

MISKOLCI EGYETEM DOKTORI (PhD) TÉZISFÜZETEI



HATVANY JÓZSEF INFORMATIKAI TUDOMÁNYOK
DOKTORI ISKOLA

ALGORITMUSOK VÉLETLEN PARAMÉTEREI

ÉS

PROGRAMOK VÉLETLEN TESZTELÉSE

Készítette:

Dr. univ. Nagy Ferenc

OKLEVELES MATEMATIKUS,

AKI (PhD) FOKOZAT ELNYERÉSÉRE PÁLYÁZIK

DOKTORI ISKOLA VEZETŐ

Prof. Dr. Tóth Tibor

A MŰSZAKI TUDOMÁNY DOKTORA

TUDOMÁNYOS VEZETŐ

Dr. habil. Galántai Aurél

AZ MTA DOKTORA

Miskolc, 2011

TARTALOMJEGYZÉK

1. BEVEZETÉS.....	3
1.1 A DOLGOZAT FELÉPÍTÉSE.....	3
1.2 IRODALMI ÁTTEKINTÉS.....	5
1.3 A KUTATÁS CÉLJA.....	6
2. ÚJ TUDOMÁNYOS EREDMÉNYEK.....	7
2.1 A DISSZERTÁCIÓ RÖVID ÁTTEKINTÉSE.....	7
2.2 TÉZISEK.....	12
2.2.1 ELSŐ TÉZIS.....	12
2.2.2 MÁSODIK. TÉZIS.....	13
2.2.3 HARMADIK TÉZIS.....	14
2.2.4 NEGYEDIK TÉZIS.....	15
2.3 TOVÁBBI TERVEK.....	17
3. AZ ÉRTEKEZÉS TÉMAKÖRÉBEN KÉSZÍTETT SAJÁT PUBLIKÁCIÓK.....	18
4. HIVATKOZOTT IRODALOM.....	19
SUMMARY.....	22

1. BEVEZETÉS

Az értekezés témája kettős. Egyrészt foglalkozik az algoritmusok randomizációjának, véletlenszerűsítésének a kérdésével, leszűkítve azt az algoritmus inputjának randomizálására, másrészt azzal a fontos problémával, amely az implementált algoritmus helyességvizsgálatát, tesztelését érinti. Szemléltetésül a szimmetrikus, pozitív definit mátrixok legnagyobb sajátértékének és egy hozzátartozó sajátvektorának (a sajátaltér egy eleme) meghatározására szolgáló hatvány módszert választottuk. A módszer tanulmányozása révén egyrészt jobban megértettük a módszer lényegét, hatását, melynek révén sikerült újabb tulajdonságokat kimutatni, valamint a helyesség ellenőrzéséhez értékes információkat nyerni. Az értekezésben javasoltak egy része általánosan is alkalmazható, mint vezérely, más része viszont az algoritmusok sokszínűsége miatt nem vihető át, de utat mutat.

A disszertációval kapcsolatos cikkek, publikációk, konferencia előadások, kiadványok, munkák önálló kutatás eredményei, és maga a disszertáció ezek összegzéseként került összeállításra.

1.1. A DOLGOZAT FELÉPÍTÉSE

Az értekezés első, bevezető fejezetében röviden szólunk a témáról, a motivációról, megindokoljuk a hatvány módszernek, mint szemléltető algoritmusnak a kiválasztási okát. Ismertetjük az értekezés szerkezeti felépítését, megemlítjük a hozzá szükséges előzetes ismeretek körét.

A második fejezet az algoritmus és a program fogalmát és viszonyát taglalja olyan szinten, amelyre a továbbiakban szükség van.

A harmadik fejezet témája a programhelyesség. Megemlítjük a programhelyesség formai megközelítését, vizsgálatának módjait és az azokban az alkotó, de hibákat is

elkövethető ember elkerülhetetlen közreműködése miatt előfordulható hibák, pontatlanságok hatását.

A negyedik fejezet tárgyalja az algoritmusok randomizálásának a kérdését. A randomizáció módot ad a programhelyesség vizsgálatára, pontosabban annak tesztelésére, ellenőrzésére statisztikai-információelméleti alapokra helyezkedve. Ebben a fejezetben írjuk le röviden az információelméleti inakkurancia fogalmának felhasználhatóságát. Alátámasztásképpen itt foglaljuk össze ennek a gondolatnak az előzményét, amely lehetővé tette, hogy a Cauchy valószínűség-eloszlás paramétereinek a becslésére új, hatékony, iteratív eljárást adhassunk.

Az értekezés ötödik fejezete és annak alfejezetei foglalkoznak a tulajdonképpeni hatvány módszer tanulmányozásával. Röviden ismertetjük a hatvány módszert, majd a randomizáció egyfajta hatásait vizsgáló Gianna M. Del Corso [17] által írt cikk eredményeit, amely kiinduló pontja lett a disszertációnak. A fejezetben a hatvány módszer iterációi során kapott vektorsorozat viselkedését tanulmányozzuk praktikus vonatkozásban. Konvergencia esetén a sorozat tagjai egyre „közelebb” kerülnek a legnagyobb sajátértékhez tartozó sajátaltérhez. A közelséget a vektor és az altér által bezárt szöggel, pontosabban annak tangensével jellemezzük. Ezen praktikus mérőszám nagyságának valószínűségére, valószínűségi eloszlásfüggvényére adunk éles alsó és felső becsléseket, korlátokat. Az eredményeket konkrét szimulációs számításokon keresztül szemléltetjük. Ezt további eredmények követik. Ilyen a vizsgált mérőszámnak, mint valószínűségi változó eloszlásfüggvényének egy lokális tulajdonsága, amelyre azért volt szükség, mert komoly technikai nehézség lép fel magának az eloszlásfüggvénynek a kiszámításakor, a zárt alakú felírásakor. Érdekes korrelációs kapcsolatok feltárásának eredményei következnek ezután, melyekre elegáns és nagyrészt elemi bizonyítást sikerült adni. A fejezetet a programtesztelésre adott információelméleti megközelítés sugalmazta elv alkalmazása zárja. Az alkalmazás a hatvány módszer egy lépésének a vizsgálata egy helyes, valamint egy helytelen implementáció esetére. Az eredmények MATLAB-ban kifejlesztett programok statisztikai mintákra végzett futtatásainak összegzéseként kapott táblázatokban láthatók.

