
tethetjük [8, 22], amelyen a t időtengellyel párhuzamos sávok képviselik a processzorokat és a sávokba bejelöljük a foglaltsági intervallumokat a $Q^{(i)}$ job-folyam i sorszámának a feltüntetésével, a tétlenségi időszakaszokat pedig sraffozással jelöljük (lásd 1. ábra).

Legyen $\tau^A[a, b)$ az $a \leqq t < b$ intervallumban a P_A processzor foglaltsági intervallumainak összhossza és legyen $\tau^A(x) = \tau^A[0, x)$. Ekkor az ütemterv hatékonyságát a

$$\gamma = \lim_{t \to \infty} \frac{\tau^A(t)}{t} \tag{1.1}$$

határértékkel definiáljuk. Egy ütemterv hatékonyságának meghatározását fogjuk az ütemterv értékelésének mondani.

A $Q^{(i)} = \{(A_{ij}, B_{ij}),\ j = 1, 2, \ldots\}$ job-folyam sok gyakorlati esetben alkalmas igények modellezésére ütemezési problémákban. A τ_i^X igények j-től való függetlensége is sok esetben megengedhető feltevés, vagy közelítés. Amikor a τ_i^X valószínűségi változó, egy sztochasztikus modellel van dolgunk. Ezt vizsgálta ARATÓ [6] és TOMKÓ [93]. A τ_i^X determinisztikus esetét tudomásunk szerint eddig nem vizsgálták, legalábbis eredményeket nem publikáltak.



1. ábra. *Gantt-diagram*

Szabályos job-folyammal modellezhető igen sok esetben egy számítógépen futtatott program a központi egység (CPU) iránti igényét tekintve (A-task-ok), amelyek között egyéb tevékenységekre (adatmozgatás stb.: B-task) van szükség. Ilyen modellel írható le egy dialógus távállomás működése, amelynél az A-task-ok a központi gép iránti igényeket (tranzakciókat) képviselik, a B-task-ok pedig a gondolkodási és gépelési időt. Egy számítógép-hálózatban rendszeres időközönként (B-task-ok) vezérlő és ellenőrző információkat kell generálni és küldeni a csomópontokra, amelyek állandó időt igényelnek (A-task-ok). Így írható le azonban bármely olyan rendszer, amelyben objektumok rendszeres időközönként (B-task) azonos időigényű ellenőrző-szabályozó tevékenységet igényelnek egy apparátustól (A-task), amely nélkül működésük késleltetődik.

Jóllehet mi a szabályos job-folyamok ütemezésének legegyszerűbb esetét, a job-folyam pár ütemezését vizsgáljuk, a fellépő nehézségek nem triviálisan leküzdhetők és következtetni engednek több job-folyam ütemezési problémájának nehézségeire. A kidolgozott módszerek esetleg továbbfejleszthetők egyéb ütemezési problémák vizsgálatára is, bár erre nem teszünk kísérletet e munkánkban.

2. Előkészítés

Ebben a fejezetben a szerző [92] tanulmányának a második fejezetére utalással röviden felidézzük azokat a fogalmakat és tételeket, amelyekre a továbbiakban szükségünk lesz és hivatkoznunk kell. Ez biztosítja, hogy a továbbiakban a [92] tanulmányra történő utalás már nélkülözhető legyen.

2.1. Racionális közelítések és koincidencia feladatok (KIF)

A szabályos job-folyam pár ütemterveinek értékelésénél gyakran fogunk találkozni olyan feladattal, hogy egy

$$0 \leqq By - Ax \prec a \tag{2.1}$$

alakú egyenlőtlenség legkisebb $B \geqq 1$, $A \geqq 0$ egész számpár megoldását kell meghatározni, ahol $x, y \geqq 0$ és $a \geqq 0$ adott valós számok és a $\prec$ szimbólum a $<$, vagy $\leqq$ relációjel valamelyike. Ha $x > 0$, akkor a (2.1) ekvivalens a

$$0 \leqq B\xi - A \prec \alpha \tag{2.2}$$

egyenlőtlenséggel, ahol

$$\xi = \frac{y}{x}, \quad \alpha = \frac{a}{x}.$$

A (2.2) egyenlőtlenség ekvivalens még a

$$0 \leqq \xi - \frac{A}{B} \prec \frac{\alpha}{B}, \tag{2.3}$$

egyenlőtlenséggel is, amely egy racionális közelítési feladat speciális megfogalmazása: a legkisebb $B > 0$ nevezőjű A/B racionális tört keresendő, amelyre a (2.3) teljesül. A racionális közelítés feladata hagyományosan a mindenütt sűrű

$$\frac{A}{B},\ B > 0,\ A, B \text{ relatív prím egészek}$$

racionális számokkal való közelítés előírt feltételeknek eleget tevő hibával, vagyis egy ξ valós szám közelítése $\omega \doteq A/B$-vel úgy, hogy a

$$\delta_\omega = \xi - \omega \tag{2.4}$$

hiba (esetleg előírt előjelű és) adott korlát alatt maradjon, azaz

$$|\delta_\omega| \prec \alpha$$

legyen, ahol α függhet a B értékétől. Adott $\alpha(B)$ esetén az ω közelítés existenciája, a közelítések számossága, a δ_ω hiba tulajdonságai a jellegezetes problémák [61]. *Legjobb* (bal, ill. jobb oldali) *közelítésnek* nevezik [61, 82] az $\omega = A/B$ közelítést, ha bármely $\omega' = A'/B'$ közelítés δ_ω hibája csak akkor lehet δ_ω-nál nem-nagyobb (esetleg azonos előjelet megkövetelve!), ha $B' > B$.

Ha a (2.4) helyett a

$$\Delta_\omega = B\delta_\omega = B\xi - A \tag{2.5}$$

számot tekintjük az $\omega \doteq (B, A)$ közelítés hibájának [60, 91], akkor a legjobb közelítések halmaza más [61]. Ezeket rendszerint nem is legjobb közelítéseknek nevezik [82]. Hincsin [61] megkülönböztetésül I., ill. II. típusú közelítésnek nevezi a δ_ω, ill. Δ_ω kritérium szerinti közelítéseket, mi pedig a Δ_ω kritérium szerinti „legjobb közelítést" nevezzük *legjobb közelítő megoldásnak* tekintettel arra, hogy a $By - Ax = 0$, vagy $B\xi - A = 0$ egyenletek $\omega = (B, A)$ közelítő megoldásairól van szó [60, 82].

Legyen Ω_ξ a legjobb közelítő megoldások halmaza a ξ számhoz növekvő B érték szerint sorbarendezve. A (2.2), ill. (2.3) egyenlőtlenség legkisebb $B \geqq 1$, $A \geqq 0$ megoldását ekkor az Ω_ξ első ω eleme szolgáltatja, amelyre $\Delta_\omega \prec \alpha$, ill. $\delta_\omega \prec \alpha/B$, ha ilyen egyáltalán létezik. Ugyanígy a

$$0 \leqq \xi - \frac{A}{B} \prec \alpha \tag{2.6}$$

egyenlőtlenség legkisebb $B > 0$ nevezőjű megoldása az Ω_ξ azon első eleme, amelyre $\delta_\omega \prec \alpha$, ahol most Ω_ξ a legjobb közelítések halmaza növekvő B nevező szerint rendezve.

Ismeretes [60, 61, 81, 82], hogy a legjobb közelítéseket és közelítő megoldásokat a $\xi = y/x$ szám szabályos lánctörtfejtése útján határozhatjuk meg. Erre a következő pontban visszatérünk.

A racionális közelítések problémájának fellépését a szabályos job-folyam párok ütemezésének vizsgálatánál nyilvánvalóvá fogja tenni az ütemtervek értékelése és az alább definiálásra kerülő koincidencia feladat később felszínre kerülő kapcsolata, valamint a racionális közelítések és speciális koincidencia feladatok alább megmutatkozó kapcsolata.

Legyen X és Y két eseménysorozat (folyam), amelyeknél az X folyamban x időközönként, az Y folyamban z pillanattól kezdve y időközönként következik be esemény. Kérdés, melyik az első eseménypár megadott sorszámok után, amelynek tagjai időben megadott közelségbe kerülnek egymáshoz, azaz kielégítenek előírt koincidencia feltételeket. Az ilyen típusú feladatot *koincidencia feladatnak (KIF)* nevezhetjük.

Az X folyamban az A-adik esemény az Ax pillanatban, az Y folyamban a B-edik esemény a $z+By$ pillanatban következik be. A két esemény időbeli „távolsága"

$$\Delta_\omega = By + z - Ax, \tag{2.7}$$

ahol az ω index az $\omega = (B, A)$ nem-negatív egész számpárra utal. Ha a koincidencia feltétel a Δ_ω „hibára" előírt intervallum, akkor a *KIF* nem más, mint egy

$$a \prec_B By + z - Ax \prec_J b, \quad \omega \geqq \omega_0 \tag{2.8}$$

egyenlőtlenség(pár) legkisebb $\omega^* \doteq (B^*, A^*) \geqq (B_0, A_0) \doteq \omega_0$ megoldásának keresése, ahol x, y, z, a, b adott valós számok $(b \geqq a)$, A_0, B_0 adott egészek és $\prec_B$ és $\prec_J$ szimbólumok adott $\prec$, vagy $\leqq$ jelek. Ezt az általános *KIF*-et mindig visszavezethetjük bizonyos standardizált *KIF*-re [92], amelynek x, y, z, a, b, A_0, B_0 paraméterei bizonyos praktikus korlátozásoknak eleget tesznek. A (2.1) egyenlőtlenség a (2.8) speciális esete, amelynél $a = z = 0$, $\prec_B = \leqq$, $\prec_J = \prec$ és $b = a$. A legkisebb $B \geqq 1$, $A \geqq 0$ megoldás keresése ekvivalens egy speciális *KIF*-fel, amelynél $\omega_0 = (1, 0)$. Ez a példa mutatja, hogy a *KIF* megoldása a legjobb közelítő megoldások $\Omega_{y/x}$ halmazának eleme.

Látni fogjuk majd, hogy bizonyos ütemtervek értékelése ekvivalens bizonyos speciális *KIF*-ek megoldásával, vagy arra visszavezethetők, ezért az ütemtervek értékelése a *KIF*-ek megoldhatóságán és a megoldás meghatározásán múlik. Ezért

szükséges a *KIF*-ek megoldási kérdéseinek vizsgálata. A rövidebb írásmód kedvéért használjuk a következő rövidítéseket:

B	bal oldali,	*E*	egyoldali,	*F*	feladat,
J	jobb oldali,	*K*	közelítés,	*KI*	koincidencia,
KM	közelítő megoldás,	*L*	legjobb,	*M*	megoldás.

Definiáljuk $B>0$, A, B egészek esetére az $\omega=\frac{A}{B}$ racionális számot, mint törtet, közelítésnek (K), az $\omega=(B, A)$ egész párat pedig közelítő megoldásnak (KM).

Egy adott ξ valós számhoz ω bal oldali (B), ill. jobb oldali (J), közös néven egyoldali (E) K és KM, ha $\delta_\omega \geqq 0$ és $\Delta_\omega \geqq 0$, ill. $\delta_\omega \leqq 0$ és $\Delta_\omega \leqq 0$. $B>0$ miatt δ_ω és Δ_ω mindig azonos előjelűek és egyszerre nullák.

DEFINÍCIÓ: Az ω' a ξ-hez $L(B$, ill. $J)K$, illetve $L(B$, ill. $J)KM$, ha

$$(\operatorname{sgn}\delta_\omega \geqq 0, \text{ ill. } \operatorname{sgn}\delta_\omega \leqq 0 \text{ és}) \text{ bármely } \omega \neq \omega'$$

esetén

$$(\text{amelyre } \operatorname{sgn}\delta_\omega = \operatorname{sgn}\delta_{\omega'})$$

$$|\delta_\omega| \leqq |\delta_{\omega'}|, \quad \text{illetve} \quad |\Delta_\omega| \leqq |\Delta_{\omega'}|$$

kizárólag

$$B > B'$$

esetén lehetséges.

A definícióból azonnal következik, hogy B, A relatív prím számpár, ha a $B=1, A=0$ párat is annak tekintjük. A definíció ekvivalens azzal, hogy

$$B \leqq B'$$

esetén feltétlenül

$$|\delta_\omega| > |\delta_{\omega'}|, \quad \text{illetve} \quad |\Delta_\omega| > |\Delta_{\omega'}|.$$

Az *LBK*-k és az *LJK*-k együtt a *LEK*-ek és az *LBKM*-ek és az *LJKM*-ek együtt az *LEKM*-ek. A definícióban a zárójelben szereplő feltételektől eltekintve (azokat elhagyva) az *LK*-k, illetve *LKM*-ek definícióját kapjuk.

Láttuk, hogy a

$$\begin{aligned} 0 &\leqq By - Ax < \alpha \\ -\alpha &< By - Ax \leqq 0 \\ |By - Ax| &< \alpha \end{aligned} \tag{2.9}$$

speciális *KIF*-ek megoldása egy Ω_ξ halmaz legelső ω eleme, amelyre $0 \leqq \Delta_\omega < \alpha$, $-\alpha < \Delta_\omega \leqq 0$, ill. $|\Delta_\omega| < \alpha$ teljesül, ahol Ω_ξ az $L(B$, ill. $J)KM$-ek halmaza. E halmazokat a szabályos lánctörtfejtés segítségével kaphatjuk meg.

2.2. *Szabályos lánctörtfejtés és a KIF megoldása*

Bármely $\xi, |\xi| < \infty$, valós számnál legyen

$$\xi_0 = \xi \quad \text{és} \quad b_0 = [\xi_0].$$

Ha $\{\xi_0\}=0$, akkor

$$\xi_0 = [b_0] \text{ a } \xi \text{ szabályos lánctörtfejtése.}$$

Ha $\{\xi_0\}>0$, akkor legyen

$$\xi_1 = \frac{1}{\{\xi_0\}} \quad \text{és} \quad b_1 = [\xi_1].$$

Ha $\{\xi_1\}=0$, akkor

$\xi_0=[b_0, b_1]$ a ξ szabályos lánctörtfejtése.

Ha $\{\xi_1\}>0$, akkor az eljárás folytatható és generálhatók a $\xi_2, \xi_3, \ldots$ és $b_2, b_3, \ldots$ sorozatok tagjai mindaddig, amíg valamely ξ_n-re $\{\xi_n\}=0$ be nem következik.

Ha $\{\xi_n\}=0$, akkor $\xi_0=[b_0, b_1, \ldots, b_n]$ a ξ szabályos lánctörtfejtése.

Ha ez a feltétel egyetlen n-re sem következik be, akkor a

$$\xi_0 \doteq [b_0, b_1, \ldots, b_\nu, \ldots]$$

végtelen sorozat a ξ szabályos lánctörtfejtése. A $b_0, b_1, \ldots$ egészeket nem-teljes hányadosoknak nevezik. Ezek száma legyen N. Véges n esetén $N=n+1$, egyébként N végtelen. A lánctörtfejtési algoritmus a legnagyobb közös osztó meghatározására szolgáló euklideszi algoritmus egy módosítása. Az algoritmusból következik, hogy

$$\xi_\nu > 1, \quad b_\nu \geqq 1, \quad 1 \leqq \nu < N, \quad \text{és} \quad \xi_n = b_n \geqq 2, \quad \text{ha } N \text{ véges},$$

továbbá az

$$\omega_\nu = [b_0, b_1, \ldots, b_\nu] = b_0 + \cfrac{1}{b_1 + \cfrac{1}{b_2 + \ddots \cfrac{}{b_{\nu-1} + \cfrac{1}{b_\nu}}}}$$

teljes hányadosokra

$\xi=\omega_\nu$, ha N véges és $\xi=\lim_{\nu\to\infty} \omega_\nu$, ha N végtelen. Igazolható, hogy

$$\xi_\nu = [b_\nu, b_{\nu+1}, \ldots] \quad \text{és} \quad \xi = [b_0, b_1, \ldots, b_{\nu-1}, \xi_\nu]$$

ha $0 \leqq \nu < N$.

Definiálva az

$$A_{-2} = 0, \quad B_{-2} = 1, \quad A_{-1} = 1, \quad B_{-1} = 0$$

$$A_\nu = b_\nu A_{\nu-1} + A_{\nu-2}, \quad B_\nu = b_\nu B_{\nu-1} + B_{\nu-2}, \quad \nu = 0, 1, \ldots, n \tag{2.10}$$

mennyiségeket (egészek), teljes indukcióval igazolható, hogy

$$\omega_\nu = \frac{A_\nu}{B_\nu}, \quad 0 \leqq \nu < N \tag{2.11}$$

és

$$\xi = \frac{A_n}{B_n} \text{ véges } N \text{ és } \xi = \lim_{\nu\to\infty} \frac{A_\nu}{B_\nu} \text{ végtelen } N \text{ esetén.}$$

A lánctörtek elméletével és a lánctörtfejtéssel kapcsolatos ismeretek PERRON [82] és HINCSIN [61] könyvében találhatók. Bizonyítható, hogy minden racionális ξ véges és minden irracionális ξ végtelen szabályos lánctörtbe fejthető egyértelműen a fentebb definiált algoritmussal.

Igazolhatók egyszerűen a következő legalapvetőbb összefüggések:

(2.12) (A_ν, B_ν), $(A_{\nu-1}, B_\nu)$, $(A_\nu, B_{\nu-1})$ relatív prím párok

(2.13) $$A_\nu B_{\nu-1} - A_{\nu-1} B_\nu = (-1)^{\nu-1}, \quad \nu = -1, 0, 1, \ldots$$

(2.14) $$\delta_\nu \doteq \xi - \omega_\nu = \frac{(-1)^\nu}{B_\nu B'_{\nu+1}}, \quad 0 \leqq \nu < N-1$$

(2.15) $$\Delta_\nu \doteq B_\nu \xi - A_\nu = \frac{(-1)^\nu}{B'_{\nu+1}}, \quad 0 \leqq \nu < N-1,$$

ahol

(2.16) $$B'_{\nu+1} = \xi_{\nu+1} B_\nu + B_{\nu-1} = \xi_{\nu+1} B'_\nu, \quad 0 \leqq \nu < N-1.$$

Nevezzük a (2.11) alatti elemekkel az

$$\Omega^* = \{\omega_\nu, \ 0 \leqq \nu < N\}$$

véges, vagy végtelen rendezett halmazt a ξ főközelítései halmazának. Ennek elemei váltakozva ξ bal, ill. jobb oldalán helyezkednek el. Az $\omega_0, \omega_2, \ldots, \omega_{2k}, \ldots$ páros indexűek BK-k, az $\omega_1, \omega_3, \ldots, \omega_{2k+1}, \ldots$ páratlan indexűek JK-k. Bizonyítható [82, 92], hogy Ω^* éppen az LKM-ek halmaza, kivéve az $\omega_0 = (1, b_0)$ számpárt, amely csak $b_1 > 1$ esetén LKM. Az $\omega_\nu \doteq (B_\nu, A_\nu)$ számpárok B_ν nevezője monoton nő és a Δ_ν hiba monoton csökken ν növekedésével. Legyen

$$\Omega_0^* \doteq \{\omega_i^*, \ i = 1, 2, \ldots\} = \{\omega_0 | b_1 > 1\} \cup \{\omega_\nu, \ 1 \leqq \nu < N\}$$

az LKM-ek halmaza.

Definiáljuk még a $0 \leqq c \leqq b_\nu$, $\nu = 0, 1, \ldots$, indexekre az

(2.17) $$A_{\nu,c} = cA_{\nu-1} + A_{\nu-2}, \quad B_{\nu,c} = cB_{\nu-1} + B_{\nu-2}$$

egészeket. Bizonyítható, hogy

(2.18) $(A_{\nu,c}, B_{\nu,c})$, $(A_{\nu,c}, A_{\nu-1})$, $(B_{\nu,c}, B_{\nu-1})$ relatív prím párok,

(2.19) $$A_{\nu-1} B_{\nu,c} - A_{\nu,c} B_{\nu-1} = (-1)^\nu,$$

(2.20) $$\delta_{\nu,c} \doteq \xi - \frac{A_{\nu,c}}{B_{\nu,c}} = \frac{(-1)^\nu (\xi_\nu - c)}{B'_\nu B_{\nu,c}},$$

ahol $0 \leqq c \leqq b_\nu$, $0 \leqq \nu < N$. Az

$$\omega_{\nu,c} = \frac{A_{\nu,c}}{B_{\nu,c}}$$

törteket mellék-közelítő törteknek nevezik [82]. Páros ν esetén $\omega_{\nu,c}$ BK, páratlan ν esetén $\omega_{\nu,c}$ JK a ξ-hez.

Legyen $\bar{c}_0 = b_0$, ha $b_1 > 1$, és $\bar{c}_\nu$, $0 < \nu < N$, az a legkisebb 1 és b_ν közötti $\bar{c}$ egész szám, amelyre még

$$|\delta_\nu| \leqq |\delta_{\nu,\bar{c}}| < |\delta_{\nu-1}|$$

teljesül. Legyen

$$\Omega^- \doteq \{\omega_i^-, \quad i = 1, 2, \ldots\} \quad \text{az} \quad \omega_{2k,c}, \quad 0 \leqq c \leqq b_{2k} - 1, \quad k = 1, 2, \ldots,$$

baloldali közelítések növekvő nevező szerint rendezett halmaza,

$$\Omega^+ \doteq \{\omega_i^+,\ i = 1, 2, \ldots\} \text{ az } \omega_{2k+1,c},\quad 1 \leqq c \leqq b_{2k+1},\quad k = 0, 1, \ldots,$$

jobboldali közelítések növekvő nevező szerint rendezett halmaza és

$$\Omega^\pm = \{\omega_i^\pm,\ i = 1, 2, \ldots\} \text{ az } \omega_{0,b_0},\ \text{ha } b_1 > 1 \text{ és az}$$

$$\omega_{v,\bar{c}},\quad \bar{c}_v \leqq \bar{c} \leqq b_v,\quad v = 1, 2, \ldots,$$

közelítések növekvő nevező szerint rendezett halmaza. Legyen

$$\Omega = \Omega^- \cup \Omega^+$$

az összes közelítések növekvő nevező szerint rendezett halmaza. Véges N esetén legyenek

$$\Omega_0^-,\quad \Omega_0^+ \quad \text{és} \quad \Omega_0^\pm$$

az $\omega_n = A_n/B_n = \xi$ racionális törttel kiegészített megfelelő halmazok. Ilyenkor a halmazok végesek és ω_n az utolsó elemük. Bizonyítható, hogy mindegyik halmaz elemeinek hibája fokozatosan csökken, amint a nevező növekszik. Ugyancsak bizonyítható [92], hogy Ω_0^- az *LBKM*-ek és egyben *LBK*-k, Ω_0^+ az *LJKM*-ek és egyben *LJK*-k halmaza és $\Omega_0^\pm$ pedig az *LK*-k halmaza.

A későbbi hivatkozás érdekében az alapvető eredményeket foglaljuk egy tételbe.

2.1. TÉTEL: A

$$(\alpha) \qquad 0 \leqq By - Ax \prec \alpha$$

$$(2.21)\quad (\beta) \qquad -\alpha \prec By - Ax \leqq 0$$

$$(\gamma) \qquad |By - Ax| \prec \alpha$$

egyenlőtlenségeknek mindig van megoldásuk, ha $\alpha > 0$, vagy $\alpha = 0$, $\prec = \leqq$ és $\xi = y/x$ racionális szám, egyébként nincs megoldás. Az $(\alpha)-(\gamma)$ egyenlőtlenségek reprezentálta koincidencia feladatok $\omega^* \geqq (1, 0)$ megoldása (amikor a fentiek szerint létezik), rendre az Ω_0^-, Ω_0^+, illetve Ω_0^* első eleme, amelyre a megfelelő egyenlőtlenség teljesül, és így *LBKM*, *LJKM*, ill. *LKM*.

Bizonyítás: Utalunk a [92] tanulmány 2.1. korollárriumára és a 2.7. tételére (2.3., ill. 2.6. pont) amelyekből állításunk azonnal következik.

A (2.21) reprezentálta *KIF*-eknek a 2.1. tétel meghatározta ω^* megoldását egyszerű algoritmusok szerint lehet előállítani a $\xi > 0$ és $\alpha \geqq 0$, és $\prec$ megadása esetén. Ha $\alpha = 0$ és $\prec = <$, vagy $\prec = \leqq$ és ξ irracionális, akkor nincs megoldás. Ha $\alpha = 0$, $\prec = \leqq$ és $\xi = A/B$, $B > 0$, A, B relatív prím egészek, racionális, akkor $\omega^* = \omega_n = A_n/B_n = A/B$, amelyet, ha nem ismerünk, a lánctörtfejtés és a (2.10) rekurzív formulák segítségével határozhatunk meg. Ezt a [92] tanulmány 2.6. pontjában *R*-algoritmus néven formálisan is megfogalmaztuk. Ha $\alpha > 0$, akkor a lánctörtfejtés algoritmusán és a (2.10) és (2.17) formulákon alapuló algoritmusokkal, a (2.14) és (2.20) hiba-formulák felhasználásával, könnyen szerkeszthetünk formális algoritmusokat ω^* meghatározására. Ezek azonnal szolgáltatják az ω^* megoldás Δ^* hibáját is. A [92] tanulmányban ezeket rendre *B(M)KIFM*-algoritmus, *J(M)KIFM*-algoritmus és *K(M)KIFM*-algoritmus néven találhatjuk az ottani 2.6. pontban. Ezekre az algoritmusokra a későbbiek során is fogunk utalni.

2.3. Néhány definíció és jelölés

Célszerű néhány definíció és jelölés bevezetése, amelyek egyszerűsítéseket tesznek lehetővé a további tárgyalásban.

Vezessük be a

$$\prec \text{ és } \succ$$

vagy megkülönböztető index alkalmazásával a

$$\prec_i \text{ és } {}_i\succ$$

szimbólumokat a $\prec$ és $\leqq$, illetve $\succ$ és $\geqq$ relációjelpárok közös jelölésére. Legyen

$$\prec' \text{ és } '\succ$$

a $\prec$, ill. $\succ$ alternatív értéke, vagyis pl.

$$\prec = \leqq \text{ esetén } \prec' = \prec.$$

Az ${}_i\succ$ szimbólum a $\prec_i$ szimbólum értékének ellentettjét jelöli, vagyis pl.

$$\prec_i = \prec \text{ esetén } {}_i\succ = \succ.$$

Ezeket alternatív relációjeleknek nevezhetjük. Az

$$a \prec b \text{ és } b \succ a,$$

valamint az

$$a \prec b,\ a' \succ b \text{ és } b \prec' a$$

relációk ekvivalensek. E mesterkéltnek tűnő szimbólumok hasznát már láttuk és később is tapasztaljuk majd.

Vezessük be a

$$\lceil \text{ és } \rceil$$

jeleket a (és [kezdő, illetve) és] vég-zárójel pár közös jelzésére. $b \geqq a$ esetén az

$$\lceil a\ , b \rceil$$

forma jelölhet egy, bármely oldalról egymástól függetlenül nyílt vagy zárt intervallumot ($a = b$ esetén bármely oldalról nyílt intervallum üres vagy tilos). A $\lceil$ és $\rceil$ jelek megfelelnek a $\prec$, ill. $\succ$ jeleknek.

Egy x valós szám *egészrésze,* illetve *törtrésze* legyen a szokásos jelölésnek megfelelően

$$[x], \text{ illetve } \{x\}.$$

Néha használjuk majd tetszőleges α, β, γ, valós számokra az

$$\alpha \equiv \beta \pmod{\gamma}$$

általános kongruencia jelölést és fogalmat, ami azt jelenti, hogy

$$\left\{\frac{\alpha}{\gamma}\right\} = \left\{\frac{\beta}{\gamma}\right\}.$$

Azt mondjuk, hogy γ osztója α-nak, ha

$$\left\{\frac{\alpha}{\gamma}\right\} = 0, \quad \text{azaz} \quad \alpha \equiv 0 (\mathrm{mod}\, \gamma).$$

Bevezetünk még négy speciális lépcsős függvényt. Legyen a z valós argumentumra

$$f_{<}(z), \quad f_{\leqq}(z), \quad f_{>}(z), \quad f_{\geqq}(z)$$

rendre a *z-nél kisebb, illetve nem-nagyobb maximális, ill. a z-nél nagyobb, illetve nem-kisebb minimális egész* szám. Mind a négy függvény kifejezhető a $[z]$ egészrész függvénnyel, monoton nem-fogyó és balról, vagy jobbról folytonos egész értékű lépcsős függvény. Érvényesek az alábbi összefüggések.

Az $f_{<}(z)$ függvény balról folytonos és

$$z-1 \leqq f_{<}(z) = -[-z]-1 < z. \tag{2.22a}$$

Az $f_{\leqq}(z)$ függvény jobbról folytonos és

$$z-1 < f_{\leqq}(z) = [z] \leqq z. \tag{2.22b}$$

Az $f_{>}(z)$ függvény jobbról folytonos és

$$z < f_{>}(z) = [z]+1 \leqq z+1. \tag{2.22c}$$

Az $f_{\geqq}(z)$ függvény balról folytonos és

$$z \leqq f_{\geqq}(z) = -[-z] < z+1. \tag{2.22d}$$

Igen fontos közös tulajdonsága ezeknek a függvényeknek, hogy az

$$f.(z)-z$$

függvény periodikus egységnyi periódushosszal és ezzel ekvivalensen

$$f.(z \pm n) = f.(z) \pm n, \quad n \geqq 0 \quad \text{egész}. \tag{2.23}$$

A $\prec$ és $\succ$ alternatív relációjelek használatával felírhatók még a következő összefüggések is. A (2.22a—d) ekvivalensek a

$$z-1 \prec' f_{\prec}(z) \prec z, \quad z+1 \,'\succ f_{\succ}(z) \succ z \tag{2.22}$$

relációkkal, továbbá következnek az alábbiak.

$$\begin{aligned} f_{\succ}(z) &= f_{\prec'}(z)+1, & f_{\prec}(z) &= f_{'\succ}(z)-1 \\ f_{\succ}(z) &= -f_{'\succ}(-z)+1, & f_{\prec}(z) &= -f_{\prec'}(-z)-1 \\ f_{\succ}(z) &= -f_{\prec}(-z), & f_{\prec}(z) &= -f_{\succ}(-z). \end{aligned} \tag{2.24}$$

Ezeket az összefüggéseket a (2.22) és (2.23.) összefüggésekkel együtt direkt utalás nélkül is gyakran használni fogjuk.

Alkalmazzuk még az

$$\omega = (A, B)$$

jelölést a B és A egész számok alkotta számpár, vektor és az A/B tört jelölésére egyaránt és a konkrét tartalma a mindenkori értelméből derül ki. Mint vektor és számpár, legyen definíció szerint

$$\omega=\omega' \text{ ekvivalens } B=B' \text{ és } A=A'$$

$$\omega\leqq\omega' \text{ ekvivalens } B\leqq B' \text{ és } A\leqq A'$$

$$\omega<\omega' \text{ ekvivalens } \omega\leqq\omega' \text{ és } B<B', \text{ vagy } A<A'$$

legalább egyike teljesül.

Néha használjuk a programozási technikából kölcsönzött

$$x := \varphi(x)$$

jelölést az

$$x' = \varphi(x), \quad x = x'$$

változótranszformáció-sorozatra (párra), ahol az x és x' változók (vektorok), $\varphi(\cdot)$ egy kifejezés(vektor) és $x=x'$ az identitás (átjelölés).

3. Definíciók és általános tételek

Ebben a fejezetben az 1.2. pontban definiált szabályos job-folyam pár ütemezési problémáját vizsgáljuk általánosságban.

A 3.1. pontban bevezetjük a konfiguráció és ütemterv fogalmát, valamint egy ütemterv leírásának, megadásának módját. Bevezetünk számos fogalmat (állapot, szituáció, szakasz, kritikus szituáció stb.), amelyekkel a továbbiakban dolgozunk, és megállapodásokat teszünk a megengedhető ütemtervekkel kapcsolatban.

A 3.2. pontban különféle szempontokat, tulajdonságokat definiálunk, amelyek alkalmasak ütemtervek jellemzésére és osztályozására. Ilyenek az összefüggőség, a szorosság, a következetesség, a periodicitás. Lemmákat bizonyítunk a tulajdonságok feltételeire vonatkozóan.

A 3.3. pontban definiáljuk a dominancia fogalmát és bizonyítunk egy tételt következetes ütemtervek dominanciájának feltételeire. Ezután a domináns döntés és a gazdaságos és természetes döntés és ütemterv fogalmát definiáljuk. Bizonyítjuk az ütemtervek dominanciáját és egy lemmába foglaljuk az ütemtervek alapvető tulajdonságait.

A 3.4. pont a gazdaságos és a természetes ütemtervek további elemzését tartalmazza és bizonyítja számos tulajdonságukat, amelyeket a későbbiekben használni fogunk. Ezek elsősorban a kritikus szituációkkal és a határozott szakaszokkal kapcsolatosak. Döntő az első kritikus szituáció léte, helye és jellege, amelyek meghatározását *KIF*-ek megoldásának meghatározására vezetjük vissza (3.14. lemma). Részletesen tanulmányozzuk az elfajult konfigurációk ütemterveit, mert egyéb konfigurációk ütemterveit később gyakran erre vezethetjük vissza.

A 3.5. pontban a gyakorlatilag fontos speciális, prioritásos ütemterveket definiáljuk.

3.1. Szabályos job-folyam pár megengedhető ütemtervei

Az 1.3. pontbeli definíció szerint legyen $Q=(Q^{(1)}, Q^{(2)})$ egy *job-folyam pár* és álljanak rendelkezésre a $\mathcal{P}=(P_A, P_{B1}, P_{B2})$ *processzorok*. A job-folyamok szabályosak és igényeikre vezessük be az alábbi alternatív jelöléseket:

$$\tau_{ij}^A = \tau_i^A = \eta_i, \quad \tau_{ij}^B = \tau_i^B = \vartheta_i, \quad i = 1, 2, \quad j = 1, 2, \ldots.$$

Magára a job-folyam párra is alternatív jelöléseket fogunk használni:

$$Q = (Q^{(1)}, Q^{(2)}) \doteq \{C_{1j}, C_{2j}\} \doteq \{(A_{1j}, B_{1j}), (A_{2j}, B_{2j})\} \doteq (C_1;\ C_2) \doteq$$
$$\doteq (A_1;\ B_1;\ A_2;\ B_2).$$

Sok esetben az igények jelképezik a job-folyam-párt, ilyenkor használjuk a

$$Q = (\tau_1^A;\ \tau_1^B;\ \tau_2^A;\ \tau_2^B) \doteq (\eta_1;\ \vartheta_1;\ \eta_2;\ \vartheta_2)$$

jelöléseket. Ezt a négy igényt nevezzük a job-folyam pár *paramétereinek*. A paraméterek egy adott értékegyüttesét *konfigurációnak* nevezzük. Ha bármelyik paraméter értéke megváltozik, új konfiguráció jön létre.

Vezessük be még a következő jelöléseket is:

$$\tau_1 = \tau_1^A + \tau_1^B = \eta_1 + \vartheta_1; \quad \tau_2 = \tau_2^A + \tau_2^B = \eta_2 + \vartheta_2;$$
$$\tau^A = \tau_1^A + \tau_1^A = \eta_1 + \eta_2; \quad \tau^B = \tau_1^B + \tau_2^B = \vartheta_1 + \vartheta_2.$$

Modellünkben bármelyik paraméter bármilyen *nem-negatív véges* valós értéket felvehet. Az összes konfigurációk tere, amelyet *konfigurációtérnek* nevezünk és $\mathcal{Q}$-val fogunk jelölni, az R^4 négydimenziós euklideszi tér

$$R_+^4 = \{\eta_1 \geqq 0,\ \vartheta_1 \geqq 0, \quad \eta_2 \geqq 0,\ \vartheta_2 \geqq 0\}$$

nem-negatív tizenhatoda.

Valamely paraméter 0 értéke azt jelenti, hogy a job-folyam párban a megfelelő task-típus igénye 0, de nem jelenti annak hiányát. Az ilyen *konfigurációt elfajultnak* nevezzük. A 0 igényű task-típust *elfajult task*-nak mondjuk. Ezzel modellezhetünk egy gyakorlatilag elhanyagolhatóan kicsiny igényű, de elengedhetetlen tevékenységet. Ha a $Q^{(i)}$ job-folyam mindkét task-ja elfajult, $Q^{(i)}$ *degenerált job-folyam*. Degenerált *konfiguráció* az, amelyiknél legalább egyik job-folyam degenerált.

Speciálisan legyen a $Q=0$, *nulla konfiguráció* az, amelynél mindkét job-folyam degenerált. Az elfajult konfigurációk megengedése nehezíti az ütemezési problémák tárgyalását, azonban általános konfigurációk ütemezését sokszor ilyen konfigurációk ütemezésére vezethetjük vissza [92].

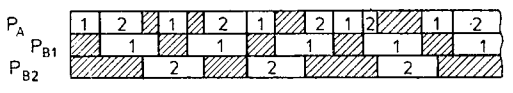
A Q job-folyam pár képezi az ütemezési problémánk igény oldalát. Az erőforrás oldal a $\mathcal{P}=(P_A, P_{B1}, P_{B2})$ processzor-hármas. Az *ütemezés* a processzorok hozzárendelése az igényekhez úgy, hogy egyidőben minden processzor legfeljebb egy igényt szolgáljon ki és minden igény csak a Q job-folyam pár sorozatban előtte álló igények kiszolgálása után szolgálható ki. Az ütemezési feltételektől függ, hogy a kiszolgálás megszakítható-e. *Lehetséges* minden olyan ütemezés, amely ezeknek a feltételeknek eleget tesz. Az ütemezéskor a 0 igényű task-ok ütemezésének pillanata és sorrendje is pontosan meghatározott. Az esetleges egyéb feltételeknek is eleget tevő ütemezéseket nevezzük *megengedhetőnek*. Ilyen feltételeket rövidesen definiálni

fogunk. Az ütemezés lehet egy döntési folyamat, vagy egy ütemterv, amely minden pillanatra megszabja, hogy melyik processzor melyik igényt szolgálja ki, illetve tétlen-e.

Az *ütemezési stratégia* az ütemezési tevékenység, vagy ütemterv elkészítés folyamatát egyértelműen meghatározó *szabály*, amely egy döntési előíráson, vagy algoritmuson keresztül valósul meg.

Jelölje egy $Q \in \mathcal{Q}$ konfigurációra $\mathcal{R}(Q)$ a megengedhető ütemtervek halmazát. Az $\mathcal{R} = \bigcup_{Q \in \mathcal{Q}} \mathcal{R}(Q)$ halmaz a *megengedhető ütemtervek tere.*

Egy ütemtervet általában a *Gantt-diagrammal* szemléltethetünk (lásd 2. ábra), amelyen a processzoroknak egy-egy időtengellyel párhuzamos sáv felel meg. A processzorok sorrendje mindig P_A, P_{B1}, P_{B2}, ezért ennek jelzését elhagyjuk. Az időegység jelölése is nélkülözhető általában, hiszen az ütemezés szempontjából csupán az igények relatív értéke lényeges. A tétlen szakaszokat sraffozott sávok jelölik. Speciálisan 0 igényt egyetlen vonal jelezhet. „Csatlakozó" igényeknél az ilyen ábrázolás nem egyértelmű, ilyenkor a *Gantt-diagramon* egyéb jelzések is szükségesek lehetnek.



2. ábra. *Job-folyam pár ütemtervének Gantt-diagramja*

A $Q^{(i)}$ $(i=1, 2)$ job-folyam $C_{ij}=(A_{ij}, B_{ij})$ $(j=1, 2, \ldots)$ task-párjának kiszolgálási igényét a job-folyam j-edik *igényciklusának* nevezzük. Ennek hossza $\tau_i \geqq 0$ (degenerált esetben $\tau_i=0$). Egy $R=R(Q)$ ütemtervben legyen $s(X)$ és $f(X)$ az X task, vagy $X=(A, B)$ task-pár kiszolgálásának első, ill. utolsó pillanata *(start-finish)*. Az $[s(C_{ij}), f(C_{ij})]$ intervallumot a $Q^{(i)}$ job-folyam j-edik *kiszolgálási ciklusának* nevezzük. Ennek hossza

$$t_{ij} = f(C_{ij}) - s(C_{ij}) \geqq \tau_i \geqq 0.$$

Az ütemezés időben játszódik le és bármely *fázisában* definiálhatjuk az *ütemezés* $\sum$' *állapotát,* mint a kiszolgálás addigi menetét, amely megfelel az ütemterv egy $[0, t)$ szakaszának. A $\sum$ matematikai megadása elfajult igények, valamint 0 tartalmú kiszolgálások esetén nem magától értetődő. Feltételezzük, hogy egy t pillanatbeli $\sum$ állapot egyértelműen meghatározza az összes $t' < t$ korábbi pillanatokbeli állapotot is. A $\sum$ állapot tulajdonképpen események sorozata, amelyek között az előidejűség nem feltétlenül jelent eltérő bekövetkezési időpontot. Egy t pillanatban több esemény bekövetkezhet, amelyek az állapotok egy $\{\sum\}_t$ rendezett halmazát definiálják. A $\sum_1$ állapot *megelőzi* a $\sum_2$ állapotot, jelölésben $\sum_1 \prec \sum_2$, ha vagy $\sum_1$ egy korábbi halmazhoz tartozik, vagy egy halmazon belül sorrendben előbb következik. Definiáljuk a $t(\sum)$ függvényt, mint az állapot *hosszát.* $t(\sum')$ annak a $\{\sum\}_t$ halmaznak az indexe, amelyhez $\sum'$ állapot tartozik.

Az ütemezés $\{\sum\}_t$ halmazok, illetve $\sum$ állapotok rendezett sorozata (halmaza), amelyet a Q konfiguráció és S ütemezési stratégia határoz meg. Jelölje ezt $\sum_{S,Q}$. Ez ekvivalens az $\mathcal{R}$ ütemtervek terének egy $R^{(S)}(Q)$ ütemtervével. A stratégiák egy $\mathcal{S}$ osztályára definiálhatjuk az $\sum_{\mathcal{S},Q} = \bigcup_{S \in \mathcal{S}} \sum_{S,Q}$ állapothalmazt.

A $\sum$ állapotok jellemzése előtt további fogalmakat vezetünk be, majd néhány megállapodással szűkítjük a megengedhető ütemezések terét.

A $\mathscr{P}=(P_A, P_{B1}, P_{B2})$ *processzor-hármas állapotát* jellemezze

$$\alpha = (\alpha_A, \alpha_{B1}, \alpha_{B2}), \text{ ahol } \alpha_X = \begin{cases} 1, & \text{ha } X \text{ processzor foglalt} \\ 0, & \text{ha } X \text{ processzor tétlen} \end{cases}$$

vektor.

Jelölje $u \geqq 0$ az α *állapot* eddigi *tartamát* az ütemezés egy fázisában. Legyen

$$p = (u; \alpha)$$

a $\mathscr{P}$ processzor-hármas jellemzése egy fázisban. Legyen a

$$\beta = (\beta^{(1)}, \beta^{(2)}), \quad \text{ahol} \quad 0 \leqq \beta^{(i)} \leqq \tau_i, \quad i = 1, 2,$$

vektor $\beta^{(i)}$ komponensének értéke a $Q^{(i)}$ job-folyam éppen kiszolgálás alatt álló (megkezdett, de be nem fejezett) igényciklusának még hátralevő igénye minden pillanatban. Ha nincs megkezdett ciklus, akkor $\beta^{(i)}=0$. Legyen a

$$v = (v^{(1)}, v^{(2)}), \quad \text{ahol} \quad v^{(i)} = 0, 1, 2,$$

vektor $v^{(i)}$ komponense a $Q^{(i)}$ job-folyam aktuális ciklusából megkezdett task-ok száma. Ez csak akkor 0, ha nincs megkezdett task, egyébként jelzi a 0 igényű task-ok kiszolgálásának eseményét is. Ekkor a

$$q = (v; \beta)$$

vektor egyértelműen jellemzi a Q job-folyam pár lokális állapotát az ütemezés egy fázisában. Legyen

$$\sigma = (p; q)$$

vektor az ütemezés *szituációja* egy fázisban. Minden $\sum$ állapothoz tartozik egy $\sigma(\sum)$ szituáció, mint annak *utolsó* szituációja. Ugyanúgy, ahogyan a $\sum$ állapotoknak, a szituációknak is egy $\{\sigma\}_t$ rendezett halmaza tartozik az ütemezés minden t pillanatához és az ütemterv t pontjához. $\sigma_1 \prec \sigma_2$ *előidejűség* értelmezése $\sum_1 \prec \sum_2$ értelmezésével analóg. $t'=t(\sum)$ esetén $t(\sigma(\sum))=t'$ a $\sigma(\sum)$ *szituáció* (egy) fellépési helye. Amíg egy R ütemtervben nem lehet két azonos $\sum$ állapot, addig lehetséges akárhány azonos σ szituáció. Használjuk még a $\sigma[t]$ és $\sum[t, \sigma]$ jelöléseket egy σ szituáció fellépési helyének feltüntetésére, ill. egy $\sum$ állapot hosszának és utolsó szituációjának feltüntetésére.

Nevezzük egy R ütemterv *lényeges pontjainak* azokat a t' pillanatokat, amelyekben a processzorok α állapota változik (akár többször is). Ezekben a pontokban $u(t)=0$. Az $u(t)$ szakaszonként lineáris jobbról folytonos függvény. Az $\alpha(t)$, mint t függvénye „jobbról folytonos" olyan értelemben, hogy a lényeges pontok között konstans a t' lényeges pontbeli utolsó $\{\sigma\}_{t'}$-beli szituációban bekövetkező értékkel. Egy $\sigma[t']$ szituációban a $\beta^{(i)}(t')=\tau_i$, ha a σ szituációban éppen egy A_{ij} task kiszolgálása kezdődik és $\beta^{(i)}(t'+t)=\tau_i-t$, ha a $[t', t'+t)$ intervallumon megszakítás nélkül a C_{ij} igényciklus kiszolgálása folyik. Ha a C_{ij} igényciklus kiszolgálása szünetel, akkor $\beta^{(i)}(t)$ konstans. $\beta^{(i)}(t)$ függvény tehát jobbról folytonos, definíció szerint. Ha $Q^{(i)}$ degenerált, akkor $\beta^{(i)}(t) \equiv 0$. Ha egy task kiszolgálása egy processzoron a t pillanatban befejeződik, akkor arra a pillanatra a processzor már szabad és másik task ütemezhető.

