

# Urnamodellek és alkalmazásai

Szakdolgozat

Juhász Balázs

Matematika BSc

Alkalmazott matematikus szakirány

Témavezető:

Móri Tamás

Valószínűségelméleti és Statisztika Tanszék

Eötvös Loránd Tudományegyetem, Természettudományi Kar



Eötvös Loránd Tudományegyetem

Természettudományi Kar

2017

# Tartalomjegyzék

<b>Ábrák jegyzéke</b>	<b>2</b>
<b>1. Bevezetés</b>	<b>4</b>
<b>2. Urnamodellek két szín esetén</b>	<b>6</b>
2.1. Pólya–Eggenberger urna . . . . .	6
2.2. Bagchi–Pal urna . . . . .	12
<b>3. Urnamodellek <math>m</math> szín esetén</b>	<b>17</b>
<b>4. Urnamodellek alkalmazása az informatikában</b>	<b>25</b>
4.1. Bináris fa . . . . .	25
4.2. Perem-kiegyensúlyozott fa . . . . .	28
4.3. $m$ -áris fa . . . . .	30
<b>Irodalomjegyzék</b>	<b>42</b>

## Ábrák jegyzéke

1.	Pólya-Eggenberger urna az (1,1) ill. a (2,1) helyzetből indítva . . . . .	12
2.	A 6, 8, 2, 5, 1, 7, 3, 4, 10, 9 csúcsok beillesztése utáni állapot. A színes csúcsok felelnek meg az urnában található golyóknak, amik tehát a lehetséges beszúrási helyeket reprezentálják. . . . .	26
3.	Véletlen növény bináris fa négy szimulációja 3000 behelyezett elemig. Az $y$ tengely megmutatja, hogy mennyivel tér el a levelek száma az összes elem harmadától. A piros görbe az elméleti szórást ábrázolja. . . . .	28
4.	Csúcs beillesztése a perem-kiegyensúlyozott bináris keresőfába. . . . .	29
5.	A (7, 5, 2, 16, 3, 1, 11, 9, 12, 4, 14, 6, 8, 10) elemek behelyezése utáni állapot, majd a (13) beszúrása az $m$ -áris fába, ahol $m = 4$ . . . . .	30
6.	A (7, 5, 2, 16, 3, 1, 11, 9, 12, 4, 14, 6, 8, 10, 13) elemek behelyezése utáni állapot ( $m = 4$ ), az urnamodell reprezentálása színes golyókkal. . . . .	31
7.	Az $m$ -áris fa reprezentációs mátrix sajátértékeinek elhelyezkedése a komplex számsíkon, különféle $m$ értékek esetén. . . . .	34
8.	A második legnagyobb valós részű sajátérték $m$ függvényében. . . . .	35
9.	Véletlen növény $m$ -áris fa csúcsainak száma osztva a behelyezett elemek számával, különféle $m$ értékek esetén. Mindegyik esetben 10 kísérlet eredménye látható. Egy kísérlet $n = 10^5$ fába helyezett elemig tart. A szaggatott vonal a 4.8. Tételben szereplő $\frac{1}{2(H_m - 1)}$ mennyiséget jelzi. . . . .	39
10.	Az $\frac{S_n - \frac{n}{2(H_m - 1)}}{n^{\sigma_2}}$ mennyiség ábrázolva $\log n$ függvényében. Három kísérlet látható, $n = 10000000 = 10^7$ behelyezett elemig. Megfigyelhető az amplitúdóbeli randomitás, illetve a fáziseltolódás az egyes kísérletek között. . . . .	41

# Köszönetnyilvánítás

Szeretnék köszönetet mondani témavezetőmnek, Móri Tamásnak, a szakdolgozat írása közben felmerülő kérdéseim hiánytalan megválaszolásáért, a konzultációk során mondott javaslatiért, tanácsaiért, illetve a dolgozat alapos átnézéséért.

Szeretném megköszönni Varga Lászlónak a téma felvetését, mely két számomra kedves tudományágot érint, a valószínűségszámítást, illetve az informatikai adatszerkezeteket.

Emellett köszönöm szüleimnek, családomnak, barátaimnak a folyamatos támogatást, melyet egész életem alatt kaptam és kapok a mai napig.

# 1. Bevezetés

Felmerülhet bennünk a kérdés, hogy vajon miért éppen a Facebook a legelterjedtebb közösségi hálózat szerte a világon. Biztosak lehetünk benne, hogy amikor elindult, számos más, teljesen hasonló lehetőségeket kínáló szolgáltatás is elérhető volt az interneten. Miért van az, ha több hasonló tulajdonságú termék közül valamelyik sokkal népszerűbbé válik, akkor a többi versengőnek már vajmi kevés esélye van megfordítani ezt az állapotot? Dolgozatom témája – többek között – ilyen és ezekhez hasonló folyamatok modellezését kívánja bemutatni. Az urnamodellek meglepően sokféle különböző területen használhatóak. Közgazdaságtanban, statisztikai eljárásokban, genetikában, az informatika területén különféle algoritmusok illetve adatszerkezetek vizsgálatában és még számos más területen elterjedt módszerek számát véletlen folyamatok modellezésére. Több száz éve élt tudósok, többek között *Huygens*, *de Moivre*, *Laplace* és *Bernoulli* is használták műveikben lényegében a modellt. Napjainkban a Pólya-féle urnamodellel elnevezés kering a köztudatban. Ez annak köszönhető, hogy a *Pólya György (1887-1985)* híres matematikusunk vizsgálódott elsőként komolyabb szinten a témával. Ő egyébként fertőzések vizsgálatára szerette volna alkalmazni a modellt.

Tekintsük a legegyszerűbb változatát. Van egy urnánk, melyben kezdetben egy fehér és egy kék színű golyó van. Húzzunk egy golyót az urnából. Ezután visszatesszük azt, amelyiket kihúztuk, valamint beteszünk az urnába még egy olyan színűt, mint amelyet húztunk. Majd újra húzzunk, és hasonló módon cselekszünk. A modell alapvető feltétele, hogy minden benyúláskor ugyanakkora eséllyel húzzuk ki az összes golyót. Észrevehetjük, hogy a második húzáskor már nagyobb eséllyel húzzunk az egyik színűből, mint a másikkól, hiszen az egyik színből 2, míg a másikkól csak 1 golyó lesz az urnában. Már ezen egyszerű esetre is érdekes eredményt kapunk, ha feltesszük a kérdést, vajon mi lesz a golyók aránya egymillió húzás után. A válasz az, hogy a fehér golyók aránya az összes golyók számához képest az egyenletes eloszláshoz konvergál a  $[0, 1]$  intervallumon. Ez azt jelenti, hogy ugyanakkora eséllyel lesz nagyjából ugyanannyi golyó a két színből, mint hogy az egyikből nagyon sok, a másikkól pedig nagyon kevés.

Számos módon általánosíthatjuk modellünket. Bevezethetjük azt az általánosítást, hogy kezdetben nem  $1 - 1$  golyó van színekből, hanem  $W_0$  fehér és  $B_0$  kék. Ezenkívül egy másik általánosítási lehetőség, hogy egy fehér (ill. kék) golyó húzása esetén hogyan járjunk el, hány golyót adjunk hozzá vagy akár vegyünk ki az egyes színekből. A harmadik általánosítási lehetőség, hogy használhatunk két szín helyett akár többet is.  $m$  szín esetén így egy  $m \times m$ -es  $A$  mátrixsal reprezentálhatjuk a folyamatot, ahol a mátrix  $A_{i,j}$  eleme megmondja, hogy egy  $i$ . színű golyó húzása esetén hány golyót tegyünk az urnába a  $j$ . színből.

A dolgozatban az urnamodellek konvergencia-tulajdonságait szeretném vizsgálni. A mű alapját Hosam Mahmoud *Pólya Urn models* [13] című könyve képezi.

A fentebb említett általánosítási lehetőségek szerint haladunk a dolgozat során, majd pedig néhány alkalmazását vizsgáljuk meg a kapott tételeknek.

A második fejezetben a kétszínű esettel foglalkozunk. Belátjuk, hogy a Pólya-Eggenberger urna esetén a fehér golyós húzások aránya *Beta* eloszláshoz konvergál, melynek paramétereit az urna kezdeti állapota határozza meg. Megemlítjük az általánosabb, Bagchi-Pal urnához kapcsolódó határeloszlás tételt.

A harmadik fejezetben továbblépünk az  $m$  színű esetre. Bevezetünk néhány feltételt az urnákra, illetve az őket reprezentáló mátrixokra, melyek szükségesek a kívánt tételek teljesüléséhez. Érdekes eredményt fogunk kapni a határeloszlás szükségességéhez, mely a mátrix sajátértékeivel lesz kapcsolatos.