A disszertáció hatodik fejezete szűkszavúan a kutatás eredményeinek a felsorolását tartalmazza konkrétan hivatkozva az eredményeket megfogalmazó tételekre.

A hetedik fejezet egy angol nyelvű összefoglaló (Summary).

Függeléként csatoltuk a disszertációhoz a valószínűség-számításból és a matematikai statisztikából feltétlenül szükséges fogalmakat, összefüggéseket főként azoknak a helyenként nehezen megjegyezhető formulái miatt.

A disszertáció végére került az irodalomjegyzék, melyet a tesztelési táblázatok elkészítéséhez használt MATLAB program forráslistája követ mellékletként.

1.2. IRODALMI ÁTTEKINTÉS

A témakörnek óriási az irodalma. A disszertáció határterülettel foglalkozik, mert ötvözi az algoritmusok, az algoritmus implementáció-helyesség, a sajátérték, sajátvektor probléma, a randomizáció, az információelmélet, a valószínűség-számítás, a matematikai statisztika, és a szimuláció eredményeit.

Ugyanazon problémának a megoldására több algoritmus is kidolgozható, de ugyanazon algoritmusnak többféle programja, realizálása lehetséges, amelyek akár erősebben is eltérhetnek egymástól. Konkrét, érdekes esetre mutat rá például Margulis cikke [33]. A programhelyesség-ellenőrzés problémáját tárgyalják az [41], [42], [27], [13], [34] könyvek. A felmerülő nehézségek miatt továbbra is rendkívüli jelentőségük van a programhelyesség-ellenőrzés feladatát részben megoldó programteszteléseknek. A tesztfeladatok kidolgozásának problémája felveti a randomizáció kérdését. Ezekről olvashatunk [41], [27], [13] mellett [18]-ban is. Az információelméleti megfontolások, alapozások [31], [15], [23], [25]-ben megtalálhatók. Ezen fogalmakkal a matematikai statisztika eszköztára részint bővíthető, részint pedig uniformizálható. Ezen fogalmak révén eddig

megválaszolatlan kérdésekre is kaphatunk választ [12], [3]. A szinte klasszikusnak számító sajátérték, sajátvektor problémáról olvashatunk [16], [20], [22], [23], [24], [29], [35], [38], [39], [40], [43]-ban. A hatvány módszerről a legfontosabbak megtalálhatók többek között [35], [22], [39]-ben. A módszert sok szempontból tanulmányozták, példaképpen megemlíthető [20], [43], [36], [37], [16], [29], [35], [39], [21]. A randomizációval összekapcsolt változata, amely a disszertáció kiinduló pontja lett, [17]-ben található. Ennek előzménye a [30] cikk. Rokon kérdéseket, bár az inverz hatvány módszerre, tárgyal a [28] cikk. A szimuláció és a randomizáció kérdései [18]-ban technikai oldalról jól össze vannak foglalva. Az integrálásoknál jól szolgált a [19] könyv.

1.3. A KUTATÁS CÉLJA

Hihetetlenül nehéz a programhelyesség-ellenőrzési elveket a gyakorlatban egy valamennyire is összetettebb algoritmus esetén precízen keresztülvinni. Ennek a problémának a megoldása még ma sincs teljes mértékben tisztázva, és az algoritmusok rendkívüli gazdagsága és sokrétűsége, általánossága miatt ez talán nem is várható el. Ugyancsak nagyon gondos tervezést igényel a programtesztelés kidolgozása addig is, amíg nincs meg a megbízható automatizált helyesség-ellenőrzés. Egy elkészült programról azonban mégis nyilatkozni kell, hogy elfogadhatjuk-e helyesnek, vagy sem, tehát döntést kell hoznunk. Bármilyen is a döntésünk, hibázhatunk. A kutatás célja, miként a disszertáció címe is, kettős. Egyrészt a randomizáció viselkedését kívántuk tanulmányozni egy konkrét algoritmuson, hogy ezáltal magának az algoritmusnak a tulajdonságaira következtethessünk. Ezek a tulajdonságok és a kutatásuk során nyert tapasztalatok a programhelyesség-ellenőrzéskor felhasználhatók. Tanulmányaink alanyául a hatvány módszert választottuk, mivel ez a módszer újra előtérbe került éppen az informatika, a WEB-technológiák révén [14], [32]. A randomizáció segítségével és a valószínűség-számítási, matematikai statisztikai, információelméleti fogalmak és eszközök ötvözése révén egységes módszert kívántunk adni a programtesztelés kivitelezésére és a döntéshozatali mechanizmusra.

2. ÚJ TUDOMÁNYOS EREDMÉNYEK

2.1. A DISSZERTÁCIÓ RÖVID ÁTTEKINTÉSE

A disszertáció a programtesztelés egy speciális témakörével foglalkozik, amikor is egy algoritmus számítógépen realizált programjának a helyességével kapcsolatban kell állást foglalnunk, döntést hoznunk. Statisztikai, információelméleti alapú tesztelési elvet javasol, amelynek révén, ha a helyességet bizonyítani nem is tudjuk, de legalább a nem helyesség erős gyanúját elérjük, ha a program valóban nem helyes. Emellett helyes program esetén lehetőleg tekintsük azt helyesnek. Az értekezés megírásakor nem elhanyagolható szempont volt az anyagnak az oktatási folyamatba történő esetleges beillesztésének az igénye.