Egy 0 igényű task ütemezésekor a processzor még ugyanarra a pillanatra felszabadul. Ezért lehetséges többek között több esemény ugyanabban a pillanatban. A fentiek alapján beszélhetünk $\sigma(t)$ függvényről, amely a lényeges pontokban többértékű is lehet. Azonban $\sigma(t)$ folytonos a lényeges pontok között és jobbról folytonos a lényeges pontokban olyan értelemben, hogy egyik értéke a jobb oldali határérték. Ugyanakkor $\sigma(t)$ az $u(t)$ növekedése miatt sosem konstans a lényeges pontok között. A $\sigma(t)$ tulajdonságaiból és a szituáció definíciójából következik, hogy ha akár egy $R(Q)$, akár különböző $R(Q)$, $R'(Q)$ ütemtervek t_1 és t_2 pontjaiban $\sigma(t_1)=\sigma(t_2)$, akkor t_1 és t_2 pontok előtt azonos távolságban vannak az első megelőző lényeges t_1', ill. t_2' pontok és azokban az utolsó megelőző szituációkra $\sigma'(t_1')=\sigma'(t_2')$, és t_1 és t_2 vagy lényeges pontok, vagy köztük legalább egy lényeges pont van.

Definiáljunk most néhány *jellegzetes szituációt.*

Legyen $\sigma_0 \doteq 0$ az a szituáció, amelynek minden komponense 0. Legyen $\sum \doteq 0$ minden ütemterv előtti szituáció, amelyre $t(\sum)=0$ és legyen definíció szerint $\sigma(\sum)=0$, ha $\sum=0$. A $\{\sigma\}_0$ sorozatnak σ_0 mindig az első szituációja. σ_0 lényeges pontokban $t>0$ mellett is felléphet. Ha σ_0 utolsó szituáció t' pontban, akkor van olyan $\varepsilon>0$ pozitív szám, hogy a $(t', t'+\varepsilon)$ intervallumon $\sigma(t)$ minden komponense 0, kivéve $u(t)$ komponenset, amely lineárisan nő.

Azt mondjuk, hogy a t pontban a $Q^{(i)}$ job-folyam *támasztja* a $Q^{(3-i)}$ job-folyamot, vagy a $Q^{(3-i)}$ job-folyam *támaszkodik* a $Q^{(i)}$ job-folyamra és ott a β_i-*szituáció* lép fel $(i=1, 2)$, ha valamely A_{ij}, $A_{3-i,j}$ task-párra $f(A_{ij})=s(A_{3-i,j})=t$ teljesül. A β_1- és β_2-szituációkat együtt *β-szituációknak* nevezzük és azt mondjuk, hogy $Q^{(1)}$ és $Q^{(2)}$ *támaszkodnak.* A t pontban az X és Y task-ok *csatlakoznak*, pontosabban az Y csatlakozik az X-hez, ha $f(X)=s(Y)=t$. Egy β_i-szituációban a $Q^{(i)}$ job-folyam *késlelteti* a $Q^{(3-i)}$ job-folyamot, ha a t fellépési pontja előtti utolsó B_{3-i}-task-ra $f(B_{3-i})<t$, azaz korábban fejeződött be $f(A_i)$-nél.

Legyen $R \doteq R(Q) \in \mathcal{R}$ egy $Q \in \mathcal{Q}$ konfiguráció megengedhető ütemterve. Nevezzük az R ütemterv egy *szakaszának* az ütemterv bármely két $\sigma[t] \prec \sigma'[t']$ szituációja közötti szakaszát és használjuk erre a

$$\Delta \sum_R \doteq \Delta \sum_R [t, \sigma, t', \sigma'] = \Delta \sum [t, \sigma, t', \sigma']$$

jelöléseket.

A $\sigma[t]$ szituációt a $\Delta\sum_R$ *alsó határszituációjának* nevezzük és nem számítjuk hozzá a szakaszhoz. A $\sigma'[t']$ szituációt a szakasz *felső* határszituációjának nevezzük és hozzászámítjuk a szakaszhoz. A t, ill. t' a szakasz *alsó*, ill. *felső határpontja* és $0 \leqq t \leqq t' \leqq \infty$. $t'-t$ a szakasz hossza. Speciálisan $\Delta\sum_R[0, 0, t', \sigma']=\sum_R[t', \sigma']$ az R állapota és $\Delta\sum_R[t, \sigma, \infty, \cdot] \doteq \Delta\sum_R[t, \sigma]$ az ütemterv egy *utolsó szakasza.* $\sum_R[\infty, \cdot]=\Delta\sum_R[0, 0]=\Delta\sum_R[0, 0, \infty, \cdot]=R(Q)$. Ha $t=t'$ és σ és σ' a $\{\sigma\}_t$ halmaznak ugyanaz az eleme, akkor legyen definíció szerint

$$\Delta \sum_R [t, \sigma, t, \sigma] = \varnothing$$

az üres halmaz. Ha azonban a $\{\sigma\}_t$ halmaznak egynél több eleme van és a $\sigma, \sigma' \in \{\sigma\}_t$ különböző elemeket reprezentálnak, akkor

$$\Delta \sum_R [t, \sigma, t, \sigma'] \neq \varnothing,$$

függetlenül attól, hogy σ és σ' értéke azonos-e, vagy nem (azonosak csak 0 értékkel lehetnek, ha legalább az egyik job-folyam degenerált). Ha σ és σ' értéke azonos, σ, akkor

$$\Delta \sum_R = \Delta \sum_R [t, \sigma, t', \sigma], \quad t' \geqq t,$$

egy *visszatérési szakasz,* pontosabban a σ szituáció egy visszatérési szakasza. Ilyenkor σ az R ütemterv egy *visszatérő szituációja.* A $t'-t$ hossz a σ *visszatérési ideje.* Ha t' a minimális $t' \geqq t$ pont, ahol $\sigma \in \{\sigma\}_{t'}$, de σ új előfordulása, akkor $t'-t$ a σ *első* visszatérésének ideje.

Nem feltétlenül ugyanannak az R ütemtervnek a

$$\Delta \sum_1 = \Delta \sum_1 [t_1, \sigma_1, t_1', \sigma_1'] \quad \text{és} \quad \Delta \sum_2 = \Delta \sum_2 [t_2, \sigma_2, t_2', \sigma_2']$$

két szakaszát *kongruensnek* mondjuk, ha $t_2'-t_2=t_1'-t_1$, $\{\sigma\}_{t_1+t} = \{\sigma\}_{t_2+t}$, $0<t<t_2'-t_2$, és az alsó és felső határpontokban a szakaszokhoz tartozó szituációk halmazai egyenlők.

A két szakasz *részben kongruens,* ha van olyan σ'' szituáció, amely mindkettőnél alsó határszituáció, vagy a szakasz szituációja és az ezutáni szakaszaik már kongruensek. A két szakasz kongruenciáját

$$\Delta \sum_1 = \Delta \sum_2,$$

részben kongruenciáját

$$\Delta \sum_1 \approx \Delta \sum_2$$

jelöli. Az utóbbi esetben

$$\Delta \sum [t_1'', \sigma'', t_1', \sigma_1'] = \Delta \sum [t_2'', \sigma'', t_2', \sigma_2'].$$

Mindkettő magában foglalja a

$$\sigma_1' = \sigma_2'$$

egyenlőséget.

Vezessük be szakaszokra a *csatolás* műveletét. A $\Delta \sum_2 [t_2, \sigma_2, t_2', \sigma_2']$ *csatolható* a $\Delta \sum_1 [t_1, \sigma_1, t_1', \sigma_1']$ szakaszhoz, ha t_1' és t_2 végesek és $\sigma_1' = \sigma_2$. Legyen ennek jelölése

$$\Delta \sum_1 \oplus \Delta \sum_2 = \Delta \sum [t, \sigma, t', \sigma'],$$

amelynek eredménye egy $\Delta \sum$ szakasz, ahol $t=t_1$, $\sigma=\sigma_1$, $t'=t_1'+t_2'-t_2$ és $\sigma'=\sigma_2'$. A $\Delta \sum$ egy rendezett állapothalmaz a $\Delta \sum_1$ és $\Delta \sum_2$ rendezett halmazok egyesítése úgy, hogy $\Delta \sum_1$ minden állapota megelőzi $\Delta \sum_2$ minden állapotát és $\Delta \sum_1$ és $\Delta \sum_2$ elemeinek egymás közötti relációja változatlan marad. A csatolás műveletével ütemterv szakaszokból új ütemtervet konstruálhatunk ugyanarra a Q konfigurációra.

Korábban láttuk, hogy ha két szituáció megegyezik, akkor azok egyenlő távolságra vannak egy-egy lényeges ponttól, amelyekben legalább az utolsó szituációk megegyeznek. Ebből következik, hogy bármely visszatérő szituáció létezésekor létezik lényeges pontbeli visszatérő szituáció is. Egy R ütemterv első visszatérő szituációja mindig lényeges pontbeli szituáció.

A hatékonyság szempontjából vizsgálva a lehetséges és megengedhető ütemezéseket, bizonyos osztályokat célszerű eleve kirekeszteni, mivel olyanok, amelyek a gyakorlatban érdektelenek és velük azonos, vagy nagyobb hatékonyságú egyéb ütemezés mindig létezik. Ezzel a megokolással alább két megállapodás formájában bizonyos megszorításokat teszünk a megengedhető ütemezésekre.

1. Megállapodás: Pozitív igény csak pozitív idejű kiszolgálásra és véges sok intervallumra ütemezhető.

Ezzel a megszorítással elejét vesszük annak, hogy egy $\tau>0$ igényű task-ot végtelen sokszor megszakítva és 0 pillanatokra ütemezzünk, ami gyakorlatilag kivihetetlen, elvileg pedig felesleges lehetőség lenne. E megállapodásból következik, hogy ha egy $\tau>0$ (hátralevő) igényű task-ot a t pillanatra ütemezünk, akkor van olyan $\varepsilon>0$, hogy a kiszolgáló processzor a $[t, t+\varepsilon)$ intervallumban foglalt és megszakítás nélkül a szóban forgó task-ot szolgálja ki. A $t+\varepsilon$ pillanatban a task igénye $0\leqq\tau-\varepsilon<\tau$. A megállapodásból az is következik, hogy megengedhető ütemtervben a lényeges pontok végesben nem torlódhatnak. E megállapodás nem zárja ki a 0 igényű task-ok ütemezését. Azokra vonatkozóan az alábbi megszorítással élünk.

2. Megállapodás: A P_A processzoron egyazon degenerált job-folyam (0 igényű) task-jai nem ütemezhetők közvetlenül egymás után; közöttük a másik job-folyam A-task-jait kell ütemezni.

Ez is egy természetes megállapodás, amely elejét veszi annak, hogy egy degenerált job-folyam az ütemezésből kiszorítsa a másik job-folyamot. A nulla konfigurációra ez azt jelenti, hogy a két degenerált job-folyam A-task-jai csak felváltva ütemezhetők.

Az 1. és 2. megállapodások eredményeként, a $Q=0$, nulla konfiguráció kivételével, egy $R(Q)\in\mathscr{R}$ ütemterv bármely t pontjában (az ütemezés t pillanatában) a $\{\sum\}_t$ állapothalmaz és a $\{\sigma\}_t$ szituációhalmaz egyaránt véges. A megállapodásokból az is következik, hogy $\{\sigma\}_t$ halmaznak nem lehet két azonos eleme, kivéve, ha $Q^{(i)}$ job-folyam valamelyike degenerált. Bármely nem lényeges pontban a $\{\sigma\}_t$ egyetlen $\sigma(t)$ szituáció. A továbbiakban csak az 1. és 2. megállapodásoknak is eleget tevő ütemezéseket és ütemterveket tekintjük megengedhetőnek. E két megállapodás lényegesen csökkenti az $\mathscr{R}(Q)$ számosságát, amely azonban még így is nagy marad. A további szűkítésére $\mathscr{R}(Q)$-nak később kerül sor.

Egy S *stratégiát* tekinthetünk egy $s(\sum')$ *döntésfüggvénynek* is, amelynek eredményeként meghatározottá válik a $\sum'$ állapotot követő $\sum''$ állapot, vagy az állapotok egy $\{\sum\}$ halmaza (sorozata). A $\{\sum\}$ halmazt teljesen meghatározza annak $\sum''$ utolsó eleme, tehát írhatjuk, hogy $\sum''=s(\sum')$ az S stratégia egy jellemzése. A $\sum''$ állapot éppen a $\Delta\sum[t', \sigma', t'', \sigma'']$ szakasz szituációit határozza meg, ahol $t'=t(\sum')$, $\sigma'=\sigma(\sum')$, $t''=t(\sum'')$, $\sigma''=\sigma(\sum'')$. A $\Delta\sum$ szituációkban, ill. $\sum'\prec\sum\prec\sum''$ állapotokban az S stratégia szerinti döntés $s(\sum')$ által meghatározott. Egy $\mathscr{S}=\{S\}$ stratégia-osztály vonatkozásában egy $\sum\in\sum_{\mathscr{S},Q}$ állapotban az *ütemezés meghatározott,* ha van olyan $\sum'\prec\sum$, $\sum'\in\sum_{\mathscr{S},Q}$ állapot, hogy az $s(\sum')$ döntés minden $S\in\mathscr{S}$ esetén meghatározza az $s(\sum)$ döntést.

Ha egy $\sum'\in\sum_{\mathscr{S},Q}$ állapotban a $\sum''=s(\sum')$ döntések a különböző $S\in\mathscr{S}$ stratégiáknál különbözők lehetnek, akkor a $\sum'$ *állapotot kritikusnak* nevezzük az $\mathscr{S}$ osztály vonatkozásában. A megfelelő $\sigma'=\sigma(\sum')$ utolsó szituációkat *kritikus szituációknak,* és kritikus állapotokban hozható döntéseket *kritikus döntéseknek* nevezzük.

A kritikus szituációk fellépési helyei a *kritikus pontok.* Ha az $s(\sum)$ döntés csak a $\sigma(\sum)$ utolsó szituációtól függ, használjuk az $s(\sigma)$ jelölést. Ha fel kívánjuk tüntetni a döntés helyét is, az $s(\sum[t])$, vagy $s(\sigma[t])$ jelölést használjuk; ha csak a helyét jelöljük, $s[t]$ jelölést használhatunk.

3.2. *Megengedhető ütemtervek osztályozása*

Az alábbiakban néhány szempont szerint jellemezzük az $R \doteq R(Q) \in \mathcal{R}(Q)$ megengedhető ütemterveket és az azokat generáló stratégiákat. Bizonyítunk néhány lemmát azok tulajdonságaira vonatkozóan.

Először is vezessük be az 0_i, $i=1, 2$ jelölést a $Q=0$ nulla konfiguráció két ütemtervére, amelyekben a $Q^{(1)}$ és $Q^{(2)}$ degenerált job-folyamokat a $t=0$ pontban ütemezzük mégpedig a 2. megállapodásnak megfelelően felváltva $Q^{(1)}$ és $Q^{(2)}$ ciklusait. 0_i ütemtervben az ütemezést a $Q^{(i)}$ job-folyam első C_{i1} ciklusával kezdjük.

Azt mondjuk, hogy az S stratégia rendelkezik a K tulajdonsággal, ha bármely $Q \in \mathcal{Q}$ konfigurációra alkalmazva az $R^{(S)}(Q)$ ütemterv rendelkezik a K tulajdonsággal. Vagyis egy stratégia tulajdonságait az általa generált ütemtervek közös tulajdonsága határozza meg. Egy S stratégiát elvileg definiálhatunk azáltal is, hogy minden $Q \in \mathcal{Q}$ konfigurációra megadjuk az $R^{(S)}(Q)$ ütemtervet. Az S tulajdonságait ekkor ezek közös tulajdonságai jellemzik. Ha egy S stratégia által generált egyetlen $R^{(S)}(Q)$ ütemterv nem rendelkezik K tulajdonsággal, akkor S nem rendelkezik ezzel a tulajdonsággal. Ez nem zárja ki azt, hogy $R^{(S)}(Q')$ ütemterv rendelkezzék ezzel a tulajdonsággal. Az alábbiakban néhány fontos tulajdonságot definiálunk.

Az S stratégiát *megszakításosnak (preemptive)* nevezzük, ha megengedi pozitív igény ütemezését több diszjunkt intervallumba, azaz kiszolgálásának megszakítását.

Az S staratégiát *összefüggőnek (non-preemptive)* nevezzük, ha minden task kiszolgálása megszakítás nélkül befejeződik, ha egyszer elkezdődött. 0 idejű megszakítás sincs megengedve.

Az S stratégia *szoros,* ha egy processzor soha nem tétlen, ha lenne legalább egy task, amely rá ütemezhető lenne. Ez azt jelenti, hogy a task kiszolgálásra kész és ütemezése megengedett lenne.

Az S stratégia *következetes,* ha azonos szituációval végződő állapotokban, a döntés mindig azonos szituációsorozatot eredményez, ha a döntés megengedett. Vagyis $\sigma(\sum_1')=\sigma(\sum_2')$ esetén a $\sum_1''=s(\sum_1')$ és $\sum_2''=s(\sum_2')$ állapotok olyanok, hogy a $\sum_1'$ és $\sum_1''$ és $\sum_2'$ és $\sum_2''$ közé eső $\Delta\sum_1$ és $\Delta\sum_2$ szakaszok kongruensek. Ha $\sigma(\sum_1'')=\sigma_1''$ és $\sigma(\sum_2'')=\sigma_2''$, akkor $\sigma_1''=\sigma_2''$ és $t(\sum_1'')=t_1''$, $t(\sum_2'')=t_2''$, $t(\sum_1')=t_1'$, $t(\sum_2')=t_2'$, $\sigma(\sum_1')=\sigma(\sum_2')=\sigma$ jelölésekkel

$$\Delta \sum [t_1', \sigma, t_1'', \sigma_1''] = \Delta \sum [t_2', \sigma, t_2'', \sigma_2''].$$

Ez alól kivétel csak akkor lehet, ha az azonos szituációban az azonos szakaszokat generáló döntés nem volna megengedett a 2. megállapodás szerint. Ilyenkor első esetben akármilyen más döntés hozható, de utána minden ilyen szituációban következetesen ugyanaz a döntés.

Ha egy szoros stratégia generálta $R(Q)$ ütemterv a 2. megállapodás nélkül nem lenne szoros, akkor *kvázi-szorosnak* is mondhatjuk. Ugyanígy egy következetes stratégia generálta $R(Q)$ ütemterv *kvázi-következetes,* ha a 2. megállapodás nélkül nem lenne következetes. Ilyen esetek csak degenerált konfigurációnál fordulnak elő.

Egy $\mathcal{S}=\{S\}$ *stratégia-osztályt következtesnek* nevezünk, ha bármely $Q \in \mathcal{Q}$ konfiguráció mellett a kritikus $\sum' \in \sum_{\mathcal{S},Q}$ állapotokban azonos a kritikus döntések készlete és az kizárólag a $\sigma(\sum')$ utolsó (kritikus) szituációtól függ.

Az R és R' ($R, R' \in \mathcal{R}$) megengedhető ütemterveket akkor tekintjük *azonosaknak*, ha $\sum_R = \sum_{R'}$, azaz állapotaik halmaza ugyanaz. Ehhez $R, R' \in \mathcal{R}(Q)$, valamely $Q \in \mathcal{Q}$-ra, nem szükséges feltétel. R és R' azonosságát jelölje $R = R'$.

Az R' és R'' megengedhető ütemterveket *lényegében* azonosaknak nevezzük, ha van R'-nek egy $\sum_{R'}[t', \sigma]$ és R''-nek egy $\sum_{R''}[t'', \sigma]$ állapota úgy, hogy attól kezdve a két ütemterv azonos. Vagyis $\Delta \sum_{R'}[t', \sigma] = \Delta \sum_{R''}[t'', \sigma]$, az utolsó szakaszok megegyeznek. Az R' ütemterv szituációhalmaza a $t'+t$ pontban azonos az R'' ütemterv $t''+t$ pontbeli szituációhalmazával minden $t>0$ mellett: $\{\sigma\}_{t'+t} = \{\sigma\}_{t''+t}$. Az ütemtervek azonossága és lényegében azonossága a szakaszok kongruenciájának és részben kongruenciájának kiterjesztése teljes ütemtervekre.

Egy R ütemtervet *periodikusnak* nevezünk, ha van olyan véges hosszúságú $\sum'_R$ állapota, hogy attól kezdve az állapotok $\sigma(\sum)$ utolsó szituációi periodikusan visszatérnek. A legelső ilyen $\sum'_R$ állapotot az R *előperiódusának* nevezzük és annak legyen $T_R = t(\sum'_R)$ a hossza. A $\sum'_R$ utáni szituációk közös minimális visszatérési idejét tekintjük az R *periódushosszának*, amelyet p_R fog jelölni. Az ütemterv bármely T_R utáni p_R hosszúságú szakasza az R egy periódusa. Nyilván $T_R \geqq 0$ és $p_R \geqq 0$. A T_R az R-nek mindig lényeges pontja. A 2. megállapodás következménye, hogy $p_R = 0$ kizárólag a $Q=0$ nulla konfiguráció ütemterveinél fordulhat elő. Könnyű látni, hogy ennek $R \approx 0_i$, $i=1$ vagy 2, szükséges és elegendő feltétele. A definícióból következik, hogy a periodikus ütemtervnél

$$\{\sigma\}_{t+kp_R} = \{\sigma\}_t, \quad t > T_R, \quad k = 1, 2, \ldots \tag{3.1}$$

A T_R és p_R a periodikus R *ütemterv jellemzői.*

Most néhány lemmát bizonyítunk:

3.1. LEMMA: Ha $R \approx R'$, akkor a két ütemterv hatékonysága azonos:

$$\gamma_R = \gamma_{R'}, \quad \text{ha} \quad R \approx R'.$$

Bizonyítás: Legyen $t \gg \max(t_R, t_{R'})$ és legyen $t' = t_{R'} - t_R + t$. Ekkor t és t' egyszerre tartanak ∞-hez és $t - t_R = t' - t_{R'}$. Mivel R-nek a $[t_R, t)$ és R'-nek a $[t_{R'}, t')$ szakaszai azonosak, ott $\tau^A_R(t_R, t) = \tau^A_{R'}(t_{R'}, t')$ is teljesül. Azonban $\tau^A_R(t_R, t) = \tau^A_R(t) - \tau^A_R(t_R)$ és hasonlóan R'-re. $\tau^A_R(t_R) \leqq t_R$ és $\tau^A_{R'}(t_{R'}) \leqq t_{R'}$, korlátosak. Ezekből következően

$$\gamma_R \doteq \lim_{t\to\infty} \frac{\tau^A_R(t)}{t} = \lim_{t\to\infty} \frac{\tau^A_R(t) - \tau^A_R(t_R)}{t + t_{R'} - t_R} =$$

$$= \lim_{t\to\infty} \frac{\tau^A_{R'}(t') - \tau^A_{R'}(t_{R'})}{t} = \lim_{t'\to\infty} \frac{\tau^A_{R'}(t')}{t'} \doteq \gamma_{R'}.$$

3.2. LEMMA: Ha $R \approx R'$, akkor R és R' csak egyszerre lehetnek periodikusak, azonos periódusokkal.

Bizonyítás: Az állítás következik az $R \approx R'$ lényegében azonosság és a periodicitás definíciójából.

3.3. LEMMA: Ha az $R = R(Q)$ ütemtervnek nincs visszatérő szituációja, akkor R következetes.

Bizonyítás: Az állítás közvetlen következménye a következetes ütemterv definíciójának, hiszen nem lévén R-ben két azonos szituáció, a megfelelő állapotokban a döntések azonosságának kérdése fel sem merülhet.

3.4. LEMMA: Periodikus $R=R(Q)$ ütemterv minden periódusában bármelyik $Q^{(i)}$, $i=1, 2$, job-folyamnak egész μ_i számú igényciklusa kerül kiszolgálásra és minden periódusban ugyanannyi.

Bizonyítás: A $\sum'_R$ előperiódus után bármely periódus elején és a rákövetkező periódus elején azonos a szituáció és így annak $q=(v, \beta)$ komponense is. Ez azt jelenti, hogy a periódus mentén $q^{(i)}=(v^{(i)}, \beta^{(i)})$ vagy végig azonos volt, amikor is $Q^{(i)}$ job-folyam $\mu_i=0$ periódusa került kiszolgálásra, vagy változott. A $\beta^{(i)}(t)$ függvény azonban a $(0, \tau_i]$ értékkészlet tartományban folytonosan változik, kivéve a $0 \to \tau_i$ ugrásokat, vagyis ugyanazt az értéket csupán egy teljes ciklusnyi kiszolgálás után veheti fel újra. Így a kiszolgált ciklusok száma μ_i egész. Ez $i=1, 2$ mellett külön-külön igaz. Ha különböző periódusokban μ_i nem lenne azonos, akkor a periódusok nem lehetnének kongruens szakaszok.

A lemmában szereplő μ_1 és μ_2 egészeket a periódus *ciklusszámainak* nevezzük. Ezek is a *periodikus ütemtervek jellemzői* a T_R és p_R jellemzők mellett. További jellemzőként definiáljuk a periodikus ütemterv P_A-*foglaltságá*ként a P_A processzor aktivitási idejét egy periódusban. Jelölje ezt a_R. Ez nyilván minden periódusban azonos és

$$a_R = \mu_1\eta_1+\mu_2\eta_2. \tag{3.2}$$

Egy periódus P_A-kihasználtsága nyilván a_R/p_R.

3. Megállapodás: Periodikus ütemtervnél $a_R=0$ mellett legyen mindig $a_R/p_R=0$, p_R értékétől függetlenül.

Ez a megállapodás praktikus célt szolgál csupán, hogy a $p_R=0$ esetet ne kelljen kivételként kezelni.

3.5. LEMMA: Periodikus R ütemterv γ_R hatékonyságát a

$$\gamma_R = \frac{a_R}{p_R} \tag{3.3}$$

formula szolgáltatja.

Bizonyítás: A nulla konfiguráció ütemterveinél az (1.1) definíció szerint $\gamma_R=0$. A 3. megállapodás biztosítja, hogy (3.3) szerint ugyanezt kapjuk $p_R=0$ esetben is. Ha Q nem a nulla konfiguráció, akkor $p_R>0$ és a (3.3) jobb oldala értelmezve van. Legyen $t \gg T_R$ és legyen a_0 az R P_A-foglaltsági ideje a $[0, T_R)$ előperiódusban.

Nyilván $a_0 \leqq T_R$, korlátos. Legyen

$$n = \left[\frac{t-T_R}{p_R}\right].$$

Ekkor $t=T_R+np_R+p'_R$, ahol $0 \leqq p'_R < p_R$, korlátos és $\tau_R^A(t)=a_0+na_R+a'_R$, ahol $0 \leqq a'_R \leqq a_R$, a P_A-foglaltság a $[T_R+np_R, t)$ szakaszon, korlátos. t és n egy-

szerre tartanak végtelenhez, ezért

$$\gamma_R = \lim_{t\to\infty} \frac{\tau_R^A(t)}{t} = \lim_{n\to\infty} \frac{a_0 + na_R + a'_R}{l_R + np_R + p'_R} = \frac{a_R}{p_R}.$$

3.6. LEMMA: A nulla konfiguráció szoros ütemtervei csak a 0_1 és a 0_2, amelyek következetesek és periodikusak $p_R=0$ periódushosszal.

Bizonyítás: A lemma azonnal következik a szoros ütemterv definíciójából.

3.7. LEMMA: A $Q=0$ nulla konfiguráció következetes R ütemterve mindig periodikus $\sum_R=0$ előperiódussal, $p_R=\max(s(B_{11}), s(B_{21})) \geqq 0$ periódushosszal és $\mu_1=\mu_2=1$ ciklusszámmal. $p_R=0$ akkor és csak akkor, ha $R=0_i$ $(i=1, 2)$.

Bizonyítás: A 2. megállapodás szerint a $Q=0$ nulla konfiigurációnál az A_{1j} és A_{2j} task-ok a P_A processzorra csak felváltva ütemezhetők. Így az első A_{i1} task megválasztása meghatározza az A-task-ok ütemezésének sorrendjét. Egy B_{ij} task sosem ütemezhető az A_{ij} előtt és $A_{i,j+1}$ a B_{ij} előtt. E feltételek mellett legyen $A_{11}, B_{11}, A_{21}, B_{21}$ ütemezve, és legyen B_{i1} az utolsó. Legyen $s(X_{i1})=$ $=f(X_{i1})=t'_{X_{i1}}$ az X_{i1} task ütemezési pillanata. A $t'_{B_{i1}}$ pillanatban visszatér a $\sigma_0=0$ szituáció és a következetesség miatt innen kezdve az ütemtervben az összes szituációnak ismétlődnie kell a $\sigma[0]=\sigma_0$ szituációtól kezdve. Vagyis valóban $T_R=0$ és $p_R=\max(s(B_{11}), s(B_{21}))$ jellemzőkkel R periodikus.

3.8. LEMMA: Következetes R ütemterv akkor és csak akkor periodikus, ha van visszatérő szituációja.

Bizonyítás: Periodikus ütemterv T_R utáni minden szituációja definíció szerint visszatérő. Legyen R egy következetes ütemterv. Ha $Q=0$, akkor a 3.7. lemma szerint periodikus. Legyen tehát $Q\neq 0$. Tegyük fel, hogy σ az R visszatérő szituációja. Ha $Q^{(1)}$ vagy $Q^{(2)}$ degenerált, akkor könnyű belátni, hogy következetes ütemtervnek van $\sigma[t_1]=\sigma[t_2]$ $t_1<t_2$, visszatérő szituációja is, ha pedig egyik jobfolyam sem degenerált, akkor a visszatérő szituációra csak $t_1<t_2$ lehet két fellépési pontra. Ekkor t_1 és t_2 előtt az azonos távolságra levő első lényeges pontokban is azonos szituációk vannak. Ugyanakkor a következetesség miatt t_1 és t_2 után is azonos távolságra azonos $\{\sigma\}_{ii}$ szituációhalmazoknak kell lenniük. Így az ott hozott döntések azonossága folytán a következő lényeges pontokig ismét azonos szituációszakaszok következnek. Ily módon véges döntés után a t_1 előtti lényeges pontból a t_2 előtti lényeges pontba érünk és közben a t_2 utáni szakaszon a szituációk teljesen azonosaknak bizonyulnak a t_1 utáni szakaszon levőkkel.

Vagyis a $\Delta\sum[t_1, \sigma, t_2, \sigma]=\Delta\sum[t_2, \sigma, 2t_2-t_1, \sigma]$ azonosság (kongruencia) teljesül. Teljes indukcióval a $t\geqq t_1$ szakasz periodicitását így bizonyítottuk. Ebből az állítás következik, hiszen akkor T_R véges előperiódus biztosan létezik.

3.9. LEMMA: Következetes periodikus R ütemterv periódusának minden szituációja különböző, kivéve esetleg az olyan konfigurációt, amelyben van degenerált job-folyam. Az utóbbinál σ_0 kétszer felléphet.

Bizonyítás: Legyen R egy következetes periodikus ütemterv $p_R\geqq 0$ periódushosszal. Ha $p_R=0$, akkor a 3.7. lemma szerint $R=0_i$ $(i=1, 2)$. Ebben σ_0 kétszer lép fel minden periódusban. Legyen $p_R>0$. Tegyük fel, hogy σ szituáció egy periódusón belül visszatér. Ha a visszatérés ugyanabban a pontban történik, akkor

valamelyik job-folyamnak degeneráltnak kell lennie és σ csak $\sigma_0=0$ lehet. Legyen tehát $t_2>t_1$ a σ két fellépési pontja egy perióduson belül. Vagyis $T_R\leqq t_1<t_2<t_1+p_R$. A 3.8. lemma bizonyítása szerint ekkor R periodikus $p'_R=t_2-t_1<p_R$ periódushosszal is, ami ellentmond p_R definíciójának.

3.10. LEMMA: Bármely $R\in\mathcal{R}$ ütemterv visszatérési szakaszán a job-folyamok egész számú igényciklusa kerül kiszolgálásra.

Bizonyítás: A bizonyítás teljesen analóg a 3.4. lemma bizonyításához, ahol tulajdonsképpen azt használtuk ki, hogy minden periódus egy visszatérési szakasz.

3.3. Dominancia tételek

Az előző pontban definiáltuk a megengedhető ütemtervek és stratégiák néhány lehetséges tulajdonságát és velük kapcsolatos néhány lemmát. Most ezeket felhasználjuk arra, hogy a megengedhető ütemtervek terét szűkítsük úgy, hogy a szűkített osztályban mindig legyen optimális ütemterv. Erre a dominancia elvet használjuk.

A $Q\in\mathcal{Q}$ konfiguráció $R'(Q)$ *ütemterve dominálja* az $R(Q)$ ütemtervét, $R'(Q)$ *dominánsa* $R(Q)$-val szemben, ha

$$\gamma_{R'}(Q)\geqq\gamma_R(Q).$$

Az S' *stratégia dominálja* az S stratégiát, ha bármely $Q\in\mathcal{Q}$ konfiguráció esetén $R^{(S')}(Q)$ dominálja $R^{(S)}(Q)$-t.

Az ütemtervek egy $\mathcal{R}'(Q)\subset\mathcal{R}(Q)$ *osztálya dominánsa,* ha bármely $R\in\mathcal{R}(Q)$ ütemtervhez van $R'\in\mathcal{R}'(Q)$ ütemterv úgy, hogy R' dominálja R-et. Legyen $\mathcal{R}'(Q)\subset\mathcal{R}''(Q)\subset\mathcal{R}(Q)$. Az ütemtervek $\mathcal{R}'(Q)$ osztálya dominánsa $\mathcal{R}''(Q)$-*val szemben,* $\mathcal{R}'(Q)$ dominálja $\mathcal{R}''(Q)$-t, ha bármely $R''\in\mathcal{R}''(Q)$-hoz van $R'\in\mathcal{R}'(Q)$ úgy, hogy R' dominálja az R''-t. Ha az $\mathcal{R}''(Q)$ osztályt a K tulajdonság határozza meg, akkor a fenti tulajdonságú $\mathcal{R}'(Q)$ osztályt *K-dominánsnak* mondjuk.

A $Q\in\mathcal{Q}$ konfiguráció $R^*(Q)$ *ütemterve optimális,* ha $R^*(Q)\in\mathcal{R}(Q)$ és

$$\gamma_{R^*}(Q)\geqq\gamma_R(Q) \quad \text{minden } R\in\mathcal{R}(Q) \text{ mellett.}$$

Legyen $\mathcal{R}''(Q)\subset\mathcal{R}(Q)$ a K tulajdonságú ütemtervek osztálya. Az $R^*(Q)$ ütemterv *K-optimális,* ha $R^*(Q)\in\mathcal{R}''(Q)$ és $\gamma_{R^*}(Q)\geqq\gamma_{R''}(Q)$ minden $R''\in\mathcal{R}''(Q)$ mellett. Ha $\mathcal{R}'(Q)$ osztály dominánsa és $R^*(Q)\in\mathcal{R}'(Q)$ ütemtervre $\gamma_{R^*}(Q)\geqq\gamma_{R'}(Q)$ minden $R'\in\mathcal{R}'(Q)$ mellett, akkor $R^*(Q)$ optimális. Ezt a tényt használja ki a *dominancia elv:* optimális ütemtervet elegendő dominánsosztályon belül keresni, K-optimális ütemtervet elegendő K-dominánsosztályon belül keresni. A dominancia elv azt is jelenti, hogy optimális ütemezés keresésekor mindig kizárhatunk a vizsgálatból olyan ütemezéseket, amelyeknél nem rosszabb ütemterv minden konfigurációnál marad vizsgálatunk körében.

Vezessük be a $Q^{(i)}$ job-folyam P_A*-igényének mértékeként* a

$$\gamma^{(i)}=\tau_i^A/\tau_i\leqq 1 \quad (=0, \text{ ha } \tau_i=0) \tag{3.4}$$

mennyiséget, ha $\tau_i>0$. Legyen definíció szerint $\gamma^{(i)}=0$, ha $\tau_i=0$, azaz $Q^{(i)}$ degenerált. Nyilvánvalóan $\gamma^{(i)}\leqq 1$, hiszen $\tau_i=\tau_i^A+\tau_i^B\geqq\tau_i^A$.

Egy bármilyen lehetséges $R(Q)$ ütemterv γ_R hatékonyságára igaz a

$$\gamma_R \leqq \min(1, \gamma^{(1)}+\gamma^{(2)}) \tag{3.5}$$

becslés, amely γ_R és $\gamma^{(i)}$ mennyiségek definíciója alapján nyilvánvaló.

Az (1.1) definícióból következik, hogy bármely R ütemterv hatékonyságát annak nem korlátos szakasza határozza meg. Ha R és R' ütemtervek csak véges szakaszokon térnek el egymástól, akkor hatékonyságuk azonos a 3.1. lemma szerint. Ezért egy $R(Q)$ ütemterv bármely véges szakaszát megváltoztathatjuk úgy, hogy érvényes (elfogadható) ütemterv maradjon, ezzel hatékonysága nem változik meg. Logikailag ide kívánkozik a következő

3.1. TÉTEL: A következetes ütemtervek osztálya domináns.

Bizonyítás: Az állítás következik a [92] tanulmányban bizonyított azon tényből, hogy az alábbiakban definiálásra kerülő következetes gazdaságos ütemtervek szűkebb osztálya is domináns.

A dominancia elv alkalmazásához bevezetünk egy fogalmat, a domináns döntés fogalmát, majd annak segítségével két domináns ütemezési stratégia-osztályt definiálunk.

Egy $\sum$ állapotban az $s'(\sum)$ *döntés domináns* az $s(\sum)$ döntéssel szemben, ha az $s'(\sum)$ döntés megengedett és mellette a folyamatban levő, vagy az éppen következő kiszolgálási ciklus egyike sem végződik később, mint $s(\sum)$ döntés esetén. Ha $f'(C_i)$ és $f(C_i)$, $i=1,2$ jelöli az $s'(\sum)$, illetve $s(\sum)$ melletti ciklusvégződéseket, akkor $f'(C_i) \leqq f(C_i)$, $i=1,2$. Ha $f'(C_i)$ és $f(C_i)$ nem lennének meghatározottak $s'(\sum)$, illetve $s(\sum)$ által, akkor a lehetséges további meghatározó döntések melletti minimális értéket kell számításba venni. Ha $f'(C_i) < f(C_i)$, de $f'(C_{3-i}) > f(C_{3-i})$, akkor $s(\sum)$ és $s'(\sum)$ egyike sem dominálja a másikat.

Egy $\sum$ állapotban az $s'(\sum)$ *döntést gazdaságosnak* nevezzük, ha nincs olyan más döntés, amely dominálná.

Egy $\sum$ állapotban az $s'(\sum)$ *döntést természetesnek* nevezzük, ha összefüggő és nincs más olyan összefüggő döntés, amely dominálná. Nyilvánvalóan nem minden természetes döntés gazdaságos és nem minden gazdaságos döntés természetes. A gazdaságos döntés lehet megszakításos, vagy összefüggő is, a gazdaságos döntések együttesen bármely egyéb döntéseket dominálnak. A természetes döntések együttesen dominálnak bármely egyéb összefüggő döntést.

A dominancia, gazdaságosság és természetesség definíciója olyan, hogy az csak a $\sigma(\sum)$ utolsó szituációtól függ és nem függ a teljes $\sum$ állapottól, feltéve, hogy $\sum$ mentén a döntések dominánsak. Fontos megjegyezni, hogy a domináns döntés mindig *megengedett*, ami azt jelenti, hogy az 1. és 2. megállapodásokat és egyéb természetes feltételt (pl. egy processzor egyszerre csak egy task-ot szolgálhat ki stb.) nem sért. Így egy döntés tulajdonképpen *kvázidomináns,* ha pl. a 2. megállapodás miatt késleltet degenerált job-folyam ütemezést.

Nevezzük *gazdaságosnak* azt az $R'(Q)$ *ütemtervet,* amelyben minden $\sum$ állapotban (azaz minden szituációban) gazdaságos ütemezési döntés van. Nevezzük *természetesnek* azt az összefüggő *ütemtervet,* amelyben minden $\sum$ állapotban (azaz minden szituációban) természetes ütemezési döntés van.

3.2. TÉTEL: A természetes ütemtervek összefüggő-dominánsak, a gazdaságos ütemtervek pedig dominánsak.

Bizonyítás: Legyen $R(Q)$ egy tetszőleges megengedhető összefüggő ütemterv. Ha $R(Q)$-nak bármelyik $\sum$ állapotában az $s(\sum)$ döntés nem természetes, helyettesítsük azt egy $s'(\sum)$ természetes döntéssel, amely azt dominálja. Ezzel $f(C_i)$ pontok legalább egyikét előbbre hozzuk. Az így nyert új $R(Q)$ dominálja a régit, hiszen az ütemezési ciklusok előbbre hozása útján bármely $[0, t)$ intervallumon P_A igénybevétele növekszik, ily módon az (1.1) határérték nem csökkenhet. Ha $R(Q)$-nál $t=0$ ponttól indulva a döntéseket ténylegesen természetesekkel cseréljük fel, meg is konstruálhatjuk $R(Q)$ egy domináns $R'(Q)$ természetes ütemtervét. Teljesen analóg módon bizonyítható a gazdaságos ütemtervek dominanciája bármely $R(Q)$ megengedhető ütemtervvel szemben.

A gazdaságos és természetes ütemterveknek megfelelően analóg módon definiáljuk a *gazdaságos* és *természetes ütemezés* és *stratégia* fogalmát.

A természetes és gazdaságos ütemezések erősen szűkítik a megengedhető ütemezések terét, amelyben az optimális ütemezést keresni kell. A következő lemma ezt bizonyítja.

3.11. LEMMA: Egy *gazdaságos* ütemterv

(a) mindig szoros (kváziszoros),

(b) abban egyetlen megszakított A-task sem szakíthat meg másik A-task-ot,

(c) az ütemtervet teljesen meghatározzák azokban a $\sigma[t]$ szituációkban hozott ütemezési döntések, amelyekben $\beta(t)=(\beta^{(1)}(t), \beta^{(2)}(t))$ komponensei közül vagy mindkettő 0 (σ_0), vagy az egyik 0, a másikra pedig a $\vartheta_i<\beta^{(i)}(t)<\tau_i$ reláció teljesül (σ_{3-i}); ezek a kritikus szituációk; a döntés ekvivalens annak eldöntésével, hogy az adott pillanatban melyik job-folyamot ütemezzük a P_A processzorra; jelölje e döntéseket s_1 és s_2; ezek a kritikus döntések.

Egy *természetes* ütemtervet teljesen meghatároznak azokban a $\sigma[t]$ szituációkban hozott ütemezési döntések, amelyekben $\beta(t)$ komponensei közül vagy mindkettő 0 (σ_0), vagy az egyik 0, a másikra pedig $\beta^{(i)}(t)\leqq\vartheta^i$ és $0<\beta^{(i)}(t)<\eta_{3-i}$ relációk teljesülnek (σ_{3-i}); ezek a kritikus szituációk; ekkor a döntés a σ_0 szituációban annak eldöntésével ekvivalens, melyik job-folyamot ütemezzük először P_A-ra (s_1 és s_2 döntések), a σ_{3-i} kritikus szituációban pedig annak eldöntésével, hogy a $Q^{(3-i)}$ job-folyamot ütemezzük-e azonnal P_A-ra (s_{3-i} döntés), vagy hagyjuk a P_A processzort tétlenül a következő $f(C_i)$ pontig, ahol a $Q^{(i)}$ job-folyamot ütemezzük P_A-ra (s_0 döntés); ezek a kritikus döntések.

A kritikus szituációk lényeges pontokban lépnek fel és ott — de csak ott — a job-folyamok késleltetik egymást.

Bizonyítás: A gazdaságos ütemtervekre egyszerre bizonyítjuk az (a), (b) és (c) állításokat és (c)-vel együtt a természetes ütemtervekre vonatkozó állítást is úgy, hogy fokozatosan kizárunk nem domináns döntéseket.

Nyilvánvalóan nem lehet egy $\sum[t, \sigma]$ állapotban domináns olyan döntés, hogy a teljes $\mathscr{P}$ processzorhármas tétlen legyen valamely $t'>t$ pillanatig. Ebből az is következik, hogy minden gazdaságos és természetes ütemterv a $t=0$ pontban valamelyik A-task ütemezésével kezdődik.

Nem lehet domináns egy $f(A_i)$ pontban olyan döntés, amely szerint a megfelelő B_i-task nem azonnal ütemeződne, mert ez hátráltatná az $f(C_i)$ ciklusvégződést. Tehát gazdaságos és természetes ütemtervekben a B-task-ok az A-task-okhoz csatlakozóan ütemeződnek. Ugyancsak nem lehet domináns B_i task megszakítása.

Ebből következik, hogy kritikus szituáció csak $f(B_i)$ pontokban léphet fel, amíg $\beta^{(i)}(t)=0$, $i=1$, vagy 2. Ilyenkor három eset lehetséges:

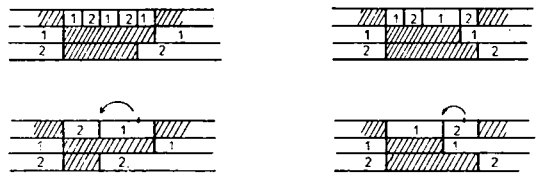
(α) egy B_{3-i}-task is ugyanabban a pillanatban fejeződött be,

(β) egy B_{3-i}-task kiszolgálás alatt áll,

(γ) egy A_{3-i}-task áll kiszolgálás alatt.