A negyedik fejezetben az előző pontban kapott eredményeket használjuk fel különböző adatszerkezetek tulajdonságainak vizsgálatára. Normális határeloszlást kapunk a véletlen növe bináris fában a levelek számára, az AVL fához hasonló perem-kiegyensúlyozott fában a cserék számára, illetve az  $m$ -áris fában a csúcsok számára. Az  $m$ -áris fa esetében érdekes eredményt kapunk, ugyanis a harmadik fejezetben kapott tételek csak az  $m \leq 26$  esetben használhatóak, és valóban, látni fogjuk, hogy  $m > 27$  esetén nem lesz igaz az a tétel, ami  $m \leq 26$  esetén igaznak bizonyult.

A dolgozat során több esetben is szimulációk segítségével kipróbáltam a kapott eredményeket. A szimulációkat Matlab programmal készítettem. A szimulációk forráskódja megtalálható a Függelékben.

## 2. Urnamodellek két szín esetén

### 2.1. Pólya–Eggenberger urna

Először tehát feltesszük, hogy az urnában kétféle színű golyó van, kék és fehér. Elsőként a Pólya-Eggenberger urnával szeretnék foglalkozni. Minden esetben az adott folyamatot a bevezetésben is említett *színszám*  $\times$  *színszám* méretű mátrixszal fogjuk reprezentálni. Ez a Pólya-Eggenberger urna esetében a következő:

$$\mathbf{A} = \begin{pmatrix} s & 0 \\ 0 & s \end{pmatrix}$$

Ez azt jelenti, hogy amennyiben fehér golyót húzunk, akkor  $s$  db fehér golyót teszünk az urnába, ha kéket húzunk, akkor pedig  $s$  db kéket. Ennek tehát speciális esete a bevezetésben említett legegyszerűbb eset, amikor  $s = 1$ .

Bevezetünk néhány jelölést:

- $W_0, B_0$ : a fehér, ill. kék golyók száma kezdetben,
- $W_n, B_n$ : a fehér, ill. kék golyók száma  $n$  húzás után,
- $S_n$ : a golyók száma az urnában  $n$  húzást követően,
- $W_n^*, B_n^*$ : a fehér, ill. kék golyós húzások száma  $n$  húzás után.

Célunk meghatározni, hogyan viselkedik az egyik szín általi húzások aránya, tehát  $\frac{W_n^*}{n}$ , ehhez először azt kell meghatároznunk, hogy mi a valószínűsége annak, hogy  $k$  fehérrel húzunk  $n$  húzásból.

**2.1. Állítás.** *Legyen  $W_n^*$  a fehér húzások száma  $n$  húzást követően a Pólya-Eggenberger urnából. Ekkor*

$$\begin{aligned} P(W_n^* = k) &= \binom{n}{k} \frac{W_0(W_0 + s) \dots (W_0 + (k-1)s) B_0(B_0 + s) \dots (B_0 + (n-k-1)s)}{S_0(S_0 + s) \dots (S_0 + (n-1)s)} \\ &= \binom{n}{k} \frac{\prod_{j=0}^{k-1} (W_0 + js) \prod_{j=0}^{n-k-1} (B_0 + js)}{\prod_{j=0}^{n-1} S_j}. \end{aligned}$$

Az állítás bizonyítása egy érdekes tulajdonságra is rámutat.

**Bizonyítás.** Tekintsünk egy adott húzási sorrendet a színek alapján, ami megfelelő, tehát  $k$  fehér és  $n - k$  kék színű húzást tartalmaz. Jelölje a fehér golyós húzások sorszámait  $1 \leq i_1 < i_2 < \dots < i_k \leq n$ . Írjuk fel annak a valószínűségét, hogy pontosan ez lesz a húzások sorrendje:

$$\begin{aligned}
& \frac{B_0}{S_0} \cdot \frac{B_0 + s}{S_1} \cdot \dots \cdot \frac{B_0 + (i_1 - 2)s}{S_{i_1-2}} \cdot \frac{W_0}{S_{i_1-1}} \\
& \cdot \frac{B_0 + (i_1 - 1)s}{S_{i_1}} \cdot \dots \cdot \frac{B_0 + (i_2 - 3)s}{S_{i_2-3}} \cdot \frac{W_0 + s}{S_{i_2-1}} \\
& \cdot \dots \cdot \\
& \cdot \frac{B_0 + (i_{k-1} - (k - 1))s}{S_{i_{k-1}}} \cdot \dots \cdot \frac{B_0 + (i_k - k - 1)s}{S_{i_k-2}} \cdot \frac{W_0 + (k - 1)s}{S_{i_k-1}} \\
& \cdot \frac{B_0 + (i_k - k)s}{S_{i_k}} \cdot \dots \cdot \frac{B_0 + (n - k - 1)s}{S_{n-1}} = \\
& = \frac{\prod_{j=0}^{n-k-1} (B_0 + js) \prod_{j=0}^{k-1} (W_0 + js)}{\prod_{j=0}^{n-1} S_j}.
\end{aligned}$$

Az egyenlet soronként egy-egy kihúzott fehér golyók közti kék golyókat mutatja. Ahogy az egyenletből is látszik, ez a valószínűség nem függ az  $i_1, i_2, \dots, i_k$  indexektől, tehát feltéve, hogy  $k$  fehéret húzunk az első  $n$  kísérletből, bármely  $k$  elemű kombinációnak egyenlő esélye van, hogy ők legyenek a fehér golyók. Mivel  $\binom{n}{k}$ -féleképpen tudjuk kiválasztani a  $k$  db indexet, ezért az állítást beláttuk. ■

Ezután rátérhetünk a várható értékre.

**2.2. Állítás.** Legyen  $W_n^*$  a fehér húzások száma  $n$  húzást követően a Pólya-Eggenberger urnából. Ekkor

$$\mathbb{E}[W_n^*] = \frac{W_0}{S_0} n.$$



**Bizonyítás.** A várható érték definícióját felírva:

$$\mathbb{E}[W_n^*] = \sum_{k=0}^{\infty} k P[W_n^* = k] = \sum_{k=0}^{\infty} k \binom{n}{k} \frac{\prod_{j=0}^{k-1} (W_0 + js) \prod_{j=0}^{n-k-1} (B_0 + js)}{\prod_{j=0}^{n-1} S_j}.$$

Használjuk fel a következő, jól ismert azonosságot:

$$k \binom{n}{k} = n \binom{n-1}{k-1}.$$

Behelyettesítve az egyenletbe, illetve a nevezőt picit átalakítva:

$$\begin{aligned} \mathbb{E}[W_n^*] &= n \sum_{k=1}^n \binom{n-1}{k-1} \frac{\prod_{j=0}^{k-1} (W_0 + js) \prod_{j=0}^{n-k-1} (B_0 + js)}{\prod_{j=0}^{n-1} (S_0 + js)} \\ &= n \frac{W_0}{S_0} \sum_{k=1}^n \binom{n-1}{k-1} \frac{\prod_{j=1}^{k-1} (W_0 + js) \prod_{j=0}^{n-k-1} (B_0 + js)}{\prod_{j=1}^{n-1} (S_0 + js)} \\ &= n \frac{W_0}{S_0} \sum_{l=0}^{n-1} \binom{n-1}{l} \frac{\prod_{j=0}^{l-1} ((W_0 + s) + js) \prod_{j=0}^{(n-1)-l-1} (B_0 + js)}{\prod_{j=0}^{n-2} ((S_0 + s) + js)} \end{aligned}$$

Tekintsük az utolsó összeget, illetve azt a kísérletet, amikor  $(n-1)$ -szer húzunk a kezdetben  $W_0 + s$  fehér és  $B_0$  kék golyót tartalmazó Pólya-Eggenberger urnából. Az összegben pontosan a kihúzott fehér golyók száma szerinti valószínűségek vannak összegezve, tehát a szumma értéke 1. Ezzel igazoltuk az állítást. ■

Megemlítjük, hogy ezzel az urnában lévő fehér golyók számának várható értékét is meghatározhatjuk:

$$W_n = sW_n^* + W_0 \implies \mathbb{E}[W_n] = s\mathbb{E}[W_n^*] + W_0 = \frac{W_0}{S_0} sn + W_0$$

Tehát a várható értéket az határozza meg, hogy kezdetben milyen arányban vannak a golyók.

A következőkben a  $\frac{W_n^*}{n}$  határértékét szeretnénk megvizsgálni. Azt tehát már tudjuk, hogy a várható értéke egyenlő  $\frac{W_0}{S_0}$ -al, viszont ennél többet is megtudhatunk.

A dolgozatban rendre fel fogjuk használni a  $\Gamma$  függvény különböző tulajdonságait. Ezeket a következő lemmák tartalmazzák.