Az első négy fejezetben a legfontosabb alapfogalmakat taglaljuk. Az algoritmust úgy fogjuk fel, mint egy függvényt, transzformációt, amely valamely $x \in X$ elemhez (input adatahoz) valamely $z \in Z$ elemet (output adatot) rendel hozzá. Egy $A : X \rightarrow Z$ leképezés. A program az absztrakt algoritmus realizálása. A realizálás hardveren történő megvalósítása az implementáció, az algoritmus implementálása. A program elkészítése a specifikáció alapján kell, hogy történjen. Az algoritmusához egy programot rendelünk. A specifikáció maga a leképezés, az input/output elemek egymáshoz rendelése, amit valamilyen módon meg kell adni, le kell írni. Kérdés, hogy a program végül is valóban olyan kiszámítási szabályt ad-e meg, amely az eredeti specifikációt valósítja meg, vagy nem. A specifikáció röviden egy

$$\begin{aligned} &\{x \mid \varphi(x)\} \\ &\{(x, z) \mid \psi(x, z)\} \end{aligned}$$

úgynevezett *előfeltétel* és *utófeltétel* páros, ahol $\varphi(x)$ és $\psi(x, z)$ predikátumok. *Teljesen helyesnek* mondjuk a P programot, ha minden olyan x inputra, amelyre $\varphi(x) = igaz$, a program befejeződik, és a $P(x)$ eredményre teljesül, hogy

$\psi(x, P(x)) = igaz$. Kérdéses, hogy a predikátumok felírása helyes-e, azaz valóban leírja-e a problémát és minden részletre kiterjedő-e, azaz teljes-e.

Különböző módszerek léteznek a programhelyesség ellenőrzésének kivitelezésére. Például a Floyd, a Manna, vagy a Hoare módszer [41], [27], [13] (a felsorolás persze nem teljes), amelyek a program részekre bontása által, logikai állítások közbeszúrása útján, azok igaz mivoltának belátása révén juttat el a végső helyes következtetésig. Ezek a módszerek általában nem kevés munkát igényelnek. Nehéz összeállítani egy bonyolult, vagy „trükkös” programhoz a közbeszúrandó állításokat. A randomizáció erre a problémára próbál statisztikai választ adni azáltal, hogy leszűkíti a kérdést a praktikus programtesztelésre és a tesztkészítés kérdését statisztikai problémával helyettesíti. Az algoritmus, az implementáció X input halmazából valószínűségi mértékteret alakítunk ki egy rajta bevezetett valószínűség-eloszlás révén, vagy rajta értelmezett valószínűségi változó által. A program implementációjának helyességvizsgálatát eloszlások eltéréseinek vizsgálatával helyettesítjük egy helyesnek feltételezett implementáció és a vizsgált esetleg helytelen implementáció outputja eloszlásainak összehasonlításával. Az összehasonlításra az információelméleti entrópia, inakkurancia és diszkrimináló információ fogalmait használjuk. A véletlenített, randomizált ξ inputra az outputok $\xi_p = A(\xi)$, $\eta_k = P(\xi)$, ahol A az algoritmust, P az implementált programot szimbolizálja. A ξ input helyett ténylegesen véletlenszerűen generált statisztikai mintát alkalmazunk. Az inakkurancia formulája az input randomizálása révén ekkor:

$$T(\eta_k \| \xi_p) = \int_R \log \frac{1}{f_{\xi_p}(x)} dF_k(x) = \sum_{i=1}^k \frac{1}{k} \log \frac{1}{f_{\xi_p}(a_i)}$$

Itt a_i jelöli a k elemű a posteriori minta (P outputja a tényleges implementáció esetén) i -dik elemét, η_k az implementáció által szolgáltatott output empirikus eloszlást, ξ_p pedig az a priori (a helyes implementáció esetén várt) eloszlást. Azt kell statisztikailag eldönteni, hogy ez a szám szignifikánsan eltér-e az a priori eloszlás entrópiáját megadó számtól. A fogalmak használhatóságával kapcsolatban

szemléltetésként felidézük a Cauchy eloszlás paraméterbecslésére általunk korábban kidolgozott módszert.

A hatvány módszer tulajdonságait az ötödik fejezetben tárgyaljuk. A módszert az $n \times n$ -es szimmetrikus, pozitív definit \mathbf{A} mátrixok esetére korlátozzuk. A *hatvány módszer* az \mathbf{A} mátrix legnagyobb sajátértékét (jelölésben λ_1) és egy hozzátartozó normált sajátvektor meghatározását célozza meg. Ez az $\mathbf{Ax} = \lambda\mathbf{x}$ egyenlet egy nemtriviális megoldásának megkeresését jelenti, ahol ismeretlenek az \mathbf{x} vektor és a λ skalár. A módszer algoritmusát tömören:

1. Válasszunk egy tetszőleges $\mathbf{x}^{(0)} \neq \mathbf{0}$ vektort.
2. Képezzük az $\mathbf{y}^{(m+1)} = \mathbf{Ax}^{(m)}$, $\mathbf{x}^{(m+1)} = \frac{1}{\|\mathbf{y}^{(m+1)}\|} \cdot \mathbf{y}^{(m+1)}$, $m = 0, 1, 2, \dots$ vektorokat.
3. Megállunk, ha az \mathbf{x} vektorok már gyakorlatilag nem változnak.