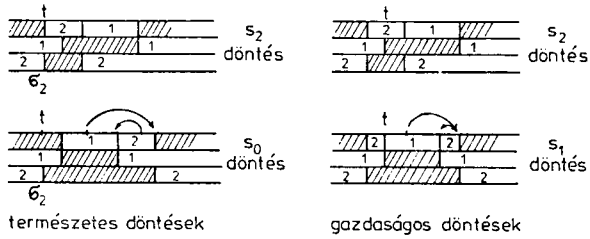
Az (α) esetben $\sigma=\sigma_0=0$ és egy A_1- és egy A_2-task egyszerre kész az ütemezésre. Két domináns döntés lehet: s_1 és s_2, hogy A_1- vagy A_2-task-ot ütemezünk-e először.

Mind gazdaságos, mind természetes ütemtervben csak összefüggő ütemezés lehet domináns, amint a 3. ábra szemlélteti.



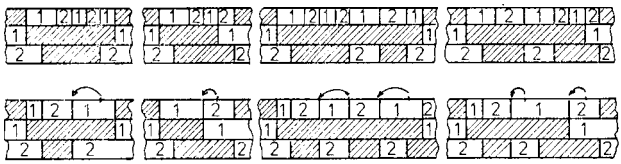
3. ábra

A (β) esetben, amikor $\beta^{(i)}(t)=0$, $0<\beta^{(3-i)}(t)\leqq\vartheta_{3-i}$, mindkét ütemtervnél domináns döntés A_i azonnali ütemezése akkor, ha $\eta_i\leqq\beta^{(3-i)}(t)$, így a szituáció nem kritikus. Ha azonban $\eta_i>\beta^{(3-i)}(t)$, akkor két domináns összefüggő ütemezési döntés lehetséges, tehát a szituáció a természetes ütemezésnél kritikus (σ_i). Ezt a 4. ábra szemlélteti. Gazdaságos ütemezésnél az A_{3-i} ütemezése domináns és a szituáció nem kritikus. Ebből következik, hogy gazdaságos ütemterv feltétlenül szoros.



4. ábra

A (γ) esetben, amikor $\beta^{(i)}(t)=0$, $\vartheta_{3-i}<\beta^{(3-i)}(t)<\tau_{3-i}$, a természetes ütemtervben, ahol megszakítás nem megengedett, a szituáció nem kritikus: a folyamatban levő A_{3-i}-task kiszolgálása folytatódik. A gazdaságos ütemtervben két domináns döntés van: A_{3-i} folytatása befejezésig, vagy megszakítása A_i kiszolgálására (s_{3-i}, ill. s_i döntés). Hogy egyik sem dominálja a másikat, azt a fenti illusztráció mutatja. Az 5. ábra igazolja, hogy egy megszakító task megszakítása vagy egy megszakított task által megszakítás nem lehet gazdaságos (domináns). Mindig csak az előbb kezdődött A-task megszakítása lehet gazdaságos a később kezdődött által. A megszakítás azonban soha nem domináns a nem megszakítással szemben, mert bár



5. ábra

a megszakító task ciklusának befejezését előbbre hozza, a megszakított task ciklusának befejezését viszont késlelteti.

A kritikus szituációkban a $\mathscr{P}$ processzorhármas állapota mindig változik, tehát azok lényeges pontokban lépnek fel. A kritikus döntések eredményezte késleltetés következik abból, hogy mindegyik döntésnél valamelyik job-folyam egy ciklusa később fejeződhet csak be, mint a másik kritikus döntésnél. Minden kritikus döntés meghatározza az először ütemezendő A-task-ot, amelynek kiszolgálása nem szakítható meg. Annak befejezése után viszont bármely domináns döntés egyértelmű, ami azt jelenti, hogy egyéb döntésnél egyik ciklusvégződés sem csökkenne, így egyik ciklus sem lehet késleltetve.

A 3.11. lemma alapján a *gazdaságos és természetes ütemtervek jellegzetességeit* a következőképpen foglalhatjuk össze.

A gazdaságos ütemtervek olyan kvázi szoros ütemtervek, amelyekben lehetnek megszakítások, de nincsenek felesleges kölcsönös megszakítások. Mindig létezik összefüggő gazdaságos ütemterv is, ami egyben természetes ütemterv. A gazdaságos ütemtervet meghatározzák az alábbi kritikus szituációkban hozott (kritikus) döntések:

$$\sigma_0\colon \beta^{(1)}(t) = \beta^{(2)}(t) = 0,$$

$$\sigma_{1,1}\colon \beta^{(1)}(t) = 0, \quad \vartheta_2 < \beta^{(2)}(t) < \tau_2,$$

$$\sigma_{2,1}\colon \vartheta_1 < \beta^{(1)}(t) < \tau_1, \quad \beta^{(2)}(t) = 0.$$

A σ_0 szituáció értéke csak 0 lehet, a $\sigma_{1,1}$ és $\sigma_{2,1}$ típusú szituációk értéke azonban nem feltétlenül ugyanaz minden előforduláskor. A kritikus döntések s_1 és s_2. A σ_0 szituációban egyik sem, a $\sigma_{1,1}$-szituációban az s_1, a $\sigma_{2,1}$-szituációban az s_2 döntés megszakítást jelent.

A természetes ütemtervek összefüggőek, de nem feltétlenül szorosak. Mindig létezik szoros természetes ütemterv is, ami egyben gazdaságos ütemterv. A természetes ütemtervet meghatározzák az alábbi kritikus szituációkban hozott (kritikus) döntések:

$$\sigma_0\colon \beta^{(1)}(t) = \beta^{(2)}(t) = 0,$$

$$\sigma_{1,0}\colon \beta^{(1)}(t) = 0, \quad 0 < \beta^{(2)}(t) \leqq \vartheta_2, \quad 0 < \beta^{(2)}(t) < \eta_1,$$

$$\sigma_{2,0}\colon 0 < \beta^{(1)}(t) \leqq \vartheta_1, \quad 0 < \beta^{(1)}(t) \leqq \eta_2, \quad \beta^{(2)}(t) = 0.$$

A σ_0 szituáció értéke mindig 0, de a $\sigma_{1,0}$ és $\sigma_{2,0}$ típusú szituációk értéke többféle lehet többszöri előforduláskor. A kritikus döntések a σ_0 szituációban s_1 és s_2, a $\sigma_{i,0}$ szituációban pedig s_i és s_0.

Fontos megjegyezni azt, hogy mind a gazdaságos, mind a természetes ütemezésnél, ha egy kritikus szituációban nem a lemma szerinti kritikus döntést hozzuk, az ütemterv megszűnik gazdaságos, illetve természetes ütemterv lenni és további szakaszára a lemma állításai nem érvényesek.

A 3.2. tétel és 3.11. lemma következménye az alábbi

3.3. TÉTEL: A szoros (megszakításos) ütemtervek osztálya domináns.

Bizonyítás: Mivel a gazdaságos ütemtervek osztálya domináns és minden gazdaságos ütemterv szoros, a szoros ütemtervek tágabb osztálya szükségképpen domináns.

Nyitott kérdés egyelőre, hogy az összefüggő szoros ütemtervek osztálya domináns-e az összefüggő ütemtervekkel szemben.

3.4. Gazdaságos és természetes ütemtervek

A gazdaságos és a természetes ütemtervek fontosságát a 3.2. tétel szerinti domináns tulajdonsága jelenti. Ezek tulajdonságainak vizsgálata, a 3.11. lemmán túl is szükséges, mert bár szűkebb osztályt alkotnak a megengedhető ütemtervek terében, mégis nem nyilvánvaló az optimális ütemtervek megkeresése.

A gazdaságos és a természetes ütemezés fogalma könnyen általánosítható lenne kettőnél több szabályos job-folyamra és nem szabályos job-folyamokra. A kritikus szituációk létezése még nem determinisztikus ütemezési stratégiák és sztochasztikus job-folyamok esetén is egyszerűsítené az ütemezés problémáját.

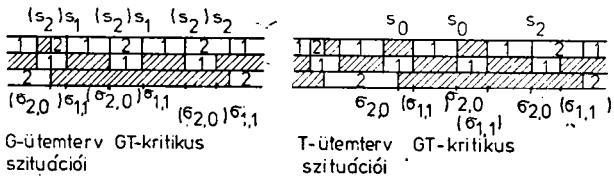
A gazdaságos és a természetes ütemtervek tulajdonságai sok vonatkozásban hasonlók, vagy analógiát mutatnak, ezért a két osztályt egy bizonyos határig célszerű együtt tárgyalni. Az írás rövidítése érdekében célszerű bizonyos rövidítéseket bevezetni.

Nevezzük *GT ütemtervnek* azt az $R=R(Q)$ ütemtervet, amely *vagy* gazdaságos, *vagy* természetes. Használjuk a *G-ütemterv* elnevezést is egy gazdaságos ütemtervre és a *T-ütemterv* elnevezést is egy természetes ütemtervre. Nevezzük *G-kritikus szituációknak* a *G*-ütemtervek $\sigma_0, \sigma_{1,1}, \sigma_{2,1}$ típusú szituációit és *T-kritikus szituációknak* a *T*-ütemtervek $\sigma_0, \sigma_{1,0}, \sigma_{2,0}$ típusú szituációit. Ezeket együtt nevezzük *GT-kritikus szituációknak*. Egy *GT*-ütemtervben annak tényleges típusától függetlenül nemcsak a *saját kritikus szituációk,* amelyek típusa saját, hanem a másik nem saját kritikus szituációk is felismerhetők, jóllehet nem pontosan a 3.11. lemma szerinti jellemzői vannak.

Nem kritikus szituációja egy *GT*-ütemtervnek egy σ_0 állapot, ha ott a 2. megállapodás miatt az ütemezés egyértelműen meghatározott. E fontos tényt sokszor ki kell majd használnunk. Ettől eltekintve a σ_0-szituációk egyszerre saját kritikus szituációi mind a *G*-, mind a *T*-ütemterveknek.

A *G*-ütemtervekben minden $\sigma_{i,1}[t]$ saját kritikus szituáció előtt $\Delta t \doteq \min(\tau_{3-i} - \beta^{(3-i)}(t), \vartheta_i)$ idővel olyan szituáció található, amely egy *T*-ütemtervben $\sigma_{3-i,0}$ típusú kritikus szituáció lenne. Most azonban a $t'=t-\Delta t$ pontban a $\beta^{(3-i)}(t')=0$ feltétel nem okvetlen teljesül, amint ezt a 6. ábrán illusztráljuk.

A *T*-ütemtervekben minden $\sigma_{i,0}[t]$ kritikus szituáció után $\Delta t \doteq \beta^{(i)}(t)$ idővel olyan szituáció lép fel, amely egy *G*-ütemtervben $\sigma_{3-i,1}$ típusú kritikus szituáció lenne. Most azonban a $t'=t+\Delta t$ pontban $\beta^{(i)}(t')=0$ is lehetséges, amint ezt a 6. ábra mutatja.

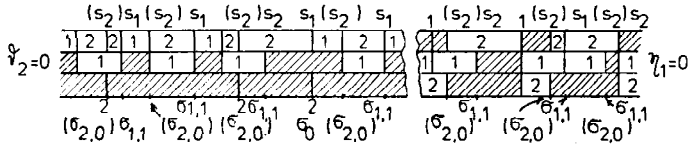


6. ábra

A *GT*-kritikus szituáció elnevezést terjesszük ki a nem saját, de mégis felismerhető szituációkra, amelyek a másik típusú ütemtervnél lennének kritikus szituációk. A *GT*-kritikus szituációk, mint látjuk, párban lépnek fel, amelyek σ_0 esetén egybeesnek, egyébként

$$\sigma_{1,0} \quad \text{és} \quad \sigma_{2,1}$$
$$\sigma_{2,0} \quad \text{és} \quad \sigma_{1.1}$$

alkotnak párokat és a $\sigma_{i,0}$ mindig megelőzi a $\sigma_{3-i,1}$ típusú *GT*-kritikus szituációt. $t=t(\sigma_{i,0})$ és $t'=t(\sigma_{3-i,1})$ esetén $t'-t \leqq \min(\vartheta_i, \eta_{3-i})$, amint ez a fenti illusztrációkon ellenőrizhető. A $\sigma_{i,0}$ és $\sigma_{3-i,1}$ *GT*-kritikus szituációpár egymásutáni lényeges pontban lép fel, amint erről szemlélet alapján könnyen meggyőződhetünk. (Lásd a 6. és 7. ábra illusztrációit.)



7. ábra *Összetartozó GT-kritikus szituációpárok.*

A *GT*-ütemtervek a (saját) kritikus szituációkban hozott (kritikus) döntésekkel egyértelműen meghatározottak. Mégpedig minden nem-kritikus szituációban a (domináns) döntés egyértelműen meghatározott az utolsó kritikus döntés által. Így többek között a nem saját *GT*-kritikus szituációkban is egyértelműen meghatározott a döntés az utolsó kritikus döntés által. Nem *GT*-kritikus szituációkban viszont a döntések egy *G*- és egy *T*-ütemtervben is azonosak, ha kölcsönösen „elfogadható" (domináns) döntés van a nem saját kritikus szituációkban is. Ezeket a tényeket precízebben kifejezi a következő 3.12. lemma.

Korábbi definíciónkkal összhangban nevezzük egy *GT*-ütemterv saját kritikus szituációi határolta szakaszait az ütemterv *határozott szakaszainak*. Ha van utolsó kritikus szituációja, akkor az utána kezdődő szakaszát szintén határozott szakasznak tekintjük. (Utolsó határozott szakasz.) Egy *G*-ütemterv határozott szakaszát nevezzük *G-szakasznak*, egy *T*-ütemterv határozott szakaszát *T-szakasznak*, ha a megkülönböztetés szükséges. A határozott szakaszok tulajdonságait foglalja össze a következő lemma. Előrebocsátjuk, hogy egy határozott szakasz hossza lehet 0 is. Ilyenkor egyetlen pontbeli szituációsorozat alkotja.

3.12. LEMMA: Egy tetszőleges $Q \in \mathcal{Q}$ konfiguráció mellett

(I) Q bármely *GT*-ütemtervénél,
(1) határozott szakaszokon az ütemezés következetes,

(2) a határozott szakaszt egyértelműen meghatározza az alsó (kritikus) határszituációjában hozott (kritikus) döntés,
(3) minden pontbeli utolsó szituáció egyértelműen meghatározza a határozott szakasz további részét; az alsó határpontbeli a teljes szakaszt,
(4) egy határozott szakasz csak akkor tartalmazhat egyenlő szituációkat, ha utolsó határozott szakasz (speciálisan a teljes ütemterv); ilyenkor az periodikus,
(5) ha két határozott szakasznak van egyenlő szituációja, akkor egyik sem lehet az utolsó határozott szakasz;

(II) Q mind gazdaságos, mind természetes ütemterveinél
(6) azonos kritikus szituációkban azonos kritikus döntések kongruens határozott szakaszt generálnak,
(7) ha két határozott szakasznak van egyenlő szituációja, akkor azok részben kongruensek; egyenlő első szituáció esetén kongruensek;

(III) Q bármilyen GT-ütemterveinél
(8) ha két határozott szakasznak van egyenlő szituációja, akkor az első ilyen pár lényeges pontokban van,
(9) ha két határozott szakasznak van egyenlő szituációja, akkor attól kezdve a két szakasz azonos mindaddig, amíg valamelyik be nem fejeződik; ha bármelyik szakasz egy utolsó határozott szakasz, akkor a másik is az és lényegében azonos GT-ütemtervek utolsó szakaszai.

Bizonyítás: Ha Q a nulla konfiguráció, akkor csak az 0_i, $i=1, 2$, a GT-ütemtervek, amelyekre az állítások triviálisak, vagy érdektelenek. Tegyük fel tehát, hogy $Q \neq 0$. Az (1)—(9) állítások (tulajdonságok) bizonyítása rendre a következő.

(1) következik abból, hogy nem-kritikus szituációkban a domináns döntés egyértelműen meghatározott a szituáció által bármely GT-ütemezésnél,

(2) következik a 3.11. lemmából,

(3) következik abból, hogy bármely pontban az utolsó szituáció már a döntés eredménye, tehát azonossága a döntés azonosságát is jelenti,

(4) következik abból, hogy egy ütemterv lényeges pontjai nem torlódhatnak végesben, ha $Q \neq 0$, és azonos szituációk vagy lényeges pontokban vannak, vagy legalább egy lényeges pont elválasztja azokat. Az (1) tulajdonság folytán azonban, ha egy σ szituáció egy határozott szakaszon egyszer visszatér, akkor végtelen sokszor vissza kell térnie. Ezért a szakasz végtelen hosszú és így csak utolsó szakasz lehet. A következetesség miatt ez a szakasz ekkor periodikus is, amint azt a 3.8. és 3.9. lemmákhoz hasonlóan egyszerűen beláthatjuk,

(5) következik a (3) tulajdonságból és abból a tényből, hogy egy GT-ütemtervnek csak egy utolsó határozott szakasza lehet,

(6) következik a (2) tulajdonságból,

(7) következik a (3) tulajdonságból,

(8) következik abból, hogy az alsó határszituációk lényeges pontok és egyenlő szituációpár előtti lényeges pontokban mindig vannak egyenlő szituációk,

(9) következik a (3) tulajdonságból, valamint abból, hogy nem-GT-kritikus szituációban a gazdaságos és a természetes döntés azonos és a szituáció által egyértelműen meghatározott. Emiatt ugyanis az egyenlő szituáció után addig a két szakasznak azonosnak kell lennie, amíg valamelyikben egy kritikus döntés be nem következik. Ez azonban csak a határozott szakasz végén történhet meg. Ha az egyik szakasz utolsó határozott szakasz, akkor abban GT-kritikus szituáció az egyenlő szituáció

után nem léphet fel. Ugyanis akkor a szomszédos lényeges pontban a másik minőségű GT-kritikus szituáció is fellépne és valamelyik saját lenne, ahol a szakasznak be kéne fejeződnie. Ekkor azonban a másik határozott szakasz sem fejeződhet be végesben, hisz az csak közös GT-kritikus pontban volna lehetséges. Így valóban mindketten utolsó határozott szakaszok, amelyek szükségképpen részben kongruensek. Ekkor azonban a két GT-ütemterv, amelynek utolsó szakaszaik, véges kezdeti szakaszokon kívül megegyezik. Ez ekvivalens a lényegében azonosság definíciójával.

E lemma segítségével könnyű igazolni a következő tételt.

3.4. TÉTEL: Annak szükséges és elegendő feltétele, hogy egy GT ütemterv következetes legyen az, hogy a kritikus szituációkban következetes legyen. Elegendő, ha az azonos típusú kritikus szituációkban a döntések azonosak a 2. megállapodás figyelembevételével.

Bizonyítás: Definíció szerint nyilvánvaló, hogy az ütemterv következetességéhez szükséges, hogy minden egyenlő kritikus szituációkban a kritikus döntések is azonosak legyenek a 2. megállapodásnak megfelelő módosítással. Ha egyenlő kritikus szituációkban ugyanazt a döntést hozzuk, akkor a szituációk után kezdődő határozott szakaszok első szituációi biztosan egyenlők lesznek. A 3.12. lemma (7) állítása szerint ekkor az ilyen határozott szakaszok kongruensek. A 3.12. lemma (1) állítása szerint viszont a határozott szakaszokon belül az ütemezés mindig következetes. Ha az azonos típusú kritikus szituációkban azonos kritikus döntést hozunk, akkor ezzel az egyenlő szituációkban hozott döntések eleve egyenlőek, azaz következetesek. Az ütemterv tehát mindkét esetben következetes.

Könnyű látni, hogy a σ_0 szituációban valamelyik A-task összefüggő ütemezése a domináns döntés. Ez azt jelenti, hogy minden GT-ütemterv összefüggő ütemezéssel kezdődik, vagyis a gazdaságos ütemtervek is összefüggőek legalább az első kritikus szituációig, de ha az σ_0, akkor még tovább. Hiszen minden ütemezés a $\sigma_0=0$ szituáció után kezdődik a $t=0$ pontban, így a G-ütemtervek is. Ez azt jelenti, hogy az összes GT-ütemterv első határozott szakasza mindössze négyféle lehet: kettő-kettőféle a G-, illetve T-ütemterveknél az $s(0)=s_1$ és s_2 döntésnek megfelelően. Azonos $s(0)=s_i$ döntés mellett a 3.12. lemma (9) állítása szerint a G- és T-ütemtervek első határozott szakaszai megegyeznek az első T-kritikus pontig (ha ilyen van) és csupán abban térnek el, hogy a G-szakasz tovább folytatódik a következő lényeges pontbeli G-kritikus szituációig (ha az nem σ_0, amikoris egybeesnek). A T-szakasz kritikus szituációi a σ_0, $\sigma_{1,0}$ és $\sigma_{2,0}$ típusok lehetnek és ennek megfelelően az azonos döntéssel kezdődő G-szakasz kritikus szituációja σ_0, $\sigma_{2,1}$, ill. $\sigma_{1,1}$ feltéve, hogy létezik kritikus szituációjuk. Látni fogjuk, hogy az első határozott szakaszok nem feltétlenül végesek. Az alábbiakban a GT-ütemtervek első határozott szakaszainak tulajdonságait vizsgáljuk, amelyből fontos törvényszerűségeket tudunk majd levezetni a teljes ütemtervekre, elsősorban a természetes ütemezésekre.

Egy $Q\in\mathcal{Q}$ konfiguráció megengedhető ütemterveinek halmazát $\mathcal{R}(Q)$ jelöli. Legyen $\mathcal{R}^{(GT)}(Q)$ a GT-ütemtervek részhalmaza. Ezen belül legyen $\mathcal{R}^{(G)}(Q)$ és $\mathcal{R}^{(T)}(Q)$ a gazdaságos, illetve természetes ütemtervek halmaza. Az utóbbi kettő nem diszjunkt, mert minden konfigurációnak van olyan ütemterve, amely egyszerre gazdaságos és természetes (összefüggő és szoros). $\mathcal{R}^{(GT)}$, $\mathcal{R}^{(G)}$ és $\mathcal{R}^{(T)}$ halmazok mindegyike felbomlik azonban két-két diszjunkt részre, amelyeket az $s(0)=s_1$,

illetve $s(0)=s_2$ kritikus döntéssel kezdődő ütemtervek alkotnak. A felbontások legyenek ennek megfelelően

$$\mathcal{R}^{(GT)} = \mathcal{R}^{(1)} \cup \mathcal{R}^{(2)}, \quad \mathcal{R}^{(G)} = \mathcal{R}^{(G1)} \cup \mathcal{R}^{(G2)}, \quad \mathcal{R}^{(T)} = \mathcal{R}^{(T1)} \cup \mathcal{R}^{(T2)}.$$

A 3.12. lemma következménye többek között az, hogy bármelyik GT-ütemtervet egyértelműen jellemez az egymás utáni kritikus szituációban hozott kritikus döntések sorozata. Az $\mathcal{R}^{(Ta)}$, $a=1, 2$, az összes lehetséges $\{s\}$ megszámlálható döntéssorozattal ekvivalens, amely $s(0)=s_a$ döntéssel kezdődik és minden további döntés egyik lehetséges kritikus döntés az előző kritikus döntés meghatározta T-kritikus szituációban.

Hasonlóan $\mathcal{R}^{(Ga)}$, $a=1, 2$, ekvivalens azon $\{s\}$ döntéssorozatokkal, amelyek s_a-val kezdődnek és minden további döntés egyik lehetséges kritikus döntés az előző döntés meghatározta G-kritikus szituációban.

Vizsgáljuk annak feltételét, hogy egy GT-ütemterv egy $t' \geqq 0$ véges pontjában az első kritikus szituáció fellépjen. Lényegében elegendő az első T-kritikus szituáció helyét meghatározni. Megállapításaink szerint ez csak az $s(0)$ első kritikus döntéstől függ. Legyen tehát T'_a, $a=1, 2$, az $\mathcal{R}^{(a)}(Q)$ halmazban az első T-kritikus szituáció helye és legyen σ'_a maga a szituáció. $T'_a \leqq \infty$ pontig az ütemezés összefüggő és szoros (hiszen természetes és gazdaságos egyszerre), ezért az $\mathcal{R}^{(a)}(Q)$ halmazban $\tau_1\tau_2>0$ feltétel mellett

$$f(C_{a,j}) = j\tau_a, \quad f(C_{3-a,j}) = \eta_a + j\tau_{3-a}, \quad j = 0, 1, 2, \ldots \tag{3.6}$$

lesznek a ciklusvégződések a T'_a pont előtt. A kritikus szituáció is ciklusvégződésnél lép fel és arra is érvényes (3.6). Mivel azok mindig párban lépnek fel és egyik a $Q^{(1)}$, a másik a $Q^{(2)}$ ciklusvégződése, jelölheti k_1 és k_2 a ciklusok sorszámát, amelyek végződésénél a GT-kritikus szituációk fellépnek. Tudjuk, hogy a két GT-kritikus szituáció távolsága a $\sigma_{i,0}$, $\sigma_{3-i,1}$ pár esetén

$$|\Delta t_i| \leqq \min(\eta_i, \vartheta_{3-i}), \quad i = 1, 2. \tag{3.7}$$

A továbbiakban célszerű átmenetileg kizárni bizonyos konfigurációkat a vizsgálatból és a kizáró feltételeket később fokozatosan feloldani.

Tegyük fel egyelőre, hogy

$$0 < \eta_i \leqq \vartheta_{3-i}, \quad i = 1, 2 \tag{3.8}$$

feltétel teljesül. Ekkor a 3.11. lemmában a $\sigma_{i,0}$ feltételeként szereplő $0<\beta^{(3-i)}(t)<\eta_i$ relációból automatikusan következik a $0<\beta^{(3-i)}(t)\leqq\vartheta_{3-i}$ feltétel teljesülése. Mivel $\beta^{(i)}(t)=0$ mindig egy $t=f(C_i)$ ciklusvégződéssel ekvivalens, ezért a GT-kritikus szituációk feltételei a ciklusvégződések relatív viszonyával is kifejezhetők és sematikusan az alábbi módon írhatók:

$$\begin{aligned}
\sigma_0&: && f(C_1) = f(C_2),\\
\sigma_{1,0}&: && f(C_2)-\eta_1 < f(C_1) < f(C_2),\\
\sigma_{2,0}&: && f(C_1)-\eta_2 < f(C_2) < f(C_1),\\
\sigma_{1,1}&: && f(C_2) < f(C_1) < f(C_2)+\eta_2,\\
\sigma_{2,1}&: && f(C_1) < f(C_2) < f(C_1)+\eta_1.
\end{aligned}$$

Az egyenlőtlenségekben mindig a középen álló ciklusvégződés adja a kritikus szituáció helyét, a másik ciklusvégződés pedig a *GT*-kritikus szituáció-párjának a helyét. A relációkból látható, hogy a $\sigma_{i,0}$ és $\sigma_{3-i,1}$ szituáció típusok feltétele ugyanaz, csupán a helyét reprezentáló ciklusvégződés cserélődik. Ez is igazolja a *GT*-kritikus szituációpárok létezését. A feltételek még az alábbi lemma szerint is felírhatók.

3.13. LEMMA: A (3.8) feltétel mellett akkor lép fel *GT*-kritikus szituáció, ha valamely $k_1 \geqq 0$, $k_2 \geqq 0$, $k_1 + k_2 > 0$ egészekre

$$f(C_{2,k_2}) - \eta_1 < f(C_{1,k_1}) < f(C_{2,k_2}) + \eta_2 \tag{3.9}$$

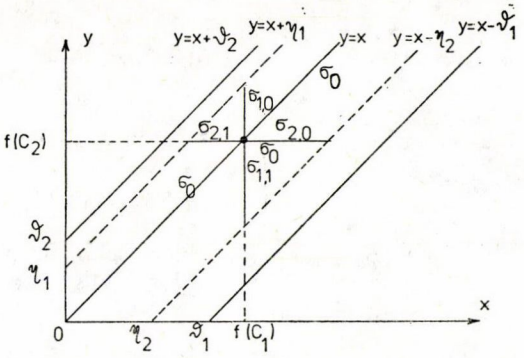
reláció kialakul. A *GT*-kritikus szituáció típusa

$$\begin{aligned} &\sigma_0, && \text{ha} && f(C_{1,k_1}) = f(C_{2,k_2}), \\ &\sigma_{1,0} \text{ és } \sigma_{2,1}, && \text{ha} && f(C_{1,k_1}) < f(C_{2,k_2}), \\ &\sigma_{2,0} \text{ és } \sigma_{1,1}, && \text{ha} && f(C_{1,k_1}) > f(C_{2,k_2}). \end{aligned}$$

Ha a (3.8) nem teljesül, akkor $\eta_i = 0$ esetén a (3.9) megfelelő oldalán $<$ helyett $\leqq$ jel írandó és $\eta_i > \vartheta_{3-i}$ esetén a (3.9) megfelelő oldalán η_i helyett ϑ_{3-i}, $<$ helyett pedig $\leqq$ jel írandó.

Bizonyítás: Ha (3.8) nem teljesül, akkor az állítást közvetlenül a 3.11. lemma alapján láthatjuk be. Az $x-y$ koordinátarendszerben $f(C_1)$-et az x, $f(C_2)$-t az y tengelyen ábrázolva, az egyes *GT*-kritikus szituációk tartományait sematikusan a 8. ábra szerint szemléltethetjük.

Ez az ábra a (3.8) feltétel mellett szemléltet. Ha azzal ellentétben valamelyik $\eta_i = 0$, akkor az ábra $y = x$ és $y = x + \eta_1$, vagy $y = x - \eta_2$ egyenesei egybeesnek, a (3.9) egyenlőtlenség megfelelő oldalán a $<$ jel helyett $\leqq$ jelet kell alkalmazni és az egyenlőségnek a σ_0 szituáció felel meg. Ha viszont $\vartheta < \eta_1$, vagy $\vartheta_1 < \eta_2$ viszony áll fenn, akkor az $y = x + \vartheta_2$ és $y = x + \eta_1$ vagy az $y = x - \eta_2$ és $y = x - \vartheta_1$ egyenesek helyzete megfordul és ismét az $y = x$ egyeneshez közelebbi egyenes fogja határolni a „kritikus sávot". Ekkor (3.8)-ban is el kell végezni a megfelelő szerepcserét és a $<$ jel helyett $\leqq$ jelet kell írni. Az ábrán feltüntetett szituációtípusok egyébként fix $f(C_1)$ mellett $f(C_2)$ függvényében, fix $f(C_2)$ mellett pedig $f(C_1)$ függvényében ábrázolják azt a ciklusvégződést, amely a kritikus szituációt eredményezi.



8. ábra. *GT-kritikus szituációk tartományai*

A 3.13. lemma egy átfogalmazása az alábbi

3.13′. LEMMA: Bármely $Q \in \mathcal{Q}$ konfiguráció *GT*-ütemterveinél az $f(C_1)$, $f(C_2)$ pontpárban *GT*-kritikus szituációpár lép fel, ha a

$$\Delta t_i \doteq f(C_i) - f(C_{3-i})$$

jelöléssel

(3.10) $$-\min(\eta_i, \vartheta_{3-i}) \prec_i \Delta t_i \prec_{3-i} \min(\eta_{3-i}, \vartheta_i), \quad (i = 1, 2)$$

teljesül, ahol

$$\prec_i = \begin{cases} <, & \text{ha } 0 < \eta_i \leqq \vartheta_{3-i}, \\ \leqq, & \text{egyébként.} \end{cases}$$

A GT-kritikus szituációk típusa:

$$\begin{aligned} &\sigma_0, && && \text{ha} \quad \Delta t_i = 0, \\ &\sigma_{i,0} && \text{és} \quad \sigma_{3-i,1}, && \text{ha} \quad \Delta t_i < 0, \\ &\sigma_{3-i,0} && \text{és} \quad \sigma_{i,1}, && \text{ha} \quad \Delta t_i > 0. \end{aligned}$$

Ezen alapszik a következő lemma.

3.14. LEMMA: Tetszőleges nem degenerált $Q \in \mathcal{Q}$ konfiguráció *összes* $R \in \mathcal{R}^{(a)}(Q)$ ($a=1$ vagy 2) *GT-ütemtervénél egyszerre,* akkor és csak akkor létezik kritikus szituáció, ha az

(3.11) $$\eta_a - \min(\eta_a, \vartheta_{3-a}) \prec_a B_a\tau_a - A_a\tau_{3-a} \prec_{3-a} \eta_a + \min(\eta_{3-a}, \vartheta_a)$$

egyenlőtlenségnek, ahol $i=1, 2$ mellett

(3.12) $$\prec_i = \begin{cases} <, & \text{ha } 0 < \eta_i \leqq \vartheta_{3-i}, \\ \leqq, & \text{egyébként,} \end{cases}$$

létezik $\omega_a=(B_a, A_a)$ megoldása az

(3.13) $$\omega_a \geqq (1, 0)$$

feltétel mellett. Ebben az esetben a T-ütemtervek σ_a' első kritikus szituációjának a helye

(3.14) $$T_a' = \min(B_a'\tau_a, \ \eta_a + A_a'\tau_{3-a}),$$

ahol $\omega_a'=(B_a', A_a')$ a (3.11) legkisebb megoldása a (3.13) feltétel mellett. σ_a' egyben a G-ütemtervek első GT-kritikus szituációja. A G-ütemtervek σ_a'' első saját kritikus szituációja azok

(3.15) $$T_a'' = \max(B_a'\tau_a, \eta_a + A_a'\tau_{3-a}) = T_a' + |\Delta_a' - \eta_a|$$

pontjában lép fel, ahol

(3.16) $$\Delta_a' \doteq B_a'\tau_a - A_a'\tau_{3-a}.$$

A σ_a' és σ_a'' szituációk típusa és értéke a $\Delta_a' - \eta_a$ értéktől függ a következők szerint:

(3.17)

σ_a'	σ_a''	$\beta(t)$ értéke	Feltétel
σ_0	σ_0	$\beta^{(i)}(T_a') = \beta^{(i)}(T_a'') = 0, \quad i = 1,2$	$\Delta_a' - \eta_a = 0$
$\sigma_{a,0}$	$\sigma_{3-a,1}$	$\beta^{(3-a)}(T_a') = \tau_a - \beta^{(a)}(T_a'') = \lvert\Delta_a' - \eta_a\rvert$	$\Delta_a' - \eta_a < 0$
$\sigma_{3-a,0}$	$\sigma_{a,1}$	$\beta^{(a)}(T_a') = \tau_{3-a} - \beta^{(3-a)}(T_a'') = \Delta_a' - \eta_a$	$\Delta_a' - \eta_a > 0$

Ha a (3.11)-nek nincs a (3.13) feltétel melletti megoldása, akkor $\mathscr{R}^{(a)}$ egyetlen R_{a0} ütemtervet tartalmaz, amely összefüggő, szoros és következetes.

Bizonyítás: A ciklusvégződések (3.6) alatti formuláit a 3.13′. lemma (3.10) egyenlőtlenségébe helyettesítve kapjuk a (3.11) egyenlőtlenséget. A (3.14) és (3.15) értékek egyszerűen adódnak a kritikus szituációk definíciójából és a (3.17) is könnyen belátható. Az utolsó állítás következik a 3.12. lemma (1) és (2) állításaiból.

Megjegyzés: A (3.11) egyenlőtlenség legkisebb $\omega'_a \geqq (1, 0)$ megoldása egy tipikus koincidencia feladat megoldását jelenti, amelynek kérdéseivel az előző fejezetben foglalkoztunk.

A (3.6) formulák érvényüket vesztik az első kritikus szituáció után, amikor valamelyik job-folyam biztosan késleltetést szenved. Ugyancsak érvénytelen a (3.6) akkor, ha valamelyik job-folyam degenerált és a 2. megállapodás miatt is késleltetés léphet fel. Ezért a degenerált esetet külön kell vizsgálni.

Vezessük be a σ_a^* jelölést az $\mathscr{R}^{(a)}$ elemeinek első kritikus, vagy β-típusú szituációjára, amely a $\sigma_0[0]$ kezdő szituáció és a $t'_a = \eta_a$ pontban esetlegesen fellépő nem-kritikus pontbeli β_a-szituáció után fellép. A fellépési pontját jelölje T_a^*. Nevezzük ezt a *GT*-ütemterv *jellegzetes szituációjának.*

Vezessük be az R_{a0} jelölést arra az esetre, amikor $\mathscr{R}^{(a)}$ elemeinek nincs kritikus szituációja és az $s(0) = s_a$ döntés az ütemtervet egyértelműen meghatározza a 3.12. lemma (1) és (2) állításai szerint. $\tau_1\tau_2 > 0$ esetén a 3.14. lemmával összhangban, (de $\tau_1\tau_2 = 0$ esetén is) R_{a0} összefüggő, szoros és következetes ütemterv.

Nevezzük az ütemtervek $\mathscr{R}_1$ és $\mathscr{R}_2 \subset \mathscr{R}$ *osztályait lényegében azonosaknak,* ha bármely $R \in \mathscr{R}_1$ ütemtervhez van pontosan egy $R' \in \mathscr{R}_2$ ütemterv úgy, $R \approx R'$, azaz lényegében azonosak egymással. Legyen e tény jelölése $\mathscr{R}_1 \approx \mathscr{R}_2$. Másként fogalmazva $\mathscr{R}_1 \approx \mathscr{R}_2$ akkor, ha $\mathscr{R}_1$ és $\mathscr{R}_2$ között egy-egyértelmű megfeleltetés létesíthető úgy, hogy a megfelelő elemek lényegében azonosak egymással.

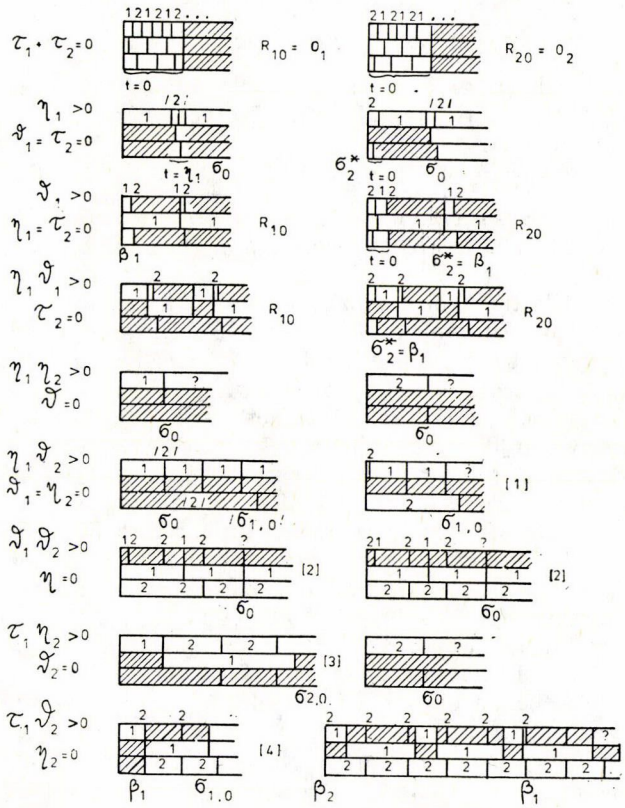
A továbbiakban is a 3.14. lemmához hasonlóan olyan tényeket fogunk igazolni nem degenerált konfigurációkra, amelyek degeneráltakra nem feltétlenül állnak fenn. Sokszor az elfajult eseteket is külön kell ellenőrizni. Ezt megkönnyíti, ha áttekintjük az *elfajult konfigurációk GT-ütemterveinek* tulajdonságait.

A $Q = (\eta_1; \vartheta_1; \eta_2; \vartheta_2)$ konfiguráció négy paramétere közül legalább egy 0 értékű $2^4 - 1 = 15$-féleképpen lehetséges. Ezeket az eseteket a 9. ábra szemlélteti és az 1. táblázatban tekintjük át.

1. TÁBLÁZAT

Elfajult konfigurációk GT ütemterveinek jellemzői

Eset	t'_1	T'_1	σ'_1	T_1^*	σ_1^*	$\mathfrak{R}^{(1)}$	t'_2	T'_2	σ'_2	T_2^*	σ_2^*	$\mathfrak{R}^{(2)}$	Megjegyzes
$\tau_1+\tau_2=0$	0	—	—	0	β_2	R_{10}	0	—	—	0	β_1	R_{20}	$R_{10} \approx R_{20}$
$\eta_1 > 0$	η_1	η_1	σ_0	—	—		0	η_1	σ_0	η_1	σ_0		$\mathscr{R}^{(1)} \approx \mathscr{R}^{(2)}$
$\vartheta_1 > 0$	0	—	—	—	—	R_{10}	0	—	—	0	β_1	R_{20}	$R_{10} \approx R_{20}$
$\eta_1\vartheta_1 > 0$	η_1	—	—	—	—	R_{10}	0	—	—	η_1	β_1	R_{20}	$R_{10} \approx R_{20}$
$\eta_1\eta_2 > 0$	η_1	η_1	σ_0	—	—		η_2	η_2	σ_0	—	—		
$\eta_1\vartheta_2 > 0$	η_1	η_1	σ_0	—	—		0	[1]	[1]	—	—		[1]
$\vartheta_1\vartheta_2 > 0$	0	[2]	[2]	—	—	(R_{10})	0	[2]	[2]	—	—	(R_{20})	[2] $(R_{10} \approx R_{20})$
$\tau_1\eta_2 > 0$	η_1	[3]	[3]	—	—		η_2	η_2	σ_0	—	—		[3]
$\tau_1\vartheta_2 > 0$	η_1	[4]	[4]	[4]	[4]		0	[4]	[4]	[4]	[4]		[4]



9. ábra. *Elfajult konfigurációk GT-ütemtervei*

[1] σ_2 kritikus szituáció biztosan létezik és

$$\sigma_2' = \begin{cases} \sigma_0, & \text{ha } \vartheta_2 \equiv 0 \pmod{\eta_1}, \\ \sigma_{1,0}, & \text{egyébként.} \end{cases}$$

[2] $\mathscr{R}^{(1)}$ és $\mathscr{R}^{(2)}$ osztályban egyszerre van kritikus szituáció, ha a $k_1\tau_1 - k_2\tau_2 = 0$ egyenletnek van nem-triviális megoldása, azaz ϑ_1 és ϑ_2 racionálisan összefüggnek. Ekkor $\sigma_1' = \sigma_2' = \sigma_0$ és $T_1' = T_2' = k_1\tau_1 = k_2\tau_2$, ahol $k_1 > 0$, $k_2 > 0$ relatív prímek. Ellenkező esetben $\mathscr{R}^{(1)}$ és $\mathscr{R}^{(2)}$ elemeinek nincs kritikus szituációja és mindkét osztály egyetlen R_{a0} elem, amelyekre $R_{10} \approx R_{20}$.

[3] Kritikus szituáció biztosan létezik éspedig az első

$$\sigma_1' = \begin{cases} \sigma_0, & \text{ha } \vartheta_1 \equiv 0 \pmod{\eta_2} \\ \sigma_{2,0}, & \text{egyébként.} \end{cases}$$

[4] Kritikus szituáció létezik $\mathcal{R}^{(a)}$ elemeinél, ha a (3.11)-nek van (3.13) melletti megoldása. $\mathcal{R}^{(1)}$ és $\mathcal{R}^{(2)}$ elemeinek egyszerre létezik kritikus szituációja és $\sigma_1'=\sigma_2'$, akkor ha $a=1$ vagy 2 mellett a

$$\Delta_a^* = B_a^*\tau_a - A_a^*\tau_{3-a} = \eta_1$$

egyenletnek van $\omega_a^*<\omega_a'$, $\omega_a^*\geqq(1,0)$ megoldása. Ekkor $\mathcal{R}^{(a)}$ elemeinél a $T_a^*=B_a^*\tau_a$ pontban β_{3-a}-szituáció lép fel.

Az 1. táblázat első oszlopában a nem hivatkozott paraméterek értéke 0.

Az első négy sorban $\tau_1\tau_2=0$, vagyis Q degenerált, az utolsó öt sorban azonban $\tau_1\tau_2>0$, vagyis Q csak elfajult.

A továbbiakban részletesebben vizsgáljuk a GT-ütemtervek tulajdonságait a (3.11) egyenlőtlenség, valamint a

$$0 \leqq B_a\tau_a - A_a\tau_{3-a} \leqq \eta \tag{3.18}$$

és

$$0 \leqq B_a\tau_a - A_a\tau_{3-a} \lessdot_{3-a} \eta_a + \min(\eta_{3-a}, \vartheta_a) \tag{3.19}$$

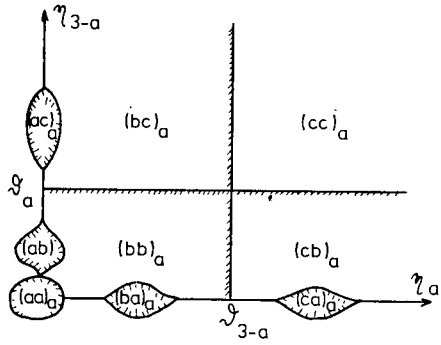
egyenlőtlenségek legkisebb $\omega_a\geqq(1,0)$ feltétel melletti megoldásainak segítségével. A (3.18) és (3.19) két baloldali homogén koincidencia feladat $(B(M)KIF)$, amelyekkel az előző fejezetben foglalkoztunk. A megoldás exisztenciáját a 2.1. tétel tisztázza. Eszerint a $0\leqq\Delta_a 2\alpha$ megoldása mindig létezik, ha $\alpha>0$, vagy τ_a és τ_{3-a} racionálisan összefüggők. Az $\alpha>0$ esetben a megoldást a $B(M)KIFM$-algoritmus, $\alpha=0$ esetben pedig az R-algoritmus szolgáltatja (lásd 2.2. pont).