**2.3. Lemma. (Stirling-approximáció a  $\Gamma$  függvényre)** *Legyen  $\Gamma$  a szokásos módon definiált gamma függvény,  $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$ . Ekkor  $\forall a, b \in \mathbb{R}$ -re:*

$$\frac{\Gamma(x+a)}{\Gamma(x+b)} = x^{a-b} + O(x^{a-b-1}) \quad (\text{ha } x \rightarrow \infty).$$

**2.4. Lemma. (Alsó, illetve felső becslés a  $\Gamma$  függvényre)** *Legyen  $0 < \lambda < 1$ , ekkor minden  $x > 0$ -ra a következő egyenlőtlenség áll fent:*

$$x^{1-\lambda} \leq \frac{\Gamma(x+1)}{\Gamma(x+\lambda)} \leq (x+1)^{1-\lambda}.$$

A gamma függvények további tulajdonságai, illetve a fenti tulajdonságok részletesebb tárgyalása megtalálható az [11] cikkben.

**2.5. Megjegyzés.** A  $\Gamma$  függvény egy másik hasznos tulajdonságát is sokszor ki fogjuk használni, miszerint:

$$\Gamma(x) = (x-1)\Gamma(x-1) \quad \forall x > 1, x \in \mathbb{R}.$$

Az előkészítések után rátérhetünk a tételre.

**2.6. Tétel. (Eggenberger és Pólya, 1923)** *Legyen  $W_n^*$  a fehér húzások száma  $n$  húzást követően a Pólya-Eggenberger urnából. Ekkor:*

$$\frac{W_n^*}{n} \xrightarrow{d} \beta\left(\frac{W_0}{s}, \frac{B_0}{s}\right),$$

ahol  $\beta$  a Beta eloszlást jelöli.

**Bizonyítás.** Írjuk fel a már többször is felírt valószínűséget, annak az esélyét, hogy  $k$ -szor húzunk fehéret. A szorzatban lévő összes tényezőt osszuk le  $s$ -sel, hogy aztán a megjegyzésben szerepelt formulát többször egymás után használva, illetve a binomiális együtthatót is átalakítva egy olyan formulát kapunk, amelyben csak  $\Gamma$  függvények

szorzata, valamint hányadosa fog maradni:

$$\begin{aligned}
\mathbb{P}(W_n^* = k) &= \binom{n}{k} \frac{\prod_{j=0}^{k-1} (W_0 + js) \prod_{j=0}^{n-k-1} (B_0 + js)}{\prod_{j=0}^{n-1} (S_0 + js)} = \\
&= \frac{n!}{k!(n-k)!} \cdot \frac{\prod_{j=0}^{k-1} \left(\frac{W_0}{s} + j\right) \prod_{j=0}^{n-k-1} \left(\frac{B_0}{s} + j\right)}{\prod_{j=0}^{n-1} \left(\frac{S_0}{s} + j\right)} = \\
&= \frac{\Gamma(n+1)}{\Gamma(k+1)\Gamma(n-k+1)} \cdot \frac{\Gamma(\frac{W_0}{s} + k)}{\Gamma(\frac{W_0}{s})} \cdot \frac{\Gamma(\frac{B_0}{s} + n - k)}{\Gamma(\frac{B_0}{s})} \cdot \frac{\Gamma(\frac{S_0}{s})}{\Gamma(\frac{S_0}{s} + n)}
\end{aligned}$$

Írjuk fel ennek megfelelően  $\frac{W_n^*}{n}$  eloszlásfüggvényét a  $0 \leq x \leq 1$  helyen:

$$\begin{aligned}
\mathbb{P}\left(\frac{W_n^*}{n} \leq x\right) &= \mathbb{P}(W_n^* \leq nx) \\
&= \sum_{k=0}^{\lfloor nx \rfloor} \frac{\Gamma(n+1)}{\Gamma(k+1)\Gamma(n-k+1)} \cdot \frac{\Gamma(\frac{W_0}{s} + k)}{\Gamma(\frac{W_0}{s})} \cdot \frac{\Gamma(\frac{B_0}{s} + n - k)}{\Gamma(\frac{B_0}{s})} \cdot \frac{\Gamma(\frac{S_0}{s})}{\Gamma(\frac{S_0}{s} + n)} \\
&= \frac{\Gamma(\frac{S_0}{s})}{\Gamma(\frac{W_0}{s})\Gamma(\frac{B_0}{s})} \sum_{k=0}^{\lfloor nx \rfloor} \left( \frac{\Gamma(k + \frac{W_0}{s})}{\Gamma(k+1)} \cdot \frac{\Gamma(n - k + \frac{B_0}{s})}{\Gamma(n - k + 1)} \cdot \frac{\Gamma(n+1)}{\Gamma(n + \frac{S_0}{s})} \right)
\end{aligned}$$

Az egyszerűbb számolás érdekében tegyük fel, hogy kezdetben kevés golyó van az urnában, vagyis, hogy  $\frac{B_0+W_0}{s} < 1$  teljesül, így pedig használhatjuk a 2.4. Lemmát, hogy alsó, illetve felső becslést nyerjünk  $\frac{W_n^*}{n}$  eloszlására. (A tétel természetesen a feltevés nélkül is igaz, ekkor egy másik, a 2.4. Lemmához teljesen hasonló egyenlőtlenséget kellene használnunk.) Az alsó, illetve felső becslésről belátjuk, hogy ugyanoda tartanak, így pedig megkapjuk  $\frac{W_n^*}{n}$  eloszlását.

Alsó becslés:

$$\begin{aligned}
& \mathbb{P}\left(\frac{W_n^*}{n} \leq x\right) \\
& \geq \frac{\Gamma\left(\frac{S_0}{n}\right)}{\Gamma\left(\frac{W_0}{s}\right)\Gamma\left(\frac{B_0}{s}\right)} \sum_{k=0}^{\lfloor nx \rfloor} (k+1)^{\frac{W_0}{s}-1} (n-k+1)^{\frac{B_0}{s}-1} n^{1-\frac{B_0+W_0}{s}} \\
& = \frac{\Gamma\left(\frac{S_0}{n}\right)}{\Gamma\left(\frac{W_0}{s}\right)\Gamma\left(\frac{B_0}{s}\right)} \frac{1}{n} \sum_{k=0}^{\lfloor nx \rfloor} \left(\frac{k+1}{n}\right)^{\frac{W_0}{s}-1} \left(\frac{n-k+1}{n}\right)^{\frac{B_0}{s}-1} \\
& = \frac{\Gamma\left(\frac{S_0}{n}\right)}{\Gamma\left(\frac{W_0}{s}\right)\Gamma\left(\frac{B_0}{s}\right)} \left(\frac{n+2}{n}\right)^{\frac{W_0+B_0}{s}-1} \frac{1}{n+2} \sum_{k=0}^{\lfloor nx \rfloor} \left(\frac{k+1}{n+2}\right)^{\frac{W_0}{s}-1} \left(\frac{n-k+1}{n+2}\right)^{\frac{B_0}{s}-1} \\
& \xrightarrow{n \rightarrow \infty} \frac{\Gamma\left(\frac{W_0+B_0}{n}\right)}{\Gamma\left(\frac{W_0}{s}\right)\Gamma\left(\frac{B_0}{s}\right)} \int_0^x t^{\frac{W_0}{s}-1} (1-t)^{\frac{B_0}{s}-1} dt
\end{aligned}$$

A kapott határérték pontosan a  $\beta\left(\frac{W_0}{s}, \frac{B_0}{s}\right)$  eloszlásfüggvénye az  $x$  helyen. Lássuk a másik becslést.

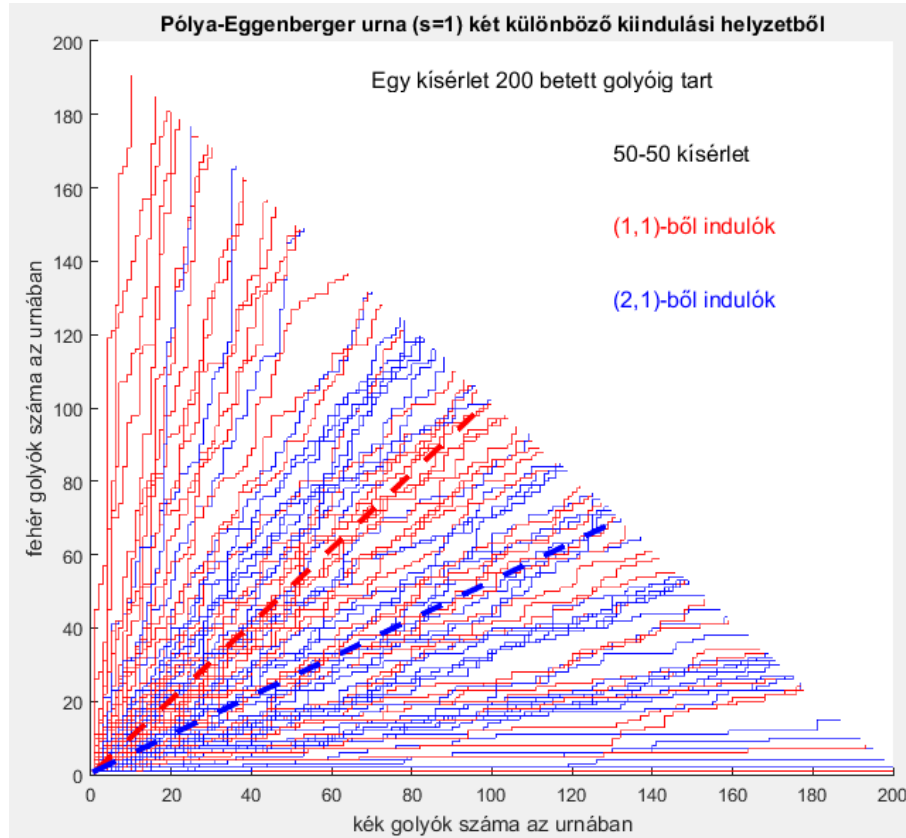
Felső becslés:

$$\begin{aligned}
& \mathbb{P}\left(\frac{W_n^*}{n} \leq x\right) \\
& \leq \frac{\Gamma\left(\frac{S_0}{n}\right)}{\Gamma\left(\frac{W_0}{s}\right)\Gamma\left(\frac{B_0}{s}\right)} \sum_{k=0}^{\lfloor nx \rfloor} k^{\frac{W_0}{s}-1} (n-k)^{\frac{B_0}{s}-1} (n+1)^{1-\frac{B_0+W_0}{s}} \\
& = \frac{\Gamma\left(\frac{S_0}{n}\right)}{\Gamma\left(\frac{W_0}{s}\right)\Gamma\left(\frac{B_0}{s}\right)} \frac{1}{n+1} \sum_{k=0}^{\lfloor nx \rfloor} \left(\frac{k}{n+1}\right)^{\frac{W_0}{s}-1} \left(\frac{n-k}{n+1}\right)^{\frac{B_0}{s}-1} \\
& = \frac{\Gamma\left(\frac{S_0}{n}\right)}{\Gamma\left(\frac{W_0}{s}\right)\Gamma\left(\frac{B_0}{s}\right)} \left(\frac{n}{n+1}\right)^{\frac{W_0+B_0}{s}-1} \frac{1}{n} \sum_{k=0}^{\lfloor nx \rfloor} \left(\frac{k}{n}\right)^{\frac{W_0}{s}-1} \left(\frac{n-k}{n}\right)^{\frac{B_0}{s}-1} \\
& \xrightarrow{n \rightarrow \infty} \frac{\Gamma\left(\frac{W_0+B_0}{n}\right)}{\Gamma\left(\frac{W_0}{s}\right)\Gamma\left(\frac{B_0}{s}\right)} \int_0^x t^{\frac{W_0}{s}-1} (1-t)^{\frac{B_0}{s}-1} dt
\end{aligned}$$

Ugyanazt a határértéket kaptuk, ezzel igazoltuk az állítást. ■

Értelmezzük a kapott eredményt! Amint az már a várható értékből is kiderült, a keresett eloszlást a kezdeti helyzet határozza meg. Megjegyezném, hogy az  $s = 1, W_0 = 1, B_0 = 1$  esetben visszkapjuk a már bevezetésben is említett eredményt, hiszen  $\beta(1, 1) = \mathbf{U}(0, 1)$ , ahol  $\mathbf{U}$  az egyenletes eloszlást jelöli.

A Pólya-Eggenberger urnát szerettem volna szimulálni, aminek az eredménye a 1. ábrán (forrásfájl: 4.3.) látható  $s = 1$  esetben, ez tehát azt jelenti, hogy olyat rakok az urnába, amelyet húztam. Kétféle kísérletet végeztem. Először 1 kék és 1 fehér golyó volt kezdetben az urnában (piros vonalak az ábrán), majd 2 kék és 1 fehér golyóval indítottam a folyamatot (kék vonalak az ábrán). Minden kísérlet 200 behelyezett golyóig tart, és mindkét kísérletet 50-szer futtattam. A vastag szaggatott vonalak az 50 kísérlet átlageredményét jelzik.



1. ábra. Pólya-Eggenberger urna az  $(1,1)$  ill. a  $(2,1)$  helyzetből indítva

## 2.2. Bagchi–Pal urna

A következőkben szeretnénk általánosabb esetet vizsgálni, ahol az urnát meghatározó mátrix a következő alakú:

$$\mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}.$$

Ez tehát azt jelenti, hogy ha fehér golyót húzunk, akkor  $a$  fehér és  $b$  kék golyót helyezünk az urnába, ha pedig kéket húzunk, akkor  $c$  fehér és  $d$  kéket.

Néhány feltételt fel kell tennünk az urna viselkedésével kapcsolatban.

Az első, hogy a mátrix minden sorában a számok összege legyen egyenlő, azaz  $a + b = c + d = R$ . Ez azt jelenti, hogy minden húzás után a golyók száma az urnában ugyanannyival nő.

A második feltétel, hogy legyen az urna fenntartható. Egy urna fenntartható, ha minden lehetséges húzási sorrend esetén tudjuk folytatni a húzási folyamatot. Ennek ellenkezője nyilván akkor következhetne be, ha lennének a mátrixnak negatív elemei, ami tehát azt jelenti, hogy kiveszünk az urnából valahány golyót valamelyik színből. Például, ha egy oszlopban két negatív szám szerepel, akkor garantált, hogy nem fenntartható az urna, hiszen minden egyes húzás után csökken a golyók száma az egyik színből. Egyszerű számolás után megkaphatjuk a szükséges feltételeit annak, hogy egy urna fenntartható. Ettől a dolgozatban eltekintenek. E vizsgálódás megtekinthető [13] 3. fejezetében.

**2.7. Állítás.** *Legyen  $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$  a Bagchi-Pal urnát reprezentáló mátrix. Ekkor az urna fenntartható, ha a következő feltételek teljesülnek:*

- $b \geq 0, c \geq 0,$
- ha  $a < 0,$  akkor  $a$  osztója  $W_0$ -nak és  $c$ -nek,
- ha  $d < 0,$  akkor  $d$  osztója  $B_0$ -nak és  $b$ -nek.

Ezenkívül kizárunk néhány további esetet. Kizárjuk, hogy  $a = c,$  hiszen ekkor az urna viselkedése nem függ a húzott golyótól, hiszen ugyanaz történik minden húzás után.

Kizárjuk továbbá, a  $b = c = 0$  esetet, illetve amikor  $b$  vagy  $c$  közül az egyik egyenlő 0-val. Az első esetben a Pólya-Eggenberger urnát kapnánk vissza. A második eset viszonylag más megközelítést kíván, *Gouet* foglalkozott vele [7], [8] cikkeiben, martingálokat használva kapott normális határeloszlást a  $W_n/n$  hányadosra.

Belátjuk a fehér golyók számának várható értékéről szóló tételt, majd kimondunk egy normális határeloszlástételt bizonyítás nélkül.

**2.8. Tétel. (Bagchi és Pal, 1985)** *Legyen  $W_n$  a fehér golyók száma  $n$  húzás után a Bagchi-Pal urnából, melynek reprezentációs mátrixa  $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ . Legyen a konstans sorösszeg  $R$ . Ekkor:*

$$\mathbb{E}[W_n] = \frac{c}{b+c}Rn + o(n).$$

**Bizonyítás.** Felírhatjuk a következőt:

$$\mathbb{P}(W_{n+1} = W_n + a | W_n) = \frac{W_n}{S_n},$$

$$\mathbb{P}(W_{n+1} = W_n + c | W_n) = 1 - \frac{W_n}{S_n}.$$

Írjuk fel a feltételes várható értékét  $W_{n+1}$ -nek  $W_n$  feltételre:

$$\begin{aligned} \mathbb{E}[W_{n+1} | W_n] &= \frac{W_n}{S_n}(W_n + a) + \left(1 - \frac{W_n}{S_n}\right)(W_n + c) \\ &= \frac{W_n^2}{S_n} + \frac{W_n a}{S_n} + W_n + c - \frac{W_n^2}{S_n} - \frac{W_n c}{S_n} = \left(1 + \frac{a - c}{S_n}\right)W_n + c. \end{aligned}$$

Amiből következik, hogy

$$\mathbb{E}[W_{n+1}] = \left(1 + \frac{a - c}{S_n}\right)\mathbb{E}[W_n] + c.$$

A következőkben a tervünk egy  $Z_n$  sorozat bevezetése, amelyből vissza tudjuk kapni a  $W_n$  sorozatot és amelyről iteratív tulajdonsága miatt beláthatjuk, hogy  $o(n)$  nagyságrendű, újra a 2.3 Lemmát használva. Ezáltal  $W_n$  várható értékét kaphatjuk meg eredményül.