A 3. pontban nem konkretizáljuk, hogy a „gyakorlatilag nem változnak” mit jelent, számos praktikus megállási feltétel ismeretes. Például megállunk, ha valamely kicsiny $\varepsilon > 0$ esetén a kettes, azaz az euklideszi normát használva fennáll, hogy

$$\frac{\|\mathbf{Ax}^{(m)} - \rho^{(m)}\mathbf{x}^{(m)}\|_2}{\|\mathbf{x}^{(m)}\|_2} \leq \varepsilon, \quad \text{ahol} \quad \rho^{(m)} = \frac{\mathbf{x}^{(m-1)T} \mathbf{Ax}^{(m-1)}}{\mathbf{x}^{(m-1)T} \mathbf{x}^{(m-1)}} \quad \text{az úgynevezett Rayleigh}$$

hányados [34]. A módszert a tetszőleges helyett véletlen kiinduló vektorral indítjuk. A véletlen az n -dimenziós egységgömb felszínén egyenletes eloszlást jelent, vagy ami a hatvány módszer esetében ezzel egyenértékű, az n -dimenziós térben standard normális eloszlást. A fejezetben alsó és felső becsléseket adunk az iteráció egyes lépéseiben a saját altértől való előírt nagyságúnál nem nagyobb eltéréssel kapcsolatos valószínűségre.

Egy vektort akkor mondunk egy *altér φ -szögekörnyezetéhez tartozónak*, ha a vektor és az altér által bezárt szög kisebb, mint az előírt $\varphi > 0$ szám. A célunk az lesz, hogy kiszámítsuk, vagy becslést adjunk arra, hogy m lépés megtétele után mennyi annak a valószínűsége, hogy a véletlen vektorral indított hatvány módszer a λ_1 alterének az

előírt φ -szögkörnyezetébe tartozik, azaz φ -nél kisebb szöget bezáró vektort szolgáltat. Az itt megfogalmazott problémához hasonló, de bonyolultabb kezelésű esettel foglalkozik Del Corso [17] cikke, amely elindítója lett a jelen vizsgálódásnak. Röviden a cikk lényege a következő. Jelöljük a szöget α -val. A cikk az alábbi módon bevezet egy $e_m^{ran}(A, p)$ *randomizált hibának* nevezett számértéket, mérőszámot: $e_m^{ran}(A, p) = \left[\int_{S_n} |\sin \alpha^{(m)}(x)|^p \mu(dx) \right]^{1/p}$, ahol S_n az egységgömb felszíne, μ a valószínűségi mérték, m az iterációs index, n a mátrix mérete, dimenziószám. és $1 \leq p \leq \infty$. Ez a randomizált hiba egyfajta várható értéke a távolságnak, pontosabban a normalizált iterált vektor és a sajátaltér közötti euklideszi távolságnak, mint az S_n -nen értelmezett valószínűségi változónak az L_p normája. A szerző ennek a viselkedését vizsgálja erős analitikus eszközök bevetésével. Kimutatja, hogy ez a hiba mindenképpen függeni fog a sajátértékek eloszlásától, főként a legnagyobb és a második legnagyobb sajátérték hányadosától. A klasszikus numerikus irodalomban ez a tény az 50-es évek óta ismert más analitikus megfontolások alapján. A cikk különböző formulákat mutat be esetszétválasztással a sajátérték multiplicitások, iterációs szám, p érték és a dimenziószám függvényében a randomizált hiba alsó és felső korlátaira. Megmutatja, hogy nem érhető el éles becslés a sajátértékek hányadosától függetlenül.

Mi a disszertációban azonban nem várható értékekre, hanem praktikusán valószínűségekre vagyunk kíváncsiak. Jelölje k a legnagyobb sajátérték multiplicitását ($k < n$), és legyen $\mu_{k+1} = \frac{\lambda_{k+1}}{\lambda_1}$ a második legnagyobb és a legnagyobb sajátérték hányadosa. A hatvány módszer szerinti transzformáció m lépése után az $\alpha^{(m)}$ szög tangense:

$$\tan(\alpha^{(m)}) = \frac{\sqrt{\mu_{k+1}^{2m} x_{k+1}^{(0)2} + \mu_{k+2}^{2m} x_{k+2}^{(0)2} + \dots + \mu_n^{2m} x_n^{(0)2}}}{\sqrt{x_1^{(0)2} + x_2^{(0)2} + \dots + x_k^{(0)2}}}$$

A tangens függvény használata azért célszerű, mert egészen kis szögek esetén gyakorlatilag megegyezik a szöggértékkel, nagyobb ugyan annál, de nagyságrendileg

azonos, másrészt felülről korlátozza a szög szinuszát, amely az euklideszi távolság értéke. Vezessük be az