A vizsgálatok áttekintésének megkönnyítésére bontsuk a $\mathcal{Q}$ paraméterteret részekre az alábbi feltételek szerint $i=1$ és 2 mellett:

(a) $\eta_i = 0$,

(b) $0<\eta_i\leqq\vartheta_{3-i}$,

(c) $0\leqq\vartheta_{3-i}<\eta_i$.



10. ábra. *A Q résztartományai*

Az η_a értékét az x-tengelyen, az η_{3-a} értékét az y-tengelyen ábrázolva, $\vartheta_a\vartheta_{3-a}>0$ esetén a Q tartományait a 10. ábra szemlélteti. Azon $(xy)_a$, $x, y=a,b,c$ jelzi az egyes tartományokat. Az ábrából látható, hogy $\vartheta_i=0$ esetén a (b) jelű tartományok eltűnnek. Az egyes tartományokban a (3.11) egyenlőtlenség speciális formáját a 2. táblázat szerinti mátrixba foglaltuk.

3.15. LEMMA: Ha a nem-degenerált $Q\in\mathcal{Q}$ konfigurációnál valamely $a=1$ vagy 2 mellett a (3.18) $B(M)KIF$ legkisebb $\omega_a^*\geqq(1,0)$ megoldására

$$\Delta_a^* = \eta, \tag{3.20}$$

2. TÁBLÁZAT

A (3.11) *egyenlőtlenség speciális esetei*

		(a)	(b)	(c)
		$\eta_{3-a}=0$	$0<\eta_{3-a}\leqq\vartheta_a$	$0\leqq\vartheta_a<\eta_{3-a}$
(a)	$\eta_a=0$	$\Delta_a=0$	$0\leqq\Delta_a<\eta_{3-a}$	$0\leqq\Delta_a\leqq\tau_a$
(b)	$0<\eta_a\leqq\vartheta_{3-a}$	$0<\Delta_a\leqq\eta_a$	$0<\Delta_a<\eta$	$0<\Delta_a\leqq\tau_a$
(c)	$0\leqq\vartheta_{3-a}<\eta_a$	$\eta_a-\vartheta_{3-a}\leqq\Delta_a\leqq\eta_a$	$\eta_a-\vartheta_{3-a}\leqq\Delta_a<\eta$	$\eta_a-\vartheta_{3-a}\leqq\Delta_a\leqq\tau_a$

akkor

(a) $\mathscr{R}^{(a)}$ elemeinek

$$T_a^*=B_a^*\tau_a=\eta+A_a^*\tau_{3-a}$$

pontjában jellegzetes szituációja

$$\sigma_a^*=\sigma_{3-a}[t'_{3-a}]=\begin{cases}\beta_{3-a}, & \text{ha}\quad 0\leqq\eta_a\leqq\vartheta_{3-a},\quad \vartheta_{3-a}>0\\ \sigma'_{3-a}, & \text{egyébként,}\end{cases}$$

ahol $\sigma_{3-a}[t'_{3-a}]$ az $\mathscr{R}^{(3-a)}$ elemeinek $t'_{3-a}=\eta_{3-a}$ pontbeli szituációja;

(b) $$\mathscr{R}^{(1)}\approx\mathscr{R}^{(2)};$$

(c) az összes *GT*-ütemtervnek egyszerre létezik vagy nem létezik kritikus szituációja.

(d) Ha létezik kritikus szituáció, akkor az első *GT*-kritikus szituáció minden ütemtervnél egyenlő és a (3.11) és (3.18) egyenlőtlenségek megoldásai között az alábbi összefüggések állnak fenn:

$$B'_a=B_a^*+A'_{3-a},\quad A'_a=A_a^*+B'_{3-a},$$

$$\omega'_a=\omega_a^*+\bar{\omega}'_{3-a}, \tag{3.21}$$

$$\Delta'_a=\Delta_a^*-\Delta'_{3-a},\quad T'_a=T_a^*+T'_{3-a}-\eta_{3-a},$$

ahol

$$\bar{\omega}'_{3-a}=(A'_{3-a},B'_{3-a}).$$

(e) Ha nem létezik kritikus szituáció, akkor $\mathscr{R}^{(i)}$ egyetlen R_{i0} elem és $R_{10}\approx R_{20}$.

Bizonyítás: $\eta_a-\min(\eta_a,\vartheta_{3-a})\geqq 0$ és $\eta_a+\min(\eta_{3-a},\vartheta_a)\leqq\eta$ következtében a (3.11) minden (3.13) feltétel melletti megoldása kielégíti a (3.18)-at is. Ezért a (3.18) egyenlőtlenség legkisebb $\omega_a^*\geqq(1,0)$ megoldására $\omega_a^*<\omega'_a$, ha ω'_a a (3.11) legkisebb $\omega'_a\geqq(1,0)$ megoldása. $\tau_1\tau_2>0$ következtében $T_a^*\leqq T'_a$ pontig a (3.6) biztosan igaz. Ezért a (3.20) ekvivalens egy

$$f(C_a)=f(C_{3-a})+\eta_{3-a} \tag{3.22}$$

egyenlőséggel valamely C_a és C_{3-a} ciklusokra.

Ha $\tau_a<\eta$, azaz $\vartheta_a<\eta_{3-a}$, akkor a (3.11) alakját a 2. táblázat (*xc*) oszlopában találjuk és annak legkisebb (3.13) melletti megoldása $\omega'_a=(1,0)$. Erre azonban

$$0<\Delta'_a=\tau_a<\eta,$$

ezért a (3.20) egyetlen $\omega_a^*\geqq(1,0)$ számpárra sem teljesülhet.

Ha $\tau_a \geqq \eta$, azaz $\eta_{3-a} \leqq \vartheta_a$, akkor a (3.22) ekvivalens azzal, hogy a $T_a^* = f(C_a)$ pontban a β_{3-a} szituáció, vagy egy T-kritikus szituáció lép fel. Azonban ugyanez a szituáció lép fel minden $R' \in \mathscr{R}^{(3-a)}$ ütemterv $t'_{3-a} = \eta_{3-a}$ pontjában is. A 3.12. lemma (9) állítása szerint ekkor a $\mathscr{R}^{(a)}$ elemei a T_a^* ponttól az első GT-kritikus szituációig ugyanolyanok, mint $\mathscr{R}^{(3-a)}$ elemei a t'_{3-a} ponttól az első GT-kritikus szituációig, ha ilyen szituáció létezik. Ez egyszerre igaz.

A GT-kritikus szituációk azonosak és $T'_a - T_a^* = T'_{3-a} - t'_{3-a} \geqq 0$ az összes GT-ütemtervnél. Ha ilyen szituáció nem létezik, akkor az összes GT-ütemterv lényegében azonos, márpedig T_a^*, ill. t'_{3-a}-től $\mathscr{R}^{(a)}$, ill. $\mathscr{R}^{(3-a)}$ elemei is azonosak. Ebből következik, hogy $s(0) = s_a$, $a = 1, 2$ egyetlen R_{a0} ütemtervet határoz meg. $R_{10} \approx R_{20}$ a fentiekből nyilvánvaló. Ha létezik kritikus szituáció, akkor viszont $\mathscr{R}^{(a)}$ és $\mathscr{R}^{(3-a)}$ tartalmazzák az összes lehetséges $\{s\}_{s(0)=s_a}$ kritikus döntéssorozatok szerinti ütemterveket, amelyek a két halmaznál azonosak. Ez következik abból, hogy az első kritikus szituációk mindkét halmazban azonosak és így a 3.12. lemma szerint a további határozott szakaszok záró kritikus szituációikkal rendre meghatározottak ugyanazon kritikus döntésekkel. Ezért valóban $\mathscr{R}^{(1)} \approx \mathscr{R}^{(2)}$. A $\sigma'_a = \sigma'_{3-a}$ következtében (3.17)-ből $\Delta'_a - \eta_a = -(\Delta'_{3-a} - \eta_{3-a})$, amiből $\Delta_a^* = \eta_a + \eta_{3-a}$ folytán $\Delta'_a = \Delta_a^* - \Delta'_{3-a}$ következik. Az ω-ra és a (B, A)-ra vonatkozó összefüggések ebből már következnek.

1. Megjegyzés. A 3.15. lemma feltétele nem teljesülhet, ha a (3.11) és (3.18) egyenlőtlenségek azonosak, vagy legalábbis a jobb oldalaik azok. A 2. táblázat szerint így az $(xa)_a$ esetek kizárhatók. Az $(xc)_a$ esetekben viszont az $\omega'_a = (1, 0)$ mindig a (3.11) megoldása, amelyre $\Delta'_a = \tau_a < \eta$, ezért a 3.15. lemma feltétele nem teljesülhet. A 3.15. lemma feltétele azonban teljesül, ha (3.20)-nak van megoldása az $(xb)_a$ esetekben, amikor

$$0 < \eta_{3-a} \leqq \vartheta_a.$$

A lemma szimmetriája miatt ez ekvivalens a

$$0 < \eta_1 \leqq \vartheta_2 \quad \text{vagy} \quad 0 < \eta_2 \leqq \vartheta_1$$

feltétellel.

2. Megjegyzés. A (3.21) összefüggések következtében, ha a 3.15. lemma feltétele teljesül $(xy)_a$ esetben, akkor a $(3.11)_a$ helyett megoldhatjuk a $(3.11)_{3-a}$ egyenlőtlenséget is, amelynél azonban az $(yx)_{3-a}$ lesz a megfelelő eset. Így tehát a

$$0 \leqq \Delta_1 < \eta_2 \quad \text{és} \quad 0 < \Delta_2 \leqq \eta_1,$$

valamint az

$$\eta_1 - \vartheta_2 \leqq \Delta_1 < \eta_1 + \eta_2 \quad \text{és} \quad 0 < \Delta_2 \leqq \tau_2$$

egyenlőtlenségek megoldása egyszerre létezik és köztük a (3.21) összefüggések érvényesek.

3. Megjegyzés. A $(bc)_{3-a}$ esetben speciálisan a $(3.11)_{3-a}$ megoldása $\omega'_{3-a} = (1, 0)$, amit (3.21)-ben figyelembe véve

$$\omega'_a = (B_a^*, A_a^* + 1)$$

lesz a $(3.11)_a$ megoldása, ha a ω_a^* eleget tesz a 3.15. lemma feltételeinek, tehát

$$B_a^* \tau_a - A_a^* \tau_{3-a} = \eta_a + \eta_{3-a}.$$

A $(bc)_{3-a}=(cb)_a$ esetben azonban teljesül

$$\tau_{3-a} < \eta \leqq \tau_a$$

reláció, ezért a (3.20) egyenlet legkisebb megoldása csak

$$\omega_a^* = \left(1, \left[\frac{\vartheta_a}{\tau_{3-a}}\right]\right)$$

lehet, amiből

$$\omega_a' = \left(1, \left[\frac{\vartheta_a}{\tau_{3-a}}\right]+1\right)$$

megoldás adódik. (3.20) megoldásának feltétele ebben az esetben nyilvánvalóan a

$$\vartheta_a \equiv \eta_{3-a} \pmod{\tau_{3-a}}$$

teljesülése.

4. Megjegyzés. A (3.20) bekövetkezhet egyszerre $a=1$ és 2 mellett is. Ilyenkor azonban a 3.12. lemma (4) állítása alapján belátható, hogy $\mathscr{R}^{(1)}$ és $\mathscr{R}^{(2)}$ mindegyike egyetlen, kritikus szituációt nem tartalmazó, R_{a0} periodikus ütemtervből áll. Ilyen tulajdonságú például a $Q=(1; 5; 1; 3)$ konfiguráció.

3.16. LEMMA: Ha bármely nem degenerált $Q\in\mathscr{Q}$ konfigurációnál valamelyik $i=1$ vagy 2 mellett

(3.23) $$0 \leqq \vartheta_{3-i} < \eta_i$$

fennáll, akkor a (3.11) egyenlőtlenség (3.13) feltétel melletti megoldása $a=1, 2$ mellett egyaránt létezik és a megfelelő (3.19) $B(M)KIF$ megoldása szolgáltatja (ω_a'').

Bizonyítás: Ha $0\leqq\vartheta_a<\eta_{3-a}$, akkor a (3.11) egyenlőtlenségnek a kívánt megoldása $\omega_a'=(1, 0)$ mindig létezik, amint ez a 2. táblázat $(xc)_a$ utolsó oszlopából látható. Ez egyben a (3.19) megoldása is.

Ha $0\leqq\vartheta_{3-a}<\eta_a$, akkor (3.11) alakját a 2. táblázat $(cy)_a$ utolsó sorában találjuk, de a $(cc)_a$ esetben ismét $\omega_a'=(1, 0)=\omega_a''$. A $(ca)_a$ és $(cb)_a$ esetekben még $0\leqq\eta_{3-a}\leqq\vartheta_a$ is teljesül. Ekkor (3.11) alakja

(3.11′) $$\eta_a-\vartheta_{3-a} \leqq \Delta_a \prec_{3-a}\eta$$

és (3.19) alakja

(3.19′) $$0 \leqq \Delta_a \prec_{3-a}\eta.$$

$\eta_a>0$ következtében a 2.1. tétel szerint ennek mindig létezik pozitív megoldása. Belátjuk, hogy legkisebb $\omega_a''\geqq(1, 0)$ megoldására nem lehet

(∗) $$0 \leqq \Delta_a'' < \eta_a-\vartheta_{3-a}.$$

Ebből következik, hogy (3.11′) és (3.19′) legkisebb megoldása mindig létezik és azonos. Ehhez elegendő megmutatni, hogy bármely a (∗)-nak eleget tevő $\omega_a''\geqq(1, 0)$ megoldáshoz létezik $\omega_a<\omega_a''$, amely kielégíti (3.11′) egyenlőtlenséget és $\omega_a\geqq(1, 0)$.

Definiáljuk a

$$K = f_{\geqq}\left(\frac{\eta_a-\Delta_a''}{\tau_{3-a}}\right) \quad \text{és} \quad A_a^* = A_a''-K$$

egészeket. A $B''_a\tau_a - A''_a\tau_{3-a} \doteq \Delta''_a$ definícióból $(*)$ miatt $A''_a\tau_{3-a} = B''_a\tau_a - \Delta''_a \geqq$ $\geqq \tau_a - \Delta''_a > \vartheta_{3-a} \geqq 0$, ezért

$$A''_a \geqq \frac{\eta_a - \Delta''_a}{\tau_{3-a}} > 0, \quad \text{vagyis} \quad A''_a \geqq K > 0.$$

Így

$$0 \leqq A^*_a < A''_a.$$

Legyen $\omega^*_a = (B''_a, A^*_a)$. Ekkor $\Delta^*_a = \Delta''_a + K\tau_{3-a}$. $z \leqq f_{\geqq}(z) < z+1$ felhasználásával K-nál,

$$\eta_a - \vartheta_{3-a} \leqq \Delta^*_a < \eta_a + \tau_{3-a}$$

relációt nyerjük. Itt vagy

$$\eta_a - \vartheta_{3-a} \leqq \Delta^*_a \prec_{3-a} \eta$$

amikor is $\omega^*_a < \omega''_a$ megoldása (3.11')-nek, vagy

$$\eta \prec'_{3-a} \Delta^*_a < \eta_a + \tau_{3-a}.$$

Ez utóbbi esetben

$$\eta \prec'_{3-a} \Delta''_a + K\tau_{3-a}$$

relációból $(*)$ felhasználásával

$$K \geqq \frac{\eta - \Delta''_a}{\tau_{3-a}} = \frac{\eta_a - \Delta''_a}{\tau_{3-a}} + \frac{\eta_{3-a}}{\tau_{3-a}} > \frac{\vartheta_{3-a}}{\tau_{3-a}} + \frac{\eta_{3-a}}{\tau_{3-a}} = 1,$$

vagyis $K \geqq 2$ adódik. Legyen $\omega_a = (B''_a, A''_a - K + 1) < \omega''_a$. ω_a-ra $\Delta_a = \Delta^*_a - \tau_{3-a}$, ezért a fentiekből

$$\eta - \tau_{3-a} \prec'_{3-a} \Delta_a < \eta_a$$

adódik, amiből $\eta - \tau_{3-a} = \eta_a - \vartheta_{3-a}$ és $\eta_a \leqq \eta$ miatt bármely $\prec_{3-a}$ érték mellett következik

$$\eta_a - \vartheta_{3-a} \leqq \Delta_a \prec_{3-a} \eta.$$

Vagyis $\omega_a < \omega''_a$ megoldása (3.11')-nek.

3.17. LEMMA: Ha a nem-degenerált $Q \in \mathcal{Q}$ konfigurációnál valamelyik $a=1$ vagy 2 mellett a (3.11)-nek nincs megoldása, de a (3.19) megoldása létezik és arra

(3.24) $$\Delta''_a = 0,$$

akkor R_{a0} periodikus

(3.25) $$p_a = B''_a\tau_a = A''_a\tau_{3-a}$$

periódushosszal.

(3.26) $$\sigma^*_a = \beta_a, \quad T^*_a = \eta_a + p_a,$$

és az $\eta_a + kp_a$, $k = 0, 1, \ldots$ pontokban a β_a-szituáció periodikusan fellép. A lemma feltétele csak

(3.27) $$0 < \eta_a \leqq \vartheta_{3-a}, \quad 0 \leqq \eta_{3-a} \leqq \vartheta_a$$

mellett teljesülhet.

Bizonyítás: A lemma feltétele csak akkor teljesülhet, ha (3.11) és (3.19) nem ugyanaz, ezért a 2. táblázat szerinti $(ay)_a$ sor eleve kiesik a lehetőségek közül. A feltétel akkor sem teljesül, ha a (3.11)-nek van megoldása. A 3.16. lemma következményeként a 2. táblázat $(cy)_a$ sora is kiesik. Kiesik továbbá a $(bc)_a$ eset is, amikor $\omega_a'=(1, 0)$ a (3.11) megoldása. Elegendő tehát a $(ba)_a$ és $(bb)_a$ esetekkel foglalkozni. Ekkor éppen a (3.27) feltételek állnak fenn.

A lemma feltevése mellett a 3.14. lemma szerint $\mathscr{R}^{(a)}$ valóban egyetlen R_{a0} összefüggő, szoros és következetes ütemterv. Ekkor (3.6) érvényes és a (3.24) azt jelenti, hogy valamely C_a és C_{3-a} ciklusokra

$$f(C_a) = f(C_{3-a}) - \eta_a$$

teljesül, vagyis az

$$f(C_{3-a}) = A_a'' \tau_{3-a} + \eta_a$$

pontban egy A_a-task végződik és egy A_{3-a}-task indul és ezért ebben a pontban éppen egy β_a-szituáció alakul ki. (3.27) miatt az A_{3-a}-task biztosan ütemeződik, de ugyanez a szituáció lép fel a $t_a'=\eta_a$ pontban is.

Így a β_a-szituáció visszatérő és a 3.8. lemma szerint R_{a0} periodikus. A periódushossz a β_a-szituáció első visszatérési ideje, vagyis (3.25). A (3.26) ebből már következik.

Megjegyzés. A (3.24) és (3.25) ekvivalens azzal, hogy a τ_a és τ_{3-a} racionálisan összefüggők. Racionálisan független τ_1 és τ_2 mellett ugyanis a (3.27) miatt mindig van (3.19)-nek olyan ω_a megoldása, amelyre $\Delta_a>0$, így ω_a megoldása (3.11)-nek is. Hiszen ekkor $\eta_a+\min(\eta_{3-a}, \vartheta_a)>0$ és a 2.1. tétel a megoldást garantálja.

A 3.14—3.17. lemmák a *GT*-ütemtervek tulajdonságainak feltételeit tartalmazzák a (3.11), (3.18) és (3.19) egyenlőtlenségek megoldásai segítségével. A feltételek nem közvetlenül a Q konfiguráció paramétereire vonatkoznak (közvetve természetesen igen). A következő lemma részben pótolja ezt a hiányt. További közvetlen kapcsolatokat is kimutathatnánk.

3.18. LEMMA: Egy $Q\in\mathscr{Q}$ konfiguráció $R\in\mathscr{R}^{(a)}$ *GT*-ütemtervének akkor és csak akkor nincs kritikus szituációja, ha

(i) $\eta = 0$, $\vartheta_1\vartheta_2 > 0$ és ϑ_1, ϑ_2 racionálisan függetlenek,

(ii) $\tau_1 = \tau_2 = 0$,

(iii) $\tau_i = 0$, $\vartheta_{3-i} > 0$, $i = 1$ vagy 2,

(iv) $\eta_a > 0$, $\tau_1\tau_2 > 0$, τ_1, τ_2 racionálisan összefüggők, és

(3.28) $$\eta \prec'_{3-a} q,$$

ahol $$\frac{\tau_1}{\tau_2} = \frac{k_1}{k_2}, \quad k_1 > 0, \quad k_2 > 0 \quad \text{relatív prímek},$$

(3.29) $$q = \frac{\tau_1}{k_1} = \frac{\tau_2}{k_2},$$

$$\prec'_{3-a} = \begin{cases} <, & \text{ha} \quad \eta_{3-a} = 0, \\ \leqq, & \text{ha} \quad \eta_{3-a} > 0. \end{cases}$$

Ha még $\eta_{3-a} > 0$ is teljesül, akkor

$$\mathscr{R}^{(GT)} = \{R_{10}, R_{20}\}. \tag{3.30}$$

Bizonyítás. Az (i)—(iv) feltételek szükségességét és elegendőségét egyszerre bizonyítjuk.

Az $\eta=0$, $\vartheta_1\vartheta_2>0$ eset a 2. táblázat $(aa)_a$ esetének felel meg, amikor a (3.11) alakja $\Delta_a=0$ egyenlet. A 3.14. lemma szerint az R ütemtervnek pontosan akkor van kritikus szituációja, ha $\Delta_a=0$-nak van $\omega_a \geqq (1, 0)$ megoldása, amely ϑ_1 és ϑ_2 racionális összefüggésével ekvivalens. $\eta=0$, $\vartheta_1\vartheta_2>0$ esetben tehát (i) szükséges és elegendő feltétel.

Az $\eta=0$, $\vartheta_1\vartheta_2=0$ esetben a GT-ütemterveket egyértelműen meghatározza az $s(0)=s_a$ kezdeti döntés és a 2. megállapodás. Így soha nincs kritikus szituáció. Ugyanez a helyzet a $\tau_i=0$, $\eta_{3-i}\vartheta_{3-i}>0$ esetben is. A $\tau_i=0$, $\vartheta_{3-i}=0$, $\eta_{3-i}>0$ esetben azonban bármelyik $s(0)=s_a$ döntés mellett a GT-ütemterv $t'=\eta_{3-i}$ pontjában a σ_0 kritikus szituáció lép fel, amelyben a döntést a 2. megállapodás sem határozza meg (lásd az 1. táblázatot és a 9. ábrát). Ezzel bebizonyítottuk, hogy a $\tau_1\tau_2=0$ degenerált esetekben az (ii) és (iii) feltételek szükségesek és elegendők ahhoz, hogy egyetlen GT-ütemtervnek se legyen kritikus szituációja.

A továbbiakban még azt kell bizonyítanunk, hogy az $\eta>0$, $\tau_1\tau_2>0$ nem-degenerált esetekben a (iv) feltétel szükséges és elegendő $\mathscr{R}^{(a)}$ elemeire vonatkozóan. A 3.14. lemma szerint a szükséges és elegendő feltétel az, hogy a $(3.11)_a$ egyenlőtlenségnek ne legyen $\omega_a \geqq (1, 0)$ megoldása. Mivel az $\eta_a>0$ következtében a (3.19) egyenlőtlenségnek mindig van megoldása, a 3.17. lemma szerint a $(3.11)_a$-nak csak (3.27) mellett nem biztos, hogy van megoldása, azaz a 2. táblázat $(ba)_a$ és $(bb)_a$ eseteiben. Ekkor a $(3.11)_a$ alakja

$$0 \prec \Delta_a \prec_{3-a} \eta. \tag{3.31}$$

Tehát $R \in \mathscr{R}^{(a)}$ ütemtervnek akkor és csak akkor nincs kritikus szituációja, ha (3.27) teljesül és (3.31)-nek nincs $\omega_a \geqq (1, 0)$ megoldása.

A (3.31)-nek akkor és csak akkor nincs megoldása, ha egyetlen $l \geqq 0$ egészhez sincs olyan $k \geqq 1$ egész, hogy

$$0 \prec k\tau_a - l\tau_{3-a} \prec_{3-a} \eta,$$

azaz

$$0 \prec k - l\frac{\tau_{3-a}}{\tau_a} \prec_{3-a} \frac{\eta}{\tau_a}$$

teljesülne. Vagyis bármely $l \geqq 0$ mellett minden $k \geqq 1$ egészre

$$\text{vagy} \quad k - l\frac{\tau_{3-a}}{\tau_a} \leqq 0, \quad \text{vagy} \quad \frac{\eta}{\tau_a} \prec'_{3-a} k - l\frac{\tau_{3-a}}{\tau_a}$$

következik be.

Legyen $k_l = \left[l\frac{\tau_{3-a}}{\tau_a}\right]$. Akkor a $z-1 \prec [z] \leqq z$ reláció felhasználásával azonnal adódik, hogy $k_l \geqq 1$ esetén minden $1 \leqq k \leqq k_l$ mellett $k - l\tau_{3-a}/\tau_a \leqq 0$, bármilyen

$k_l \geqq 0$ esetén minden $k \geqq k_l + 2$ mellett $k - l\tau_{3-a}/\tau_a > 1$, és $k = k_l + 1$ $(\geqq 1)$ esetén $0 < k - l\tau_{3-a}/\tau_a \leqq 1$. Feltételünk tehát ekvivalens azzal, hogy

$$k = \left[l \frac{\tau_{3-a}}{\tau_a} \right] + 1$$

mellett teljesül az

$$\frac{\eta}{\tau_a} \prec'_{3-a} k - l \frac{\tau_{3-a}}{\tau_a}$$

reláció. Ez azonban ekvivalens az

$$\frac{\eta}{\tau_a} \prec'_{3-a} 1 - l \left\{ \frac{\tau_{3-a}}{\tau_a} \right\}, \quad \text{azaz}$$

$$(*) \qquad \left\{ l \frac{\tau_{3-a}}{\tau_a} \right\} \prec'_{3-a} \frac{\vartheta_a - \eta_{3-a}}{\tau_a} = 1 - \frac{\eta}{\tau_a}$$

relációval. Így az $R \in \mathcal{R}^{(a)}$ ütemtervnek akkor és csak akkor nincs kritikus szituációja, ha (3.27) és *minden l-re* $(*)$ teljesül.

Könnyen bizonyítható (l. pl. [92]-ben a 2.3. pontban), hogy a $k\tau_a - l\tau_{3-a}$ alakú számok mindenütt sűrűek, ha τ_a és τ_{3-a} racionálisan függetlenek. Ez azt jelenti, hogy ekkor a $k - l\frac{\tau_{3-a}}{\tau_a}$ számok is mindenütt sűrűek és így az

$$x_l \doteq \left\{ l \frac{\tau_{3-a}}{\tau_a} \right\}, \quad l = 0, 1, \ldots$$

számok a [0, 1) intervallumban mindenütt sűrűek. (Sőt, ismeretes [68], hogy az (x_l) sorozat egyenletes eloszlású a [0, 1) bármely részintervallumán.) Ezért irracionális τ_{3-a}/τ_a esetén a $(*)$ reláció nem teljesülhet minden $l \geqq 0$ mellett, hiszen végtelen sokszor az $\left(1 - \frac{\eta}{\tau_a}, 1\right)$ intervallumba is belép az x_l sorozat. Ebből következik, hogy τ_1 és τ_2 racionális függősége szükséges feltétel.

Egyszerűen megmutatható (l. pl. [92] 2.3. pont), hogy

$$\frac{\tau_{3-a}}{\tau_a} = \frac{k_{3-a}}{k_a}, \quad k_a > 0, \quad k_{3-a} \text{ relatív prímek}$$

esetén az x_l sorozat periodikusan végigfutja a k/k_a, $0 \leqq k \leqq k_a - 1$ osztópontjait a [0, 1) intervallumnak, tehát ezeken egyenletes eloszlású. Ezért $(*)$ csak akkor teljesülhet minden $l \geqq 0$ mellett, ha a $(k_a - 1)/k_a$ utolsó osztópontra is teljesül, hogy

$$\frac{k_a - 1}{k_a} \prec'_{3-a} 1 - \frac{\eta}{\tau_a},$$

azaz

$$\frac{\eta}{\tau_a} \prec'_{3-a} \frac{1}{k_a},$$

azaz

$$\eta \prec'_{3-a} \frac{\tau_a}{k_a} \doteq q$$

teljesül.

Ezzel a (iv) és (3.27) feltételek együttes szükségességét és elegendőségét bizonyítottuk. A (3.27) azonban a (3.28) mellett redundancia, hiszen $k_1>0, k_2>0$ miatt (3.29)-ből $q\leqq\min(\tau_1, \tau_2)$ és így $\eta=\eta_1+\eta_2\leqq\min(\tau_1, \tau_2)$, amiből $\eta_i\leqq\vartheta_{3-i}$, $i=1, 2$ következik. Ezzel együtt a $\prec_{3-a}$ relációjel (3.12) alatti definíciójából következik a $\prec'_{3-a}$ jel (3.29) alatti formulája is.

Végül a $Q^{(1)}$ és $Q^{(2)}$ job-folyamok szimmetrikus szerepéből a (3.28) és (3.29)-ben, következik, hogy ha még $\eta_{3-a}>0$, azaz $\eta_1\eta_2>0$ teljesül, akkor $\mathscr{R}^{(a)}$ és $\mathscr{R}^{(3-a)}$ egyaránt egyetlen ütemterv és (3.30) igaz.

1. Megjegyzés. A 3.17. lemma szerint a (iv) esetekben, a 9. ábra szerint viszont a (ii)—(iii) esetekben R_{a0} periodikus.

2. Megjegyzés. A (iv) feltétel szükségességéből következik, hogy $\eta_a=0$ esetben az $\eta_{3-a}>0$ és $\tau_1\tau_2>0$, valamint (3.28) feltétel mellett is kell lennie $R\in\mathscr{R}^{(a)}$-ban kritikus szituációnak. Ez az eset a 2. táblázat $(ab)_a$ esete, amelynél (3.11)-nek racionálisan összefüggő τ_a, τ_{3-a} esetén a 2.1. tétel szerint mindig van megoldása. Legyen például $\eta_1=0$ és $\eta_2>0$. Ekkor $a=2$-re (iv) teljesül, így R_{20}-nak nincs kritikus szituációja. Ezért a

$$0\prec B_2\tau_2-A_2\tau_1\leqq\eta_2$$

egyenlőtlenségnek (2. táblázat $(ba)_2$ mező) nincs megoldása. Azonban $k_1\tau_2-k_2\tau_1=0$ teljesül.

A $0\leqq B_1\tau_1-A_1\tau_2\prec\eta_2$ egyenlőtlenségnek ekkor a megoldása csak $B_1=k_2, A_1=k_1$ lehet. Ugyanis, ha lenne kisebb megoldása, akkor annak hibája legalább akkora lenne, mint a $\xi=\tau_1/\tau_2$ szám bal oldali legjobb pozitív hibájú közelítése. Ennek hibája azonban a (2.20) formula szerint pontosan $\frac{1}{k_2}$ nagyságú. Ezért (B_1, A_1) helyébe ezt a közelítést téve, a hiba $\frac{\tau_2}{k_2}=q$ lesz. Ha tehát (3.28) teljesül, akkor az $(ab)_1$ esetben (3.11)-nek nem lehet $\Delta_a>0$ tulajdonságú megoldása, mert a legkisebb pozitív hibájú közelítő megoldásra is $\Delta_a=q$, márpedig $\eta_2\prec'_1 q$. De (3.30) szerint $\prec'_1=\prec$, így $\eta_2<q$ a (3.28) feltétel, ezért $\Delta_a>\eta_2$. Vagyis $\eta_a=0$, $\eta_{3-a}>0$ esetén (iv) többi feltétele mellett $R\in\mathscr{R}^{(a)}$-nak, $\eta_a>0$ és $\eta_{3-a}=0$ és (iv) többi feltétele mellett $R\in\mathscr{R}^{(3-a)}$-nak a σ_0 az első visszatérő szituációja, hiszen $\Delta'_a=0$, ill. $\Delta'_{3-a}=0$ a (3.11) legkisebb megoldására.

3. Megjegyzés. A 3.18. lemma egyszerű lehetőséget nyújt kritikus szituáció nélküli *GT*-ütemtervekkel rendelkező Q konfigurációk konstruálására. A degenerált esetektől eltekintve az alábbi módon kaphatunk ilyen konfigurációt.

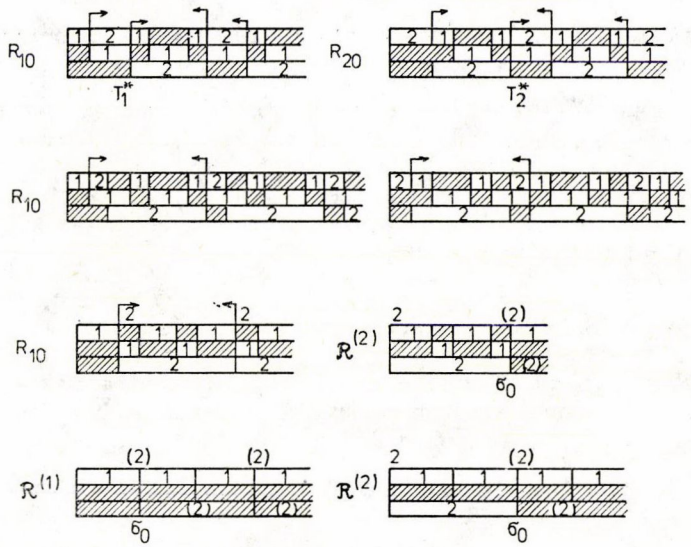
— válasszunk tetszőleges k_1, k_2 pozitív relatív prím egész számpárt,
— válasszunk egy $q>0$ valós számot,
— válasszuk meg η_1 paramétert $0<\eta_1<q$ feltétellel,
— válasszuk meg az η_2 paramétert a $0\leqq\eta_2\leqq q-\eta_1$ feltétellel,
— képezzük ϑ_1 és ϑ_2 paramétereket a $\tau_1=k_1q$, $\tau_2=k_2q$, $\vartheta_1=\tau_1-\eta_1$, $\vartheta_2=\tau_2-\eta_2$ formulákkal.

Az így nyert $Q=(\eta_1; \vartheta_1; \eta_2; \vartheta_2)$ konfigurációnál az $R\in\mathscr{R}^{(1)}$ ütemtervre teljesülnek a 3.18. lemma (iv) feltételei, ezért $\mathscr{R}^{(1)}$ azonos R_{10}-lal, amelynek nincs kritikus szituációja, összefüggő, szoros, következetes és periodikus, amelyben a β_1-

szituáció visszatérő (3.17. lemma). $\eta_2=0$ esetén $\mathscr{R}^{(2)}$ elemeinek σ_0 kritikus szituációja lesz. η_1 és η_2 szerepe természetesen felcserélhető. Az alábbi 3. táblázat néhány tipikus példát mutat a (iv) feltételt kielégítő, fenti módon konstruált konfigurációra. A 11. ábrán mutatjuk be azek ütemterveinek kezdő szakaszait.

Mind a négy konfigurációnál

$$k_1 = 1, \quad k_2 = 2, \quad q = 3, \quad \tau_1 = 3, \quad \tau_2 = 6.$$



11. ábra. *Kritikus szituáció nélküli ütemtervek*

3. TÁBLÁZAT

Kritikus szituáció nélküli konfigurációk

Sorszám	η_1	ϑ_1	η_2	ϑ_2	σ_1^*	σ_1'	p_1	σ_2^*	σ_2'	p_2	Megjegyzés
1.	1	2	2	4	β_2	—	6	β_1	—	6	$R_{10} \approx R_{20}$
2.	1	2	1	5	—	—	6	—	—	6	$R_{10} \not\approx R_{20}$
3.	2	1	0	6	—	—	6	—	σ_0	—	$\eta_2 = 0$
4.	3	0	0	6	—	σ_0	—	—	σ_0	—	$\eta_1 = q$, $\eta_2 = 0$, (iv) nem teljesül

A 3.13.—3.18. lemmák feltárják a *GT*-ütemtervek számos sajátosságát, amelyek azonban így nehezen tekinthetők át. Ezért célszerű a főbb eredményeket egy tételbe foglalni.

3.5. TÉTEL: Bármely $Q \in \mathscr{Q}$ konfiguráció *GT*-ütemterveire igazak az alábbi megállapítások.

(A) Az $\mathscr{R}^{(a)}(Q)$ halmaz elemeinek a σ_a^* jellegzetes, a σ_a' első *GT*-kritikus szituációja, valamint az $\mathscr{R}^{(Ga)}$ halmaz elemeinek σ_a'' első saját kritikus szituációja

minden elemre egyszerre létezik, vagy nem létezik. Ha létezik, minden elemre azonos, ugyanazon pontban lép fel és az $s(0)=s_a$ első kritikus döntéssel meghatározott. Fellépési helyeikre a $0 \leqq t'_a \leqq T^*_a \leqq T'_a \leqq T''_a \leqq \infty$ érvényes. Az első kritikus pontig az azonos típusú elemek kongruensek.

(B) Az $\mathscr{R}^{(GT)}(Q)$ halmaz összes elemére egyszerre létezik, vagy nem, a kritikus szituáció, és $\mathscr{R}^{(1)} \approx \mathscr{R}^{(2)}$ ha

(i) $\eta = 0$, $\vartheta_1 \vartheta_2 > 0$ és ϑ_1, ϑ_2 racionálisan függetlenek,

(ii) $\tau_1 \tau_2 = 0$,

(iii) valamelyik $i = 1, 2$ mellett

$$0 < \eta_{3-i} \leqq \vartheta_i \tag{3.32}$$

és a

$$0 \leqq \Delta_i \leqq \eta$$

egyenlőtlenség legkisebb $\omega_i^* \geqq (1, 0)$ megoldására

$$\Delta_i^* = \eta$$

teljesül.

Az $\mathscr{R}^{(1)}$ és $\mathscr{R}^{(2)}$ megfelelő lényegében azonos elemei a T_i^*, ill. $t'_{3-i} = \eta_{3-i}$ pontjuktól megegyeznek egymással. A T_i^* értéke megfelelően a fenti eseteknek

(i) $T_i^* = 0$, $i = 1, 2$,

(ii) $T_i^* = \eta_{3-i}$, ha $\tau_i = 0$,

(iii) $T_i^* = B_i^* \tau_i = A_i^* \tau_{3-i} + \eta$, ahol $\omega_i^* = (B_i^*, A_i^*)$.

Az (ii) és (iii) esetekben a T_i^* és t'_{3-i} pontokban fellépő közös szituáció a σ_i^* jellegzetes szituáció, amelynek típusa az alábbi:

$$\sigma_i^* = \begin{cases} \sigma_0, & \text{ha} \quad \vartheta_{3-i} = 0, \\ \beta_{3-i}, & \text{ha} \quad \vartheta_{3-i} > 0, \quad 0 \leqq \eta_i \leqq \vartheta_{3-i}, \\ \sigma_{i,0}, & \text{ha} \quad 0 < \vartheta_{3-i} < \eta_i. \end{cases} \tag{3.33}$$

Ha még (3.32) mellett $0 \leqq \vartheta_{3-i} < \eta_i$ és $\vartheta_i \equiv \eta_{3-i} \pmod{\tau_{3-i}}$, akkor $\omega_i^* = (1, [\vartheta_i/\tau_{3-i}])$, $T_i^* = \tau_i$, $\sigma_i^* = \beta_{3-i}$.

(C) Az $\mathscr{R}^{(a)}$ egyetlen R_{a0} ütemtervből áll, amely összefüggő, (kvázi) szoros és (kvázi) következetes, ha

(i) $\eta = 0$, $\vartheta_1 \vartheta_2 > 0$ és ϑ_1, ϑ_2 racionálisan függetlenek,

(ii) $\tau_1 = \tau_2 = 0$,

(iii) $\tau_i = 0$, $\vartheta_{3-i} > 0$ valamelyik $i = 1, 2$ mellett,

(iv) $\eta_a > 0$, $\tau_1 \tau_2 > 0$, τ_1, τ_2 racionálisan összefüggők és

$$\eta \prec'_{3-a} q = \tau_1/k_1 = \tau_2/k_2, \quad \text{ahol} \quad k_1, k_2 > 0$$

relatív prímek.

Az R_{a0} az (i) eset kivételével periodikus és periódushossza megfelelően

(ii) $p_a = 0$,

(iii) $p_a = \tau_{3-i}$,

(iv) $p_a = B_a'' \tau_a = A_a'' \tau_{3-a}$, ahol $\omega_a'' = (B_a'', A_a'')$

a $\Delta_a = 0$ egyenlet legkisebb $\omega_a \geqq (1, 0)$ megoldása.

Az R_{a0}-ban jellegzetes visszatérő szituáció

(iii) a β_{3-i}-szituáció az $\eta_{3-i} + k\tau_{3-i}$, $k = 0, 1, \ldots$ pontokban,

(iv) a β_a-szituáció az $\eta_a + kp_a$, $k = 0, 1, \ldots$ pontokban.

Az $\mathscr{R}^{(GT)}(Q) = \{R_{10}, R_{20}\}$ az (i)—(iii) esetekben, valamint a (iv) esetben $\eta_1 \eta_2 > 0$ mellett.

Bizonyítás. A bizonyítás lényegében a 3.14.—3.18. lemmákra történő hivatkozással történik.

(A) Definíció szerint $T_a^* \leqq T_a'$, ezért a σ_a^* közös volta $\mathscr{R}^{(a)}$ elemeire következik a 3.12. lemma (2) állításából. A σ_a' és σ_a'' azonossága és azonos helye $\mathscr{R}^{(a)}$ elemeinél, következik a 3.14. lemmából.

(B) Az $\mathscr{R}^{(1)} \approx \mathscr{R}^{(2)}$ következik a σ_i^* közös jellegzetes szituáció létéből, amely az (ii) esetben az 1. táblázatból olvasható ki, (iii) esetben pedig a 3.15. lemma állítása. A (3.32) feltétel a 3.15. lemma utáni 1. megjegyzéssel áll összhangban. A (3.33) következik a 3.15. lemmából. Az (i) esetben σ_i^* jellegzetes szituáció nem létezik, azonban a $t = 0$ pontbeli $\{\sigma\}_0$ halmaz utolsó eleme azonos, ezért $T_i^* = 0$ ponttól R_{10} és R_{20} megegyeznek.

A (B) alatti utolsó állítást a 3.15. lemma utáni 3. megjegyzésben bizonyítottuk.

(C) Az első állítás azonos a 3.18. lemmával. A periodicitás következik a 3.17. lemmából, illetve degenerált esetben leolvasható a 9. ábrákról. A visszatérő β-szituáció bizonyítása ugyanez. Az $\mathscr{R}^{(GT)}$ kételemű volta az (i)—(iii) esetekben az 1. táblázatból, az (iv) esetben a 3.18. lemmából következik.

3.5. Prioritásos ütemtervek

A gyakorlatban alkalmazott ütemezési stratégiák sokszor prioritásos jellegűek, amelyeknél „kritikus szituációkban" az igények prioritása határozza meg az ütemezési döntést (nem feltétlenül egyértelműen).

Job-folyamatok ütemezésénél sztochasztikus esetben szintén elsősorban prioritásos stratégiákat vizsgáltak tüzetesebben (l. ARATÓ [6] és TOMKÓ [93] munkákat). A prioritás ilyenkor *osztály*, nem dinamikus prioritás. Ez a teljes ütemezés folyamán változatlan.

A továbbiakban mi is elsősorban prioritásos ütemezéseket fogunk vizsgálni. Ez azonban nem szakad el eddigi vizsgálatainktól, mert a *GT*-ütemtervek speciális egyedeit szolgáltatják az egyes konfigurációknál.

Kétféle prioritást vizsgálunk. Az egyiknél a prioritás *csak* annyi *előnyt* jelent, hogy üres P_A processzorra a várakozó igények közül a magasabb prioritásút ütemezzük először megszakítás nélkül. A másik *prioritás megszakító*, ami azt jelenti,

hogy amikor a magasabb prioritású job-folyam igényli a P_A processzort, azt megkapja olyan áron is, hogy a másik job-folyam kiszolgálását felfüggesztjük, ha folyamatban volt. Mivel job-folyam párról van szó, mindössze két prioritási osztály van és mindegyikben egy job-folyam. Vagyis összesen négyféle prioritásos ütemezése lehet bármely $Q \in \mathcal{Q}$ konfigurációnak.

A *prioritásos ütemezés* elvéből következik azonban, hogy az *szoros,* mindig ütemezni kell szabad processzorra, ha van kiszolgálásra kész igény. Ez azt is jelenti, hogy minden szituációban domináns döntést kell alkalmazni. Így a prioritásos ütemezés eredménye mindig *GT*-ütemterv. A *GT*-ütemezésnél éppen a kritikus szituációk azok, amelyekben két alternatív döntés közül csak a valamelyik job-folyam késleltetése árán lehet választani. A természetes ütemezésnél a $\sigma_{i,0}$ típusú kritikus szituációkban a szorosság egyértelművé teszi a megteendő kritikus döntést. A döntés egyértelműen s_i. σ_0 szituációkban a prioritás dönti el, melyik kritikus döntést hozzuk — mindenesetre következetesen. A gazdaságos ütemezés eleve szoros. Ott az abszolút prioritás (megszakító) azt eredményezi, hogy $\sigma_0, \sigma_{i,1}$, $i=1,2$ típusú kritikus szituációkban egyaránt és következetesen ugyanazt a $Q^{(a)}$ job-folymot kell ütemezni: azt, amelyiknek prioritása van. A prioritásos ütemezések tehát *következetesek* is a 3.4. tétel szerint.

Ezzel máris igazoltunk egy igen fontos tételt.