Legyen

$$Z_n := W_n - \frac{c}{b + c}S_n.$$

Ekkor felírva  $Z_{n+1}$  várható értékét:

$$\begin{aligned} \mathbb{E}[Z_{n+1}] &= \mathbb{E}[W_{n+1}] - \frac{c}{b + c}S_{n+1} \\ &= \left(1 + \frac{a - c}{S_n}\right)\mathbb{E}[W_n] + c - \frac{c}{b + c}(S_n + a + b) \\ &= \left(1 + \frac{a - c}{S_n}\right)\mathbb{E}[W_n] - \frac{c}{b + c}S_n + \frac{c(b + c) - c(a + b)}{b + c} \\ &= \left(1 + \frac{a - c}{S_n}\right)\mathbb{E}[W_n] - \frac{c}{b + c}S_n - \frac{c(a - c)}{b + c} \\ &= \left(1 + \frac{a - c}{S_n}\right)\left(\mathbb{E}[W_n] - \frac{c}{b + c}S_n\right) \\ &= \left(1 + \frac{a - c}{S_n}\right)\mathbb{E}[Z_n]. \end{aligned} \tag{1}$$

Mivel tudjuk, hogy

$$\mathbb{E}[W_n] = \frac{c}{b+c} S_n + E[Z_n],$$

ezért a tétel igazolásához elég belátnunk, hogy  $\mathbb{E}[Z_n] = o(n)$ .

Az előbb belátott iteratív tulajdonság miatt felírhatjuk a következőt:

$$\mathbb{E}[Z_n] = Z_0 \prod_{i=0}^{n-1} \left(1 + \frac{a-c}{S_i}\right).$$

Vizsgáljuk meg a szorzatot. A célunk, hogy végül a Stirling-formula segítségével tudjunk becsülni:

$$\begin{aligned} \prod_{i=0}^{n-1} \left(1 + \frac{a-c}{S_i}\right) &= \frac{S_0 + a - c}{S_0} \cdot \frac{S_0 + R + a - c}{S_0 + R} \cdot \dots \cdot \frac{S_0 + (n-1)R + a - c}{S_0 + (n-1)R} \\ &= \left(\frac{S_0 + a - c}{R} \cdot \dots \cdot \frac{S_0 + (n-1)R + a - c}{R}\right) \left(\frac{1}{\frac{S_0}{R}} \cdot \frac{1}{\frac{S_0+R}{R}} \cdot \dots \cdot \frac{1}{\frac{S_0+(n-1)R}{R}}\right) \\ &= \frac{\Gamma\left(n + \frac{S_0+a-c}{R}\right)}{\Gamma\left(\frac{S_0+a-c}{R}\right)} \cdot \frac{\Gamma\left(\frac{S_0}{R}\right)}{\Gamma\left(n + \frac{S_0}{R}\right)}. \end{aligned}$$

Így tehát használva 2.3. Lemmában bemutatott formulát:

$$\mathbb{E}[Z_n] = \left(W_0 - \frac{c}{b+c} S_0\right) \frac{\Gamma\left(n + \frac{S_0+a-c}{R}\right) \Gamma\left(\frac{S_0}{R}\right)}{\Gamma\left(\frac{S_0+a-c}{R}\right) \Gamma\left(n + \frac{S_0}{R}\right)} = O(n^{(a-c)/R}).$$

Mivel  $b, c > 0$ , ezért  $a - c < R$ , így  $\mathbb{E}[Z_n] = o(n)$ . Tehát igazoltuk az állítást. ■

Most már rátérhetünk  $W_n^*$  aszimptotikus eloszlására. A határeloszlásnak feltételei lesznek, amelynek szükségessége a dolgozat következő fejezetében válik világosabbá.

A normális határeloszlástételt nem bizonyítom. A bizonyítás megtalálható *Chern, Hwang és Tsai*[5] cikkében.  $W_n^*$  momentumainak konvergenciáit vizsgálva halad bizonyításuk.

**2.9. Tétel. (Bagchi és Pal, 1985)** *Legyen  $W_n$  a fehér golyók száma  $n$  húzás után az Bagchi-Pal urnából. Legyen a konstans sorösszeg  $R$ . Legyen továbbá  $\rho = \frac{a-c}{R} > \frac{1}{2}$ .*

*Ha  $\rho < \frac{1}{2}$ , akkor*

$$\frac{W_n - \frac{c}{b+c} Rn}{\sqrt{n}} \xrightarrow{d} \mathcal{N}\left(0, \frac{bcR(a-c)^2}{(b+c)^2(R-2(a-c))}\right).$$



Ha  $\rho = \frac{1}{2}$ , akkor

$$\frac{W_n - \frac{c}{b+c}Rn}{\sqrt{n \ln n}} \xrightarrow{d} \mathcal{N}\left(0, \frac{bc}{K}\right),$$

ahol  $\mathcal{N}(\mu, \sigma^2)$  a  $\mu$  várható értékű,  $\sigma^2$  szórásnégyzetű normális valószínűségi változót jelöli.

Érdekes, hogy a két esetben különböző normálást kell használjunk, hogy nem elfajuló határeloszlást kapjunk.

Szeretnék néhány szót szólni a még nem vizsgált,  $\rho > \frac{1}{2}$  esetről, illetve arról, miért is ilyen fontos ez a  $\rho$  hányados.

**2.10. Megjegyzés.** Tekintsük a Bagchi-Pal urna azon változatát, amikor  $b = c$ . Ekkor persze következik, hogy  $a = d$ . (Ezt az esetet az irodalom *Bernard Friedman* urnának nevezi.) Legyen  $\rho = \frac{a-c}{R} > \frac{1}{2}$ . Ekkor:

$$\frac{W_n - B_n}{n^\rho} \xrightarrow{d} \beta\left(\frac{W_0}{a}, \frac{B_0}{a}\right).$$

Érdekes, hogy teljesen más eloszlást kapunk. Ráadásul az előbbi két esetben az urnát reprezentáló mátrix elemei határozzák meg az eloszlást, míg itt a kezdeti állapot a döntő fontosságú. A megjegyzésbe leírtak bizonyítása megtalálható *Freedman* [6] cikkében.

**2.11. Megjegyzés.** Ha az  $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$  mátrixra igaz, hogy  $a + b = c + d$ , vagyis a sorösszeg konstans, akkor a mátrix sajátértékei  $a - c$  és  $a + b (= R)$ .

**Bizonyítás.** Az  $(1, 1)^\top$  vektor jobboldali sajátvektora az  $(a + b)$  sajátértéknek, míg  $(1, -1)$  baloldali sajátvektora  $(a - c)$ -nek. ■

Visszatekintve a kimondott tételre, azt láthatjuk, hogy az eloszlás a két sajátérték hányadosától függ, egészen pontosan akkor kaptunk normális határeloszlást, amikor a nagyobb sajátérték legalább akkora, mint a kisebbik sajátérték kétszerese. Valami nagyon hasonló feltételt fogunk kapni az általánosított,  $m$  színű esetben is, a következő fejezetben.

### 3. Urnamodellek $m$ szín esetén

A következőkben 2 szín helyett  $m$  színű golyók szerepelnek urnánkban. Néhány feltételt kikötünk az urnával kapcsolatban, amelyek bevezetése után szép tételekhez juthatunk az urna hosszútávú viselkedését illetően.

Jelölje  $A_{i,j}$  azt, hogy az  $i$ . golyó húzása esetén hány golyót teszünk az urnába a  $j$ . színből.

**3.1. Definíció.** *Egy rendszert többszínű urnamodellrendszernek nevezünk, ha a reprezentációs mátrixa:*

$$\mathbf{A} = \begin{pmatrix} A_{1,1} & A_{1,2} & \dots & A_{1,m} \\ A_{2,1} & A_{2,2} & \dots & A_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ A_{m,1} & A_{m,2} & \dots & A_{m,m} \end{pmatrix},$$

Indexeljük a mátrix sajátértékeit a valós részük nagysága szerint:

$$\Re\lambda_1 \geq \Re\lambda_2 \geq \dots \geq \Re\lambda_m.$$

Most bevezetünk néhány feltételt, melyek szükségesek lesznek később a konvergenciatétel teljesüléséhez.

**3.2. Definíció.** *Egy többszínű urnamodell kiterjesztett urnamodellrendszernek nevezünk, ha teljesülnek az alábbi feltételek:*

(i) *Az urnarendszer fenntartható.*

(ii) *A mátrix sorösszege konstans:  $\sum_{j=1}^n A_{i,j} = \text{const.}$*

(iii) *A  $\lambda_1$ -hez tartozó baloldali sajátvektor minden koordinátája szigorúan pozitív.*

(iv)  *$\Re\lambda_2 < \frac{1}{2}\Re\lambda_1$ .*

**3.3. Megjegyzés.** Szeretném megemlíteni, hogy még ennél is tovább általánosítható az urnamodell, hogyha a fentebb említett  $\mathbf{A}$  mátrix elemei (egészértékű) valószínűségi változók. Ekkor a további vizsgálódáshoz várható értékek mátrixára kell feltennünk, hogy konstans legyen a sorösszeg. Szükséges továbbá, hogy a valószínűségi változók véges szórásnégyzetűek legyenek, illetve, hogy az egy sorban található valószínűségi változók függetlenek. Ezen feltételek mellett a fejezetben található hosszabb számolás teljesen hasonlóan levezethető.