$$F_{n,k}^{(m)}(x) \stackrel{\text{def}}{=} P(\tan(\alpha^{(m)}) < x)$$

jelölést. Ez a függvény a szög tangensének valószínűségi eloszlásfüggvénye. A tárgyalás során az $F_{n,k}^{(m)}(x)$ -re adunk alsó és felső éles becsléseket a statisztikából ismert $(n-k, k)$ szabadságfokú F -eloszlás négyzetgyöke és a k szabadságfokú t -eloszlás abszolút értékének eloszlásfüggvényei által. Sztochasztikus konvergencia bizonyítást adunk a hatvány módszer algoritmusára. Szimulációs futtatásokon keresztül kis dimenziószám esetére szemléltetjük a konkrétan kiszámított formulák viselkedését. Az $F_{n,k}^{(m)}(x)$ eloszlásfüggvény kiszámítása technikailag nagyon nehézkes, ezért kis szögértékekre megadjuk az $F_{n,k}^{(m)}(x)$ függvény aszimptotikus növekedési rendjét, amely polinomiálisnak bizonyul a dimenziószám és a multiplicitás függvényében. Ezen eredményeket követi az iterált vektorok koordinátái közötti korrelációs kapcsolatok tanulmányozása. Külön bizonyítást adunk a kétdimenziós esetre és az n -dimenziós esetre. Megállapítást nyer mindkét esetben, hogy a páronkénti korrelációk a $-1, 0, 1$ számok valamelyikéhez konvergálnak az első iterációs lépés eredményétől függően, amennyiben a legnagyobb sajátvektor multiplicitása egy. Kétdimenziós esetben még a konvergencia sebességének linearitását is sikerül kimutatni. A bizonyításból az is kiderül, hogy ha a jelzett típusú konvergencia nem észlelhető, akkor biztosan nem egy a multiplicitás. A fejezet végén egy programtesztelés eredményét mutatjuk be, amelyben a mátrixszorzás algoritmusát helyesen implementáltuk, összehasonlítva azzal, amikor hibásan implementáltuk. A hiba jellege az, hogy a sorok, oszlopok skaláris szorzatainak kiszámításánál a rekesz nullázását a legelső nullázás kivételével elhagyjuk, a nullázást rossz helyre tesszük a kódban. Tipikus sorcsere a kódban programhiba. A teszteléskor a hiba milyenségét természetesen nem vehetjük és nem is vesszük figyelembe. A nagymintás statisztikai próba következtében a futtatási eredmények révén kapott döntésekkel zárjuk a fejezete, amelynek számértékei meggyőzőek. A számításokhoz szükséges formulák egyike a helyes implementáció esetén az output entrópiája

$$\begin{aligned}
H(\eta) &= \int_{R^m} p(\mathbf{x}) \log \frac{1}{p(\mathbf{x})} d\mathbf{x} = \int_{R^m} p(\mathbf{x}) \left(\log \left(\sqrt{(2\pi)^n \cdot \det(\mathbf{S})} \cdot e^{\frac{1}{2}\mathbf{x}^T \mathbf{S}^{-1} \mathbf{x}} \right) \right) d\mathbf{x} = \\
&= \frac{1}{2} [n \cdot \log(2\pi) + \log \det(\mathbf{S}) + n],
\end{aligned}$$

ahol $\mathbf{S} = \mathbf{A}\mathbf{A}^T$, valamint a helyes és a helytelen implementáció outputjai közötti inakkurancia becslése

$$\hat{T}(\eta_q \| \eta) = \sum_{i=1}^k \frac{1}{k} \cdot \log \frac{1}{p(\mathbf{x}_i)} = \frac{1}{k} \cdot \sum_{i=1}^k \log \frac{1}{p(\mathbf{x}_i)} = \frac{1}{k} \cdot \sum_{i=1}^k \log \left(\sqrt{(2\pi)^n \cdot \det(\mathbf{S})} \cdot e^{\frac{1}{2}\mathbf{x}_i^T \mathbf{S}^{-1} \mathbf{x}_i} \right),$$

ahol \mathbf{x}_i az implementáció outputja az input mintának megfelelően.

Az eredmények összefoglalása a hatodik, az angol nyelvű összefoglalás (Summary) a hetedik fejezetbe került. Függelék szól a fontos statisztikai alapformulákról. Ezt irodalomjegyzék követi, majd pedig csatoltuk a tesztelést végző program MATLAB-os forráskódját.

2.2. TÉZISEK

A kutatások eredményeit négy tézisben foglaltuk össze. A eredmények magyar, orosz és angol nyelven kerültek publikálásra. Impakt faktorról rendelkező nemzetközi folyóirat is van közöttük.

2.2.1. ELSŐ TÉZIS

A hatvány módszer randomizálása esetén valószínűségi alsó és felső becslés adható az iterációs vektor és a saját altér közötti szögére. A becslések élesek olyan értelemben, hogy van eset, amikor egyenlőség áll fenn. Belőlük következik a módszer sztochasztikus konvergenciája.

Ezen tézis alapja a disszertáció 5. fejezetében általam bebizonyított 5.3.2., 5.3.3. és az 5.3.4. tétel, amelyek megfogalmazásai alább következnek.

5.3.2. Tétel: *Alsó és felső valószínűségi becslés*

Annak a valószínűségére, hogy a szög tangense az m -dik lépésben már ε alatt marad, fennáll az $F_{n,k}^{(0)}(\varepsilon / \mu_{k+1}^m) \leq F_{n,k}^{(m)}(\varepsilon) \leq F_{|t_k|}(\varepsilon / c_{n,k}^{(m)})$

egyenlőtlenség, ahol $c_{n,k}^{(m)} = \sqrt{\frac{\sum_{j=1}^{n-k} \mu_{k+j}^{2m}}{k \cdot (n-k)}}$.

5.3.3. Tétel: *A becslés élességének jellemzése*

Ha csak kétféle sajátérték van, azaz $\lambda_{k+1} = \lambda_{k+2} = \dots = \lambda_n$ is fennáll, akkor az 5.3.2. tételbeli formulában mindkét egyenlőtlenség helyén egyenlőség van.

5.3.4. Tétel: *A hatvány módszer sztochasztikus konvergenciája*

A hatvány módszer által szolgáltatott vektorok sorozata standard normális eloszlású kezdővektor esetén sztochasztikusan konvergál a legnagyobb sajátérték saját alterének valamely vektorához. Konvergencia esetén a sajátvektor rajta van a kezdővektornak a saját altérre vett merőleges vetületvektora hatásvonalán, továbbá a legnagyobb sajátérték előáll az iteráció során tapasztalt nyújtási tényezők mértani közepe limeszeként.

Az ezzel kapcsolatos kutatások eredményeit a [4], [5], [6], [8] cikkekben publikáltam, konferencián ismertettem.