3.6. TÉTEL: Minden prioritásos ütemterv szoros és következetes *GT*-ütemterv.

Definiáljuk most formálisan is a négy prioritásos ütemezést. Legyen az S_{a0} stratégia összefüggő, szoros és következetes ütemezés a $Q^{(a)}$ job-folyam előnyével, $a=1,2$; $S_{a,3-a}$ stratégia megszakításos, szoros és következetes ütemezés a $Q^{(a)}$ job-folyam prioritásával, $a=1,2$; S_{ab}, $a=1,2$, $b=0,3-a$, ezek közös jelölése; $R_{ab} \doteq R_{ab}(Q)=S_{ab}(Q)$ a $Q \in \mathcal{Q}$ konfiguráció ütemezésének és ütemtervének alternatív jelölései az S_{ab} stratégia alkalmazása esetén; X_{ab} az R_{ab} ütemterv valamely X jellemzőjének megkülönböztető jelölése (így pl. γ_{ab} a hatékonyság, p_{ab} a periódushossz stb.).

Bármely $Q \in \mathcal{Q}$ konfigurációra így a következő négy *prioritásos ütemtervet* definiáltuk:

R_{10}: *összefüggő szoros* és következetes ütemterv a $Q^{(1)}$ job-folyam előnyével σ_0 szituációkban;
R_{20}: *összefüggő szoros* és következetes ütemterv a $Q^{(2)}$ job-folyam előnyével σ_0 szituációkban;
R_{12}: *megszakításos szoros* és következetes ütemterv a $Q^{(1)}$ job-folyam prioritásával $\sigma_0, \sigma_{i,1}$ szituációkban;
R_{21}: *megszakításos szoros* és következetes ütemterv a $Q^{(2)}$ job-folyam prioritásával $\sigma_0, \sigma_{i,1}$ szituációkban.

E négy ütemterv alkotja a Q prioritásos ütemterveinek halmazát. Ez sokszor két elemre csökken az alábbi korollárium szerint.

3.1. KOROLLÁRIUM: Ha az $R^{(a)}$ egyetlen R_{a0} ütemterv, akkor R_{a0} éppen a prioritásos ütemterv a $Q^{(a)}$ job-folyam előnyével, vagy prioritásával. Ebben az esetben $R_{a,3-a}(Q)=R_{a0}(Q)$ $(a=1,2)$.

Bizonyítás: Következik a 3.6. tételből, mely szerint prioritásos ütemterv csak *GT*-ütemterv lehet.

A 3.6. tétel jelentősége az, hogy aszerint a prioritásos ütemtervek az $\mathscr{R}^{(GT)}$ halmaz elemei, ezért rendelkeznek a GT-ütemtervek közös tulajdonságaival. Pontosabban az R_{a0} ütemterv egy $\mathscr{R}^{(a)}$-beli T-ütemterv és az $R_{a,3-a}$ ütemterv egy $\mathscr{R}^{(a)}$-beli G-ütemterv.

Ezért értelme van a prioritásos ütemtervek olyan jellemzőiről beszélni, amelyek a GT-ütemtervek sajátjai: kritikus szituáció, határozott szakasz stb. Valójában már nincs kritikus szituáció és a teljes R_{ab} *ütemterv meghatározott az* $s(0)=s_a$ *első döntés által,* de struktúrájának elemzésekor hasznos, mint GT-ütemtervet felfogni és annak strukturális elemeire hivatkozni. Egyébként hangsúlyozzuk, hogy

az R_{a0} olyan T-ütemterv, amelynél a kritikus döntések következetesen
σ_0 szituációban: s_a,
$\sigma_{i,0}$ szituációban: s_i, $i=1, 2$;
az $R_{a,3-a}$ olyan G-ütemterv, amelynél a kritikus döntések következetesen mind a σ_0, mind a $\sigma_{i,1}$ szituációkban: s_a.

IRODALOM

[1] ABRAMSON, N., KUO, F. F. (Ed.), *Computer-Communication Networks* (Prentice-Hall, 1973).
[2] ACKOFF, R. L. (Ed.), *Progress in Operations Research* (Wiley, 1961).
[3] ADIRI, I., „Queueing models for multiprogrammed computers" in: *Proc. Symp. Computer-Communication Networks and Teletraffic* (Polytechnic Press, New York, 1972) 441—448.
[4] AHO, A. W., HOPCROFT, J. E., ULLMAN, J. D., *The Design and Analysis of Computer Algorithms* (Addison-Wesley, 1974).
[5] ALLEN, F. E., *Program Optimization* (Pergamon Press, 1969).
[6] ARATÓ, M., „Diffusion approximation for multiprogrammed computer systems" in: *Comp. et Math. with Appl.* (Pergamon Press, 1975) 315—326.
[7] ASHOUR, S., „An experimental investigation and comparative evaluation of flow-shop scheduling techniques", *Oper. Res.* **18** (1970) 541—549.
[8] BAKER, K., *Introduction to Sequencing and Scheduling* (Wiley, 1974).
[9] BARRON, D. W., *Computer Operating Systems* (Chapman and Hall Ltd., 1971).
[10] BASKETT, F., MUNTZ, R. R., „Queueing network models with different classes of customers" in: *Proc. 6th Ann. IEEE Int. Conf.* (1972) 205—209.
[11] BELADY, L. A., „A study of replacement algorithms for a virtual-storage computer", *IBM Syst. J.* **5** (1966) 78—101.
[12] BELLMAN, R., „Mathematical aspects of scheduling theory", *J. Soc. Ind. and Appl. Math.* **4** (1956) 168—205.
[13] BHARUCHA-REID, A. T., *Elements of the Theory of Markov Processes and Their Applications* (McGraw-Hill, 1960).
[14] BORODIN, A., MUNRO, I., *The Computational Complexity of Algebraic and Numerical Problems* (Am. Elsevier, 1975).
[15] BRINCH-HANSEN, P., *Operating Systems Principles* (Prentice-Hall, 1973).
[16] BRUCKER, P., LENSTRA, J. K., RINOOY KAN, A. H. G., *Complexity of Machine Scheduling Problems* (Math. Centrum, Tech. Rep. BW 43/75, Amsterdam, 1975).
[17] BRUNO, J., COFFMAN, E. G., SETHI, JR. and R., „Scheduling independent tasks to reduce mean finishing time", *Commun. of the ACM* **17** (1974) 382—387.
[18] BRUNO, J., SETHI, R., *On the Complexity of Mean Flow-Time Scheduling* (Pennsylvania State Univ., Comp. Sc. Dpmt. Tech. Rep., 1975).
[19] BUZEN, J. „Queueing Network Models of Multiprogramming", Ph. D. Thesis. Harward Univ., Cambridge, 1971.
[20] CAMPBELL, H. G., DUDEK, R. A., SMITH, M. L., „A heuristic algorithm for n job m machine sequencing problem", *Man. Sc.* **16** (1970) B630—B637.
[21] CHARLTON, J. M., DEATH, C. C., „A method of solution for general machine-scheduling problems", *Oper. Res.* **18** (1970) 689—707.
[22] CLARK, W., *The Gantt Chart* (Sir Isaac Pitman et Sons, London, 1952).

[23] COFFMAN, E. G., „Studying multiprogramming systems with the queueing theory", *Datamation* **13** (1967) June.
[24] COFFMAN, E. G. (Ed.), *Computer and Job-Shop Scheduling Theory* (Wiley Intersc., 1976).
[25] COFFMAN, E. G., DENNING, JR. and P. J., *Operating Systems Theory* (Prentice-Hall, 1973).
[26] COHEN, J. W., *The Single Server Queue* (North Holland, 1969).
[27] COLIN, A. J. T., *Bevezetés az operációs rendszerek tanulmányozásába* (KSH Kiadó, Budapest, 1976).
[28] CONWAY, R., MAXWELL, L., MILLER, L., *Theory of Scheduling* (Addison-Wesley, 1967).
[29] COOK, S. A., „The complexity of theorem-proving procedures" in: *Proc. 3d ACM Symp. on Theory of Comp.* 1971, 151—158.
[30] DENNING, P. J., „The working set model for program behavior", *Commun. of the ACM* **11** (1968) 323—333.
[31] DENNING, P. J., „Virtual memory", *ACM Computing Surveys* **2** (1970) 153—189.
[32] DENNIS, J. B., „Segmentation and the design of multiprogrammed computer systems", *J. of the ACM* **12** (1965) 589—602.
[33] DEO, N., *Graph Theory with Applications to Engineering and Computer Science* (Prentice-Hall, 1974).
[34] DIJKSTRA, E. W., „The structure of the THE multiprogramming system", *Commun. of the ACM* **11** (1968) 341—356.
[35] ELMAGHRABY, S. E. (Ed.), *Symposium on the Theory of Scheduling and its Applications* (Springer, 1973).
[36] GAREY, M. R., JOHNSON, D. S., „Complexity results for multiprocessor scheduling under resource constraints", *SIAM J. on Computing* **4** (1975) 397—411.
[37] GAREY, M. R., JOHNSON, D. S., „Scheduling tasks with nonuniform deadlines on two processors", *J. of the ACM* **23** (1976) 461—467.
[38] GAREY, M. R., JOHNSON, D. S., „Two-processor scheduling with start-times and deadlines", *SIAM J. on Computing* **6** (1977) 416—426.
[39] GAREY, M. R., JOHNSON, D. S., SETHI, R., „Complexity of flowshop and jobshop scheduling", *Math. Oper. Res.* **1** (1976) 117—129.
[40] GAVER, D. P., „Probability models for multiprogramming computer systems", *J. of the ACM* **14** (1976) 423—438.
[41] GAVER, D. P., SHEDLER, G. S., „Approximate models for processor utilization in multiprogrammed computer systems", *SIAM J. on Computing* **2** (1973) 183—192.
[42] GAVER, D. P., SHEDLER, G. S., „Processor utilization in multiprogramming systems via diffusion approximations", *Oper. Res.* **12** (1973) 569—576.
[43] GIFFLER, B., THOMPSON, G. L., „Algorithms for solving production-scheduling problems", *Oper. Res.* **8** (1960) 483—503.
[44] GIGLIO, R. J., WAGNER, H. M., „Approximate solution to the three-machine scheduling problem", *Oper. Res.* **12** (1964) 305—324.
[45] GILL, J., „Computational complexity of probabilistic Turing machines", *SIAM J. on Computing* **6** (1977) 675—695.
[46] GONZALEZ, JR. M. J., SOH, J. W., „Periodic job scheduling in a distributed processor system", *IEEE Trans. on Aerospace and Electr. Syst. AES*—**12** (1976) 530—535.
[47] GONZALEZ, T., IBARRA, O. H., SAHNI, S., „Bounds for LPT schedules on uniform processors", *SIAM J. on Computing* **6** (1977) 155—166.
[48] GRAHAM, R. L., „Bounds on multiprocessing timing anomalies", *SIAM J. on Appl. Math.* **17** (1969) 416—429.
[49] GREENBERG, H. H., „A branch-bound solution to the general scheduling problem", *Oper. Res.* **16** (1968) 353—361.
[50] HELLER, J., „Some numerical experiments for an $M \times J$ flow-shop and its decision-theoretical aspects", *Oper. Res.* **8** (1960) 178—184.
[51] HOARE, C. A. R., PERROTT, R. M. (Ed.), *Operating Systems Techniques*, Proc. of a Seminar, Queen's Univ. Belfast (Academic Press, 1972).
[52] IGNALL, E., SCHRAGE, L., „Application of the branch-and-bound technique to some flow-shop scheduling problems", *Oper. Res.* **13** (1965) 400—412.
[53] *Information Processing* '74. *Vol.* 2: *Software Vol.* 3: *Mathematical Aspects of Information Processing* IFIP Congress 74, Preprints (North-Holland, 1974).
[54] JACKSON, J. R., „An extension of Johnson's results on job-lot scheduling", *Naval Res. Logist. Quart.* **3** (1965) 3.
[55] JACKSON, J. R., „Networks of waiting lines", *Oper. Res.* **5** (1957) 518—521.

[56] JOHNSON, S. M., „Discussion: Sequencing in jobs on two machines with arbitrary time lags", *Man. Sc.* **5** (1959) 3.
[57] JOHNSON, S. M., „Optimal two- and three-stage production schedules with set-up times included", *Naval Res. Logist. Quart.* **1** (1954) 61—68 and in: [79] pp. 13—20.
[58] KARP, R. M., „On the computational complexity of combinatorial problems", *Networks* **5** (1975) 45—68.
[59] KARP, R. M., „Reducibility among combinatorial problnms", in [76], pp. 85—103.
[60] KHINTCHINE, A., „Einige Sätze über Kettenbrüche mit Anwerdung auf die Theorie der diophantischen Approximationen", *Math. Annalen* **92** (1924) 115—125.
[61] KHINTCHINE, A., *Kettenbrüche* (Leipzig, 1956).
KHINTCHIN, A., *Continued Fractions* (Groningen, 1963 and Univ. of Chicago Press, 1964).
HINCSIN, A. JA., *Cepnüe drobi* (Moszkva, 1935, 1942).
[62] KLEINROCK, L., *Communication Nets: Stochastic Message Flow and Delay* (McGraw-Hill, 1964 and Dover Publ., 1972).
[63] KLEINROCK, L., MUNTZ, R. R., „Multilevel processor-sharing queueing models for time-shared systems", in: *Proc. 6th Int. Teletraffic Congr.* 1970, 341/1—341/8.
[64] KNAPP, E. O., „Solution of the *n*-Job *m*-Machine Scheduling Problem with Intraservice Queue Constraints", M. S. Thesis. Dpmt. of Eng. Syst., Univ. of California, 1969.
[65] KNUTH, D. E., *The Art of Computer Programming. Vol. 1: Fundamental Algorithms* (Addison-Wesley, 1968).
[66] KOBAYASHI, H., „Application of the diffusion approximation to queueing networks 1: Equilibrium queue distributions", *J. of the ACM* **21** (1974) 316—328.
[67] KOBAYASHI, H., „Application of the diffusion approximation to queueing networks 2: Nonequilibrium distributions and applications to computer modeling", *J. of the ACM* **21** (1974) 459—469.
[68] KUPIERS, L., NIEDERREITER, H., *Uniform Distribution of Sequences* (Wiley, 1974).
[69] LABETOULLE, J., „Some theorems on real time scheduling", in: *Computer Architectures and Networks. Modelling and Evaluation.* IRIA Workshop 1974 (North-Holland, 1974) 285—298.
[70] LEWIS, P. A. W., SHEDLER, G. S., „A cyclic-queue model of system overhead in multiprogrammed computer systems", *J. of the ACM* **18** (1971) 199—220.
[71] LIU, C. L., LAYLAND, J. W., „Scheduling algorithms for multiprogramming in a hard real-time environment", *J. of the ACM* **20** (1973) 46—61.
[72] MAEKAWA, M., „Queueing models for computer systems connected by a communication line", *J. of the ACM* **24** (1977) 566—582.
[73] MAHL, R., „An Analytical Approach to Computer Systems Scheduling", Ph. D. dissert. Dpmt. of El. Eng., Univ. of Utah, 1970.
[74] MANNE, A. S., „On the job-shop scheduling problem", *Oper. Res.* **8** (1960) 219—223 and in: [79] pp. 187—192.
[75] MCKINNEY, J. M., „A survey of analytical time-sharing models", *ACM Computer Surveys* **1** (1969) 2, 105—116.
[76] MILLER, R. E., THATCHER, J. W. (Ed.), *Complexity of Computer Computation* (Plenum Press, New York, 1972).
[77] MITTEN, L. G., „Sequencing *n* jobs on two machines with arbitrary time lags", *Man. Sc.* **5** (1959) 293—298.
[78] MUNTZ, R. R., COFFMAN, E. G. JR., „Preemptive scheduling of real time tasks on multiprocessor systems", *J. of the ACM* **17** (1970) 324—338.
[79] MUTH, J. F., THOMPSON, G. L., *Industrial Scheduling* (Prentice-Hall, 1963).
[80] PARMELEE, R. P., PETERSON, T. I., TILLMAN, C. C., HARTFILD, D. J., „Virtual storage and virtual machine concepts", *IBM Syst. J.* **11** (1972) 118—130.
[81] PERRON, O., *Irrationalzahlen* (Berlin, Leipzig, 1921).
[82] PERRON, O. *Die Lehre von den Kettenbrüchen. Bd. 1: Elementare Kettenbrüche* (Teubner, Stuttgart, 1954).
[83] *Proc. of Computer Science Conf.*, Székesfehérvár, 1973.
[84] SASIENI, M., ET AL., *Operations Research Methods and Problems* (Wiley, 1959).
[85] SERLIN, O. „CPU scheduling in a time critical environment", *ACM Operating System Review* 1970.
[86] SERLIN, O., „Scheduling of time critical processes", in: *SJCC AFIPS Conf. Proc. Vol.* **40** (1972) 925—932.
[87] SISSON, R. L., „Methods on sequencing in job shops. A review", *Oper. Res.* **7** (1959) 10—29.

[88] SOH, J. W., „Scheduling Strategies for Periodic Jobs in a Multiprocessor Environment", Ph. D. dissert., Northwestern Univ. Evanston, 1974.
[89] SPECKER, E., *Seminar über die Komplexität von Entscheidungsproblemen* (Springer, 1976).
[90] SPIRN, J. R., *Program Behavior: Models and Measurements* (Elsevier, 1977).
[91] SZÜSZ, P., *Über die metrische Theorie der diophantischen Approximation* (Akad. Kiadó, Budapest, 1958).
[92] TANKÓ, J., Szabályos job-folyam párok ütemezésének vizsgálata (*MTA Számítástechnikai és Automatizálási Kutató Intézet, Tanulmányok* **82, 83,** Budapest, 1978).
[93] TOMKÓ, J., „Processor utilization study", in: *Comp. et Math. with Appl. 1* (Pergamon Press, 1975) 337—344.
[94] TRAUB, J. F. (Ed.), *Algorithms and Complexity. New Directions and Recent Results. Proc. Conf. on* (Academic Press, 1976).
[95] ULLMAN, J. D., „Complexity of sequencing problems", in: [24] pp. 139—164.
[96] ULLMAN, J. D., „NP-complete scheduling problems", *J. of Comp. Syst. Sc.* **10** (1975) 384—393.
[97] WATSON, R., *Timesharing System Design Concepts* (McGraw-Hill, 1970).
[98] WISMER, D. A., „Solution of the flow shop scheduling problem with no intermediate queues", *Oper. Res.* **20** (1972) 689—697.
[99] YOURDON, E., *Design of On-Line Computer Systems* (Prentice-Hall, 1972).

(Beérkezett: 1978. szeptember 4.)

TANKÓ JÓZSEF
MTA SZÁMÍTÁSTECHNIKAI ÉS AUTOMATIZÁLÁSI KUTATÓ INTÉZET
1250 BUDAPEST, URI U. 49.

DOMINATING SCHEDULES OF STEADY JOB-FLOW PAIRS

J. TANKÓ

This article is dealing with the special non-finite deterministic scheduling problem of steady job-flow pairs. The scheduling problem is to service two permanently renewing demand series by a single processor and to maximize the utilization of the processor.

In the first chapter we give a brief account of the literature dealing with scheduling problems and models in general and introduce the new model of the steady job-flow. In the second chapter we outline some preparatory knowledge necessary to discuss our scheduling problems later on. We are dealing with the continued fraction expansion and the task of solving so called coincidence problems. These first two chapters are the very shortened versions of a study of the author [92]. In the third chapter we define the concept of the steady job-flow pair the accuring concepts of scheduling steady job-flow pairs and classify the feasible schedules, the scheduling strategies and, by means of dominance theorems, reduce the set of schedules the optimal schedule must be searched in.

Alkalmazott Matematikai Lapok **4** (1978) 327—357

A ROBUSZTUS BECSLÉSEKRŐL

KERÉKFY PÁL
Budapest

A statisztikai becslések felhasználási területein előforduló megfigyelésekről megállapítható, hogy egyes kivételes esetektől eltekintve, hibaeloszlásuk nem normális, bár a nagy számok törvénye alapján ezt általában feltételezik. A valamely paraméteres modell alapján létrehozott optimális becslések általában nagyon érzékenyek a modelltől való eltérésekre. Ezek a problémák vezettek arra, hogy a 60-as évek közepén fellendült az ilyen eltérésekkel szemben robusztus becslések kutatása.

A bevezetésben ismertetjük a különböző statisztikusok által kitűzött kutatási irányokat és a *Prohorov-távolságon* alapuló modell céljait és felépítését. A 2. fejezet előkészíti a továbbiakat, a felhasznált fogalmakat ismerteti. A 3. fejezet F. HAMPEL definícióját és az ezzel kapcsolatos tételeket tartalmazza. Az $(X, \mathcal{A})$ mérhető tér $\mathcal{F}$ valószínűségi mértékein értelmezett T becslés pontosan akkor robusztus és konzisztens, ha T folytonos ($\mathcal{F}$-en a gyenge konvergencia topológiáját tekintve) (3.4. tétel). Definiáljuk a hatásgörbét és a katasztrófapontot. A 4. fejezet R. BERAN *Hellinger-differenciálhatóságon* alapuló definícióját és ennek következményeit ismerteti. Ezek után egyes becslések elemzése következik az egydimenziós esetben. Megvizsgáljuk a hatásgörbét, a katasztrófapontot, a becslés szórását és normalitását a normális és a Cauchy-eloszlás, valamint különböző szennyezett normális eloszlások esetén. Az összehasonlító vizsgálatok alapján a *Hampel-féle M-becslések* használata látszik célszerűnek.

1. Bevezetés

Miért szükségesek a robusztus becslések?

A statisztikai becsléselméletben gyakran feltesszük, hogy a megfigyelt mennyiségek független valószínűségi változók, közös F_{θ_0} eloszlással, amely egy meghatározott $\{F_\theta: \theta \in \Theta\}$ eloszláshalmazba tartozik. A paramétertér általában az $\mathbf{R}^k$ k-dimenziós euklideszi tér egy részhalmaza. Megpróbáljuk θ_0 értékét a megfigyelések alapján becsülni, azaz keresünk egy leképezést, amely az összes megfigyelések halmazát Θ-ba képezi le és θ_0-hoz közeli értékeket vesz fel nagy valószínűséggel, ha F_{θ_0} a valódi eloszlás. Az ilyen felépítést hívjuk paraméteres modellnek; szokásos példája a normális modell, amely feltételezi, hogy a valós értékű megfigyelések normális eloszlásúak ismeretlen várható értékkel és ismert vagy ismeretlen szórásnégyzettel. A becsülendő paraméter a várható érték vagy a (várható érték, szórásnégyzet) pár.

A közeli múltig a statisztikai becsléselmélet főleg olyan becslési eljárásokat (becsléseket) keresett, amelyek valamilyen értelemben optimálisak, ha a feltételezett paraméteres modell pontosan leírja a megfigyelések eloszlását. Sajnos ezek a modellek szinte sohasem igazak; bizonyos egyszerű diszkrét esetektől eltekintve, a valódi eloszlás sohasem egyezik meg pontosan a feltételezett paraméteres modellben szereplő eloszlások egyikével sem, még akkor sem, ha az értékeket előállító eljárást pontosan leírja a modell. Ennek okait négy fő csoportba sorolhatjuk:

(1) nagy hibák előfordulása: egy értéket nem pontosan másoltak le, rosszul olvasták le a mérőeszközről, vagy valami mást mértek (pl.: egy másik populáció egy elemét),
(2) a mérések korlátozott pontossága, kerekítések,
(3) ha az előző hatásokat sikerül is elég alacsony szinten tartani, gyakran előfordul, hogy a valódi eloszlás jelentősen különbözik a paraméteres modellben levőktől. Sokszor maga a modell is csak közelítőleg érvényes, vagy a paraméter változik az idő során,
(4) hosszú és „független" csillagászati és vegyészeti adatsorozatok is jelentős korrelációt mutatnak.

Ezekre az eltérésekre vannak példák az irodalomban. HAMPEL [5, 7] sok adatot idéz, ezek alapján 5—10% nagy hiba inkább szabálynak látszik, mint kivételnek. Még a nagy és igen pontos csillagászati és geodéziai minták is a normális eloszlásnál hosszabb „farok"-kal rendelkező hiba eloszlást mutatnak.

Ezek a hibák általában nem hagyhatók figyelmen kívül a gyakorlatban, mivel még nagyon enyhe és észrevehetetlen eltérések is teljesen elronthatják az „optimális" becslés viselkedését. Pusztán pontosan elvégzett kerekítések hatására megtörténik, hogy a becslés szórásnégyzete $1/n^2$ helyett $1/n$ rendben csökken[1].

A megfelelő megoldás természetesen függ a vizsgálat jellegétől és céljától. Egyes esetekben csak bizonyos kiugró *(outlier)* értékeket kell felismernünk. Máskor nem maguk a kiugró értékek az érdekesek, de meg kell próbálni tanulmányozni őket, hogy elkerülhetők legyenek, hogy közelebb kerüljünk az „ideális" eloszláshoz. A robusztus becsléseket akkor használjuk, ha a kiugró értékeket nem akarjuk vizsgálni, minden érdeklődésünk az adatok zömére irányul — és a célunk egy módszer létrehozása, amely jól működik a modellen kiugró értékek és más eltérések esetén is. A robusztus becslések alkalmazhatók mind a kicsi, alig észlelhető, mind a nagy eltérések esetén (akkor is, ha meg akarjuk tartani a modellt, bár tudjuk, hogy nagy és ismeretlen típusú hibák („szennyeződések", *contaminations*) lépnek fel). A különbség csak az, hogy az utóbbi esetben „még robusztusabb" becsléseket fogunk alkalmazni, még akkor is, ha ezek nem olyan jók az eredeti paraméteres modellen.

A robusztus becslések fő megközelítési módja az, hogy létrehozunk egy „modellt" a paraméteres modelltől való eltérésekre és olyan becsléseket keresünk, melyek jól megfelelnek ennek a modellnek és közelítőleg optimálisak.

TUKEY [15] két normális eloszlás keverékeit vizsgálta, ahol az egyiknek (a szennyezőnek) ugyanaz a várható értéke, mint a másiknak, de a szórása nagyobb. Modelljében a hiba eloszlása

$$F(x) = (1-\varepsilon)\Phi(x)+\varepsilon\Phi\left(\frac{x}{3}\right), \quad 0 \leqq \varepsilon \leqq 0,1$$

volt. Ha $\varepsilon=0,1$, akkor az átlag hatékonysága[2] kb. 70%, kisebb a mediánénál. A 6%-os levágott átlag hatékonysága minden ε esetén nagyobb 95%-nál. Érdemes megjegyezni, hogy a normális eloszlás esetén az átlagos eltérés a szórásnak 88%-os

[1] Lásd [5, A—3. old.].
[2] Hatékonyságon itt és a továbbiakban relatív efficienciát értünk, azaz a legkisebb szórású és a vizsgált becslés aszimptotikus szórásnégyzetének hányadosát tekintjük. Helyenként a legkisebb szórású helyett más becslést választunk viszonyítási alapnak, ezt akkor megnevezzük.

hatékonyságú becslése, azonban ha ε eléri a 0,0018 értéket, akkor az átlagos eltérés már jobb a négyzetes eltérésnél.

JEFFREYS [12, 5.7. pont] tapasztalatai szerint azonos körülmények között végzett igen gondos megfigyelések hibái 5—9 szabadsági fokú t-eloszlást követnek. Ha a körülmények nem azonosak (pl. más megfigyelő), akkor még hosszabb farkú lehet a hibák eloszlása. Rövidebb farkú eloszlások esetén általában kimutatható a korreláció a hibák között. Ha a hibát t_5, illetve t_9 eloszlásúnak tekintjük, akkor az átlag hatékonysága 80—93%, a négyzetes eltérésé 40—83%.

HUBER [10] az igen hasznos „nagy hiba" modelljében *(G-modell, gross-error-model, GEM)* megenged minden

$$(1-\varepsilon)F_\theta+\varepsilon G$$

alakú eloszlást, ahol F_θ a paraméteres modellhez tartozik és G vagy θ-ra szimmetrikus, vagy teljesen tetszőleges. Az ő $H_1(k)$ megoldása a modellen kívül is nagy jelentőségű.

Vannak tisztán matematikai és esztétikai szempontok is. A funkcionálanalízisben és alkalmazásaiban általában folytonos vagy differenciálható funkcionálokkal foglalkoznak. A matematikai statisztikában rendszeresen használt átlag, mint a valószínűségeloszlások terén definiált funkcionál, sehol sem folytonos (pl. a gyenge topológiában) és csak egy sűrű komplemeterű sűrű halmazon van értelmezve. A robusztus becslések elméletében majd folytonos (differenciálható) funkcionálokkal helyettesítjük.

Ezek után jogosan merül fel a kérdés, hogy eddig miért nem volt szükség a robusztus becslésekre. A gyakorlati statisztikusok általában többé-kevésbé intuitív módszerekkel robusztussá szokták tenni a becsléseket. Egy szokásos eljárás a kiugróan rossz értékek elhagyása, vagy ezek súlyának csökkentése, ez robusztus becslést ad. Teljes általánosságban azonban nem elégedhetünk meg az intuitív módszerekkel a következő problémák miatt:

1. Sokszor nem is olyan egyszerű kiválasztani az elhagyandó értékeket.

2. Ha a becslés hatékonyságát lényegesnek ítéljük, akkor nem folyamodhatunk egyszerű elhagyásos becslésekhez, mert ezek hatékonysága nem kielégítő. Néhány adat ennek illusztrálására: ha t_3 eloszlás várható értékét akarjuk becsülni, az elhagyásos módszerek $17A$-ra vagy $25A$-ra (def: 5. pont) vonatkoztatott relatív hatásfoka 86% alatt van, a mediáné a 80%-ot sem éri el.

3. RELLES és ROGERS végzett egy érdekes kísérletet, megvizsgálták, hogy gyakorlott statisztikusok által végrehajtott szubjektív elhagyásos módszerek milyen hatékonyságúak. Az eredmények t_3 eloszlásra 85%, normális eloszlásra 80% körül voltak.

Még nagyobb a robusztus becslések jelentősége a gépi adatfeldolgozás elterjedése miatt. Ha előzetesen nem elemezzük az adatokat, hogy a kiugróan rossz értékeket kiküszöböljük, akkor a nem robusztus becslések nagyon rossz eredményt szolgáltathatnak.

Hangsúlyozni kell, hogy vannak szituációk, amikor a robusztus becslések nem alkalmazhatók. Például, ha a vizsgálandó populáció tagjai nagyon különböznek a mért mennyiségben (eltekintve a mérési hibáktól), és ha a mennyiség átlagát akarjuk becsülni. Ilyen lehet egy jövedelmi statisztika — amelynél nem egy paraméteres modellbe akarjuk beilleszteni a megfigyeléseket, hanem valóban az átlagra vagyunk kíváncsiak.

A robusztus becslések elméletének Hampel-féle felépítése

Feltételezzük, hogy a megfigyeléseket generáló eljárás közelítőleg egy paraméteres modellel írható le, és a modell paramétereit vagy ezek egy függvényét akarjuk becsülni. Azonban tudjuk, hogy a paraméteres modell nem pontos. Ezért olyan becslést keresünk, melynek eloszlása csak enyhén változik meg, ha a megfigyelt értékek eloszlása csak kicsit tér el a modelltől.

Meg kell mondanunk, hogy milyen típusú eltéréseket engedünk meg. Az előzőekben felsorolt eltérések lesznek a megengedett eltérések. Az (1) megfelel annak, hogy egy kis tömeg tetszőlegesen helyezkedhet el, a (2) megengedi az egész tömeg elmozdulását egy kis környezetbe. Az (1) és (2) eltérés kvantitatív leírását a *Prohorov-távolság* adja[3], ez a gyenge konvergenciára vezet, így a (3) feltételt is tartalmazza. A *Prohorov-távolság* nagyon általános tereken definiálható. (Ez egy előnye a *Lévy-távolsággal* szemben, amely szintén a gyenge konvergenciára vezet. Továbbá, általában két olyan eloszlás, melyek *Prohorov-távolsága* nagy és *Lévy-távolsága* kicsi, könnyen megkülönböztethető a gyakorlatban. Általában a nagy *Prohorov-távolság* összefügg a könnyen megkülönböztethetőséggel, ha a mintatér metrikája természetes. A nagy *Kolmogorov-* vagy *variációs távolságú* eloszlások nem mindig különböztethetők meg a kerekítési hibák miatt.)

Azt fogjuk megkövetelni, hogy a becslés eloszlása (a *Prohorov-távolságra* vonatkozóan) folytonos funkcionálja legyen a valódi eloszlásnak. Azonban, ha becslésről beszélünk, akkor mindig becslések egy egész sorozatára gondolunk. Megtörténhet, hogy a megfigyelések számának növekedésével ($n \to +\infty$) a tényleges eloszlásnak egyre közelebb kell lennie a paraméteres modellhez, hogy a becslés eloszlását közel tartsuk a modellnek megfelelő eloszláshoz. Egy ilyen becléssorozat igen rosszul fog viselkedni nagy n esetén, ezért megköveteljük még, hogy a folytonosság n-ben egyenletes legyen. Ez lesz a robusztusság kvalitatív definíciója.

A kvantitatív vizsgálatok céljából célszerű lineáris megközelítést alkalmazni, azaz a funkcionál deriváltját vizsgálni (*von Mises-derivált,* ld. [18]). Ez a derivált a mintatéren értelmezett függvény és elég jól leírja a becslés lokális viselkedését. Természetes, hogy abszolút értékének felső határát használjuk a robusztusság kvantitatív mértékéül. Ez a szuprémum méri a kilógó értékek lehetséges maximális hatását, és ezt nevezzük a becslés érzékenységének *(sensitivity)*. A legtöbb esetben a derivált és az érzékenység elég jól jellemzi a becslés robusztusságát.

2. Előkészítés, felhasznált tételek, elnevezések

Mértékek távolsága

Ebben a fejezetben néhány, a mértékeken értelmezett, távolságfogalmat definiálunk.

Legyen $(X, \mathcal{A})$ mérhető tér, ahol X egy szeparábilis teljes metrikus tér és $\mathcal{A}$ a topológia által generált σ-algebra. Jelölje a tér metrikáját ϱ, legyen

$$A^\varepsilon := \{x \in X: \varrho(A, x) < \varepsilon\}$$

az $A \subset X$ halmaz nyílt környezete.

[3] L. a 3. pontot.

2.1. DEFINÍCIÓ: Az $(X, \mathcal{A})$ téren értelmezett P, Q véges mértékek *Prohorov-távolsága*

$$\pi(P, Q) = \inf \{\varepsilon: \forall A \in \mathcal{A}: P(A) \leqq Q(A^\varepsilon) + \varepsilon \quad \text{és} \quad Q(A) \leqq P(A^\varepsilon) + \varepsilon\}.$$

Az eredeti definícióval ekvivalens definíciót kapunk, ha csak zárt A halmazokra szorítkozunk. Ha a két mérték teljes tömege megegyezik, azaz $P(X) = Q(X)$, akkor elég az első egyenlőtlenséget megkövetelni. Tehát a valószínűségi mértékekre a következő ekvivalens definíció érvényes:

$$\pi(P, Q) = \inf \{\varepsilon: \forall A \in \mathcal{A}, \text{ zárt}: P(A) \leqq Q(A^\varepsilon) + \varepsilon\}.$$

Egyszerűen látható, hogy az így értelmezett π valóban metrika és $\pi \leqq 1$.

PROHOROV [17] megmutatta, hogy a π metrikában való konvergencia ekvivalens a mértékek gyenge konvergenciájával.

STRASSEN [14] 11. tételének következménye szerint

$$\pi(P, Q) \leqq \varepsilon \Leftrightarrow \quad \text{van olyan } R \text{ valószínűségi mérték } X \times X\text{-en},$$

melynek marginálisai P és Q, és $R(D(\varepsilon)) \geqq 1 - \varepsilon$. ($D(\varepsilon)$ az átló zárt ε-környezete a maximum metrikában.)

2.2. DEFINÍCIÓ: A $\mu: \mathcal{S} \to \mathbf{R}$ σ-algebrán értelmezett előjeles mérték *teljes változása*

$$V_\mu(E) = \sup\{\sum |\mu(E_i)|: X = \dot{\bigcup} E_i, E_i \in \mathcal{S}\}.$$

Ismeretes, hogy $V_\mu = |\mu| = \mu_+ + \mu_-$, ahol $\mu = \mu_+ - \mu_-$ az előjeles mérték *Jordán-felbontása*. A teljes változás norma a korlátos változású előjeles mértékek terén, $\|\mu\| = |\mu|(X) = \mu_+(X) + \mu_-(X)$.

2.1. LEMMA: Ha $\mu = P - Q$, ahol P és Q valószínűségi mértékek, akkor

$$\|\mu\| = \sup \{|\mu|(A): A \in \mathcal{A}\}.$$

Bizonyítás: Legyen $X = X_1 \dot{\cup} X_2$ az X tér *Hahn-féle felbontása* a $P - Q$ előjeles mértékre nézve (X_1 a pozitív, X_2 a negatív halmaz).

$$|P - Q|(A) = (P - Q)(X_1 \cap A) + (Q - P)(X_2 \cap A) \leqq$$

$$\leqq (P - Q)(X_1) + (Q - P)(X_2) \quad \text{és}$$

$A = X$ esetén egyenlőség áll. Ezzel az állítást bebizonyítottuk.

Látható, hogy elég zárt halmazokra tekinteni a szuprémumot:

$$\|P - Q\| = \sup \{|P - Q|(A): A \in \mathcal{A}, \text{ zárt}\}.$$

2.3. DEFINÍCIÓ: Legyen $\mathcal{F}$ a valós egyenesen értelmezett sűrűségfüggvények halmaza, az $f, g \in \mathcal{F}$ függvények *Hellinger-távolsága*

$$\|f^{1/2} - g^{1/2}\|,$$

ahol $\|\cdot\|$ az L_2-normát jelöli.

2.2. LEMMA: $f_n \xrightarrow{H} f \Leftrightarrow \int \sqrt{f_n f} \to 1$.

Bizonyítás:

$$f_n \xrightarrow{H} f \Leftrightarrow \|f_n^{1/2} - f^{1/2}\| \to 0$$

$$\|f_n^{1/2} - f^{1/2}\| = \left(\int (\sqrt{f_n} - \sqrt{f})^2\right)^{1/2} =$$

$$= \left(\int f_n + \int f - 2\int \sqrt{f_n f}\right)^{1/2} = \left(2 - 2\int \sqrt{f_n f}\right)^{1/2}.$$

Ebből már látszik az állítás.

A következőkben az itt definiált metrikák kapcsolatát vizsgáljuk. A variációs távolság közvetítő szerepet játszik majd a *Prohorov-* és a *Hellinger-metrika* között.

A Hellinger-távolság és a teljes variáció kapcsolata

2.4. DEFINÍCIÓ: Legyen μ_1 és μ_2 két mérték **R**-en és a λ σ-véges mérték egy dominálló mértéküknek, a μ_i *Radon—Nykodim-deriváltja* (sűrűségfüggvénye) p_i. Legyen $f: (0, +\infty) \to \mathbf{R}$ konvex függvény. A μ_1 és μ_2 mértékek *f-divergenciáját* definiáljuk a következő módon:

$$\mathcal{I}_f(\mu_1, \mu_2) = \int p_2(x) f\left(\frac{p_1(x)}{p_2(x)}\right) \lambda(dx)$$

Megjegyzés.

(1) $\mathcal{I}_f(\mu_1, \mu_2)$ értéke nem függ λ megválasztásától,

(2) $\mathcal{I}_f(\mu_1, \mu_2) \geqq f(1)$ és ha f szigorúan konvex 1-ben, akkor

$$\mathcal{I}_f\{\mu_1, \mu_2) = f(1) \Leftrightarrow \mu_1 = \mu_2,$$

(3) ez utóbbi megjegyzés alapján $\mathcal{I}_f(\mu_1, \mu_2) - f(1)$ alkalmas arra, hogy a μ_1 és μ_2 távolságát mérjük vele.

2.1. TÉTEL. Ha f szigorúan konvex 1-ben és $f(0)$, valamint $\lim_{\mu \to +\infty} f(u)/u$ véges, akkor a fenti $\mathcal{I}_f(\mu_1, \mu_2) - f(1)$ által generált topológia megegyezik a teljes variáció normából származtatott topológiával.

A fenti eredmények CSISZÁR IMRE [14] cikkéből ismertek.

Az előző tétel egyszerű következménye a következő

2.2. TÉTEL. A *Hellinger-távolság* által származtatott topológia a valós egyenesen értelmezett valószínűségi mértékek terén megegyezik a teljes variáció által generált topológiával.

Bizonyítás. A *Hellinger-távolság:*

$$\|g^{1/2} - h^{1/2}\|_{L_2} = \left(\int (\sqrt{g} - \sqrt{h})^2 d\lambda\right)^{1/2} =$$

$$= \left(\int g + h - 2\sqrt{gh}\, d\lambda\right)^{1/2} = \left(2 - 2\int \sqrt{gh}\, d\lambda\right)^{1/2}.$$

Legyen $f(x) = +2\sqrt{x}$, így

$$\|g^{1/2} - h^{1/2}\| = (\mathcal{I}_f(\mu_g, \mu_h) - f(1))^{1/2},$$

ahol μ_g és μ_h a g, ill. h sűrűségfüggvényű eloszlás.

Ebből már látható, hogy a két topológia megegyezik.

A Prohorov-távolság és a Hellinger-távolság kapcsolata

2.3. LEMMA: $\pi(P, Q) \leqq \|P-Q\|$.

Bizonyítás: A teljes variáció definíciója szerint

$$P(A)-Q(A) \leqq \|P-Q\|,$$

$$P(A) \leqq Q(A)+\|P-Q\| \Rightarrow \pi(P, Q) \leqq \|P-Q\|.$$

2.3. TÉTEL: Változtassuk az X tér metrikáját a következő módon: szorozzuk meg a metrikát n-nel, legyen ez ϱ_n. Ekkor a mértékek *Prohorov-távolsága* is változik (mert függ az ε-környezet mértékétől), jelölje ezt π_n $(n \geqq 1)$. Ekkor $\pi_n \to |.,.|_{\text{var}}$.

Bizonyítás.

a) Először belátjuk, hogy (π_n) monoton növő, azaz $\forall P, Q: \pi_n(P, Q) \leqq \leqq \pi_{n+1}(P, Q)$. Jelölje a ϱ_n szerinti környezeteket A_n^ε, így egyszerűen látható:

$$A_{n+1}^\varepsilon \subset A_n^\varepsilon \Rightarrow Q(A_{n+1}^\varepsilon) \leqq Q(A_n^\varepsilon).$$

A definíció szerint

$$\pi_{n+1}(P, Q) \leqq \varepsilon \Leftrightarrow \forall A \in \mathscr{A}, \quad \text{zárt:} \quad P(A) \leqq Q(A_{n+1}^\varepsilon)+\varepsilon$$

$$\Rightarrow P(A) \leqq Q(A_n^\varepsilon)+\varepsilon \Rightarrow \pi_n(P, Q) \leqq \varepsilon.$$

Tehát (π_n) monoton növő.

b) Bebizonyítjuk, hogy $\lim A_n^\varepsilon = A$.

Tudjuk, hogy $A \subset A_n^\varepsilon$ $(\forall n \in \mathbf{N})$ és így

$$A \subset \liminf A_n^\varepsilon \subset \limsup A_n^\varepsilon. \tag{2.1}$$

Lássuk be, hogy

$$\limsup A_n^\varepsilon \subset A; \tag{2.2}$$

illetve helyette:

$$A^c \subset (\limsup A_n^\varepsilon)^c.$$

Ha $x \notin A$, akkor (A zártsága miatt) $\varrho(x, A)=\delta>0$, így $\exists N: \forall n \geqq N: n \cdot \delta > > \|P-Q\| \geqq \varepsilon \Rightarrow x$ csak véges sok A_n^ε-nak eleme $\Rightarrow x \notin \limsup A_n^\varepsilon$. Ebből következik (2.2).

(2.1) és (2.2) összevetéséből

$$A = \liminf A_n^\varepsilon = \limsup A_n^\varepsilon = \lim A_n^\varepsilon.$$

c) Belátjuk, hogy $\sup_n \pi_n(P, Q) = \|P-Q\|$. Legyen $\varepsilon>0$ felső korlátja a sorozatnak, azaz olyan szám, hogy $\forall n \in \mathbf{N}: \pi_n(P, Q) \leqq \varepsilon \Rightarrow \forall n: \forall A \in \mathscr{A}$, zárt: $P(A) \leqq \leqq Q(A_n^\varepsilon)+\varepsilon$. Mivel $\lim A_n^\varepsilon = A$, a *Fatou-tételt* alkalmazva $\lim Q(A_n^\varepsilon) = Q(A)$. Így $P(A) \leqq Q(A)+\varepsilon$, azaz $P(A)-Q(A) \leqq \varepsilon$. Ebből következik, hogy $\|P-Q\| \leqq \varepsilon$. Ezek szerint a legkisebb felső korlátra is áll az egyenlőtlenség, tehát $\sup \pi_n(P, Q) \geqq \geqq \|P-Q\|$. Ezt a 2.3. lemmával összevetve kapjuk az állítást.

d) Eddig láttuk, hogy $(\pi_n(P, Q))$ monoton növő és szupremuma $\|P-Q\|$, így ez a határértéke is.

2.4. TÉTEL: Van **R**-en értelmezett valószínűségi mértékeknek olyan sorozata, amely a *Prohorov-metrikában* konvergál egy valószínűségi mértékhez, de a variációs normában nem konvergál ahhoz.

Bizonyítás: PROHOROV tétele szerint $F_n \xrightarrow{\pi} F \Leftrightarrow F_n \to F$ gyengén, ismeretes továbbá, hogy ez utóbbi ekvivalens azzal, hogy az eloszlásfüggvények sorozata pontonként konvergál F-hez, annak minden folytonossági pontjában. Meg fogjuk mutatni, hogy tetszőleges folytonos eloszlásfüggvényhez tudunk olyan eloszlásfüggvény-sorozatot konstruálni, amely egyenletesen konvergál, de a variációs távolságban nem konvergál. Legyen F folytonos eloszlásfüggvény. Ekkor F mérhető és korlátos, így a mérhető függvények alaptétele szerint létezik (F_n) mérhető lépcsősfüggvény-sorozat, amely egyenletesen konvergál F-hez és a következőképpen áll elő.