Ez a definíció Smythe [15] ötlete volt, az alábbi bizonyítások is az ő nevéhez köthetők.

Többféleképpen is próbálkoztak a történelem során az effajta általánosításaival urnamodelleknek. *Athreya* és *Karlin* [1] cikkükben például teljesen hasonló megszorításokat használtak, viszont ők feltették még azt is, hogy a mátrix diagonális elemei nem lehetnek  $-1$ -nél kisebbek.

A következő számolások célja megmutatni a kiterjesztett urnamodellek aszimptotikus tulajdonságait, a mátrix sajátértékei segítségével. Bizonyítást fogunk nyerni az előző fejezetben kimondott eredményre, mely teljesen más megközelítést használt a bizonyítás során.

**3.4. Tétel. (Smythe, 1996)** *Tekintsünk egy  $m$  színű kiterjesztett urnamodellrendszert. Legyen  $v = (v_1, v_2, \dots, v_m)$  a  $\lambda_1$  sajátértékhez tartozó bal oldali sajátvektora a reprezentációs mátrixnak. Válasszuk  $v$ -t olyannak, hogy a koordináták összege 1 legyen. (Ez megtehető, hiszen (iv) feltétel szerint a bal oldali sajátvektor minden koordinátája pozitív.) Jelölje  $C_{i,n}$  az  $i$ . színű golyók számát az urnában  $n$  húzást követően ( $i = 1, \dots, m$ ). Ekkor*

$$\frac{C_{i,n}}{n} \xrightarrow{p} \lambda_1 v_i.$$

A tétel bizonyítása során az egyszerűség kedvéért feltesszük, hogy a mátrix összes sajátértéke valós, illetve, hogy a jobboldali sajátvektorok lineárisan függetlenek.

A tétel akkor is igaz, ha a sajátértékek komplexek, viszont a bizonyítás igen bonyolulttá válik ebben az esetben, mely megtalálható Smythe [15] cikkében.

Ne feledjük el, hogy kiterjesztett urnamodellről beszélünk, tehát teljesül, hogy

$$\lambda_m \leq \dots \leq \lambda_2 < \frac{1}{2} \lambda_1.$$

Jelölje rendre  $u_1, u_2, \dots, u_m \in \mathbb{R}^m$  a  $\lambda_1, \lambda_2, \dots, \lambda_m$  sajátértékekhez tartozó jobboldali sajátvektorokat.

Legyen  $\mathcal{F}_i$  az első  $i$  húzás által generált  $\sigma$ -algebra.

Használjuk a továbbiakban a  $C_n = (C_{1,n}, C_{2,n}, \dots, C_{m,n})^\top$  jelölést, illetve a mátrix  $i$ -edik sorának  $j$ -edik elemét jelölje  $a_{i,j}$ , mely tehát azt mondja meg, hogy ha  $i$ -edik színű golyót húzunk, akkor hány golyót teszünk az urnába a  $j$ -edik színből.

Definiáljuk az  $X_n$  sorozatot a következőképpen:

$$X_n := u_2^\top C_n.$$

A bizonyítás első felében szeretnénk belátni, hogy  $\frac{X_n}{n} \xrightarrow{p} 0$ .

Megjegyezném, hogy ha  $X_n$  definíciójában nem a második, hanem bármelyik más sajátértéket íránk  $\lambda_1$ -en kívül, akkor ugyanazt az eredményt kapnánk, hiszen egyedül azt fogjuk felhasználni, hogy  $\lambda_2 < \frac{1}{2}\lambda_1$ .

Írjuk fel  $X_n$  és  $X_{n-1}$  különbségnek várható értékét  $\mathcal{F}_{n-1}$ -re nézve.

$$\begin{aligned}
\mathbb{E}[X_n - X_{n-1} | \mathcal{F}_{n-1}] &= \mathbb{E}[X_n | \mathcal{F}_{n-1}] - X_{n-1} \\
&= \mathbb{E}[u_2^\top C_n | \mathcal{F}_{n-1}] - X_{n-1} \\
&= \mathbb{E}\left[\sum_{j=1}^m u_2^{(j)} C_{j,n} | \mathcal{F}_{n-1}\right] - X_{n-1} \\
&= \sum_{j=1}^m u_2^{(j)} \mathbb{E}[C_{j,n} | \mathcal{F}_{n-1}] - X_{n-1} \\
&= \sum_{j=1}^m u_2^{(j)} \left( C_{j,n-1} + \sum_{k=1}^m a_{k,j} \frac{C_{k,n-1}}{S_n - 1} \right) - X_{n-1} \\
&= \sum_{j=1}^m u_2^{(j)} C_{j,n-1} + \sum_{k=1}^m \left( \left( \sum_{j=1}^m u_2^{(j)} a_{k,j} \right) \frac{C_{k,n-1}}{S_{n-1}} \right) - X_{n-1} \\
&= X_{n-1} + \frac{1}{S_{n-1}} \begin{pmatrix} u_2^{(1)} & \dots & u_2^{(m)} \end{pmatrix} \begin{pmatrix} a_{1,1} & a_{2,1} & \dots & a_{m,1} \\ a_{1,2} & a_{2,2} & & a_{m,2} \\ \vdots & & \ddots & \vdots \\ a_{1,m} & a_{2,m} & \dots & a_{m,m} \end{pmatrix} \begin{pmatrix} C_{1,n-1} \\ C_{2,n-1} \\ \vdots \\ C_{m,n-1} \end{pmatrix} - X_{n-1} \\
&= X_{n-1} - \frac{1}{S_{n-1}} u_2^\top A^\top C_{n-1} - X_{n-1} \\
&= \frac{\lambda_2}{S_{n-1}} X_{n-1}.
\end{aligned}$$

Definiáljuk a következőféleképpen az  $M_n$  sorozatot:

$$M_n := X_n - \left(1 + \frac{\lambda_2}{S_{n-1}}\right) X_{n-1}.$$

Az előbbi számolás miatt az  $M_n$ -ek martingáldifferenciák, azaz  $\mathbb{E}[M_n | \mathcal{F}_{n-1}] = 0$ .

Ezután definiáljuk az  $N_n$  sorozatot az  $M_i$  differenciák lineáris kombinációjaként, a becslés hibáját pedig jelölje  $\varepsilon_n$ .

$$N_n := \sum_{i=1}^n \gamma_{i,n} M_i = X_n + \varepsilon_n$$

Célunk olyan  $\gamma_{i,n}$  együtthatókat találni, hogy a hibatag  $o(n)$  nagyságrendű legyen.

**3.5. Állítás.** *Meg tudjuk úgy választani a  $\gamma_{i,n}$  együtthatókat, hogy  $\varepsilon_n = o(n)$  teljesüljön.*

**Bizonyítás.** Vegyük szemügyre az  $N_n$  összeget, írjuk fel fordított sorrendben a tagokat:

$$\begin{aligned} N_n = & \gamma_{n,n} \left( X_n - \left(1 + \frac{\lambda_2}{S_{n-1}}\right) X_{n-1} \right) + \gamma_{n-1,n} \left( X_{n-1} - \left(1 + \frac{\lambda_2}{S_{n-2}}\right) X_{n-2} \right) \\ & + \cdots + \gamma_{1,n} \left( X_1 - \left(1 + \frac{\lambda_2}{S_0}\right) X_0 \right) = X_n + \varepsilon_n. \end{aligned}$$

Legyen  $\gamma_{n,n} = 1$ , hiszen ekkor kapunk  $X_n$ -et, majd sorban állítsuk be a  $\gamma_{n-1,n}, \gamma_{n-2,n}, \dots, \gamma_{1,n}$  együtthatókat úgy, hogy  $X_{n-1}, X_{n-1}, \dots, X_1$  együtthatói az összegben az összevonások után 0 -k legyenek:

$$-\gamma_{n,n} \left(1 + \frac{\lambda_2}{S_{n-1}}\right) + \gamma_{n-1,n} X_{n-1} = 0,$$

amiből következik, hogy

$$\gamma_{n-1,n} = 1 + \frac{\lambda_2}{S_{n-1}}.$$

Írjuk fel  $X_{n-2}$  együtthatóját:

$$-\gamma_{n-1,n} \left(1 + \frac{\lambda_2}{S_{n-2}}\right) + \gamma_{n-2,n} = 0,$$

amiből következik, hogy

$$\gamma_{n-2,n} = \left(1 + \frac{\lambda_2}{S_{n-2}}\right) \gamma_{n-1,n} = \left(1 + \frac{\lambda_2}{S_{n-2}}\right) \left(1 + \frac{\lambda_2}{S_{n-1}}\right).$$