2.2.2. MÁSODIK TÉZIS

Az iterációs vektor és az altér közti szög eloszlásfüggvényének növekedési rendjét kicsiny szögekre polinomiális O (ordo) növekedés erőly jellemzi.

Ezen eredmény a disszertációnak az 5.5. alfejezetben található 5.5.1. tételként, amely így szól:

5.5.1. Tétel: Az $F_{n,k}^{(m)}(x)$ eloszlásfüggvény lokális tulajdonsága

$$\text{Kis } \varepsilon > 0 \text{ szögértékek esetén } F_{n,k}^{(m)}(\varepsilon) = O(\varepsilon^{n-k}).$$

Publikálása a [7]-ben történt meg.

2.2.3. HARMADIK TÉZIS

Az iterációs véletlen vektor koordinátáinak korrelációs együtthatói 0-hoz vagy +/-1-hez konvergálnak, ha λ_1 multiplicitása egyszeres. Kétdimenzióban a konvergencia sebessége lineáris. Ha ez a fajta konvergencia nem lép fel, akkor λ_1 multiplicitása nem egyszeres.

Ezen tézis megalapozása a disszertáció 5.6. fejezetében található 5.6.1., 5.6.2. tételként és ez utóbbihoz fűzött megjegyzésként megfogalmazva, amint az alább következik.

5.6.1. Tétel: *A koordináták korrelációjának konvergenciája*

Legyen $r^{(m)}$ a kétdimenziós esetben az iterált vektor koordinátái közötti korrelációs együttható az iteráció m -dik lépésében. Akkor

$$\lim_{k \rightarrow \infty} r^{(m)} = \text{sign}(r^{(1)}).$$

A konvergencia sebessége lineáris.

5.6.2. Tétel: *A koordináták korrelációjának konvergenciája n -dimenzióban*

A hatvány módszer iterációja során n -dimenzióban elhagyva a normalizálást a legnagyobb sajátérték egyszeres multiplicitása esetén a korrelációs $\mathbf{R}^{(m)}$ mátrixra és a páronkénti korrelációs $r_{ij}^{(m)}$ együtthatókra

teljesül, hogy $\text{rank}(\lim_{m \rightarrow \infty} \mathbf{R}^{(m)}) = 1$ és $\lim_{m \rightarrow \infty} r_{ij}^{(m)} = a$, ahol $a^3 = a$, $a \in \mathbf{R}$.

Az eredmények a [9], [11] publikációkban jelentek meg.

2.2.4. NEGYEDIK TÉZIS

A diszkrimináló információ (Kullback eltérés, I-divergencia) fogalomrendszere alkalmas lehet programhelyesség tesztelésre.

A tézis alapjául szolgálnak a Cauchy eloszlás paraméterbecslésével kapcsolatos eredmények. A programhelyesség tesztelésére vonatkozó gondolatok, eszmefuttatások a disszertációnak részint a negyedik fejezetében, a konkrét alkalmazás pedig az 5.7. fejezetben található, tételként nincs külön megfogalmazva, mivel itt egy elvről és annak realizálásáról van szó. A Cauchy eloszlás paraméterbecslésével kapcsolatos eredmények természetesen tétel formájában is testet öltöttek.

Jelöljük azt a tényt, hogy egy ξ valószínűségi változó Cauchy eloszlású (c, s) paraméterekkel a következő módon $\xi \sim C(c, s)$. A Cauchy eloszlás esetén az inakkurancia alakja $(T(c, s))$ jelölje röviden az inakkuranciát):

$T(c, s) \equiv \sum_{i=1}^n \frac{1}{n} \log \left\{ \pi s \left[1 + \left((a_i - c) / s \right)^2 \right] \right\}$. Vezessük be az alábbi jelöléseket. A

c, s paramétereket ismeretlennek, de fixnek feltételezzük, a mintaelemeket pedig $a_i, i = 1, \dots, n$ jelenti. További jelölések:

$$\begin{aligned} u_i &= (a_i - c) / s, & u_{ik} &= (a_i - c_k) / s_k, \quad i = 1, \dots, n \\ e_0 &= \frac{1}{n} \sum_{i=1}^n \frac{1}{1 + u_i^2}, & e_{0k} &= \frac{1}{n} \sum_{i=1}^n \frac{1}{1 + u_{ik}^2} \\ e_1 &= \frac{1}{n} \sum_{i=1}^n \frac{u_i}{1 + u_i^2}, & e_{1k} &= \frac{1}{n} \sum_{i=1}^n \frac{u_{ik}}{1 + u_{ik}^2} \\ e_2 &= \frac{1}{n} \sum_{i=1}^n \frac{u_i^2}{1 + u_i^2}, & e_{2k} &= \frac{1}{n} \sum_{i=1}^n \frac{u_{ik}^2}{1 + u_{ik}^2}. \end{aligned}$$

Lemma a Cauchy eloszlások közötti inakkuranciáról: Legyen $\xi_1 \sim C(c_1, s_1)$,

$\xi_2 \sim C(c_2, s_2)$, akkor $T(\xi_1 \parallel \xi_2) = \log \frac{\pi \left[(s_1 + s_2)^2 + (c_1 - c_2)^2 \right]}{s_2}$.

Lemma a Cauchy inakkurancia minimumhely jellemzésére: $T(c, s)$ minimumhelyén

$$\text{teljesül, hogy: } c = c + s \cdot \frac{e_1}{e_0}, \quad s^2 = s^2 \cdot \frac{e_2}{e_0} = s^2 \cdot \left(\frac{1}{e_0} - 1 \right).$$

A lemma alapján ezen egyenletrendszer megoldására az alábbi iterációs formulapár

$$\text{javasolható: } c_{k+1} = c_k + s_k \cdot \frac{e_{1k}}{e_{0k}}, \quad s_{k+1} = s_k \sqrt{\frac{1}{e_{0k}} - 1}, \quad k = 0, 1, 2, \dots,$$

ahol c_0 tetszőleges, s_0 pozitív valós szám.