Legyen

$$c_i^n = i2^{-n}, \quad i = 0, 1, \ldots, 2^n; \quad n = 1, 2, \ldots$$

$$H_i^n = \mathbf{R}(c_{i-1}^n > F \geqq c_i^n), \quad i = 1, 2, \ldots, 2^n$$

$$F_n = \sum_{i=2}^{2^n} c_i^n \chi_{H_i^n}.$$

Ez az előállítás az alaptétel konstruktív bizonyításában szereplő sorozat egyszerűsített és enyhén módosított formája. A sorozat tagjai eloszlásfüggvények.

Ha tekintjük a számegyenes $\mathbf{R} = \left(\bigcup_{i=1}^{2^n} \operatorname{int} H_i^n\right) \cup \left(\bigcup_{i=1}^{2^n} \mathrm{b}H_i^n\right)$ diszjunkt felosztását, akkor látható, hogy $|F_n, F|_{\mathrm{var}} \geqq 2$, mert az első unió F_n-mértéke zérus és F-mértéke 1, a második unió F-mértéke zérus és F_n-mértéke 1.

Az előző tétel és a 2.2. lemma alapján látható a

2.5. TÉTEL. A valós egyenesen értelmezett valószínűségi mértékek terén a teljes variáció topológiája finomabb, mint a *Prohorov-távolság* topológiája és a két topológia nem egyezik meg.

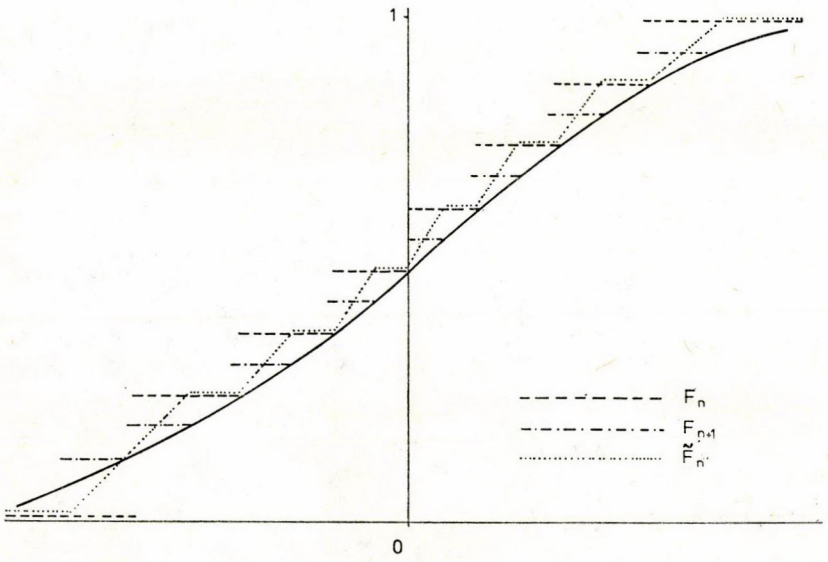
Megjegyzés. A 2.4. tételben alkalmazott konstrukció kis módosításával olyan $(\tilde{F}_n)$ sorozatot is tudunk mutatni, amely abszolút folytonos. Származtassuk $\tilde{F}_n$-et F_{n+1}-ből úgy, hogy minden második lépcsőt egy növekvő lineáris függvénnyel helyettesítünk, amely folytonosan összeköti a két szomszédos lépcsőt. A függvényt az 1. ábrán szemléltetjük.

Az ábrából látszik, hogy $\tilde{F}_n$ mindenhol F és F_n közötti értékeket vesz fel, azaz $\tilde{F}_n$ is konvergál F-hez. Az eredetihez hasonló meggondolás mutatja, hogy $|\tilde{F}_n, F|_{\mathrm{var}} \to 0$ (pozitív alsó korlátja van).

Látható, hogy ilyen módon tetszőleges sokszor differenciálható sorozatot is létrehozhatunk, ha a sarkokat megfelelően lekerekítjük.

A fenti eredményeket a 2.2. tétellel összevetve kapjuk, hogy

2.6. TÉTEL: A valós egyenesen értelmezett valószínűségi mértékek terén a *Hellinger-távolság* topológiája finomabb, mint a *Prohorov-távolság* topológiája és a két topológia nem egyezik meg. Minden abszolút folytonos valószínűségi mértékhez található olyan abszolút folytonos valószínűségi mértékekből álló sorozat, amely a *Prohorov-metrikában* konvergál hozzá, de a *Hellinger-metrikában* nem.



1. ábra

3. A robusztus becslések definíciója

Becsléssorozatok

Legyen $(X, \mathcal{A})$ mérhető tér, úgy hogy X szeparábilis teljes metrikus tér és $\mathcal{A}$ a topológia által generált σ-algebra. Legyen $\mathcal{F}$ az $(X, \mathcal{A})$ mérhető téren értelmezett összes valószínűségi mértékek halmaza. Ha $n \in \mathbf{N}$, legyen $\mathcal{F}_n \subset \mathcal{F}$ azon atomos mértékek halmaza, melyek atomjai k/n mértékűek.

Legyen $\tilde{X}^n$ az X^n tér a koordináták sorrendjétől eltekintve, azaz X^n modulo a koordináta permutációk csoportja. Ez természetes módon azonosítható $\mathcal{F}_n$-nel, hiszen minden $(x_1, \ldots, x_n)$ sorozat meghatároz egy F_n-et (az empirikus eloszlást) és fordítva, F_n meghatároz egy sorozatot, a koordináták sorrendjétől eltekintve. Ily módon $\mathcal{F}_n$ elemei az n megfigyelésből álló kísérletek kimeneteleinek tekinthetők. Itt most nem vesszük figyelembe a megfigyelt értékek sorrendjét, bár ennek a gyakorlati problémákban nagy jelentősége lehet. Ezt az egyszerűsítést megtehetjük akkor, ha független, azonos eloszlású valószínűségi változókra szorítkozunk.

Tekintsük $\mathbf{R}^k$-t a maximum metrikával és $\mathcal{F}_n$-et a *Prohorov-metrikával,* definiáljuk a becléssorozat fogalmát.

3.1. DEFINÍCIÓ. A (T_n) sorozatot *becléssorozat*nak nevezzük, ha

$$T_n : \mathcal{F}_n \to \mathbf{R}^k$$

mérhető leképezés minden $n \in \mathbf{N}$ esetén.

Megjegyzés.

1) T_n tekinthető $T_n: X^n \to \mathbf{R}^k$ leképezésnek is, figyelembe véve $\tilde{X}^n$ és $\mathscr{F}_n$ fenti azonosítását. Ily módon a definíció megfelel a szokásos felépítésnek. Minden megfigyelés n-eshez hozzárendelünk egy paraméterértéket.
2) Megadhatjuk T_n-et úgy is, hogy megadunk egy $T: \mathscr{F} \to \mathbf{R}^k$ leképezést és $T_n = T|\mathscr{F}_n$.

Robusztus becsléssorozatok

$\mathscr{F}$-en és $\mathscr{F}_n$-en a *Prohorov-metrikát* és az átala meghatározott topológiát tekintjük.

Legyenek $x_1, x_2, \ldots$ független, azonos eloszlású valószínűségi változók F eloszlással, és legyen F_n az első n által meghatározott valószínűségi mérték. A $T_n: \mathscr{F}_n \to \mathbf{R}^k$ (vagy $T_n: X^n \to \mathbf{R}^k$) leképezés indukál egy valószínűségi mértéket $\mathbf{R}^k$-n, ez T_n-eloszlása F-re vonatkozóan:

$$\alpha_F(T_n) = F^n T_n^{-1}, \quad \text{ahol} \quad T_n^{-1} \quad \text{a} \quad T_n$$

teljes inverzét jelöli[4].

3.2. DEFINÍCIÓ. Egy (T_n) becsléssorozat *robusztus* az F valószínűségi mértéken, ha

$$\forall \varepsilon > 0: \ \exists \delta > 0: \ \forall n \in \mathbf{N}, \quad \forall G \in \mathscr{F}:$$

$$\big(\pi(F, G) < \delta \Rightarrow \pi\big(\alpha_F(T_n), \alpha_G(T_n)\big) < \varepsilon\big).$$

Ha a becsléssorozat minden $F \in \mathscr{F}$ eloszlásban robusztus, akkor robusztusnak nevezzük.

Megjegyzés. Azt is mondhatjuk, hogy a (T_n) becsléssorozatot akkor nevezzük robusztusnak, ha az $F \mapsto \alpha_F(T_n)$ leképezés folytonos a *Prohorov-metrikára* vonatkozóan (n-ben egyenletesen).

3.3. DEFINÍCIÓ. A (T_n) becsléssorozat *folytonos* F-ben, ha

$$\forall \varepsilon > 0: \ \exists \delta > 0, \quad \exists n_0: \ \forall n, m \geqq n_0, \quad \forall F_n \in \mathscr{F}_n, \ F_m \in \mathscr{F}_m:$$

$$\big(\pi(F_n, F) < \delta, \ \pi(F_m, F) < \delta \Rightarrow |T_n(F_n) - T_m(F_m)| < \varepsilon\big),$$

A $T: \mathscr{F} \to \mathbf{R}^k$ funkcionál *folytonos* F-ben, ha

$$\forall \varepsilon > 0: \ \exists \delta > 0: \ \forall G \in \mathscr{F}.$$

$$\big(\pi(F, G) < \delta \Rightarrow |T(F) - T(G)| < \varepsilon\big).$$

Egyszerűen megmutatható, hogy ha (T_n) folytonos F-ben, akkor $\big(T_n(F_n)\big)$ sztochasztikusan tart egy $T_\infty(F)$ értékhez, azaz konzisztens becslése $T_\infty(F)$-nek. Ha $T_n = T|\mathscr{F}_n$, akkor T folytonosságából következik a (T_n) sorozat folytonossága.

3.1. TÉTEL. Ha a (T_n) sorozat robusztus és konzisztens $\mathscr{F}$-en, $T_n(F_n) \to T_\infty(F)$ st., akkor T_∞ folytonos $\mathscr{F}$-en. Ekkor a $\tilde{T}_n := T_\infty|\mathscr{F}_n$ sorozat robusztus és kon-

[4] Ha $H \subset \mathbf{R}^k$, akkor $(\alpha_F(T_n))(H)$ annak a valószínűsége, hogy a T_n által szolgáltatott paraméterérték a H halmazba esik:

$$F^n(\{x \in X^n: \ T_n(x) \in H\}).$$

zisztens becslése $T_\infty(F)$-nek. A $\tilde{T}_n$ funkcionálhoz tartozó, X^n-en értelmezett, pontfüggvény folytonos.

Most elégséges feltételeket adunk arra, hogy egy becléssorozat folytonos legyen valamely F helyen.

3.2. TÉTEL. Legyen (T_n) folytonos F-ben és mint X^n-en értelmezett pontfüggvény folytonos (egy nulla F^n-mértékű halmaztól eltekintve). Ekkor (T_n) robusztus F-ben.

A sorozatok folytonosságáról mondottak alapján a tételt a következőképpen is kimondhatjuk:

3.3. TÉTEL. Legyen $T: \mathscr{F} \to \mathbf{R}^k$ folytonos F-ben és $T_n = T|\mathscr{F}_n$, mint X^n-en értelmezett függgény, folytonos. Ekkor (T_n) robusztus F-ben.

A következő tételben szükséges és elégséges feltételeket adunk konzisztens becléssorozatok mindenütt való robusztusságára.

3.4. TÉTEL. Legyen $T: \mathscr{F} \to \mathbf{R}^k$ funkcionál és $T_n = T|\mathscr{F}_n$, ekkor a következő állítások ekvivalensek:

(1) T folytonos (minden F-ben),

(2) (T_n) robusztus és konzisztens,

$$T_n(F_n) \to T(F) \quad \text{st.} \quad (\forall F \in \mathscr{F}),$$

(3) $\forall K \subset \mathscr{F}$ gyengén kompakt, $\forall \varepsilon > 0: \exists \delta > 0:$

$$\forall F \in K, \quad G \in \mathscr{F}: \ (\pi(F, G) < \delta \Rightarrow |T(F) - T(G)| < \varepsilon),$$

azaz T egyenletesen folytonos minden gyengén kompakt halmazon.

A tételek bizonyítása megtalálható HAMPEL [5, 6] műveiben.

Az egyes becslések vizsgálatánál az utóbbi tételből az (1)⇔(2) ekvivalenciát fogjuk a robusztusság eldöntésére használni.

A definíció módosítása

Legyen $(X^n, \mathscr{A}_n)$ $(n \in \mathbf{N})$ az $(X, \mathscr{A})$ mérhető tér n-szeres *Descartes-szorzata* önmagával. Ha F egy valószínűségi mérték X-en, akkor F^n jelölje a szorzatmértéket X^n-en.

Legyen $\tilde{X}^n$ az X^n tér modulo a koordinátapermutációk csoportja, amely mint láttuk, természetes módon megfeleltethető $\mathscr{F}_n$-nek. Az $\mathscr{F}_n$-beli *Prohorov-távolság* által származtatott metrikát jelöljük $\tilde{X}^n$-ben is π-vel. Az $\tilde{X}^n$-on értelmezett mértékeket $\tilde{F}^n$, $\tilde{Q}_n$, ... jelöli.

3.4. DEFINÍCIÓ. Egy (T_n) becléssorozat *π-robusztus* F-ben, ha

$$\forall \varepsilon > 0: \ \exists \delta > 0: \ \forall n:$$

$$(\pi(\tilde{F}^n, \tilde{Q}_n) < \delta \Rightarrow \pi(\alpha_{F^n}(T_n), \alpha_{Q_n}(T_n)) < \varepsilon).$$

Megjegyzés. Ez a definíció megenged némi összefüggést a megfigyelt n-esek között és megengedi a valódi eloszlás kis megváltozását is. Nem követeljük meg,

hogy $\tilde{Q}_n=\tilde{G}^n$ legyen valamely $G\in\mathcal{F}$ eloszlásra, lehet egy tetszőleges ($\tilde{F}^n$-hez elég közeli) mérték $\tilde{X}^n$-on.

A π-robusztus becslések egyben robusztusak is, az állítás megfordítása nem igaz [6].

A robusztus becslésekre vonatkozó tételek érvényesek maradnak, ha „robusztus" helyett „π-robusztus"-t írunk.

A katasztrófapont

Most definiálunk egy mennyiséget, amely azt méri, hogy meddig terjed egy becslés robusztussága.

3.5. DEFINÍCIÓ. A $\delta^*((T_n), F)=\sup\{\delta\leqq 1:$ létezik a paramétertérnek olyan K kompakt részhalmaza, melyre $\pi(F, G)<\delta\Rightarrow\lim_n G(T_n\in K)=1\}$ mennyiség a (T_n) becsléssorozat *katasztrófapontja* az F eloszlás esetén.

Az így meghatározott katasztrófapont azt méri, hogy milyen messze lehet a valódi eloszlás a paraméteres modelltől (hány százalék nagy hiba fordulhat elő), hogy a becslés „ne menjen teljesen tönkre".

Az 5. fejezetben megadjuk néhány becslés katasztrófapontját.

Becslések deriváltjai

3.6. DEFINÍCIÓ. Legyen $X=\mathbf{R}$, $\mathcal{E}\subset\mathcal{F}$ csillagszerű halmaz F_0 körül. A $T:\mathcal{F}\to\mathbf{R}^k$ funkcionál *Volterra-értelemben differenciálható* F_0-ban $\mathcal{E}$-re vonatkozóan és deriváltja $T'(F_0; x)$, ha

$$\text{(1)}\qquad \frac{d}{d\varepsilon}\big(T(F_0+\varepsilon(G-F_0))\big) \quad \text{létezik, ha}$$
$$0\leqq\varepsilon\leqq 1 \quad \text{és} \quad G\in\mathcal{E},$$

$$\text{(2)}\qquad \forall G\in\mathcal{E}:\ \frac{d}{d\varepsilon}\big(T(F_0+\varepsilon(G-F_0))\big)\Big|_{\varepsilon=0}=$$
$$=\int T'(F_0; x)(G-F_0)(dx).$$

3.7. DEFINÍCIÓ. Legyen $X=\mathbf{R}$. A $T:\mathcal{F}\to\mathbf{R}^k$ funkcionál *Fréchet-értelemben differenciálható* F_0-ban és deriváltja $T'(F_0; x)$, ha $\forall G\in\mathcal{F}$:

$$T(G)-T(F_0)=\int T'(F_0; x)(G-F_0)(dx)+o(\|G-F_0\|_x),$$

ahol
$$\|G\|_x=\sup_{x\in\mathbf{R}}|G(x)-G(-\infty)|.$$

3.8. DEFINÍCIÓ. Legyen X tetszőleges szeparábilis teljes metrikus tér, $T:\mathcal{F}\to\mathbf{R}^k$. Ha a következő határérték minden $x\in X$ esetén létezik, akkor az alábbi $IC_{T,F}:X\to\mathbf{R}^k$ függvényt nevezzük a T becslés *hatásgörbéjének* (*„influence curve"*).

$$IC_{T,F}(x)=\lim\frac{1}{\varepsilon}(T((1-\varepsilon)F+\varepsilon\delta_x)-T(F)),$$

ahol δ_x az x pontra koncentrált mértéket jelöli.

Az IC a T funkcionál *von Mises-deriváltja,* ld. például [18].

A hatásgörbe interpretálható úgy, mint az x pontban levő infinitezimális tömeg hatása.

Ha egy becslés (funkcionál) *Fréchet-* vagy *Volterra-értelemben differenciálható* és az $\mathscr{E}$ halmaz tartalmazza az egy pontra koncentrált mértékeket, akkor az IC létezik és megegyezik a deriválttal. (Az egyértelműség érdekében legyen a derivált integrálja F_0 szerint nulla.)

3.9. DEFINÍCIÓ. Tegyük fel, hogy $IC_{T,F}$ létezik. A T becslés *érzékenysége („sensitivity, gross-error-sensitivity") F*-ben legyen

$$\sigma^* = \sigma^*(T, F) = (\sup |IC_{T,F}(x)|) \in \overline{\mathbf{R}}^k$$

(a koordináták mindegyike lehet $+\infty$), ahol

$$(\sup |z|) = (\sup |z_1|, \sup |z_2|, \ldots, \sup |z_k|).$$

Szemléletesen azt mondhatjuk, hogy az érzékenység annak a mértéke, hogy egy lokális zavaró hatás mennyire befolyásolhatja a becslés értékét.

A *G-modell*el való kapcsolat jól látszik. Feltételezzük, hogy a valódi eloszlás $(1-\varepsilon)F_0+\varepsilon G$ alakú, ahol F_0 a modellhez tartozik és G tetszőleges. Itt $\varepsilon\sigma^*(T, F_0)$ az ε zavaró tényezőhöz tartozó maximális hatás (közelítő értéke).

A *P-modell*el[5] való kapcsolat kevésbé szembetűnő. Feltételezzük, hogy a valódi G eloszlás kielégíti a $\pi(F_0, G) \leqq \varepsilon$ feltételt, ahol F_0 a modellnek megfelelő eloszlás. A *P*-modell megengedi, hogy a teljes tömeg egy ε-környezetben helyezkedjék el, eltekintve egy ε nagyságútól, amely teljes szabadsággal rendelkezik. A legtöbb esetben ez az utóbbi „szabad" tömeg jelent veszélyt a becslésre nézve. (A lokális szóródásnak kisebb szerepe van. Ennek hatását a *lokális-változás-érzékenység*gel *(„local-shift-sensitivity")* mérhetjük. Tegyük fel, hogy IC eleget tesz a *Lipschitz-feltételnek,* és legyen az érzékenység $\lambda^* = \sup \{|IC(x)-IC(y)|\,|x-y|^{-1}: x \neq y\}$, a legkisebb lehetséges *Lipschitz-konstans.*)

A lokális szóródást kiküszöbölhetjük a *P*-modellben a következő módon. Az X tér metrikáját megszorozzuk egy nagy konstanssal és tartunk vele a végtelenhez. Ez megváltoztatja a *Prohorov-távolságot.* Az eljárás során a *Prohorov-távolság* nem csökken és felülről korlátos, így egy határértékhez tart, mellyel egy új metrikát definiálhatunk az eloszlások terén. A 2.3. tétel szerint ez az új metrika a teljes variáció. Az így kapott teljes variációs modell (T-modell) a *P*-modell határesete. Megmutatható, hogy a T-modellben $|F_0, H| \leqq \varepsilon \Rightarrow |T(H)-T(F_0)|$ maximuma körülbelül $2\varepsilon\sigma^*(T, F_0)$.

Miért jó a definíció?

Felmerül a kérdés, hogy az előzőekben adott definíció mennyiben szolgálja az 1. fejezetben kitűzött céljainkat. Vizsgáljuk meg az ott megfogalmazott zavaró tényezők és a *Prohorov-távolság* kapcsolatát!

(1) *nagy hibák*

Legyen $x=(x_1, \ldots, x_n)$ a pontos érték, melyet meg kívánunk mérni és $\tilde{x}=(\tilde{x}_1, \ldots, \tilde{x}_n)$ a mért érték. Tegyük fel, hogy minden mérésünk bizonyos p_i való-

[5] A *Prohorov-távolságon* alapuló *Hampel-féle modell.*

színűséggel tartalmaz nagy hibát, és legyen $\varepsilon = \max p_i$. Ha most $x_i \in H \subset \mathbf{R}^k$, akkor legalább $1-\varepsilon$ valószínűséggel $\tilde{x}_i \in H$ és legfeljebb ε valószínűséggel $\tilde{x}_i \notin H$. Tehát azt kapjuk, hogy

$$F(H) \leqq \varepsilon \tilde{F}(\mathbf{R}^k \setminus H) + (1-\varepsilon)\tilde{F}(H) \leqq \varepsilon + \tilde{F}(H).$$

(F a valódi, $\tilde{F}$ a tapasztalati eloszlás).

(2) *kerekítési hibák*

Legyen $\tilde{x}$ és x a fenti. Tegyük fel, hogy az i-edik mérésnél a hiba δ_i, azaz $x_i = \tilde{x}_i + \delta_i$, és legyen $\varepsilon = \max |\delta_i|$. Tekintsük az $\mathbf{R}^k$ egy H részhalmazát, ha $x_i \in H$, akkor $\tilde{x}_i \in H^\varepsilon$. Így az $F(H) \leqq \tilde{F}(H^\varepsilon)$ összefüggést kapjuk.

Ha az (1) és (2) zavaró hatást együttesen alkalmazzuk, akkor éppen a *Prohorov-távolság* definíciójához jutunk.

(3) a modell csak *közelítőleg* érvényes.

A *Prohorov-távolság* szerinti konvergencia megegyezik a mértékek gyenge konvergenciájával, ez pedig az eloszlásfüggvények pontonkénti konvergenciájával (a folytonossági pontokban).

Mint korábban már megjegyeztük, az egyes becslések esetében azt fogjuk vizsgálni, hogy a becslést meghatározó funkcionál folytonos-e. A folytonosság implikálja a π-robusztusságot is, így a folytonos funkcionállal meghatározott becslések megengedik, hogy az azonos és független eloszlás feltételezése is csak közelítőleg teljesüljön.

4. Beran definíciója

A következőkben ismertetett felépítés a robusztusság egy más definícióját adja. BERAN a *Hellinger-metrikára* való differenciálhatóságot választotta feltételnek.

Legyenek $x_1, x_2, \ldots$ független, azonos eloszlású, valós értékű valószínűségi változók. Sűrűségfüggvényük f, melyet eltolástól eltekintve ismerünk, azaz a feladat: megkeresni azt a θ értéket, melyhez az $f_0 = f$ sűrűségfüggvény tartozik az $\{f_0 : f_0(x-\theta) = f_0(x)\}$ paraméteres modellből.

Legyen $\mathcal{F}$ az $\mathbf{R}$-en értelmezett sűrűségfüggvények halmaza. Bevezetjük a következő jelölést: ha $r : \mathbf{R} \to \mathbf{R}$ függvény és $h \in \mathbf{R}$, akkor legyen

$$r * h : \mathbf{R} \to \mathbf{R}, \quad (r * h)(x) = r(x-h).$$

4.1. DEFINÍCIÓ. A $T : \mathcal{F} \to \mathbf{R}$ funkcionál *lokációs funkcionál,* ha

$$\forall g \in \mathcal{F}, \quad h \in \mathbf{R} : \; T(g * h) = T(g) + h.$$

Tekintsük $\mathcal{F}$-et, mint az $L^2(\mathbf{R})$ *Hilbert-tér* részhalmazát, jelölje $\langle .,. \rangle$ és $\| . \|$ a *Hilbert-tér* szokásos műveleteit.

4.2. DEFINÍCIÓ. A $T : \mathcal{F} \to \mathbf{R}$ funkcionál *differenciálható* (a *Hellinger-metrikára* vonatkozóan) g-ben és deriváltja ϱ_g, ha $\varrho_g \in L^2$ és

$$\lim_{n \to +\infty} \| g_n^{1/2} - g^{1/2} \|^{-1} \left(T(g_n) - T(g) - \langle \varrho_g, \, g_n^{1/2} - g^{1/2} \rangle \right) = 0,$$

minden $(g_n) \subset \mathcal{F}$, $g_n \xrightarrow{H} g$ (*Hellinger-metrikában* konvergens) sorozatra.

Megjegyzés. A ϱ_g definíciója nem egyértelmű: ϱ_g helyett $\varrho_g+\alpha g^{1/2}$, $\alpha\in\mathbf{R}$ is megfelel. A definíciót a $\langle\varrho_g, g^{1/2}\rangle=0$ feltétellel tesszük egyértelművé.

Megmutatható, hogy a differenciálható funkcionálok folytonosak is a *Hellinger-metrikára* vonatkozóan.

Bontsuk fel $g_n^{1/2}$-et az L^2 *Hilbert-térben* $g^{1/2}$-del egyirányú és rá ortogonális komponensekre (ez a *Riesz-féle felbontási tétel* szerint egyértelműen elvégezhető), legyen

$$g_n^{1/2} = \cos(\theta_n)g^{1/2}+\sin(\theta_n)\delta_n, \tag{4.1}$$

$$\|\delta_n\| = 1, \quad \langle g^{1/2}, \delta_n\rangle = 0, \quad \cos(\theta_n) = \langle g^{1/2}, g_n^{1/2}\rangle,$$

$$\theta_n\in[0, \pi/2].$$

A differenciálhatóságot ebben a formában a következőképpen fogalmazhatjuk meg:

4.1. TÉTEL. A $T:\mathscr{F}\to\mathbf{R}$ funkcionál pontosan akkor differenciálható g-ben, ha létezik $\varrho_g\in L^2$:

$$\forall\{\theta_n:\ \theta_n\to 0, \quad \theta_n\in[0, \pi/2]\},$$

$$\forall\{\delta_n:\ \|\delta_n\| = 1, \quad \langle g^{1/2}, \delta_n\rangle = 0\}:$$

$$\lim_n \theta_n^{-1}(T(g_n)-T(g)-\theta_n\langle\varrho_g, \delta_n\rangle) = 0.$$

4.3. DEFINÍCIÓ. A $T:\mathscr{F}\to\mathbf{R}$ lokális funkcionál *Beran-értelemben robusztus* g-ben, ha differenciálható g-ben.

4.2. TÉTEL. Ha T differenciálható g-ben, akkor

(1) differenciálható a $\{g*h:\ h\in\mathbf{R}\}$ halmazon és $\varrho_{g*h}=\varrho_g*h$,
(2) ha g abszolút folytonos és a majdnem mindenütt létező deriváltjára $g'g^{-1/2}\in L^2$ áll, akkor $\langle\varrho_g, -g'/2g^{1/2}\rangle=1$.

Optimalizálási kérdések megoldásánál felmerül a következő probléma: ismerünk egy T_0 differenciálható funkcionált és létre akarunk hozni egy olyan T differenciálható funkcionált, amely egy adott $f\in\mathscr{F}$ helyen adott ϱ deriválttal rendelkezik.

A feladat megoldásához nyújt segítséget a

4.3. TÉTEL. Ha a T_0 funkcionál differenciálható és a $\varrho\in L^2$ függvény ortogonális $f^{1/2}$-re, abszolút folytonos és $\varrho'\in L^2$, akkor a

$$T(g) = T_0(g)+\int\varrho(x)g^{1/2}(x+T_0(g)-T_0(f))dx$$

funkcionál differenciálható és az f-beli deriváltja az adott ϱ függvény.

A következő tétel biztosítja, hogy ilyen kiindulási T_0 becslés található a *Huber-féle M*-becslések között. (Ezekkel részletesen az 5. fejezetben foglalkozunk.)

4.4. TÉTEL. Legyen $T_0(g)$ az $\int \psi(x-T_0(g))g(x)\lambda(dx)=0$ egyenlet megoldása. Ha ψ szigorúan monoton növő, korlátos függvény, $\lim_{+\infty}\psi>0$, $\lim_{-\infty}\psi<0$ és ψ' folytonos, korlátos függvény, akkor T_0 differenciálható $\mathscr{F}$-en és deriváltja:

$$\varrho_{0,g}(x)=\frac{2\psi(x-T_0(g))g^{1/2}(x)}{\int\psi'(x-T_0(g))g(x)\lambda(dx)}.$$

Ha még ψ kétszer differenciálható és ψ'' korlátos, akkor

$$T_0(g_n)=T_0(g)+\theta_n\langle\varrho_{0,g},\delta_n\rangle+O(\theta_n^2),$$

ahol θ_n és δ_n a $g_n^{1/2}$ (4.1) felbontásából származik.

A tétel feltételeit kielégíti $\psi=\operatorname{arc\,tan}$.

Eddig csak sűrűségfüggvényeken (azaz abszolút folytonos eloszlásokon) értelmeztük a funkcionálokat. Ha egy mintából kell a lokációs paramétert becsülnünk, akkor csak empirikus eloszlások állnak rendelkezésünkre, ezek nem abszolút folytonosak. Ez esetben a $T(g)$ értékét becsülni kell a mintából, $T(g)$ becslése $T(\hat{g}_n)$ lesz, ahol

$$\hat{g}_n(x)=\frac{1}{nc_n}\sum_{i=1}^{n}w\left(\frac{x-x_i}{c_n}\right).$$

Itt (c_n) szigorúan pozitív zérussorozat $(c_n>0, c_n\to 0)$ és $w\in\mathscr{F}$ megfelelően sima függvény, $\int w\,d\lambda=1$, $\{x_i: i=1,2,\ldots,n\}$ a minta.

4.5. TÉTEL. Ha

(1) w abszolút folytonos és w' integrálható,
(2) g folytonos,
(3) $c_n\to 0$ és $c_n n^{1/2}\to+\infty$ és
(4) T folytonos a *Hellinger-metrikában,*

akkor $T(\hat{g}_n)\to T(g)$ majdnem mindenütt.

4.6. TÉTEL. Ha

(1) $w\in L^2$, 0-ra szimmetrikus és $\int x^2w(x)\lambda(dx)<+\infty$,
(2) w abszolút folytonos és w' integrálható,
(3) g abszolút folytonos, g' abszolút folytonos és g'' korlátos,
(4) g tartója egy I véges intervallum és ezen teljesül, hogy $g\geqq\delta>0$,
(5) $c_n n^{1/2}\to+\infty$, $c_n^2 n^{1/2}\to 0$,
(6) van olyan $\varrho_g\in L^2$ függvény, hogy

$$T(g_n)=T(g)+\langle\varrho_g, g_n^{1/2}-g^{1/2}\rangle+o(\|g_n^{1/2}-g^{1/2}\|^2),$$

ha g_n a g egy elég szűk környezetében van és
(7) ϱ_g folytonos I-n és eltűnik I-n kívül,
akkor

$$n^{1/2}\big(T(\hat{g}_n)-T(g)\big)\text{ határeloszlása}$$

$$N(0,\|\varrho_g\|^2/4).$$

A 4.3., 4.4. és 4.6. tételek szerint: ha ismerünk egy alkalmas differenciálható T_0 funkcionált, akkor tudunk olyan T funkcionált konstruálni, amely robusztus és $n^{1/2}(T(\hat{g}_n)-T(g))$ aszimptotikus szórásnégyzete tetszőlegesen kicsi (feltéve, hogy a 4.6. tétel (3), (4) feltételeit teljesíti a g sűrűségfüggvény).

A $T(\hat{g}_n)$ becslés helyett tekinthetünk tetszőleges olyan $\hat{T}_0(x_1, \ldots, x_n)$ becslést, melyre

$$n^{1/2}(T_0(\hat{g})-\hat{T}_0(x_1, \ldots, x_n)) \to 0 \text{ m. m.}$$

Miután a *Beran-féle elmélet* főbb vonalait ismertettük, természetesen adódó feladatunk lenne a két elmélet kapcsolatának megvilágítása.

A *Prohorov-folytonosság*, a *Hellinger-differenciálhatóság* definícióját és a 2.6. tételt figyelembe véve azt láthatjuk, hogy általános esetben egyik értelemben vett robusztusság sem implikálja a másikat. A részletes analízis (melyek azok a becslések, amelyek csak az egyik értelemben robusztusak és melyek robusztusak mindkét értelemben, a különböző eloszlások hogyan befolyásolják ezt a kapcsolatot?) még hátra van és ez további vizsgálódásokat kíván.

A következő megállapítások bizonyos részeredmények lehetnek:

1. A *Beran-féle robusztusság* jól működik a G-modell esetén: legyen a sűrűségfüggvény $f_\varepsilon=(1-\varepsilon)f+\varepsilon g$, megmutatjuk, hogy ha $\varepsilon \to 0$, akkor $d_H(f_\varepsilon, f) \to 0$. Tudjuk, hogy

$$\|f_\varepsilon^{1/2}-f^{1/2}\| = \left(2-2\int \sqrt{(1-\varepsilon)f^2+\varepsilon fg}\right)^{1/2}.$$

Ha most az $\{f_\varepsilon\}$ halmaz L_1-korlátos, akkor *Lebesgue tétele* szerint

$$\int \sqrt{(1-\varepsilon)f^2+\varepsilon fg} \to \int f,$$

hiszen pontonként konvergál.

Az L_1-korlátosság teljesül, mert

$$0 \leqq \sqrt{(1-\varepsilon)f^2+\varepsilon fg} \leqq \sqrt{1-\varepsilon}f+\sqrt{\varepsilon}\sqrt{fg} \leqq f+\sqrt{fg},$$

hiszen $0 \leqq \varepsilon \leqq 1$, és ez a függvény L_1-beli, mert f és g L_1-beli és alkalmazható a *Hölder-egyenlőtlenség*.

2. Az átlag nem robusztus; a medián robusztus, ha a sűrűségfüggvény pozitív és folytonos a mediánjában.

5. Egydimenziós lokációs feladatok

Ebben a fejezetben régi, jól ismert és új, még minden részletre kiterjedően nem vizsgált, becslésekről lesz szó. Foglalkozni fogunk a robusztussággal és ezt több szempontból is megvizsgáljuk. Bár említünk tételeket is, nem a kvalitatív definíciók vizsgálata a fejezet célja. Ez a munka sok esetben igen hosszadalmas lenne és nem biztos, hogy az eredmény arányban állna a befektetett munkával. Ezen túlmenően a gyakorlati alkalmazhatóság szempontjából a kvalitatív robusztusság helyett a becslések kvantitatív jellemzőit érdemes vizsgálni. Az analitikus vizsgálatok mellett (vagy helyett) jól megtervezett *Monte-Carlo-vizsgálatok* eredményeit célszerű figyelembe venni a becslések összehasonlításánál. Ez azért is fontos, mert a véges mintákra nem mindig alkalmazhatók az aszimptotikus eredmények.

A vizsgált becsléseket három csoportba soroljuk:

1. *Kvantilisek súlyozott átlaga.* Ezek a becslések $T(F)=\int_0^1 F^{-1}(t)K(dt)$ alakban írhatók fel, ahol K tetszőleges 1/2-re szimmetrikus eloszlás a (0, 1) intervallumon. Ide tartoznak: a minta átlag, a minta medián, a levágott átlagok és egyéb becslések. Szemléletesen látható, hogy robusztusságuk attól függ, milyen mértékben csökken K sűrűsége az intervallum végein.

2. *R-becslések.* A becslés az

$$\int J(F(x)-F(2T-x))F(dx) = 0$$

egyenlet T megoldása. Itt J korlátos függvény, és ha emellett szigorúan monoton is akkor a megoldás egyértelműen létezik.

3. *M-becslések.* Ez a becslés az

$$\int \psi\left(\frac{x-T}{S}\right)F(dx) = 0$$

egyenlet T megoldása, ahol ψ egy valós függvény és S a szórásnégyzet becslése. S meghatározására szokás az

$$\int \chi\left(\frac{x-T}{S}\right)F(dx) = 0$$

egyenletet hozzávenni az eredeti egyenlethez, vagy az

$$S = \text{medián } |x_i - \text{medián } (x_i)|/0{,}6745$$

becslést tekinteni.

A *Monte-Carlo-eredmények* forrásai: [1] és [16]. Az előbbiben 68 becslést vizsgáltak 14 különböző eloszlásra, melyek túlnyomórészt szimmetrikusak voltak, a mintanagyság $n=5$, 10, 20 és 40 volt. Az utóbbiak 19 becslést vizsgáltak 19 eloszlásra és különböző elsőrendű autoregresszív folyamatokra, a mintanagyság minden esetben $n=20$ volt. Az eredmények mindkét esetben 1000 ismétlés átlagaként adódnak. Ez a két vizsgálat jól kiegészíti egymást és lehetőséget ad a becslések különböző helyzetekben való összehasonlítására. A tanulmányozott eloszlások felsorolása és jelölése az 1. táblázatban található.

1. TÁBLÁZAT

Jel	Eloszlás	Mintanagyság
N	standard normális	5, 10, 20, 40
C	*Cauchy-eloszlás,* medián = 0	5, 10, 20, 40
$\alpha\%\ 3N$	$(100-\alpha)\%\ N(0,1)+\alpha\%\ N(0,9)$	
	$\alpha = 5, 25$	20
	$\alpha = 10$	10, 20
$\alpha\%\ 1/U$ [6]	$(100-\alpha)\%\ N(0,1)+\alpha\%\ N(0,1)/U(-1,1)$	
	$\alpha = 25$	5, 10, 20, 40
	$\alpha = 10$	20
$\alpha\%\ 10N$	$(100-\alpha)\%\ N(0,1)+\alpha\%\ N(0,100)$	
	$\alpha = 5, 10, 15$	20

[6] $U(a, b)$ az (a, b) intervallumon egyenletes eloszlást jelöli.

A következő három tulajdonság teljesülését meg fogjuk vizsgálni az egyes becslések esetében:

(A) ha X sztochasztikusan nagyobb Y-nál, akkor

$$T(F_X) \geqq T(F_Y),$$

(B) ha $F^*(x) = F\big((x-b)/a\big)$, $a > 0$, akkor

$$T(F^*) = aT(F)+b,$$

(C) ha $F^*(x) = F(-x)$, akkor $T(F^*) = -T(F)$.

Könnyen látható, hogy a (B) és (C) feltételeket kielégítő funkcionálok a szimmetrikus eloszlásokhoz a centrumukat rendelik.

Bár az (A) feltétel igen szemléletes, a lokációs funkcionálok egy része nem teljesíti, mert a levágásos módszerek (amelyek egy bizonyos ponton túl nem veszik figyelembe az értékeket (rejection point)) megfordíthatják a rendezést.

Példa. Legyen $G(t)=t$, ha $0<t<1$ és

$$F(t) = \begin{cases} t, & \text{ha} \quad 0 < t < t_0, \\ t_0+\dfrac{B-t_0}{A-t_0}(t-t_0), & \text{ha} \quad t_0 < t \leqq t_1, \\ 1, & \text{ha} \quad t_1 < t, \end{cases}$$

ahol $$t_0 < A < B < 1, \quad t_1 = t_0+(A-t_0)(1-t_0)/(B-t_0).$$

Könnyen ellenőrizhető, hogy $G(t)<F(t)$, ha $t_0<t<1$, viszont $F^*(t)<G^*(t)$, ha $0<t<A$, ahol F^* és G^* a feltételes eloszlásfüggvények, feltéve, hogy a változó $<A$.

Ezek alapján az (A) feltételt nem fogjuk megkövetelni a becsléseinktől, hiszen ésszerűnek látszik, hogy a centrum becslése figyelmen kívül hagyja a nagyon távoli értékeket.

Kvantilisek súlyozott átlaga

Az F eloszlás p-kvantilise az $F^{-1}(p)$. Egyszerűen látható a következő

5.1. ÁLLÍTÁS. A $T_a(F)=\big(F^{-1}(a)+F^{-1}(1-a)\big)/2$, $a\in(0, 1)$ funkcionál kielégíti az (A)—(C) feltételeket, továbbá ilyen funkcionálok konvex lineáris kombinációja is megfelel ezeknek a feltételeknek.

A kvantilisek súlyozott átlaga

$$T(F) = \int_0^1 F^{-1}(t)K(dt)$$

alakban írható, ahol K tetszőleges 1/2-re szimmetrikus eloszlásfüggvény (0, 1)-en.

Fontos speciális esetek

M: *minta átlag*
K az $U(0, 1)$ eloszlásfüggvénye,

$$M(F) = \int_0^1 F^{-1} = \int_R xF(dx)$$

MED: *minta medián*
K az 1/2 pontra koncentrált eloszlásfüggvény

$$MED(F) = F^{-1}(0{,}5)$$

Mα: *α-levágott átlag*
A rendezett minta legnagyobb és legkisebb $\alpha \cdot n$ elemét hagyjuk el; ha $\alpha \cdot n$ nem egész szám, akkor a szélső elemeket súlyozzuk. A gyakorlati megvalósításnál célszerű $\alpha \cdot n$ értékét a legközelebbi egészre kerekíteni. (Vö. 5.2. tétel.)

$$M_\alpha(F) = \frac{1}{1-2\alpha} \int_\alpha^{1-\alpha} F^{-1}$$

TRI: *a Tukey-féle „trimean"*
K az 1/4, 1/2, 3/4 pontokra rendre 1/4, 1/2, 1/4 súlyokat helyező eloszlás

$$TRI(F) = 0{,}5F^{-1}(0{,}5) + 0{,}25\big(F^{-1}(0{,}25) + F^{-1}(0{,}75)\big).$$

Ebbe a csoportba tartoznak az *alkalmazkodó* (adaptive) *levágott átlagok:* ezeknél α értékét az eloszlásból határozzák meg, úgy hogy minimalizáljon egy veszteségfüggvényt. A BICKEL által módosított *Jaeckel-féle eljárás* (*BIC*) a

$$B(F, \alpha\Big(\int_\alpha^{1-\alpha} (\tilde{F}^{-1}(t))^2 dt + 2\alpha(\tilde{F}^{-1}(\alpha))^2\Big)(1-2\alpha)^{-2}$$

kifejezést minimalizálja. Itt $\tilde{F}$ az F szimmetrizáltja:

$$\tilde{F}(x) = 0{,}5\big(F(x - F^{-1}(0{,}5)) + 1 - F(F^{-1}(0{,}5) - x)\big).$$

5.1. TÉTEL. Ha a $T(F) = \int_0^1 F^{-1}(t) K(dt)$ funkcionál robusztus és konzisztens (gyengén folytonos), akkor van olyan *pozitív* α szám, hogy a $(0, \alpha)$ és az $(1-\alpha, 1)$ intervallum *K*-mértéke nulla.

Ha *K* folytonos és van olyan pozitív α szám, hogy a $(0, \alpha)$ és az $(1-\alpha, 1)$ intervallum *K*-mértéke nulla, akkor a *T* funkcionál gyengén folytonos [3].

Következmények.

1. a minta átlag nem robusztus,
2. az α-levágott átlag robusztus és konzisztens, ha $0 < \alpha < 0{,}5$.

Belátható, hogy *MED* és *TRI* akkor robusztusak, ha a megfelelő kvantilisek egyértelműen léteznek.

STIEGLER [13] mutatta meg a következő tételt:

5.2. TÉTEL. A levágott átlag pontosan akkor aszimptotikusan normális, ha az α és $1-\alpha$ minta kvantilisek egyértelműen léteznek.

STIEGLER egyúttal javasolt is olyan eljárást, amely biztosítja a kvantilisek egyértelmű létezését. Az $U(\alpha, 1-\alpha)$ eloszlásfüggvénye helyett a

$$K(u) = \begin{cases} (u-\alpha/2)2h/\alpha, & \text{ha} \quad \alpha/2 \leqq u \leqq \alpha, \\ h, & \text{ha} \quad \alpha \leqq u \leqq 1-\alpha, \\ (1-\alpha/2-u)2h/\alpha, & \text{ha} \quad 1-\alpha \leqq u \leqq 1-\alpha/2, \\ 0, & \text{egyébként} \end{cases}$$

súlyfüggvény használatát javasolta $(h=2/(2-3\alpha))$. Ebben az esetben a becslés aszimptotikusan normális.

5.3. TÉTEL. Az α-levágott átlag hatékonysága (az átlagra vonatkoztatva) legalább $(1-2\alpha)^2$. A 2. táblázatban láthatók a standard normális és az $(1-\varepsilon)N(0, 1)+$ $+\varepsilon N(0, 9)$ eloszlásokra vonatkozó alsó korlátok.