Hasonlóan folytatva adódik, hogy

$$\gamma_{i,n} = \prod_{j=i}^{n-1} \left(1 + \frac{\lambda_2}{S_j}\right).$$

A hibatag a maradék, vagyis

$$\varepsilon_n = -\gamma_{1,n} \left(1 + \frac{\lambda_2}{S_0}\right) X_0 = \prod_{j=0}^{n-1} \left(1 + \frac{\lambda_2}{S_j}\right) X_0.$$

Ne feledjük, hogy a  $\lambda_1$  egyenlő a sorösszeggel, vagyis minden húzás után  $\lambda_1$  új golyó kerül az urnába. Kihaszználva ezt a következő adódik:

$$\begin{aligned}
\varepsilon_n &= \prod_{j=0}^{n-1} \left(1 + \frac{\lambda_2}{S_0 + j\lambda_1}\right) X_0 = \prod_{j=0}^{n-1} \frac{j\lambda_1 + S_0 + \lambda_2}{j\lambda_1 + S_0} \\
&= \prod_{j=0}^{n-1} \frac{j + \frac{S_0 + \lambda_2}{\lambda_1}}{j + \frac{S_0}{\lambda_1}} = \frac{\left(\frac{S_0 + \lambda_2}{\lambda_1}\right) \left(1 + \frac{S_0 + \lambda_2}{\lambda_1}\right) \dots \left(n - 1 + \frac{S_0 + \lambda_2}{\lambda_1}\right)}{\frac{S_0}{\lambda_1} \left(1 + \frac{S_0}{\lambda_1}\right) \dots \left(n - 1 + \frac{S_0}{\lambda_1}\right)} \\
&= \frac{\Gamma\left(n + \frac{S_0 + \lambda_2}{\lambda_1}\right)}{\Gamma\left(\frac{S_0 + \lambda_2}{\lambda_1}\right)} \cdot \frac{\Gamma\left(\frac{S_0}{\lambda_1}\right)}{\Gamma\left(n + \frac{S_0}{\lambda_1}\right)} = O\left(n^{\frac{\lambda_2}{\lambda_1}}\right).
\end{aligned}$$

Mivel a feltevésünk az volt, hogy  $\lambda_2 < \frac{1}{2}\lambda_1$ , így valóban igaz, hogy  $\varepsilon_n = o(n)$ . ■

Ezután már rátérhetünk  $X_n/n$  vizsgálatára.

**3.6. Állítás.** *Tekintsük a fentebb definiált  $X_n$  sorozatot. Ekkor*

$$\frac{X_n}{n} \xrightarrow{p} 0.$$

**Bizonyítás.**

Jegyezzük meg, hogy  $\mathbb{E}[N_n] = 0$ .

Ha belátjuk, hogy  $\mathbb{D}^2[N_n/n]$  korlátos, azaz  $\mathbb{D}^2[N_n] \leq Kn$ , akkor használva a Csebisev-egyenlőtlenséget adódik, hogy

$$\mathbb{P}\left(\left|\frac{N_n}{n}\right| > \varepsilon\right) \leq \frac{Kn}{\varepsilon^2 n^2} \rightarrow 0$$

Ha pedig tudjuk, hogy  $N_n/n \xrightarrow{p} 0$ , akkor mivel beláttuk, hogy  $\varepsilon_n/n \xrightarrow{p} 0$ , ezért

$$\frac{X_n}{n} = \frac{N_n - \varepsilon_n}{n} \xrightarrow{p} 0.$$

Tehát elég belátnunk, hogy  $\mathbb{D}^2(N_n) \leq Kn$  valamilyen  $K > 0$  számra.

$$\begin{aligned}
\mathbb{D}^2[N_n] &= \mathbb{E}[N_n^2] = \left(\sum_{i=1}^n \gamma_{i,n} M_i\right)^2 \\
&= \mathbb{E}\left[\sum_{i=1}^n \gamma_{i,n}^2 M_i^2\right] + 2\mathbb{E}\left[\sum_{1 \leq i < j \leq n} \gamma_{i,n} \gamma_{j,n} M_i M_j\right]
\end{aligned}$$

Az összeg második tagja 0, hiszen a martingáldifferenciák ortogonálisak egymásra, ugyanis

$$\mathbb{E}[\gamma_{i,n} \gamma_{j,n} M_i M_j] = \mathbb{E}\left[\mathbb{E}[\gamma_{i,n} \gamma_{j,n} M_i M_j | \mathcal{F}_i]\right] = \gamma_{i,n} \gamma_{j,n} \mathbb{E}\left[M_i \mathbb{E}[M_j | \mathcal{F}_i]\right] = 0.$$

Az első tag megvizsgálása előtt becsüljük felülről  $|X_n|$ -et.

$$\begin{aligned}
|X_n| &= \left| u_2^{(1)} C_{1,n} + \cdots + u_2^{(m)} C_{m,n} \right| = \left| \sum_{j=1}^m u_2^{(j)} C_{j,n} \right| \\
&\leq \sum_{j=1}^m \left| u_2^{(j)} \right| C_{j,n} \leq \max \left( |u_2^{(1)}|, |u_2^{(2)}|, \dots, |u_2^{(m)}| \right) S_n \\
&= \max \left( |u_2^{(1)}|, |u_2^{(2)}|, \dots, |u_2^{(m)}| \right) (\lambda_1 n + S_0) \\
&\leq \max \left( |u_2^{(1)}|, |u_2^{(2)}|, \dots, |u_2^{(m)}| \right) (\lambda_1 n + \lambda_1 n) \quad \text{ha } n \text{ elég nagy} \\
&= 2 \max \left( |u_2^{(1)}|, |u_2^{(2)}|, \dots, |u_2^{(m)}| \right) \lambda_1 n = K' n.
\end{aligned}$$

Vizsgáljuk meg a  $\gamma_{i,n}$  együtthatók nagyságrendjét, hasonlóan, ahogyan korábban az  $\varepsilon_n$  hibatagot becsültük, használva a 2.3 Lemmát:

$$\begin{aligned}
\gamma_{i,n} &= \prod_{j=i}^{n-1} \frac{j\lambda_1 + S_0 + \lambda_2}{j\lambda_1 + S_0} = \prod_{j=i}^{n-1} \frac{j + \frac{S_0 + \lambda_2}{\lambda_1}}{j + \frac{S_0}{\lambda_1}} \\
&= \frac{\left(i + \frac{S_0 + \lambda_2}{\lambda_1}\right) \left(i + 1 + \frac{S_0 + \lambda_2}{\lambda_1}\right) \cdots \left(n - 1 + \frac{S_0 + \lambda_2}{\lambda_1}\right)}{\left(i + \frac{S_0}{\lambda_1}\right) \left(i + 1 + \frac{S_0}{\lambda_1}\right) \cdots \left(n - 1 + \frac{S_0}{\lambda_1}\right)} \\
&= \frac{\Gamma\left(n + \frac{S_0 + \lambda_2}{\lambda_1}\right)}{\Gamma\left(n + \frac{S_0}{\lambda_1}\right)} \cdot \frac{\Gamma\left(i + \frac{S_0}{\lambda_1}\right)}{\Gamma\left(i + \frac{S_0 + \lambda_2}{\lambda_1}\right)} \\
&= \left(\frac{n}{i}\right)^{\lambda_2/\lambda_1} + O(n^{\lambda_2/\lambda_1 - 1}).
\end{aligned}$$

Most már készen állunk az összeg kibontására:

$$\begin{aligned}
\mathbb{D}^2[N_n] &= \mathbb{E} \left[ \sum_{i=1}^n \gamma_{i,n}^2 M_i^2 \right] = \sum_{i=1}^n \gamma_{i,n}^2 \mathbb{E}[M_i^2] \\
&= \sum_{i=1}^n \gamma_{i,n}^2 \mathbb{E} \left[ \left( (X_i - X_{i-1}) - \frac{\lambda_2 X_{i-1}}{S_{i-1}} \right)^2 \right] \\
&\leq \sum_{i=1}^n \gamma_{i,n}^2 \mathbb{E} \left[ \left( |X_i - X_{i-1}| - \frac{\lambda_2 |X_{i-1}|}{S_{i-1}} \right)^2 \right] \\
&\leq \sum_{i=1}^n \gamma_{i,n}^2 \mathbb{E} \left[ \left( |u_2^\top (C_i - C_{i-1})| + \frac{\lambda_2 K'(i-1)}{\lambda_1(i-1) + S_0} \right)^2 \right] \\
(*) &\leq \sum_{i=1}^n \gamma_{i,n}^2 \mathbb{E} \left[ \left( m \max_{1 \leq i \leq m} \left( |u_2^{(i)}| \right) \max_{1 \leq i \leq m, 1 \leq j \leq m} (|a_{i,j}|) + \frac{\lambda_2 K'(i-1)}{\lambda_1(i-1) + S_0} \right)^2 \right] \\
&\leq K'' \sum_{i=1}^n \gamma_{i,n}^2 \\
&= K'' \sum_{i=1}^n \left( \left( \frac{n}{i} \right)^{\lambda_2/\lambda_1} + O(n^{\lambda_2/\lambda_1 - 1}) \right)^2 \\
&= K'' \left[ \sum_{i=1}^n \left( \left( \frac{n}{i} \right)^{2\lambda_2/\lambda_1} + O\left( \frac{n^{2\lambda_2/\lambda_1 - 1}}{i^{\lambda_2/\lambda_1}} \right) + O(n^{2\lambda_2/\lambda_1 - 2}) \right) \right] \\
&= K''' n,
\end{aligned}$$

ahol a (\*) egyenlőtlenségnél azt használtuk fel, hogy egy lépés alatt bármelyik színből maximum  $\max_{1 \leq i \leq m, 1 \leq j \leq m} |a_{i,j}|$  új golyó kerülhet az urnába.