Tétel a konvergenciáról: Az előzőben felírt iterációs formulák szerinti számsorozatok nagy minta esetén a valódi paraméter értékekhez konvergálnak. A konvergencia sebessége geometriaként jellemezhető.

Tétel a pontos megoldásról: Nagy minta esetén tetszőleges c_k -ra és $s_k > 0$ -ra fennáll:

$$c = c_k + s_k \cdot \frac{v_{1k}}{v_{0k}^2 + v_{1k}^2}$$

$$s = s_k \cdot \left(\frac{v_{0k}}{v_{0k}^2 + v_{1k}^2} - 1 \right), \quad k = 0, 1, 2, \dots$$

Tétel a minimumhely egyértelműségéről: Nagy minta esetén a minimumhely egyértelmű.

A tézishoz kapcsolódó eredmények publikálása a [1], [2], [3], [10], [12]-ben történt meg.

2.3. TOVÁBBI TERVEK

A disszertációban összegzett kutatások egyáltalán nem befejezett, kerek, lezárt témák. Számptalan kérdés maradt nyitva. Csak néhányat említünk ezek közül. A hatvány módszerrel kapcsolatban kevés hiányzik az 5.6.2. tételben azon sejtés bizonyításához, hogy a konvergencia sebessége n -dimenzióban is lineáris. Másik idevágó probléma: algoritmus kidolgozása arra, hogy amennyiben az 5.6.2. tételben nincs meg az ott leírt típusú konvergencia, azaz a legnagyobb sajátérték nem egyszeres multiplicitású, akkor a megadandó algoritmus állapítsa meg a multiplicitás értékét.

Harmadik lehetőség lehet az, hogy válogassuk ki azokat az eredményeket, amelyek esetleg érvényesek olyan mátrixokra, amelyek nem feltétlenül szimmetrikusak és pozitív definiték.

Természetesen a disszertációban leírt szemléletmódnak megfelelően más algoritmusokat is lehetne példaként vizsgálni.

A programhelyesség tesztelése terén rengeteg a nyitott kérdés. Tovább lehet lépni abba az irányba, hogy az input eloszlástól való függést, az input eloszlás megválasztásának a hatását vizsgáljuk. Tovább lehet fejleszteni, finomítani a teszt elvégzését követő döntési rendszert, a statisztikai próbát. A teszt elvégezhetőségéhez szükséges konvergencia feltételek vizsgálata akár új problémakörként is felléphet.

3. AZ ÉRTEKEZÉS TÉMAKÖRÉBEN KÉSZÍTETT SAJÁT PUBLIKÁCIÓK

1. Nagy Ferenc: *Ocenyivanyije parametrov raszpregyelenyija Kosi*, Third Conference of Program Designers, Budapest, 1987. pp: 231-238
2. Nagy Ferenc: *Az egydimenziós Cauchy eloszlás és paraméterbecslése*, NME, Matematikai Intézet Miskolc, 1988, (Az EGPO-59/86 állami szerződéses munka keretében. Témavezető: dr Klafszyk Emil)
3. Nagy Ferenc : *A Cauchy eloszlás paraméterbecslése*, Egyetemi doktori értekezés. Miskolci Egyetem, 1990.
4. Nagy Ferenc: *Valószínűségi becslések a hatvány módszer randomizálásában*, Doktori Fórum 2003. nov. 6. Miskolci Egyetem
5. Nagy Ferenc: *Two-sided probability bound estimates in the randomized power method*, MicroCAD 2004 Conference, Miskolc
6. Nagy Ferenc: *Egy eloszlásfüggvény kiszámításáról*, Doktori Fórum 2004. nov. 4-9, pp. 191-204, Miskolci Egyetem
7. Nagy Ferenc: *Egy eloszlásfüggvény lokális viselkedéséről*, Doktori Fórum 2005. nov. 9, pp. 191-204, gépész, pp. 154-156, Miskolci Egyetem
8. Nagy Ferenc: *Valószínűségi becslések sajátérték meghatározásakor*, Gazdaságmodellezési Szakértői Konferencia, Velence, 2005. jan. 27-29.
9. Nagy Ferenc: *An elementary proof for the correlation of the coordinates in the two dimensional power method*, MicroCAD 2005 Conference, Miskolc, 2005. 03. 10-11.
10. Nagy Ferenc: *A Cauchy eloszlás paraméterbecslése információelméleti megközelítésben*, IF2005 Konferencia, Debrecen, 2005. 08. 24-26. Abstract p. 175. CD teljes szöveg (**ISBN: 963 472 909 6**)
11. Nagy Ferenc: *Correlation relations in the randomized power method*, MicroCAD 2006 Conference, Miskolc, 2006. 03. 16-17, G. szekció, pp. 79-81.
12. Nagy Ferenc: *Parameter Estimation of the Cauchy Distribution in Information Theory Approach*, Journal of Universal Computer Science, vol. **12** No. **9** (2006) pp. 1332-1344. (impakt faktor: 0.338)