2. TÁBLÁZAT

Alsó korlátok a levágott átlagok hatékonyságára

α	N	5% $3N$	25% $3N$	$(1-2\alpha)^2$
0,05	0,971	1,186	1,402	0,81
0,1	0,943	1,197	1,622	0,64
0,15	0,909	1,197	1,786	0,49
0,25	0,833	1,085	1,667	0,25
0,5	0,637	0,833	1,327	0

Aszimptotikus és Monte-Carlo-eredmények

A levágott átlagok hatásgörbéje szimmetrikus eloszlásokra:

$$IC(x) = \begin{cases} (1-2\alpha)^{-1}F^{-1}(\alpha), & \text{ha} \quad x < F^{-1}(\alpha), \\ (1-2\alpha)^{-1}x, & \text{ha} \quad F^{-1}(\alpha) \leqq x \leqq F^{-1}(1-\alpha), \\ (1-2\alpha)^{-1}F^{-1}(1-\alpha), & \text{ha} \quad F^{-1}(1-\alpha) < x, \end{cases}$$

ennek megfelelően az aszimptotikus szórásnégyzet:

$$A(F) = (1-2\alpha)^{-2} \int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} x^2 F(dx) + 2\alpha\left(F^{-1}(\alpha)\right)^2.$$

A mediánra

$$IC(x) = \operatorname{sign}\left(x - F^{-1}(0{,}5)\right)/2f\left(F^{-1}(0{,}5)\right)$$

és

$$A(F) = \left(2f(F^{-1}(0{,}5))\right)^{-2}.$$

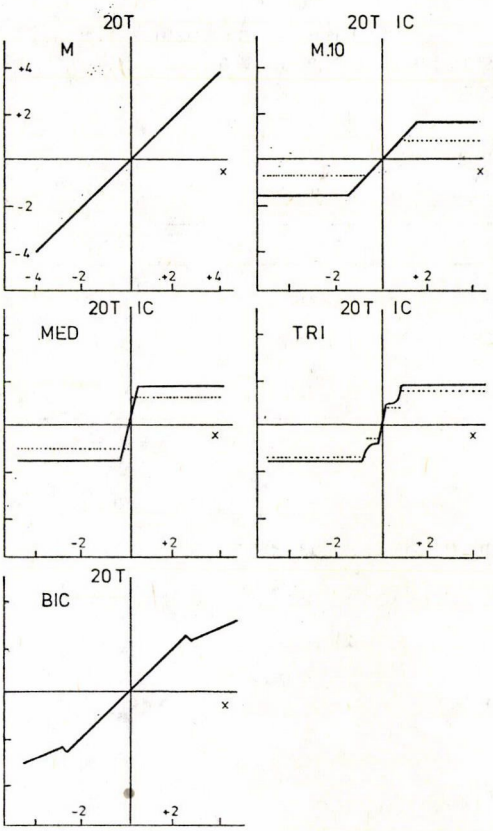
A *Tukey-féle TRI* esetén

$$IC(x) = \begin{cases} (4f(0))^{-1}, & \text{ha } 0 < x < F^{-1}(0{,}75), \\ (4f(0))^{-1} + (4f(F^{-1}(0{,}75)))^{-1}, & \text{ha } F^{-1}(0{,}75) < x, \\ -IC(-x), & \text{ha } x < 0, \end{cases}$$

$$A(F) = ((f(0))^{-2} + ((f(0))^{-1} + (f(F^{-1}(0{,}75)))^{-1})^{-2})/32.$$

A *BIC* becslés *IC*-je és aszimptotikus szórásnégyzete megegyezik a megfelelő levágott átlagéval.

A 2. ábra mutatja a hatásgörbéket és a stilizált érzékenységgörbéket. Ez utóbbiak úgy jöttek létre, hogy 19 standard normális eloszlást követő értékhez az x helyen levő huszadikat vettük hozzá és a $20T$ értéket ábrázoltuk x függvényeként. Az ábrán a hatásgörbéket pontozott vonal jelöli.



2. ábra

R-becslések

Természetesnek látszik, hogy az F eloszlás centrumának azt a F értéket tekintsük, melyre a

$$P(|X-T| \leqq x) = F(x+T) - F(-x+T)$$

valószínűségek (valamilyen átlagos értelemben) a legnagyobbak. Legyen L monoton növő, konvex, korlátos, páros függvény a $[-1, +1]$ intervallumon, $L(0)=0$. Definiáljuk a $T(F)=T$ értéket mint az

$$\int_{\mathbf{R}} (L(F(x+T) - \\ -F(-x+T)) - L(F(x) - F(-x)))\,dx$$

kifejezést maximalizáló T értéket.

Ha F-nek van f sűrűségfüggvénye és L folytonosan differenciálható, $J=L'$ korlátos és szigorúan monoton, és az integrál mögé be lehet differenciálni, akkor a feladat az

$$(5.1) \quad \int_{\mathbf{R}} J(F(x) - F(2T-x))f(x)dx = 0$$

egyenlet megoldása. Az egyenletnek pontosan egy megoldása van. Ha valamennyi fenti feltétel teljesül, akkor a becslés kielégíti az (A)—(C) feltételeket. Definiálhatjuk

$T(F)$-et a következő ekvivalens módon. Legyen $T(F)=MED(0{,}5(F^{-1}(U)+F^{-1}(V)))$, ahol (U, V) az egységnégyzeten

$$Q(u,v)=\int_0^u (L'(v-t)-L'(-t))dt/2L(1) \tag{5.2}$$

eloszlású.

5.4. TÉTEL. Az R-becslések robusztusak minden eloszlásban, melyre egyértelműen vannak definiálva ([3]). Ha F helyett az F_n empirikus eloszlásfüggvényt vizsgáljuk, akkor előfordulhat, hogy az (5.1) egyenletnek nincs megoldása. Azonban ebben az esetben is definiálhatjuk a becslést (5.2) szerint.

Az R-becslések használatánál gondot okozhat, hogy minden R-becsléshez található olyan eloszlás, melyen (az átlagra vonatkoztatott) hatékonysága zérus. Ez általában csak nagyon mesterséges eloszlásoknál fordul elő.

Speciális esetek

MED: *medián* $L(y)=|y|$,

H/L *(Hodges—Lehmann-becslés)* vagy *pszeudo-medián:* az $\int F(2T-x)F(dx)=0{,}5$ egyenlet megoldása. Interpretálható $MED(0{,}5(X_1+X_2))$-ként, ahol X_1 és X_2 független $F(x)$ eloszlású.

RN: *„Normal Scores"*, $L(y)=\Phi^{-1}(0{,}5(y+1))$, ahol Φ a standard normális eloszlásfüggvény.

Ez utóbbi két becslésre a hatékonyság alsó korlátja szimmetrikus eloszlás esetén 0,864, ill. 1. Az *RN* hatékonysága igen nagyra nőhet (az átlaghoz viszonyítva) a szennyezett normális eloszlások esetén.

A *H/L* becslés hatásgörbéje

$$IC(x)=(0{,}5-F(2T(F)-x))/\int f(2T(F)-t)f(t)dt,$$

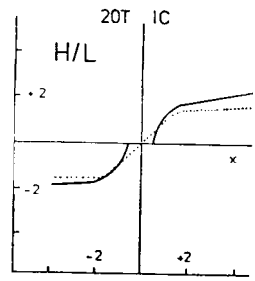
szimmetrikus eloszlás esetén az egyszerűbb

$$IC(x)=(F(x)-0{,}5)/\int f^2$$

alakot ölti. Aszimptotikus szórásnégyzete

$$A(F)=\left(\int f^2\right)^{-2}/12.$$

A 3. ábrán látható a stilizált érzékenységgörbe folytonos vonallal és az *IC* pontozott vonallal.



3. ábra

M-becslések

Legyen $x_1, \ldots, x_n$ $f(x-T)$ sűrűségfüggvénnyel rendelkező eloszlásból vett minta. A likelihood függvény logaritmusa

$$\ln L(T) = \sum_{i=1}^{n} \ln f(x_i - T) = -\sum_{i=1}^{n} \varrho(x_i - T), \qquad \text{ahol} \quad \varrho = -\ln f,$$

$$\frac{d \ln L(T)}{dT} = -\sum \frac{f'(x_i - T)}{f(x_i - T)} = -\sum \psi(x_i - T), \quad \text{ahol} \quad \psi = \varrho'.$$

A standard normális és a *Laplace-eloszlás* esetén a megfelelő ϱ és ψ függvények:

$$\varrho(x) = 0{,}5 \ln(2\pi) + x^2/2, \quad \psi(x) = x,$$

$$\varrho(x) = \ln 2 + |x|, \quad \psi(x) = \operatorname{sgn}(x).$$

A megoldások az átlag és a medián.

Az előbbiek figyelembevételével legyen az M-becslés az

$$\int \psi\left(\frac{x-T}{S}\right) = 0 \tag{5.3}$$

egyenlet T megoldása. Itt S a szórásnégyzet becslése, szokásos meghatározási módjai:

(I) az előbbi egyenlet az

$$\int \chi\left(\frac{x-T}{S}\right) = 0 \quad \text{egyenlettel} \tag{5.4}$$

kiegészítve, a kapott egyenletrendszert oldjuk meg T-re és S-re,

(II) $$S = MED(|x_i - MED(x_i)|),$$

(III) $$S = \frac{F^{-1}(0{,}75) - F^{-1}(0{,}25)}{\Phi^{-1}(0{,}75) - \Phi^{-1}(0{,}25)}, \quad \text{vagy} \quad F \text{ helyett}$$

az $\tilde{F}(x) = 0{,}5\left(F(x - F^{-1}(0{,}5)) + 1 - F(F^{-1}(0{,}5) - x)\right)$

szimmetrizáltját véve (IV).

Speciális esetek

$H_2(k)$: HUBER [10] javasolta az (5.3), (5.4) egyenletrendszer megoldását a

$$\psi(x) = \begin{cases} -k, & x < k \\ x, & -k \leqq x \leqq k \\ k, & k < x \end{cases}$$

$\chi(x) = \psi^2(x) - \int \psi^2(x) \Phi(dx)$ függvényekkel.

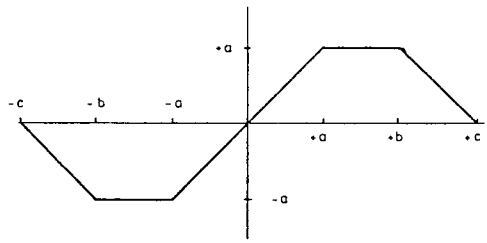
A fenti egyenletrendszer szokásos megoldása a $T=MED$ és $S=S_{III}$ kezdőértékkel való iteráció.

Látható, hogy ez a ψ függvény olyan eloszlásnak felel meg, amely a centrum közelében normális, a farkai viszont *Laplace-eloszlást* követnek.

$A_1(k)$: HAMPEL javaslata, az (5.4) egyenletet elhagyva, S értékét a (III) kifejezéssel becsüli.

$A_2(a, b, c)$: HAMPEL másik javaslata kompakt tartójú, szakaszonként lineáris ψ függvényt használ.

$$\psi(x) = \begin{cases} x, & |x| \leqq a \\ a \operatorname{sign}(x), & a < |x| \leqq b \\ \dfrac{c-|x|}{c-b}\, a, & b < |x| \leqq c \\ 0, & |x| > c \end{cases}$$



4. ábra

Az (5.4) egyenlet helyett a (II) becslést használja.

$ADA(b, c)$: az előbbi becslés, azzal az eltéréssel, hogy a értékét a mintából megbecsüljük. Legyen L a következő halmaz átlaga:

$$\{y = S_{II}|x_i - MED|^{-1}: y < 1\},$$

ezek után legyen

$$a = \begin{cases} 1, & L \leqq 0{,}44 \\ (75L-25)/8, & 0{,}44 < L \leqq 0{,}6 \\ 2{,}5, & 0{,}6 \leqq L \end{cases}$$

AMT: JAECKEL és HAMPEL javaslata alapján ANDREWS dolgozta ki.

$$\psi(x) = \begin{cases} \sin(x/2{,}1), & |x| < 2{,}1\pi \\ 0, & \text{egyébként.} \end{cases}$$

A skála becslésére (III)-at vagy (IV)-et használja.

Φ: $\psi = \Phi - 0{,}5$, a skála becslése S_{II}.

A táblázatokban a következőképpen fogunk az ide tartozó becslésekre hivatkozni:

Hxx, ill. *Axx* a $H_2(k)$, ill. $A_1(k)$ becslés a $k=x, x$ értékre, például *H*07: $H_1(0, 7)$;

$12A$: $A_2(1{,}2;\ 3{,}5;\ 8{,}0)$, $17A$: $A_2(1{,}7;\ 3{,}4;\ 8{,}5)$

$21A$: $A_2(2{,}1;\ 4{,}0;\ 8{,}2)$, $22A$: $A_2(2{,}2;\ 3{,}7;\ 5{,}9)$

$25A$: $A_2(2{,}5;\ 4{,}5;\ 9{,}5)$, ADA: $ADA(4{,}5;\ 8{,}0)$.

Ebbe a csoportba tartoznak még a *Hogg-féle becslések,* amelyek az eloszlás farkaitól függően választanak az *M*-becslések közül. A döntés alapjául a

$Q=(U(0,2)-L(0,2))/(U(0,5)-L(0,5))$ érték szolgál, ahol $U(\alpha)$, ill. $L(\alpha)$ az alsó, ill. felső $[(N+1)\alpha]$ érték átlagát jelöli.

A becslések:

$$HG1 = \begin{cases} 25A, & Q \leqq 1{,}81 \\ 21A, & 1{,}81 < Q \leqq 1{,}87 \\ 12A, & 1{,}87 < Q \end{cases}$$

$$HG2 = \begin{cases} 25A, & Q \leqq 1{,}90 \\ ADA, & 1{,}90 < Q \leqq 2{,}05 \\ 17A, & 2{,}05 < Q \end{cases}$$

$$HG3 = \begin{cases} 25A, & Q \leqq 1{,}95 \\ ADA, & 1{,}95 < Q \leqq 2{,}10 \\ 17A, & 2{,}10 < Q \end{cases}$$

$$HG4 = \begin{cases} 21A, & Q \leqq 2{,}00 \\ 12A, & 2{,}00 < Q \end{cases}$$

5.5. TÉTEL. Ha ψ korlátos és folytonos függvény és a becslést meghatározó egyenletnek egyértelmű megoldása van, akkor a megfelelő becslés konzisztens és robusztus.

A tételt folytonossági meggondolások alapján be lehet látni.

Nem érdemes általános feltételeket alkotni arra, hogy az egyenletnek mikor létezik egyértelmű megoldása — ez az eloszlástól is igen erősen függ.

Aszimptotikus és Monte-Carlo-eredmények

Az általános típusú M-becslések hatásgörbéje, szimmetrikus F eloszlás, páratlan ψ és páros χ esetén:

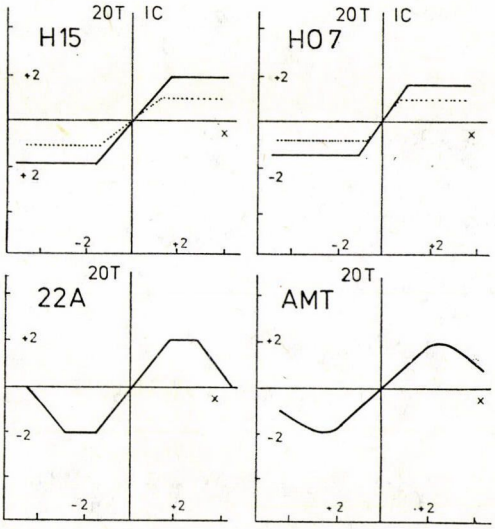
$$IC_T(x) = \psi(x/S(F))S(F)\left(\int \psi'(y/S(F))F(dy)\right)^{-1}$$

$$IC_S(x) = \psi(x/S(F))S(F)\left(\int \chi'(y/S(F))\, y/S(F)F(dy)\right)^{-1}$$

A H_2 becslésekre lényegesen egyszerűbb alakúak. A hatásgörbe megegyezik az $\alpha=F(-kS(F))$ paraméterű levágott átlagéval, az aszimptotikus szórásnégyzet

$$A(F) = \int \psi^2(x)\Phi(dx)S(F)^2(1-2\alpha)^{-2}.$$

Az 5. ábra a stilizált érzékenységgörbéket és a H_2 becslések hatásgörbéjét mutatja.



5. ábra

Összehasonlító eredmények

Monte-Carlo

A *normális eloszlás* esetén a $H20$ becslés hatékonysága gyakorlatilag nem tér el az átlagétól, a veszteség 4% alatt marad. Nagyobb mintáknál a $20A$ is ugyanígy viselkedik.

*Cauchy-eloszlás*ra az átlag gyakorlatilag alkalmazhatatlan, a medián elég jól viselkedik. Hasonlóan, vagy kicsit jobban szerepeltek az A_2 becslések — minél kisebb volt az a paraméter, annál jobban. Valamivel gyengébbek a H_2 becslések $k=1,0$ esetén, nagyobb paraméter értékekre a szórás megnő. Az ADA és AMT becslések az A_2-höz hasonlóan viselkednek.

Szennyezett normális eloszlások esetén az átlag szórása rohamosan nő. Kissé szennyezett eloszlások (5—15% $3N$, 5% $10N$) esetén a levágott átlagok jól szerepelnek, nagyobb szennyezéseknél már csak nagy (25—50) α% esetén megfelelőek. A H_2 becslések 1,5—2,0 paraméter értékeknél hasonlóan viselkednek, mint az előzőek, bár a szórásuk kisebb. Egyenletesen megbízható eredményt a *Hampel-féle becslések* adnak[7].

A H_2 becslések nagyon hasonló eredményt szolgáltatnak, ha az iterációt csak egy lépésre végezzük el. Érdekes megjegyezni, hogy a $T=M$ kezdőérték-választás jó példája annak, hogyan nem érdemes próbálkozni: ez a becslés néha még az átlagnál is rosszabb eredményeket adott. A *Tukey-féle „trimean"* jól szerepelt, szórása

[7] Ismeretes a *Hampel-féle becslések* optimális típusa, nevezetesen az, amelynek ψ függvénye egy bizonyos tangens hiperbolikusz mentén közelíti meg a nullát [9]. Az ilyen becslésekről még nem állnak rendelkezésre *Monte Carlo-eredmények*.

nem volt lényegesen nagyobb, mint az A_2 becsléseké. A *Hogg-típusú becslések* a nagyon erősen szennyezett eloszlásokon valamivel jobbak, mint az A_2 típusúak, egyébként megegyeznek a tulajdonságaik.

A 3. táblázat egyes becslések véges mintákra számított *katasztrófapont*ját mutatja. Az eljárás során $n-j$ standard normális eloszlást követő elemhez j db másikat vettek hozzá, rendre a 100, 200, ..., $100j$ helyen. A táblázatban az szerepel, hogy hány százalék zavaró adattal maradt a becslés értéke 3 alatt. Az utolsó oszlop az aszimptotikus értéket tartalmazza.

Még egy *Monte-Carlo-eredményt* közlünk, ez a becslések eloszlásáról ad némi felvilágosítást. Vizsgáljuk meg az

$$s = \frac{F^{-1}(0{,}01)}{\Phi^{-1}(0{,}01)} \Big/ \frac{F^{-1}(0{,}25)}{\Phi^{-1}(0{,}25)} \quad \text{kifejezést,}$$

a normalitási indexet. Ennek értéke közel van az 1-hez, ha az eloszlás közelítőleg normális. Nagyobb 1-nél, ha az eloszlás hosszabb farkú és kisebb 1-nél, ha rövidebb farkú, mint a normális. Az eredményeket a 4. táblázat tartalmazza.

Aszimptotikus eredmények

[8] táblázata nagyon hasznos a különböző típusú becslések összehasonlítása szempontjából. Az oszlopok jelentése:

σ^2: a becslés aszimptotikus szórásnégyzete,
$\sigma^* = \sup |IC|$: az érzékenység,
δ^*: a katasztrófapont,
λ^*: a lokális érzékenység,
ϱ: a levágási pont *(rejection point)*,
sup *GEM*: az asz. szórásnégyzet szupremuma az 5%-os szimmetrikus nagy hiba modellben,
dif *GEM*: a fenti, ha $F(x)=0{,}95\Phi(x)+0{,}05\Phi(x/k)$, $k \to +\infty$.

A becslések egy kivétellel már szerepeltek, H „*sk*" 2,71: a *Huber-féle becslés*, a $\psi(x)=x$, ha $|x| \leqq 2{,}71$ és 0 egyébként függvénnyel és az S_{II} becsléssel.

Összefoglalva az eredményeket, megállapíthatjuk, hogy az *átlag* használata nagyon rossz becsléseket eredményezhet, ha a minta eloszlása nem pontosan normális. Normális eloszlásból vett és gondosan kezelt adatok esetén (ha a szennyeződés nem nagyobb, mint 15% $3N$ vagy 5% $10N$) elég jó eredményt adnak a *10—25% levágott átlagok*, a *trimean* és a H_2 becslések. Ha a szennyeződés nagyobb, vagy már az eredeti eloszlás is kicsit bizonytalan, akkor a *Hampel-féle becslések* használata látszik célszerűnek.

Az *egylépéses eljárások* (a H_2, A_1 és A_2 becslése meghatározása egylépéses iterációval) gyakorlatilag ugyanúgy viselkednek, mint a megfelelő alapeljárások, ezért célszerű ezeket használni. A *Hogg-típusú (adaptive)* becslések nem különböznek lényegesen az alapul vett becslésektől.

3. TÁBLÁZAT *Katasztrófapontok*

	10	20	40	δ^*
M	0	0	2,5	0
M, 05	0	5	7,5	5
M, 10	10	10	10	10
M, 15	10	15	15	15
M, 25	20	25	25	25
MED	40	45	47,5	50
TRI	20	20	22,5	
BIC	10	20	20	
*H*20	10	15	17,5	
*H*17	20	20	20	
*H*15	20	25	25	26
*H*12	30	30	27,5	
*H*10	30	30	32,5	
*H*07	30	35	37,5	
*A*20	30	35	37,5	50
*A*15	40	40	40	50
25*A*	40	45	47,5	50
22*A*	40	45	47,5	50
21*A*	40	45	47,5	50
17*A*	40	45	47,5	50
12*A*	40	45	47,5	50
ADA	40	45	47,5	50
AMT	40	45	47,5	
H/*L*	30	25	27,5	29

4. TÁBLÁZAT *Normalitási indexek*, $n = 20$

	N	10% 3*N*	25% 1/*U*	10% 10*N*	*C*
M	1,00	0,97	5,82	0,85	15,14
M, 05	1,00	1,00	2,02	1,42	3,33
M, 10	1,00	1,00	1,18	0,98	1,75
M, 15	1,00	1,00	1,07	0,99	1,38
M, 25	1,00	1,01	1,04	0,99	1,29
MED	1,01	1,03	1,05	1,00	1,30
TRI	1,00	1,00	1,07	0,99	1,37
BIC	1,00	0,99	1,27	1,16	1,51
*H*20	1,00	0,99	1,27	0,96	1,65
*H*17	1,00	0,99	1,15	0,97	1,53
*H*15	1,00	0,99	1,11	0,98	1,47
*H*12	1,00	1,00	1,07	0,99	1,41
*H*10	1,00	1,00	1,06	0,99	1,35
*H*07	1,00	1,01	1,04	0,99	1,30
*A*20	1,00	1,00	1,13	0,97	1,38
*A*15	1,00	1,00	1,09	0,98	1,35
25*A*	1,01	1,00	1,09	1,00	1,31
22*A*	1,01	1,01[1]	1,07	1,01	1,34
21*A*	1,00	1,00	1,07	1,00	1,31
17*A*	1,00	1,01	1,05	1,00	1,29
12*A*	1,00	1,02	1,04	1,00	1,26
ADA	1,01	1,01	1,06	1,01	1,32
AMT	1,01	1,00	1,08	1,00	1,36
H/*L*	1,00	1,00	1,08	0,98	1,42

5. TÁBLÁZAT

Egyes kvantitatív tulajdonságok normális eloszlás esetén

Becslés	σ^2	σ^*	δ^*	λ^*	ϱ	sup *GEM*	dif *GEM*
M	1,000	∞	0,00	1,00	∞	∞	∞
RN	1,000	∞	0,24	1,00	∞	1,48	1,48
H/L	1,047	1,77	0,29	1,41	∞	1,29	1,29
Φ	1,047	1,77	0,50	1,41	∞	1,28	1,28
MED	1,571	1,25	0,50	∞	∞	1,74	1,74
M, 05	1,026	1,83	0,05	1,11	∞	1,30	1,30
M, 10	1,060	1,60	0,10	1,25	∞	1,26	1,26
M, 0668	1,037	1,73	0,07	1,15	∞	1,271	1,271
*H*15	1,037	1,73	0,26	1,15	∞	1,264	1,264
*A*15	1,037	1,73	0,50	1,15	∞	1,262	1,262
$H_1(1{,}5)$	1,037	1,73	0,50	1,15	∞	1,258	1,258
25*A*	1,026	1,86	0,50	1,10	6,41	1,35	1,07
$A_1(1{,}686)$	1,024	1,86	0,50	1,10	∞	1,28	1,28
H"sk" 2,71	1,066	2,89	0,50	∞	2,71	∞	1,10
$A_1(2{,}71)$	1,001	2,73	0,50	1,01	∞	1,52	1,52

További kérdések a lokációs feladatokban

Az előzőek alapján látszik, hogy a szennyezett normális eloszlásokra, a normálisnál hosszabb farkú eloszlásokra és ezen belül a független esetre elég jó becslésekkel rendelkezünk.

Újabb vizsgálatok azt mutatják, hogy ezek a becslések nagyon messze vannak az ideálistól a *rövid farkú eloszlások* esetén. Az egyenletes és a szennyezett egyenletes eloszlásokra mind lényegesen rosszabb eredményt ad, mint a *külső átlag* („outer mean", az átlag, a 0,25 és 0,75-kvantilisek közötti rész elhagyása után). Itt megjegyezzük, hogy JEFFREYS eredménye szerint, amelyet az 1. fejezetben már említettünk, az ilyen rövid farkú eloszlások akkor fordulnak elő, ha a hibák korreláltak. WEGMAN és CARROLL megfigyelései szerint az elektronikus adatfeldolgozás tipikusan ilyen hibaeloszlást eredményez.

WEGMAN és CARROLL alkalmazták a robusztus becsléseket *autoregressziós folyamatokra*. A vizsgált folyamatok $x_t = \varrho x_{t-1} + \varepsilon_t$, $t = 1, 2, \ldots, n$ alakúak voltak. A paraméterek: $\varrho = 0{,}2$; 0,5; 0,9, ε_t független *N*, *C*, ill. *U* eloszlású. A normális és egyenletes eloszlás esetén a $\varrho = 0{,}2$ értékre még elfogadhatóan viselkedtek a becslések. Egyébként a torzítás és a szórás nagyon megnőtt.

IRODALOM

[1] ANDREWS, D. F., BICKEL, P. J., HAMPEL, F. R., HUBER, P. J., ROGERS, W. H. and TUKEY, J. W., *Robust Estimates of Location: Survey and Advances* (Princeton University Press, Princeton, 1972).

[2] BERAN, R., „Robust location estimates", *Ann. Statist.* **5** (1977) 431—444.

[3] BICKEL, P. J. and LEHMANN, E. L., „Descriptive statistics for nonparametric models. II. Location.", *Ann Statist.* **3** (1975) 1045—1069.

[4] CSISZÁR, I., „On topological properties of *f*-divergences", *Studia Sci. Math. Hung.* **2** (1967) 329—339.

[5] HAMPEL, F. R., „Contributions to the theory of robust estimation", Ph. D. dissertation, University of California, Berkeley, 1968.
[6] HAMPEL, F. R., „A general qualitative definition of robustness", *Ann. Math. Statist.* **42** (1971) 1887—1896.
[7] HAMPEL, F. R., „Robust estimation: A condensed partial survey", *Z. Wahrscheinlichkeitstheorie werw. Geb.* **27** (1973) 87—104.
[8] HAMPEL, F. R., „The influence curve and its role in robust estimation", *J. Amer. Statist. Assoc.* **69** (1974) 383—393.
[9] HAMPEL, F. R., „Modern trends in the theory of robustness", Research Report No. 13, Fachgruppe fuer Statistik, Eidgenoessische Technische Hochschule, Zürich, 1977.
[10] HUBER, P. J., „Robust estimation of a location parameter", *Ann. Math. Statist.* **35** (1964) 73—101.
[11] HUBER, P. J., „Robust statistics: A review", *Ann. Math. Statist.* **43** (1972) 1041—1067.
[12] JEFFREYS, H., *Theory of Probability* (Clarendon Press, Oxford, 1939, 1948, 1961.).
[13] STIEGLER, S. M., „The asymptotic distribution of the trimmed means", *Ann. Statist.* **1** (1973) 472—477.
[14] STRASSEN, V., „The existence of probability measures with given marginals", *Ann. Math. Statist.* **36** (1965) 423—439.
[15] TUKEY, J. W., „A survey of sampling from contaminated distributions", *Contrib. to Prob. and Statist.* szerk. *Olkin* 448—485. (Stanford University Press, 1960.)
[16] WEGMAN, E. J. and CARROLL, R. J., „A Monte Carlo study of robust estimators of location" *Commun. Statist. — Theor. Meth.* **A6** (1977) 795—812.
[17] Прохоров, Ю. В., «Сходимость случайных процессов и предельные теоремы теории вероятностей» *Теория Вероятностей и её Применения* **1**(1956) 177—238.
[18] Филиппова, А. А., «Теорема Мизеса о предельном поведении функционалов от эмпирических функций распределения и её статистические применения» *Теория Вероятностей и её Применения* **7**(1962) 26—60.

(Beérkezett: 1978. szeptember 4.)

KERÉKFY PÁL
MTA SZÁMÍTÁSTECHNIKAI ÉS AUTOMATIZÁLÁSI KUTATÓ INTÉZET
1132 BUDAPEST, VICTOR HUGO U. 18—22.

ON ROBUST ESTIMATES

P. KERÉKFY

This paper was written to summarize the basic concepts and results of the theory of robust estimation.

TUKEY's and HUBER's important models for contaminated normal distributions are presented. Then HAMPEL's qualitative definition based on *Prohorov distance* is given, influence curve of estimates is defined as a *von-Mises derivative*. BERAN's alternative concept is partially compared with HAMPEL's one by the aid of CSISZÁR's work on *f*-divergences. Interesting *Monte-Carlo results* together with parts of asymptotic theory are given.

Alkalmazott Matematikai Lapok 4 (1978) 359—362

KÉTSZERES VALÓS GYÖKÖKKEL RENDELKEZŐ VALÓS EGYÜTTHATÓS POLINOMOK FAKTORIZÁLÁSA

VARGA GYULA
Budapest

A cikk egy, a *Newton—Raphson iterációs eljáráson* alapuló módszert ad meg valós együtthatós polinomok kétszeres valós gyökeihez tartozó gyöktényezőinek a meghatározására. A módszer alkalmazható valós együtthatós polinomok szélsőértékhelyeinek megkeresésére is.

1. Bevezetés

A *Newton—Raphson-módszer* egyik kétdimenziós változata, a *Bairstow-féle eljárás* [2] sikerrel alkalmazható valós együtthatós polinomok másodfokú tényezőkre bontására abban az esetben, ha a polinomnak nincsenek többszörös gyökei. Használatát gyorsasága és könnyen kezelhető volta teszi célszerűvé. Ha azonban a polinom többszörös gyökökkel is rendelkezik, akkor az eljárás általában nem konvergens. Vannak módszerek tetszés szerinti multiplicitású valós vagy komplex gyökök, ill. a hozzájuk tartozó gyöktényezők meghatározására, de ezek bonyolultabbak, és használatuk hosszadalmasabb, több időt vesz igénybe ([2], [1]). Bizonyos speciális tulajdonságú polinomok faktorizálására egyszerűbb eljárások is készíthetők. Egy ilyen eljárást ismertetünk az alábbiakban olyan polinomok faktorizálására, amelyeknek a valós gyökei legfeljebb kétszeres multiplicitásúak. Az eljárás éppen a kétszeres gyökökhöz tartozó gyöktényezők leválasztására szolgál. A cikkben egyéb alkalmazásai is szóba kerülnek.

2. A módszer ismertetése

Tekintsük az $a(x)$ valós együtthatós polinomot az alábbi előállításban:

$$a(x) \equiv (x-p)^2 b(p, x) + c(p)x + d(p) \tag{2.1}$$

(itt p a polinom egy kétszeres valós gyökének valamely közelítése).

Deriváljuk ezt az azonosságot p szerint:

$$0 \equiv (x-p)^2 \frac{\partial b(p,x)}{\partial p} - 2(x-p)b(p,x) + \frac{\partial c(p)}{\partial p} x + \frac{\partial d(p)}{\partial p}. \tag{2.2}$$

Az $x=p$ értéket véve a

$$\frac{\partial c(p)}{\partial p} p + \frac{\partial d(p)}{\partial p} = 0, \quad \text{azaz a} \quad \frac{\partial d(p)}{\partial p} = -\frac{\partial c(p)}{\partial p} p$$

egyenlőséget kapjuk. Behelyettesítve a (2.2)-be adódik:

$$0 \equiv (x-p)^2 \frac{\partial b(p,x)}{\partial p} - 2(x-p)b(p,x) + \frac{\partial c(p)}{\partial p}(x-p). \tag{2.3}$$

Leosztva $(x-p)$-vel és átrendezve kapjuk:

$$2b(p,x) = (x-p)\frac{\partial b(p,x)}{\partial p} + \frac{\partial c(p)}{\partial p}, \tag{2.4}$$

$x=p$-t véve kapjuk:

$$\frac{\partial c(p)}{\partial p} = 2b(p,p). \tag{2.5}$$

Célunk az, hogy a p-t úgy válasszuk meg, hogy a (2.1) polinomosztás maradéka 0 legyen. Először oldjuk meg a $c(p)=0$ egyenletet. Valamely $p^{(0)}$-ból kiindulva alkalmazzuk a *Newton—Raphson-módszert*. A (2.1)-ből megkaphatjuk a $c(p)$ függvényértékét, a (2.5)-ből pedig a derivált értékét. Így képezhetjük a

$$p^{(i+1)} = p^{(i)} - \left.\frac{c(p)}{\frac{\partial c(p)}{\partial p}}\right|_{p=p^{(i)}} \tag{2.6}$$

iterációt a $c(p)=0$ egyenlet megoldására.

Az iteráció konvergenciájára és az eljárás alkalmazhatóságára vonatkozik a

2.1. TÉTEL: Ha az alábbi két feltétel teljesül:

(i) $$|b(p^{(i)}, p^{(i)})| > K > 0 \quad (i = 0, 1, \ldots)$$

(ii) $$\exists p^*: \left.\frac{\partial a(x)}{\partial x}\right|_{x=p^*} = 0,$$

akkor a (2.6) iteráció konvergál az $a(x)$ polinom egy szélsőérték helyéhez. A konvergencia másodrendű.

Bizonyítás: Írjuk fel a (2.1) előállítást p^* segítségével:

$$a(x) \equiv (x-p^*)^2 b(p^*, x) + c(p^*)x + d(p^*). \tag{2.7}$$

x szerint deriválva adódik:

$$\frac{\partial a(x)}{\partial x} = 2(x-p^*)b(p^*,x) + (x-p^*)^2 \frac{\partial b(p^*,x)}{\partial x} + c(p^*).$$

$x=p^*$-ot behelyettesítve kapjuk:

$$\left.\frac{\partial a(x)}{\partial x}\right|_{x=p^*} = c(p^*) = 0, \tag{2.8}$$

tehát p^* a (2.6) iteráció fixpontja.

Továbbá (2.5) miatt

$$\left|\frac{\partial c(p)}{\partial p}\bigg|_{p=p^*}\right| \geqq 2K > 0,$$

tehát a *Newton—Raphson-módszerre* ismert eredmények alapján a konvergencia másodrendű.

Számítsuk ki $a(x)$ x szerinti második deriváltját:

$$\frac{\partial^2 a(x)}{\partial x^2} = 2b(p^*, x)+4(x-p^*)\frac{\partial b(p^*, x)}{\partial x}+(x-p^*)^2\frac{\partial^2 b(p^*, x)}{\partial x^2}.$$

$x=p^*$-ot behelyettesítve kapjuk:

$$\frac{\partial^2 a(x)}{\partial x^2}\bigg|_{x=p^*} = 2b(p^*, p^*) \neq 0. \tag{2.9}$$

A (2.8)-at és (2.9)-et egybevetve látjuk, hogy az $x=p^*$ az $a(x)$ polinom egy szélsőérték helye. Ezzel a tételt bebizonyítottuk.

KÖVETKEZMÉNY: Behelyettesítve az $x=p^*$-ot a (2.7)-be, kapjuk:

$$a(p^*) = d(p^*).$$

Ha tehát az iterációs eljárás végén azt találjuk, hogy $d(p^*)=0$-nak adódik, akkor

$$a(x) = (x-p^*)^2 b(p^*, x)$$

az $x=p^*$ kétszeres gyökhöz tartozó felbontást adja meg.

3. A módszer alkalmazhatósága

Az előző szakasz végén tett megállapítás a módszer alkalmazásával kapcsolatban óvatosságra int. Az iteráció segítségével kiszámított p^* csak akkor kétszeres gyöke a polinomnak, ha $d(p^*)=0$ adódik, egyébként csak szélsőérték helye. A módszert ez utóbbinak a kiszámítására is felhasználhatjuk. Nem mindegy, honnan indítjuk az iterációt, mert előfordulhat, hogy bár van egy kétszeres gyöke (amely egyben szélsőérték hely is) mégis egy másik szélsőérték helyhez konvergál. A módszer számítási igényét iterációs lépésenként a (2.1) polinomosztás és a (2.5) polinom behelyettesítési érték kiszámításának műveletigénye adja meg. A 2.1. tétel (ii) feltétele szükséges is, mert ha nincs a polinomnak szélső értéke, nyilván nem konvergálhat az iteráció. Pl. az $a(x)=x^3+x$ polinom esetén a felbontás $a(x)\equiv(x-p)^2(x+2p)+$ $+(3p^2+1)x-2p^3$. A $c(p)=3p^2+1=0$ egyenlet valós p-vel nem oldható meg.

Megjegyzések: A javasolt eljárás alkalmazásához megfelelő kezdő érték áll rendelkezésre a *Bairstow-eljárás* sikertelen befejezésekor (a *Jacobi-mátrix* közel szinguláris, így determinánsának abszolút értéke igen kicsiny).

Feladatunk egy polinom faktorizálása, s így polinomosztásra szükség van. A módszer szervezésében ez egyszerűen beépítve szerepel.

A $c(p)=0$ egyenletnek nincs többszörös gyöke, tehát a többszörös gyök problémáját megkerültük.

Az *Euklidesz-féle algoritmusban* a szukcesszív polinomosztások miatt a hibák felhalmozódhatnak, s a végén adódó többszörös faktor pontatlan, azt még valamilyen eljárással (iterációval) pontosítani kell, ezért használatát lehetőleg mellőzzük.

A módszerrel kapcsolatban az eddigiekben nyert numerikus tapasztalatok kedvezők. A módszer lineáris diszkrét rendszerek sztochasztikus analízisében már alkalmazást nyert. Programja az MTA CDC 3300 gépére készült FORTRAN nyelven, s a felhasználók rendelkezésére áll.

IRODALOM

[1] RALSTON, A., Bevezetés a numerikus analízisbe (Műszaki Könyvkiadó, Budapest, 1965).
[2] Березин, И. С. и Жидков, Н. П., Методы вычислений (Наука, Москва, 1966).

(Beérkezett: 1978. szeptember 5.)

VARGA GYULA
MTA SZÁMÍTÁSTECHNIKAI ÉS AUTOMATIZÁLÁSI KUTATÓ INTÉZET
1250 BUDAPEST, ÚRI U. 49.

FACTORIZATION OF REAL POLYNOMIALS HAVING DOUBLE REAL ROOTS

GY. VARGA

This paper gives a method based on the *Newton—Raphson iteration* for finding the factors belonging to double real roots of real polynomials. It can be applied to determine the extremum points of real polynomials too.

A külföldi szakirodalomból

A KATASZTRÓFA-ELMÉLET STRUKTURÁLIS SZINTJEI ÉS AZOK NÉHÁNY ALKALMAZÁSA A TÁRSADALOMTUDOMÁNYOKBAN ÉS A BIOLÓGIÁBAN[1]

E. C. ZEEMAN

A katasztrófa-elmélet módszerét THOM alkotta meg (l. [14]), amikor a természet modellezésére sima leképezések szingularitásait alkalmazta.

Az ilyen modelleknél gyakran többféle strukturális szint lehetséges, például a topológiai, a differenciális, az algebrai, az affin stb. Ahogy a geometriában általában a topológiai szint a legmélyebb és megszorításokat adhat a magasabb szintekre vonatkozóan, úgy az alkalmazott matematikában, ha létezik katasztrófa-szint, akkor általában az a legmélyebb és nagy valószínűséggel megszorításokat ad a magasabb szintekre vonatkozóan, például az előforduló differenciálegyenletekre, az aszimptotikus viselkedésre stb. Továbbá, a geometriában a magasabb szintek bonyolultsága alkalmazhatatlanná teheti azokat, úgy, hogy legfeljebb implicite kezelhetők, de explicit módon nem, míg az alattuk rejlő topologikus invariánsok esetleg ki is számíthatók; ehhez hasonlóan az alkalmazott matematikában a differenciálegyenletek bonyolultságuk miatt gyakran kezelhetetlenek (még számítógép alkalmazása esetén is!), úgy, hogy csak implicit módon tudjuk vizsgálni azokat és nem explicite, ugyanakkor az alattuk rejlő katasztrófa modellezhető, esetleg oly mértékben, hogy számszerű előrejelzés is adható.

Tehát a katasztrófa-elmélet két fő előnye: egyrészt gyakran a legmélyebb betekintést adja a jelenség lényegébe és egyszerű megértést tesz lehetővé, másrészt nagyon bonyolult rendszereknél, mint amilyenek például a biológiában és a társadalomtudományok területén előfordulnak, gyakran olyan esetben is lehetővé teszi a modellezést, amikor az korábban elképzelhetetlen volt. Ebben a dolgozatban különböző strukturális szintekről lesz szó, amelyeket az alapjául szolgáló katasztrófára építünk és ezeket példákkal illusztráljuk. Egyszerűség kedvéért többnyire a közismert csúcskatasztrófáról lesz szó (l. [5], [13], [14], [24]).

1. szint: Szingularitások.
2. szint: Gyors dinamika (homeosztázis).
3. szint: Lassú dinamika (fejlődés).
4. szint: Visszacsatolás.
5. szint: Zaj.
6. szint: Diffúzió.

[1] E dolgozat „Levels of structure in catastrophe theory illustrated by applications in the social and biological sciences" címmel a *Proceedings of the International Congress of Mathematicians Vancouver*, 1974 Vol. 2 533—546. oldalain jelent meg. A magyar nyelvű fordítás közléséhez a szerző hozzájárult.

A THOM osztályozásában szereplő elemi katasztrófák az első szinthez tartoznak, a 2., 3., 4. szint közönséges differenciálegyenletekkel, a 6. szint parciális differenciálegyenletekkel kapcsolatos.

1. *szint: Szingularitások.*

Először felidézzük a fő osztályozási tételt. Legyenek C és X sokaságok, $\dim C \leqq 5$, legyen $f \in C^\infty(C \times X)$. Tegyük fel, hogy f tipikus *(generic)* abban az értelemben, hogy a hozzátartozó $C \to C^\infty(X)$ grafikon metszi a $C^\infty(X)$-en értelmezett $\mathrm{Diff}(X) \times \mathrm{Diff}(R)$ csoport pályáit. (A tipikusság nyílt-sűrű a *Whitney-féle C^∞-topológiában.*) Legyen $M \subset C \times X$ a $\nabla_X f = 0$ egyenlettel definiálva, a $\chi: M \to C$ függvényt pedig indukálja a $C \times X \to C$ vetítés.

Thom tétele:
a) M a C-vel megegyező dimenziójú sokaság;
b) χ minden szingularitása ekvivalens egy elemi katasztrófával;
c) χ stabilis az f kis perturbációira nézve.

Az elemi katasztrófák száma csak a C dimenziójától függ (és nem függ az X-től):

C dimenziója:	1	2	3	4	5	6
Elemi katasztrófák száma:	1	2	5	7	11	(∞)

Az elemi katasztrófákra vonatkozó részletek találhatók például [10], [14]-ben. Az első érvényes bizonyítást MATHER ([8]) adta; további irodalom: [1], [17], [18], [28].

Megjegyzés: A szingularitások osztályozásánál $\dim C \geqq 6$ esetén végtelen sok esetet kapnánk, de a táblázat kiterjeszthető úgy, hogy csak véges sok esetet kapjunk, ha az elemi katasztrófa fogalmát az alábbiak szerint alkalmasan módosítjuk. A szingularitások egy csira-téren *(space of germs)* ható csoport pályáinak felelnek meg. (L. a 4. szintet alább.)

Speciálisan, a ∞ azért jelenik meg $\dim C = 6$-nál, mert létezik egy 6 kodimenziós réteg, amelyet 7 kodimenziós pályák levelekre osztanak. ARNOLD ([1]) a levelekreosztás kodimenzióját, ami ebben az esetben 1, a réteg modalitásának *(modality)* nevezi. Általánosabban a pályák P egy leveles rétegződését adják, amelyről ARNOLD kimutatta, hogy lokálisan véges. Az egyes kodimenziókhoz tartozó rétegek véges száma adja a táblázat kívánt kiterjesztését.

Hogy katasztrófa-elmélet létezik, annak oka az a szerencsés véletlen, hogy P 5-egyszerű, azaz P minden 5-nél nem nagyobb kodimenziójú rétege egyszerű, vagyis triviálisan levelekre osztott egyetlen levéllel. Ezek a rétegek $\dim \leqq 5$ esetén az elemi katasztrófáknak felelnek meg, ezért ez utóbbiak végesen osztályozott differenciálinvariánsok. A legtöbb alkalmazásnál $\dim \leqq 5$ elegendő, ezért nem kell aggódnunk a magasabb rétegek levelekre osztása miatt.