4. HIVATKOZOTT IRODALOM

13. ALAGIĆ, Suad; ARBIB, Michael A.: *The Design of Well Structured and Correct Programs*, Springer-Verlag, New York Heidelberg Berlin, 1978
14. BERRY, Michael W.; BROWNE, Murray: *Understanding Search Engines: Mathematical Modeling and Text Retrieval*, SIAM, Philadelphia, 1999
15. BLAHUT, Richard E.: *Principles and Practice of Information Theory*, Addison-Wesley Publishing Company, Reading, Massachusetts, 1987
16. CHATELIN, Françoise: *Eigenvalues of Matrices*, John Wiley and Sons, New York, 1993
17. Del CORSO, Gianna M.: *Estimating an eigenvector by the power method with a random start*, SIAM J. Matrix Anal. Appl., **18**, No. 4. pp. 913-937, 1997
18. DEÁK ISTVÁN: *Random number generators and simulation*, Akadémiai Kiadó, Budapest, 1990
19. DWIGHT, H. B.: *Tables of integrals and other mathematical data*, The MacMillan Company, New York, 1961
20. FADDEEV, D. K. and FADDEEVA, V. N.: *Computational Methods of Linear Algebra*, W. H. Freeman and Co., San Francisco, 1963
21. FRIEDMAN, J.: *Error bounds on the power method for determining the largest eigenvalue of symmetric, positive definite matrix*, Linear Algebra and its Applications, **280** (1998) pp. 199-216
22. GALÁNTAI AURÉL, JENEY ANDRÁS: *Numerikus módszerek*, Miskolci Egyetem Kiadó, Miskolc, 1998
23. GANTMAHER, F. R.: *Tyeorija matric*, Nauka, Moszkva, 1967
24. GELFAND, I. M.: *Lekcii po linyejnoj algebre*, Nauka, Moszkva, 1971
25. GOLDMAN, S.: *Information Theory*, Prentice Hall Inc., New York, 1953
26. GRAY, R. M.: *Entropy and Information Theory*, Springer Verlag, New York, Berlin, 1990
27. GRIES, David: *The Science of Programming*, Springer-Verlag, New York, Heidelberg, Berlin, 1981

28. JESSUP, E. R.: *A Statistical Evaluation of Inverse Iteration*, NATO ASI Series, Vol. F70, Numerical Linear Algebra, Digital Signal Processing and Parallel Algorithms, Edited by G. H. Golub and P. Van Dooren, Springer-Verlag Berlin Heidelberg, 1991
29. KRÜLOV, V. I.; BOBKOV, V. V.; MONASZTÜRNÜJ, P. I.: *Vücsiszlityelnüje metodü vüszsej matyematyiki*, tom 1, Vüsejsaja skola, Minszk, 1972
30. KUCZYŃSKI, J.; WOZŃIAKOWSKI, H.: *Estimating the largest eigenvalue by the power and Lanczos algorithms with a random start*; SIAM J. Matrix Anal. Appl., **13**(1992), pp. 1094-1122
31. KULLBACK, S.: *Information Theory and Statistics*, John Wiley & Sons. Inc. New York, 1959
32. LANGVILLE, A. N. and MEYER, C. D.: *Understanding Web Search Engine Rankings: Google's PageRank, Teoma's HITS and other ranking algorithms*, Princetown University Press, Princetown, 2005
33. MARGULIS, L. F.: *O szvjazi algoritmov Zl i prjamogo sztyepennogo metoda*, Dokl. Akad. Nauk Tadz. SSR. **32** No **4**. (1989) pp. 218-222
34. OLIVEIRA, S. & STEWART, D.: *Writing Scientific Software: A Guide to Good Style*, Cambridge University Press, New York, 2006
35. PARLETT, B. N.: *The Symmetric Eigenvalue Problem*, Prentice-Hall, Englewood Cliffs, N. J. 1980.
36. PITUK MIHÁLY: *Nonnegative iterations with asymptotically constant coefficients*, Linear Algebra and Its Applications **431**(2009) pp 1815-1824
37. PITUK MIHÁLY: *A link between the Perron-Frobenius theorem and Perron's theorem for difference equations*, Linear Algebra and Its Applications **434**(2011) pp 490-500
38. RAO, C. R.: *Linear statistical inference and its applications*, John Wiley and Sons Inc., New York-London-Sidney, 1965
39. STOYAN G., TAKÓ G.: *Numerikus módszerek I*; ELTE-TypoTex, Budapest, 1993
40. UEBERHUBER, Ch. W.: *Numerical Computation, Method, Software, and Analysis*, Springer-Verlag, Berlin, 1997

41. VARGA LÁSZLÓ, *Rendszerprogramok elmélete és gyakorlata*, Akadémiai Kiadó, Budapest, 1978
42. VARGA LÁSZLÓ, *Programok analízise és szintézise*, Akadémiai Kiadó, Budapest, 1981
43. WILKINSON, J. H.: *The Algebraic Eigenvalue Problem*, Oxford, Clarendon Press, 1988

SUMMARY

This PhD dissertation deals with the problem of randomization of algorithms and program testing. It contains some new results, especially in connection with the iterative power method for computing the largest eigenvalue of the positive definite, symmetric matrices. The input of the algorithm of the power method is randomized by generating standard, normally distributed random vectors in the n -dimensional space.

The behavior of the output of the algorithm is analysed. The first four sections play role of introductory character of the notion of algorithm, program, correctness of program, randomization of algorithms. The main results are discussed in the fifth section and its subsections. New scientific results are in the following theses.

Thesis 1: Lower and upper bounds can be formulated for the angle between the iterated vector and the eigensubspace in the randomized power method. Bounds are sharp, that is, there are cases of equality. The stochastic convergence of the method can be proved by these formulas.

Thesis 2: Assume small angle between the iterated vector and the eigensubspace. In this case the order of the growth of the distribution function of the angle can be characterized as polynomial in *ordo* sense.

Thesis 3: Assuming that the multiplicity of the largest eigenvalue is one, the correlation coefficients of the coordinates of the iterated vector converge to -1, 0, or +1. In two dimensional case the rate of convergence is linear. If this kind of convergence can not be observed then the multiplicity of the largest eigenvalue is not one.

Thesis 4: The system of the notions of the discrimination information (Kullback's divergence, I-divergence) can be suitable for testing of program correctness.