Alkalmazás. Tekintsük tárgyaknak, vagy eseményeknek egy halmazát, amellyel kapcsolatban egy ok-okozat hipotézist akarunk ellenőrizni. Először ábrázoljuk azokat az ok-okozat térben és ellenőrizzük, hogy gráfot alkotnak-e. Itt C fogja leírni az okot, X az okozatot és $f(c, x)$ annak a valószínűségét, hogy a c ok az x okozatot váltja ki.

A legvalószínűbb okozatokat a valószínűség maximumainál találjuk, ott, ahol a gradiens eltűnik: $\nabla_x f = 0$, és a *Hesse-determináns* negatív definit: $\nabla_x^2 f < 0$. Ez meghatároz egy az M-mel azonos dimenziójú G részsokaságot. Ezen G lesz a keresett ok-okozat grafikon a $C \times X$-ben. Az eseményeket egy a G körül torlódó pontfelhő reprezentálja, a torlódási sűrűség a valószínűségeloszlások szórásától függ.

Tekintsük az első két elemi katasztrófát, amelyek $\dim C \leqq 5$-nél lépnek fel. A ránc-katasztrófa a G határán következik be, de minthogy az első szintnél nincs dinamika, katasztrófa-ugrás nem következhet be, csak annyit mondhatunk, hogy a pontfelhő itt véget érni látszik.

Csúcskatasztrófa akkor lép fel, amikor a valószínűségeloszlás bimodálissá válik. Ebben az esetben a megfigyelők implicite felismerhetik a jelenség lényegét és részben ki is fejezhetik, vagy úgy, hogy megnevezik a két kimeneti módot, vagy szólás, illetve hiedelem formájában fogalmazva meg azt. A csúcskatasztrófa azonban gyakran a jelenség mélyebb vonásait világítja meg és új szintézist tesz lehetővé a megértésben. A fenti két alternatívát néhány példával szemléltetjük.

1. *Példa: Agresszió* ([22]).

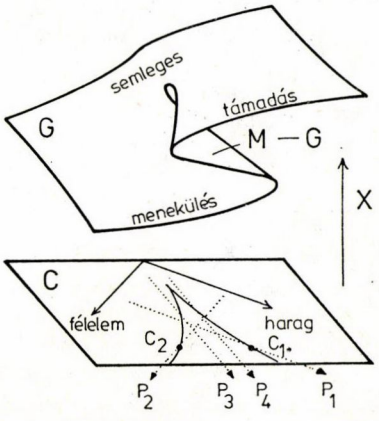
KONRAD LORENZ ([7]) szerint félelem és harag az agressziót ellentétesen befolyásoló hajtóerők. Itt a két szélsőséges viselkedési mód a támadás—menekülés, X a viselkedés 1-dimenziós spektrumát reprezentálja, amely a semlegestől a két szélsőségig változhat. A C ok 2-dimenziós és az állatban az adott pillanatban jelenlevő félelem és düh indítékok erejét reprezentálja. LORENZ megfigyelte, hogy kutyák esetén a félelem és düh koordináták leolvashatók az állatok pofájáról. (L. [7] 81. oldal.) Ha csak düh van jelen, akkor támadás következik be, ha csak félelem, akkor menekülés, ha mindkettő jelen van, akkor az okozat a két szélsőség valamelyike, de nem előrejelezhető, hogy melyik. Ezért a valószínűségeloszlás bimodálissá válik, első közelítésként azt várhatjuk, hogy a pontfelhő egy olyan ok-okozat grafikon körül torlódik, amely ekvivalens az 1. ábrán szemléltetett csúcskatasztrófával. Ehhez a példához és alkalmazásaihoz a 2. szintnél visszatérünk.

A bimodalitás további közismert, a csúcskatasztrófával modellezhető példái:

a) folyadék-gáz ([3], [11], [15]),
b) szívelernyedés—összehúzódás ([23]),
c) mániákus—depressziós ([25]),
d) galamb—héja ([5]),
e) (tőzsdei) hossz—bessz ([26] és az alábbi 7. példa).

A bimodalitást ezen eseteknél vagy ellentétes hatást kiváltó tényezők okozzák, mint például hőmérséklet és nyomás az *a)* esetben, vagy valamilyen megosztó tényező váltja ki azt, mint a feszültség a *b)* példánál, betegség a *c)*-nél, költség, illetve spekuláció a *d)*, illetve az *e)* esetben.

Következő példaként tekintsünk egy közmondást.



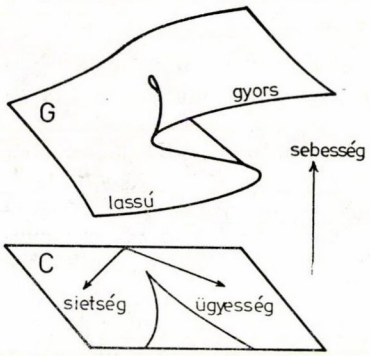
1. ábra

2. *Példa: Lassan járj, tovább érsz!*

Angliában közismert, az *Egyesült Államokban* viszont szinte ismeretlen az alábbi közmondás: *„More haste, less speed"*, azaz: Minél nagyobb a sietség, annál kisebb a sebesség. A közmondás rövid és frappáns; az ellenkezőjét fejezi ki annak, amit általában várnánk, különösen akkor, ha a dolgozó ügyes (rátermett és jól képzett) a munkaterületén.

Ebből azt a következtetést vonhatjuk le, hogy a sebesség valójában két tényezőtől függ, a sietségtől és az ügyességtől, amelyek más-más irányban hatnak. Ha ugyanis mindkét tényezőt növeljük, akkor kétféle hatás következhet be: vagy a dolgozó ügyessége lehetővé teszi számára, hogy sebességét (teljesítményét) növelje, vagy kapkodó sietsége csökkenteni fogja azt. Tehát ismét azt várhatjuk, hogy pontfelhőnk a csúcskatasztrófának megfelelő, a 2. ábrán látható görbe mentén torlódik.

Ezen modell többféle felhasználását is javasolhatjuk. Először, a pszichológiában továbbfejleszthető adott készségre vonatkozó, kvantitatív előrejelző modellé (l. az alábbi 2. szintet). Másodszor, a szociológiában szembenálló elméletek összehangolásának prototípusa lehet. A tétel szerint ugyanis a csúcskatasztrófa fellépése több jelenségnél is várható, és bár az ábra látszólag egyszerű, a finomságait nem könnyű röviden jellemezni az írott vagy beszélt nyelven. ([5], [10].) Ezért, bár ösztönösen gyakran felismerünk egy ilyen jelenséget, mégis hajlamosak vagyunk arra, hogy szavakban túlságosan leegyszerűsítsük a leírást, esetleg a figyelmet csak a váratlan hatásra irányítva. Például a „lassan járj, tovább érsz" a figyelmet a 2. ábrának csak az alsó, „lassú"-val jelölt lapjára irányítja. Hasonlóan, egy vitában összeütköző két vélemény, vagy két szembenálló szociológiai elmélet esetleg egy alapjában bimodális jelenség egyik oldalát emelheti ki, és a konfliktus néha feloldható azáltal, hogy rámutatunk, hogy a két viselkedési mód egy csúcskatasztrófa két lapjának felel meg, amelyek egy azonosan értékelt terület sima folytatásai.



2. ábra

2. *Szint: Dinamika.*

Tegyük fel, hogy az $f \in C^\infty(C \times X)$ tipikus függvényen kívül egy D dinamika is adva van az alábbi módon: jelöljük az asszociált $K \to C^\infty(X)$ függvényt $c \mapsto f_c$-vel. Ekkor $D = \{D_c\}$ az X-en értelmezett, a C-vel paraméterezett differenciálegyenletcsalád, úgy, hogy minden $c \in C$-re f_c a D_c egy *Ljapunov-függvénye*. Más szóval f_c növekszik (ill. csökken) a D_c pályái mentén, így az f_c maximumai (ill. minimumai) a D_c attraktorai. Tehát D_c gradiens-rendszer — ezt a megszorítást adja az első szint a második szintre vonatkozóan.

A G gráf most már a D attraktorait reprezentálja. Az alkalmazásoknál a pontfelhőt intuitíve most már nem a G körül statikusan tömörülve képzeljük el, hanem mint dinamikusan a G-hez áramló és aztán ott maradó pontokat. A modell

félig dinamikus, félig statikus. Néha hasznos a C paraméter-teret kontrollnak, X-et pedig állapottérnek tekinteni. Ha c-t lassan változtatjuk, akkor az x állapot folytonosan változik a G-n ameddig ez lehetséges; más szóval: az, hogy a rendszer *Thom késleltetési törvénye* ([5], [14]) szerint viselkedik, a 2. szintre vonatkozó tétel.

Amikor c áthalad a bifurkációs halmazon, akkor előfordulhat, hogy x átlépi a G halmaz ∂G határát. Ebben az esetben a dinamika x-et gyorsan átviszi a G egy másik lapjára. A „gyorsan" kifejezés arra utal, hogy a kontroll változása lassú a dinamikához képest; ez a ránc-katasztrófánál a második szinten fellépő hirtelen ugrás adta az elmélet nevét: katasztrófa-elmélet.

3. *Példa: A katasztrófa-gép* ([9], [24], [29]).

Kartonlap és két rugalmas pánt felhasználásával készített egyszerű játék, amely jól szemlélteti a katasztrófa-ugrást, ezért a témában járatlan olvasónak azt javasoljuk, hogy készítsen egyet magának. Itt az első szint f függvénye a rugalmas pántok potenciális energiája, amelyet a *Hook-törvény* ad meg, a D dinamikát pedig a *Newton-féle mozgástörvény* adja, megfelelő csillapítással, úgy hogy minimalizálja f-et.

1. *Példa:* Visszatérve első példánkhoz, láthatjuk, hogy kiterjeszthetjük azt az első szintről a második szintre. X-et ugyanis újraértelmezhetjük úgy, hogy legyen az az agy azon részének állapottere, amely a kedélyállapotot befolyásolja (ez talán a *hypothalamus*), D pedig az ehhez tartozó dinamika, amely az idegi tevékenységet reprezentálja. Ekkor D attraktorai a támadó, illetve a menekülő idegállapotot reprezentálják, amelyek alapján a viselkedési döntések születnek. Bár X dimenziója természetesen nagyon magas, következésképpen D kezelhetetlen, minthogy csupán implicit, mindazonáltal G csak 2-dimenziós. Ezért a csúcs-katasztrófa explicit modellt szolgáltathat, amely konkrét állatok esetén kvantitatívvá tehető, úgy hogy ennek alapján előrejelzés is adható. Továbbá, minthogy ez 2. szintű modell, bár a D csak implicit, az idegállapotban katasztrófa-jellegű ugrások lesznek, amelyek hirtelen támadásokat és visszavonulásokat eredményeznek. Például az 1. ábrán a P_1 pálya rögzített félelemszint mellett növekvő dühöt reprezentál, például egy sarokbaszorított kutyánál, amely egy c_1-beli hirtelen támadáshoz vezet, míg a P_2 pálya a c_2-beli megfutamodáshoz. Ugyanakkor a P_3 és P_4 pályák azt illusztrálják, hogy közeli pályák is vezethetnek eltérő viselkedéshez. Hasonlóan az emberek is, ha dühítik és ijesztik őket, kiszámíthatatlanokká válnak, nem viselkednek racionálisan, és szidalmazásból bocsánatkérésbe, hisztériából könnyekbe csaphatnak át. A példa érdekessége, hogy az agresszivitás kontrollálására adhat egy általános modellt, amely különféle fajokra változó körülmények között érvényes és betekintést adhat abba, hogy ezen kontrollok hogy alakultak ki és hogyan fejlődtek. Általánosabban, arra szolgáltat prototípust, hogy a neurológiát összekapcsoljuk a viselkedést meghatározó kedélyállapotok pszichológiájával.

2. *Példa:* Második példánkat is kiterjeszthetjük az első szintről a másodikra, mivel valamely személy teljesítményét tekintve, az illető törekvése, hogy sebességét például x-re növelje (ügyességének korlátain belül, egy bizonyos sietséget vállalva), tulajdonképpen azt jelenti, hogy van egy implicit dinamika, amely a sebességet x-hez közelíti. Az 1. ábrán látható P_1 pálya itt rögzített sietségi szint mellett növekvő

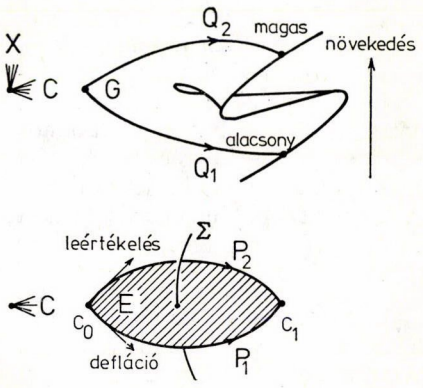
ügyességet reprezentál, amely a sebességet x-hez közelíti, mint például amikor kerékpározni tanul valaki és egy c_1 pontnál a katasztrófa hirtelen bekövetkezik, vagyis az illető egyszerre csak képes nyeregben maradni, kerékpározni. Továbbá, minél nagyobb a sietség — például egy kevésbé stabilis gép hajtásához gyorsabb reagálásra van szükség — annál nagyobb ügyességre van szükség ahhoz, hogy a katasztrófa bekövetkezzék. Ugyanakkor a P_2 pálya ennél a példánál rögzített ügyesség mellett növekvő sietséget reprezentál, mint például egy rádiós esetén, aki mind gyorsabban és gyorsabban próbál olvasni egy *Morse-kódot,* míg a c_2-nél a katasztrófa bekövetkezik, vagyis a teljesítménye hirtelen lezuhan. Nyilvánvaló, hogy minél nagyobb az ügyesség, annál nagyobb sietség lehetséges a katasztrófa bekövetkezéséig.

Általában a 2. szintet sokkal könnyebb kísérletileg ellenőrizni, mint az első szintet, mert a pontfelhő pontosabban meghatározza G-t, a katasztrófák pedig meghatározzák ∂G-t. Valahányszor egy jelenségnél a bimodalitás, az elágazás, a katasztrófa-ugrás, vagy a hiszterézis-késedelem valamelyike fellép, akkor azon jelenség csúcs-katasztrófával modellezhető és a másik három tulajdonság előrejelezhető. A csúcs-katasztrófa fontos lehet az alkalmazásoknál akkor is, amikor a kontroll-tér dimenziója nagyon magas, mint az alábbi példa mutatja.

4. *Példa: Gazdasági növekedés.*

Reprezentálja X egy gazdaság állapotterét, C az ezen gazdaságra ható külső kényszert és a kormány rendelkezésére álló belső kontroll-tényezőket. A C várhatóan nagyon magas dimenziójú lesz, úgy hogy első látásra úgy tűnik, hogy a tételnek nem sok hasznát látjuk. Valójában azonban a gazdasági fejlődésnek a C-ben egy 1-dimenziós pálya felel, amely a G-re felvetítve szintén egy 1-dimenziós pályát határoz meg, a megfelelő 1-dimenziós katasztrófák pedig a gazdasági válságok, inflációs árrobbanások stb.

A kormány előtt álló jellegezetes probléma annak realizálása, hogy míg jelenlegi politikája a c_0 kontroll ponton van, addig az elkövetkező néhány hónapban meg kell változtatnia azt és a c_1 pontba hozni, külső kényszerítő tényezők (például fizetési mérleg stb.) miatt. A kormány cselekvési szabadsága esetleg csak arra korlátozódik, hogy megválassza a c_0-ból c_1-be vezető pályát. Azonban egy ilyen választás kritikus is lehet, mint az alábbiakban megmutatjuk. Tegyük fel, hogy választhatunk a P_1 és P_2 pályák között. Egyszerűség kedvéért tegyük fel, hogy egyik pálya mentén sem következik be katasztrófa. Azt a kérdést kell megvizsgálnunk, hogy a $P_1 \cup P_2$ görbe közrefogja-e a bifurkációs halmaz valamilyen 2-dimenziós $\sum$ rétegét, mert ha igen, akkor a $P_1 \cup P_2$-t kifeszítő 2-dimenziós E lap metszi a $\sum$-t és a G-nek az E feletti része csúcskatasztrófát fog tartalmazni, mint az a 3. ábrán látható.



3. ábra

A P_1-nek és P_2-nek a G-re eső Q_1 és Q_2 vetületei divergálnak, amely jelentősen befolyásolja a növekedést, inflációt, munkanélküliséget stb.

Tegyük fel, hogy P_1 deflációt reprezentál, amelyet leértékelés követ (mint az *Egyesült Királyságban* 1967-ben), P_2 pedig ugyanezeket fordított sorrendben (mint *Franciaországban* 1968-ban). Ekkor Q_1 alacsony növekedést eredményezhet, mivel a cégek csökkentett készleteikkel nem tudják kihasználni a leértékelést; ugyanakkor Q_2 magas növekedést eredményezhet, mert a cégek raktárkészleteiket a leszűkült hazai piacról kivihetik a nemzetközi piacra és így növekedési rátájukat nem kell csökkenteniük. Tehát a közgazdászoknak nemcsak a nyilvánvalóbb 1-kodimenziós katasztrófa-problémákkal kell törődniük, hanem az elágazás és választás rejtettebb, 2-kodimenziós problémáival is.

3. *Szint: Növekedés.*

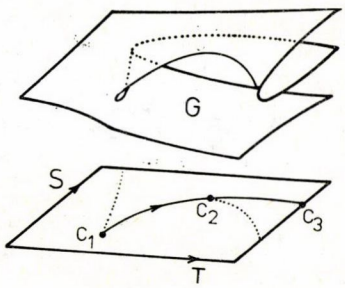
Tegyük fel, hogy az $f \in K^\infty(C \times X)$ függvényen és a D dinamikán kívül még fellép a T idő a C kontroll-tér egyik tengelyeként. Feltesszük, hogy ez a T lassú a D dinamikában fellépő időhöz képest.

5. *Példa: Embriológia.*

A katasztrófa elméletnek az embriológia területén való alkalmazásánál a THOM által adott legfontosabb eset a 3. szinttel kapcsolatos ([13], [14], [16]), ahol a $C = S \times T$ kontroll tér a téridőt, X pedig a sejt állapotát reérezentálja. Például X lehet az R^n egy korlátos, nyílt részhalmaza többezer koordinátával, amelyek a sejt különféle fizikai és kémiai paramétereit jelentik. A D dinamika reprezentálja a sejt homeosztázisát, amely gyorsan visszaviszi az egyensúlyi helyzetbe, T pedig a sejtek lassú fejlődését. Az ezzel kapcsolatos eredményekre egy példa az alábbi.

Tétel: ([27]): Valahányszor egy szövet két különböző típusúra differenciálódik, a köztük levő határfelület a szövet egyik oldalán képződik és a szöveten keresztülhaladva éri el végső helyzetét, ahol stabilizálódik. A bizonyítás a 4. ábrán szemléltetett csúcs-katasztrófa segítségével történik. S merőleges a határfelületre, 1-dimenziós. A sejtek fejlődési pályáit úgy kapjuk, hogy az idő-vonalakat a G-re felvetítjük. A határ először c_1-nél képződik, aztán hullámként terjed az S-ben a csúcs c_1c_2 íve mentén, míg végül stabilizálódik c_2-nél, ahol a csúcs érinti a c_2c_3 idővonalat. Egy ilyen hullám gyakran a gének rejtett aktivizálása és bizonyos késedelem után a fizikai megvalósulás egy második hulláma morfogenezist eredményezhet. Például [27]-ben részletes modell található gasztrulációs és neurulációs morfogenezisre kétéltűeknél.

Tér-katasztrófák. A fenti eredmény azon múlott, hogy az idő-tengely és a csúcstengely nem érintőleges, amit a tipikusságra *(genericity)* való hivatkozással indokolhatunk. Ahhoz azonban, hogy a tipikusságnak ezt a fajtáját matematikailag megalapozzuk, a klasszikus elméletet általánosítanunk kell az alábbiakban megadott módon, amelyet WASSERMANN ([19]) tér-katasztrófa elméletnek nevez. (Ő tanulmányozza a duális fogalmat is, amelyet idő-katasztrófának nevez.)



4. ábra

Jelölje E_n az $R^n \to R$ C^∞ függvények 0-beli csiráinak a csoportját, legyen m_n a maximális ideál, G_n pedig az $(R^n, 0) \to (R^n, 0)$ C^∞ diffeomorfizmusok csiráinak a csoportja. Ekkor G_n az m_n-en hat, az m_n^2-et invariánsan hagyva. A klasszikus katasztrófa-elmélet ([1], [8], [14], [17], [18]) az m_n^2 G_n általi P leveles rétegezettségét *(foliated stratification)* tanulmányozza. Az s-dimenziós elemi katasztrófákat a P s-kodimenziós rétegei adják. Minthogy P 5-egyszerű, az elemi katasztrófák $s \leqq 5$ esetén végesen osztályozott differenciálinvariánsok, amelyek függetlenek az n-től ($n \geqq 2$).

Az általánosításhoz további definíciókra lesz szükségünk. Azt mondjuk, hogy $\alpha \in G_{n+r}$ lefedi $\beta \in G_r$-et, ha $\pi\alpha = \beta\pi$, ahol $\pi: R^{n+r} \to R^r$ a vetítés. Legyen: $G_n^r = \{(\alpha, \beta) \in G_{n+r} \times G_{1+r};\ \exists \gamma \in G_r$ úgy, hogy α, β lefedi γ-t$\}$. Ekkor G_n^r az m_{n+r}-en hat, úgy, hogy $m_n^2 + m_r E_{n+r}$-et változatlanul hagyja. A térkatasztrófák elméletéhez válasszuk $r = 1$-et (az időt reprezentálva) és vizsgáljuk $m_n^2 + m_1 E_{n+1}$-nek a G_n^1 által meghatározott Q leveles rétegzettségét. Az s-dimenziós térkatasztrófákat a Q $s+1$ kodimenziós rétegei adják.

WASSERMANN ([19]) megmutatta, hogy a Q 2-egyszerű, ezért az 1-tér-katasztrófák végesen osztályozott differenciálinvariánsok, amelyek függetlenek n-től ($n \geqq 1$). Pontosan négy ilyen katasztrófa van, a 4. ábrán látható hullám c_1 kezdete, a közepe és a c_2 vége, valamint a c_2 „csendes" duálisa. Tehát a fenti tétel érvényes és mindegyiket megadja.

Q azonban nem 3-egyszerű, mivel WASSERMANN megmutatta, hogy a fecskefarkak és köldökpontok *(swallowtails, umbilics)* P rétegeit a Q nemcsak továbbrétegezi, hanem levelekre is osztja. Ezért a 2-tér szingularitásainak a száma végtelen, és bár a 2-tér katasztrófái még végesen osztályozhatók, de már nem lesznek többé differenciálinvariánsok. Némelyik közülük invariáns marad: például a ránc-felületnek a 2-térre eső csúcsvetülete. (Analóg módon a ránc-görbe 1-térre eső ránc-vetületéhez a 4. ábra c_2 pontjában.) Ezt a példát alkalmazták a kétéltűeknél a szomit-képződés modellezésére ([27]).

Minthogy THOM ([14], [16]) az embriológiában kiterjedten alkalmazta a fecskefarok, a pillangó és a köldökpont esetét, tehát fontos, hogy osztályozzuk a 2-tér és a 3-tér katasztrófáit. Matematikailag ez azt jelenti, hogy tanulmányoznunk kell a Q rétegeit egészen a 4 kodimenzióig és meg kell értenünk a differenciálhatóság elvesztésének a jellegét, amelyet levelezettségük okoz.

4. *Szint: Visszacsatolás.*

Itt feltesszük, hogy a lassú dinamikai rendszer nem olyan egyszerű, hogy egyetlen koordinátával megadható a kontrolltérben, hanem a G különböző levelein különböző irányú lehet. Tehát tulajdonképpen visszacsatolásnak tekintjük.

$$C \underset{\text{lassú visszacsatolás}}{\overset{\text{gyors dinamika}}{\rightleftarrows}} X$$

Pontosabban: az f-en és a D-n kívül legyen adva egy $F: C \times X \to TC$ C^∞-leképezés, ahol TC jelöli a C érintősokaságát és $F(c, x)$ minden $c \in C$, $x \in X$-re egy c-beli érintő. Tehát D és F együtt a $C \times X$-en egy közönséges differenciálegyenletet adnak (azzal a kikötéssel, hogy D gyors és F lassú).

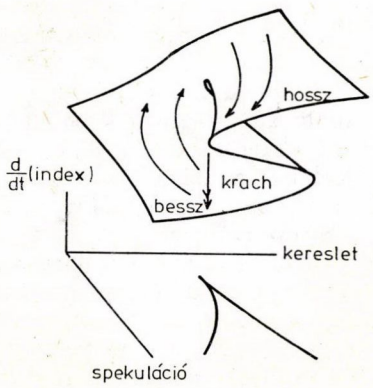
6. Példa: Szívverés és idegi impulzus

A csúcs-katasztrófával kapcsolatos visszacsatolásos közönséges differenciálegyenletek explicit példái szolgáltak modellként. Mindegyik esetben a dinamikai rendszernek stabilis egyensúlyi helyzete volt, amelyet egy „külső hatással" alkalmasan megzavarva, elindult egy katasztrófa a D-n keresztül és visszatérés az egyensúlyi helyzetbe az F-en keresztül. A szívverés esetén a visszatérésnél fellép egy második katasztrófa (relaxáció az összehúzódás után), míg az idegi impulzusnál a visszatérés sima *(repolarizáció)*.

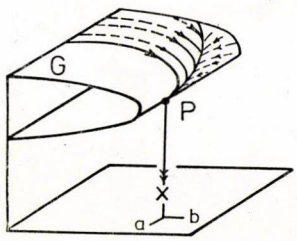
Ezeknek a modelleknek két érdekes tulajdonsága van. Az egyik, hogy a visszacsatolás nem pontosan a G-n, hanem csak közel a G-hez ad meg egy dinamikai rendszert, ahol a közelség mértéke a K (gyors/lassú) hányadostól függ. Ha $K \to \infty$, akkor a G-n egy idealizált dinamikai rendszert kapunk, ami az elektromosságtanbeli relaxációs rezgések általánosítása. Másodszor, a fentiekben a „külső hatás"-ra való utalás mutatja, hogy a modellek inadekváltak, minthogy csak közönséges differenciálegyenletekkel írják le a szívizom és az idegszelet lokális viselkedését; valójában ezeket be kellene ágyaznunk egy nagyobb parciális differenciálegyenletbe, amely a globális viselkedést írja le hullámként. Erre a problémára a 6. szintnél visszatérünk.

7. Példa: Tőzsdei adás-vétel.

A tőzsdei adás-vételek modellezésére a csúcs-katasztrófa az alábbiak szerint használható fel. A túlkereslet a normál tényező *(normal factor)*, amely a csereindexet kontrollálja, a piacon folyó spekulációs tevékenység pedig a megosztó tényező *(splitting factor)*. A D dinamika reprezentálja az index közvetlen válaszát a befektetésekre, F pedig a valamivel lassúbb visszacsatolást. Plauzibilis gazdasági feltevések olyan dinamikai rendszerhez vezetnek, ahol periódikusan változik a (tőzsdei) hossz, a gazdasági visszaesés, a (tőzsdei) bessz és a fellendülés. Ahhoz azonban, hogy ez a modell reális legyen, a zaj figyelembevételével az 5. szintre kell fejlesztenünk.



5. ábra



6. ábra

8. *Példa: Tölcsér.*

A tipikus *(generic)* alacsony dimenziójú visszacsatolási katasztrófák osztályozásánál TAKENS ([12]) a közelmúltben egy érdekes új fajtát fedezett fel, amelynek legegyszerűbb esetét tölcsérnek nevezte. Az asszociált ideálizált dinamikai rendszernél a G egyetlen P ráncpontba torkollik. A 6. ábra az alábbi explicit példát szemlélteti:

Gyors dinamika: D: $\dot{x} = -K(x^2+2b)$ K nagy konstans.

Lassú dinamika: F: $\dot{a} = 1$, $\dot{b} = 3a+4x$.

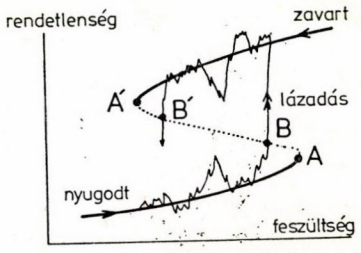
Tölcsérek fordulhatnak elő biológiai szabályozásokbál, például ha x, b reprezentálja a sejt belső önszabályozását, a pedig valamilyen — a sejten kívül felhasznált — hormon termelését, amelynél a termelési hányadosnak nagyon pontosan kell szabályozódnia.

5. *Szint: Zaj.*

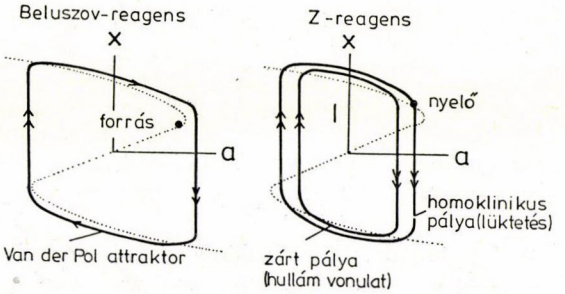
Az $\{f, D, F\}$-hez még hozzávehetünk egy sztochasztikus zajt a kontroll és a viselkedés véletlenszerű kis elmozdulásai formájában. A legtöbb zaj esetén a D dinamika az állapotot gyorsan visszaviszi a G-re, úgy, hogy az F lassú dinamikai rendszer változatlan marad, így a zajt figyelmen kívül hagyhatjuk. Két esetben azonban a zaj katasztrófát okozhat: egyik eset, amikor a kontroll-zaj eléri a bifurkációs halmazt, a másik, érdekesebb eset, amikor az állapotzaj eléri a szeparatrixot.

7. *Példa: A tőzsde példánál* a zaj a külső eseményeket és az ezek miatti piaci bizonytalanságot reprezentálja, aminek következtében recesszió léphet fel mielőtt a bifurkációs szintet elérnénk.

9. *Példa: Zendülések.* ([4].) Ez a modell P. SHAPLAND, C. HALL és M. MARRIAGE börtönpszichológusokkal és J. HARRISON statisztikussal közösen folyó kutatásaink eredménye. Egy nyilvánvaló igazsággal kezdjük. Minél nagyobb a feszültség egy intézményben, annál nagyobb a rendetlenség (zűrzavar, zavargás). Ez nemcsak olyan intézményekre vonatkozik, mint börtönök, egyetemek, céhek, vagy országok, hanem egyénekre is. A börtön adatok vizsgálata azt sugallja, hogy a feszültség (idegesség, elégedetlenség) mérhető a beteget jelentők számával, ha azt alkalmasan kisimítjuk; a zavargás pedig mérhető a fellépő incidensek komolyságára adott független becslések összevetése alapján. Az elidegenedés (vagy kommunikáció hiánya) szétválasztó tényezőnek tűnik, amely két viselkedési módot eredményezhet, amelyeket a 7. ábrán nyugodtnak és zavartnak neveztünk; az adatok azt sugallják, hogy ez mérhető a fegyelmi ügyek számával. A visszacsatolás dinamikai rendszere nyugalom idején a feszültség növekedését, (hónapokig), zavargások idején az enyhülést (napokon át) reprezentálja. A zaj az incidenseket írja le, és ha a zajszint B-nél átmetszi az AA' szeparatrixot, akkor az incidensek zendüléssé növekednek, ami katasztrófát eredményez. A börtönlakók bizonyos típusai (pl. fiatal



7. ábra



8. ábra

hosszú távra elítéltek) általában magasabb zajszintűek és ezért hajlamosabbak a zendülésre. Ha a feszültség néhány nap után kissé enyhül, akkor egy incidens B'-nél fordított irányú katasztrófát eredményezhet. Ugyanez az incidens nem biztos, hogy korábban ugyanezt eredményezte volna, ami jól mutatja a megfontoltság előnyét.

10. *Példa: Fázisátmenet.*

Amikor a zaj gyakori és a zajszint magas, az állapot időátlagban az f maximumához (vagy minimumához) tart. Ez az oka annak, hogy a folyadék-gáz fázisátmenetnél a *Van der Waals egyenletet* a *Maxwell-szabállyal* kell kiegészíteni, és nem a késleltetési szabály érvényes.

Másrészt, ha a zajszintet alacsonyan tartjuk, akkor parciális késleltetés jöhet létre, mint például a túltelített állapot a ködkamrában. A *Maxwell-szabály* statisztikus mechanikában szokásos bizonyításánál a legnagyobb változás irányában integrálunk, de minthogy ez a módszer a kritikus pont környékén nem alkalmazható, érdekes lenne, ha egy új, absztrakt bizonyítást lehetne adni, Thom tételének bizonyításához hasonlóan, ezáltal is gazdagítva a kritikus pont analízist.

6. *Szint: Diffúzió.*

Az alábbiak SHARON HINTZÉvel közös, folyamatban levő kutatásainkon alapulnak, amelyeket WINFREE ([20], [21]), továbbá KOPPEL és HOWARD ([6]) a *Zhabotinsky-reakcióval* kapcsolatos dolgozatai serkentettek. Először a matematikáról.

Legyen Y egy sokaság, g pedig egy C^∞ vektormező az Y-on. A megfelelő közönséges differenciálegyenlet:

$$\dot{y} = g(y). \tag{1}$$

Elsősorban azt vizsgáljuk, hogy milyen típusú a 4. szint által szolgáltatott differenciálegyenlet, tehát az $Y = C \times X$ és a $g = \{D, F\}$. Legyen Y egy bizonyos közeg lokális állapotainak a tere az $S \times T$ tér-időben, g pedig legyen ezen közeg reakciója. Tegyük fel továbbá, hogy a közeg nemcsak reagál, de diffundál is. Akkor [6] alapján

a közeg $y: S \times T \to Y$ globális állapota kielégíti az alábbi reakciós-diffúziós parciális differenciálegyenletet:

$$\frac{\partial y}{\partial t} = g(y) + k\Delta^2 y \tag{2}$$

ahol k egy konstans (pontosabban egy vektorköteg-leképezés

$$k: TY \to TY),$$

amely az Y különféle komponenseinek különböző diffúziós hányadosait reprezentálja. Elsősorban az érdekel bennünket, hogy a közegben képződhetnek-e stabilis periodikus hullámvonulatok, vagy stabilis lüktetések (izolált hullámok). Ha ilyenek létrejöhetnek és θ-val jelöljük a sebességet, akkor az y globális állapotot faktorizálhatjuk: $S \times T \to R \to Y$ úgy, hogy $\partial y/\partial t = \theta \dot{y}$ és $\nabla^2 y = \ddot{y}$, ahol a pontok az R szerinti deriválást jelentik. Ily módon tehát a 2 parciális differenciálegyenlet a

$$\theta \dot{y} = g(y) + k\ddot{y} \tag{3}$$

közönséges differenciálegyenletté redukálódik.

A 6. szint esetén az érdeklődés középpontjában a (3) egyenlet áll. Hasonlítsuk össze az (1) egyenlettel: ha k kicsi, akkor (3) tekinthető az (1) egy szinguláris perturbációjának, de fontos alkalmazások esetén a k nagy, tehát új módszerekre van szükség.

A (3) tekinthető például egy dinamikai rendszernek a TY-on, amelynek az Y zérus metszeten ugyanazok a fixpontjai, mint (1)-nek. Az (1) valamely attraktora lehet a (3) egy nyeregpontja és ezen nyeregpont egy homoklinikus pályája a (2) egy lüktető megoldását jelenti, ugyanakkor a (3) egy zárt pályája a (2) egyenletnek egy hullámvonulat megoldását reprezentálja. A fentiek miatt a (3) egyenlet olyan homoklinikus és zárt pályáit keressük, amelyek a (2)-re nézve stabilisak. Jelenleg erről még viszonylag keveset tudunk, még abban az esetben is, amikor g egy kanonikus elemi katasztrófát reprezentál, a lehető legegyszerűbb visszacsatolással.

6. *Példa: A szívverés és az idegi impulzus dinamikája* ([23]).

CONLEY és CARPENTER ([2]) igazolták homoklinikus és zárt pályák létezését, hátra van még a stabilitás bizonyítása és a kísérleti adatokkal való egyeztetés.

11. *Példa: Zhabotinsky reagensek.*

BELUSZOV felfedezett egy kémiai vegyületet, amely színében körülbelül percenként kétszer oszcillál, később ZHABOTINSKY és ZAIKEN észrevették, hogy ezt az oszcillációt a reagensben terjedő cirkuláris hullámvonulatok mozgatják. WINFREE ([21]) később módosította a *Beluszov-reagenst* egy kicsivel több bromid és kevesebb acid hozzáadásával, ezzel megállítva az oszcillációt. Vegyületét ZHABOTINSZKY és ZAIKEN tiszteletére Z-reagensnek nevezte és kimutatta, hogy ebben mind lüktetések, mind forgó csigavonalszerű hullámvonulatok felléphetnek. [21]-ben WINFREE olyan egyenleteket adott meg, amelyek alapján nagyon szépen megmagyarázhatók a geometriai viszonyok, de amelyek 4 szempontból kissé bírálhatók. Először, hogy

a dinamikája nem folytonos, és a nyilvánvaló módja annak, hogy a modellt differenciálhatóvá tegyük az, hogy katasztrófa-modellel közelítjük. Valóban, KOPELL és HOWARD ([6]) szerint léteznek mind gyors (a másodperc törtrészei), mind lassú (percek) reakciók, csakúgy, mint egy nagyon lassú (órák) energiaveszteség. Ennélfogva az ember általában azt várja, hogy a reakció-dinamika a 4. szinthez tartozzék. Másodszor, WINFREE egyenletei nem tükrözik a *Beluszov—Z-reagens* módosítást. Ugyanakkor ez természetesen szemléltethető a katasztrófa-elmélet segítségével úgy, hogy változtatunk egy konstanst, ami *Hopf-bifurkációt* eredményez, mint alább megmutatjuk. Harmadszor, az ő egyenleteiből ugrásszerű visszatérés következnék, mint a szívverésnél, ugyanakkor fényképei sima kék→piros visszatérést mutatnak — az idegi impulzus repolarizációjához hasonlóan — ellentétben a hirtelen vörös→kék átmenettel. Ez a jelenség jól indokolható a csúcs-katasztrófa segítségével ([20], [23]). Negyedszer, WINFREE nem bizonyítja az egzisztenciát és a stabilitást.

Mint WINFREE kimutatta ([20]), az első két bírálatra kielégítő választ ad az alábbi 2-dimenziós ránckatasztrófa-modell (l. [23], 7. és 8. ábra). Legyen $Y=R^2$, g pedig legyen az alábbi egyenletekkel definiálva:

D — gyors dinamika: $\quad \dot{x} = -(x^3-3x+a)$,
F — lassú visszacsatolás: $\quad \dot{a} = \varepsilon(x-\lambda)$,

ahol ε, λ konstansok és ε kicsi. A *Beluszov-reagensnél* válasszunk $\lambda<1$-et, míg a *Z-reagensnél* $\lambda>1$ legyen. Akkor a λ paraméter csökkentése az 1 értéknél *Hopf-bifurkációt* eredményez. A keletkező dinamikai rendszereket a 8. ábra szemlélteti, ahol a lassú mozgásnak megfelelő sokaságot a pontozott vonalak szemléltetik. A *Beluszov-reagens* esetén KOPPEL és HOWARD ([6]) egy tétele biztosítja a (3) egyenletre zárt pályák létezését és stabilitását a *Van der Pol attraktor* közelében, de csak elegendő kicsiny diffúzió esetén. A *Z-reagensre* vonatkozóan CONLEY és CARPENTER ([2]) bizonyították homoklinikus és zárt pályák létezését, (de stabilitását még nem) nagy x diffúzió esetén, feltéve, hogy az ε elegendően kicsiny. Most arra lenne szükség, hogy a két esetet együtt kezeljük és bizonyítsuk a stabilitást nagy diffúzió esetén, becslést adva az ε lassú/gyors hányadosra. Ezután az eredményt ki kellene terjeszteni a csúcs-katasztrófára, ([23], 8. példa). Végül egyenleteinket a konkrét kémiai reakciókra felírva, kvantitatív modellt kellene adni és ennek alapján a különböző hullámok sebességét előre jelezni.

IRODALOM

[1] ARNOLD, V. I., „Singularities of differentiable functions", *Proceedings of the International Congress of Mathematicians* 1974, Vancouver, Canada.
[2] CARPENTER, G. A., „Travelling wave solutions of nerve impulse equations", Thesis, University of Madison, Wis., 1974.
[3] FOWLER, D. H., „The Riemann—Hugoniot catastrophe and Van der Waal's equation", *Towards a Theoretical Biology* 4, Edinburgh University Press, 1972 (1—7).
[4] HALL, C., HARRISON, P. J., MARRIAGE, H., SHAPLAND, P. and ZEEMAN, E. C., „A model for prison disturbances", *Jour. Math. and Stat. Psychology* (megjelenés alatt).
[5] ISNARD, C. A. and ZEEMAN, E. C., „Some models for catastrophe theory in the social sciences", *Use of Models in the Social Sciences*, London, 1974.
[6] KOPELL, N. and HOWARD, L. N., „Pattern formation in the Belousov reaction", *A. A. A. S.*, 1974, *Some Math. Questions in Biology*, VIII., *Lectures on Math. in the Life Sci.*, 7., *Amer. Math. Soc. Providence, R. L.*, 1974 (201—216).
[7] LORENZ, K., *On Aggression* (Methuen, London, 1967).

[8] MATHER, J. N., „Right equivalence", *Warwick Univ.*, 1969.
[9] POSTON, T. and WOODCOCK, A. E. R., „Zeeman's catastrophe machine", *Proc. Cambridge Philos. Soc.* **74** (1973) 217—226.
[10] POSTON, T. and WOODCOCK, A. E. R., „A geometrical study of the elementary catastrophes", *Lecture Notes in Math.*, Springer-Verlag, Berlin, 1974.
[11] SHULMAN, L. S. and REVZEN, M., „Phase transitions as catastrophes", *Collecting Phenomena* **1** (1972), 43—47.
[12] TAKENS, F., „Constrained differential equations", *Dinamical Systems*, Warwick, 1974, (*Lecture Notes in Mathem.* **468**, Spriner-Verlag, Berlin 80—82).
[13] THOM, R., „Topological models in biology", *Topology* **8**, (1969) 313—335.
[14] THOM, R., *Stabilité structurelle et morphogénèse* (Benjamin, New York, 1972).
[15] THOM, R., „Phase-transitions as catastrophes", *Conf. on Stat. Mechanics*, Chicago, I11., 1971.
[16] THOM, R., „A global dynamical scheme for vertebrate embriology", *A. A. A. S.*, 1971, *Some Math. Questions in Biology*, IV. *Lectures on Math. in the Life Sci.*, **5**, *Amer. Math. Soc., Providence, R. I.*, 1973, 1—45.
[17] TROTMAN, D. J. A. and ZEEMAN, E. C., „The classification of elementary catastrophes of codimension 5", *Lecture Notes, Warwick Univ.*, 1974.
[18] WASSERMANN, G., „Stability of unfoldings", *Lecture Notes in Math.*, 393, Springer-Verlag, Berlin, 1974.
[19] WASSERMANN, G., „(r, s)-stability of unfoldings", *Regensburg Universität*, 1974.
[20] WINFREE, A. T., „Spacial and temporal organisation in the *Zhabotinsky reaction*", *Aakron Katchalsky Memorial Sympos.*, Berkeley, Calif., 1973.
[21] WINFREE, A. T., „Rotating chemical reactions", *Scientific American* **230**, (1974), 82—95.
[22] ZEEMAN, E. C., „Geometry of catastrophes", *Times Literary Supplement*, 1971, 1556—1557.
[23] ZEEMAN, E. C., „Differential Equations for heartbeat and nerve impulse", *Dynamical Systems*, Academic Press, New York, 1973, 683—741.
[24] ZEEMAN, E. C., „A catastrophe machine", *Towards a Theoretical Biology* **4**, Edinbourgh Univ. Press, 1972, 276—282.
[25] ZEEMAN, E. C., „Applications of catastrophe theory", *Manifols, Tokyo* 1973 ,Univ. Tokyo Press, 1975, 11—23.
[26] ZEEMAN, E. C., „On the unstable behaviour of stock exchanges", *J. Math. Economics* **1** (1974) 34—39.
[27] ZEEMAN, E. C., „Primary and secondary waves in developmental biology", *A.A.A.S.*, 1974, *Some Math. Questions in Biology*, VIII., *Lectures on Math. in the Life Sci.*, 7, *Amer. Math. Soc., Providence, R. I.*, 1974, 69—161.

Magyar publikációk:

[28] FARKAS, M., „Folyamatok kvalitatív vizsgálatáról", *Alkalmazott Matematikai Lapok* **2** (1976) 237—257.
[29] FARKAS, M., „A társadalmi rendszer fejlődésének katasztrófaelméleti modellje", *Magyar filozófiai Szemle* **22** (1978) 802—808.
[30] FORGÁCS, G., „Katasztrófaelmélet", *Fizikai Szemle*, 1976/9 329—333.

FORDÍTOTTA:

LŐKÖS ÁGNES
BME GÉPÉSZMÉRNÖKI KAR MATEMATIKAI TANSZÉK
1521 BUDAPEST, STOCZEK U. H ÉP. IV. EM.

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 a felelős szerkesztő címére kell beküldeni:

Prékopa András, felelős szerkesztő, MTA SZTAKI
1502 Budapest XI., Kende u. 13—17.

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 Ungelichungen«, *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: 42,— Ft

INDEX: 26498
ISSN 0133—3399

TARTALOMJEGYZÉK

INDEX

A kiadásért felel az Akadémiai Kiadó igazgatója

79-1896 — Szegedi Nyomda — F.v.: Dobó József igazgató

Műszaki szerkesztő: Marton Andor
A kézirat nyomdába érkezett: 1979 IV. 2. Terjedelem: 18,55 (A/5 ív)