Ezzel beláttuk a kívánt állítást. ■

Megjegyezném még egyszer, hogy e hosszas számolás bármelyik más sajátértékre igaz, tehát tudjuk, hogy

$$\frac{u_i^\top C_n}{n} \xrightarrow{p} 0, \quad 2 \leq i \leq m \quad \text{esetén.}$$

Most már rátérhetünk a 3.4. Tétel bizonyítására.

**Bizonyítás.** Mivel feltettük, hogy a jobb oldali sajátvektorok lineárisan függetlenek, ezért egy tetszőleges  $y \in \mathbb{R}^m$  vektort felírhatunk azok lineáris kombinációjaként:

$$y = \alpha_1 u_1 + \alpha_2 u_2 + \cdots + \alpha_m u_m.$$



Vizsgáljuk meg ekkor a  $vy$  skaláris szorzatot.

$$vy = \alpha_1 vu_1 + \alpha_2 vu_2 + \cdots + \alpha_m vu_m.$$

Vegyük észre, hogy az első tagon kívül mindegyik tag 0, hiszen  $2 \leq i \leq m$  esetén

$$\lambda_1 vu_i = vAu_i = \lambda_i vu_i$$

teljesül, és mivel tudjuk, hogy  $\lambda_1 \neq \lambda_i$  ezért szükségképpen  $vu_i = 0$ .

Így tehát

$$vy = \alpha_1 vu_1 = \alpha_1 v(1, 1, \dots, 1)^\top = \alpha_1 (v_1 + v_2 + \cdots + v_m) = \alpha_1$$

tekintve, hogy  $v$ -t úgy választottuk meg, hogy a koordináták összege 1 legyen.

Vizsgáljuk meg  $\frac{1}{n}yC_n$  határértékét:

$$\begin{aligned} \frac{1}{n}yC_n &= \frac{\alpha_1}{n}(1, \dots, 1)C_n + \frac{\alpha_2}{n}u_2^\top C_n + \cdots + \frac{\alpha_m}{n}u_m^\top C_n \\ &= \frac{1}{n}yv(C_{1,n} + C_{2,n} + \cdots + C_{m,n}) + \frac{\alpha_2}{n}u_2^\top C_n + \cdots + \frac{\alpha_m}{n}u_m^\top C_n \\ &= \frac{\lambda_1 n + S_0}{n}yv + \frac{\alpha_2}{n}u_2^\top C_n + \cdots + \frac{\alpha_m}{n}u_m^\top C_n \\ &\xrightarrow{p} \lambda_1 yv + 0 + \cdots + 0 = \lambda_1 yv. \end{aligned}$$

Mivel  $y$  tetszőleges  $\mathbb{R}^m$ -beli vektor volt, állítsuk  $y$ -t úgy, hogy az  $i$ . koordinátája legyen 1, a többi pedig 0. Ekkor a

$$\frac{C_{i,n}}{n} \xrightarrow{p} \lambda_1 v_i$$

összefüggéshez jutunk, mely pontosan a 3.4 Tétel állítása volt. ■

## 4. Urnamodellek alkalmazása az informatikában

Meglepően hangozhat elsőre, de az informatikában, adatstruktúrák aszimptotikus tulajdonságaira, azon belül is legfőképp a keresőfák vizsgálatára alkalmas modell lehet a Pólya-féle urnamodell.

A keresőfák közös tulajdonsága, hogy kezdetben üres fából indulva adjuk hozzá a csúcsokat egyenként a fához. Egy csúcsot bővíthetőnek nevezünk, ha még tudunk alá új csúcsot beilleszteni. Mindig egy bővíthető csúcsot kiválasztva adjuk hozzá az új elemet a fához, amely után a beillesztett csúcsot a felette álló gyerekének nevezzük. Van egy rendezési reláció értelmezve a csúcsok halmazán, ennek megfelelően illesztjük be az elemeket az aktuális fába.

### 4.1. Bináris fa

A bináris fa a legismertebb keresőfa. A keresőfák adatok tárolására szolgálnak. Minden adat rendelkezik egy sajátos kulccsal, mely szerint rendezni tudjuk az adatokat. A keresőfák célja, hogy gyorsan tudjunk keresni az elemek között.

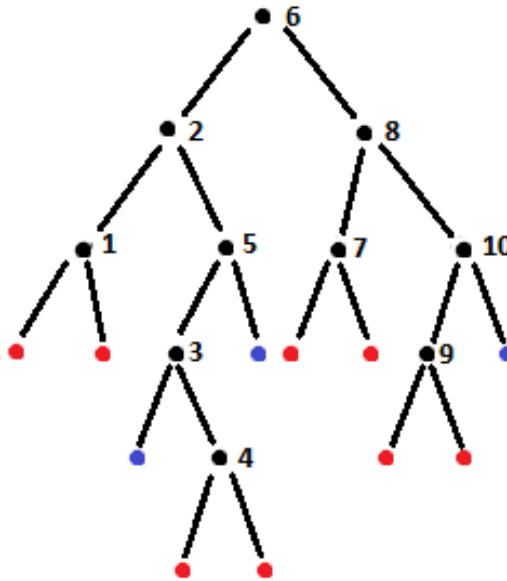
A bináris fa egy egyszerű tárolási elvet valósít meg. A legelső elem az ún. gyökérem. Ezután a következő elemeket az ún. *kisebb balra*, *nagyobb jobbra* elv szerint helyezük be a fába. Minden csúcsnak két gyereke lehet, egy jobb oldali, illetve egy bal oldali. A fában a beillesztési tulajdonság miatt minden csúcsra igaz, hogy nagyobb a tőle balra lévő leszármazott részfa összes csúcsúnál, és kisebb a tőle jobbra lévő részfa csúcsainál. Itt a kisebb, illetve nagyobb értelemszerűen a kulcsok szerint értendő.

A fák növekedéséről azt tesszük fel, hogy a beillesztendő csúcsok egy véletlen permutációja szerint illesztjük be az elemeket a fába, azaz  $n$  beillesztendő csúcs esetén mind az  $n!$  permutáció egyenlően valószínű. E feltevés a valós életben is számos esetben megfelelő lehet a modell építésekor.

Egy fa véletlen növekedését urnákkal pedig a következőképpen tudjuk szimulálni. Minden esetben az alapfeltevésünk az, hogy az összes beszúrandó hely közül egyet választunk véletlenül, és oda szúrjuk be a következő elemet. Tehát a golyók a beszúrandó helyeket fogják reprezentálni. Az, hogy mi alapján különböztetjük meg e golyókat (azaz mi alapján lesznek a színek kiosztva), az az adatszerkezettől függ.

Könnyen meggondolható, hogy az ilyen módszerrel növesztett fa pontosan megfelel annak, mint hogy egy véletlen permutációját választjuk ki a beillesztendő csúcsoknak. Az állítás bizonyítása megtalálható *Knuth* [10] könyvében.

A bináris fa esetében a korábbi fejezetekben belátott tételek segítségével becslést kaphatunk a levelek számára a fában. A fában a beszúrható helyek kétféle típusúak lehetnek. Az első típus, amikor egy olyan csúcs alá szúrjuk be az aktuális elemet, akinek nincs még gyereke, míg a második típus, amikor egy olyan csúcs alá szúrjuk be az elemünket, amelynek van gyereke. A 2. ábrán található egy példa, a 6, 8, 2, 5, 1, 7, 3, 4, 10, 9 elemek beillesztése utáni állapot látható. Az első típusú beszúrható helyek piros, míg a második típusúak kék színnel vannak jelölve.



2. ábra. A 6, 8, 2, 5, 1, 7, 3, 4, 10, 9 csúcsok beillesztése utáni állapot. A színes csúcsok felelnek meg az urnában található golyóknak, amik tehát a lehetséges beszúrási helyeket reprezentálják.

**4.1. Tétel. (Devroye, 1991)** Jelölje  $L_n$  a levelek számát egy  $n$  csúcsú véletlen bináris fában. Ekkor

$$\frac{L_n - \frac{1}{3}n}{\sqrt{n}} \xrightarrow{d} \mathcal{N}\left(0, \frac{2}{45}\right).$$

**Bizonyítás.** Jelölje  $W_n$  a piros csúcsok számát a fában (a  $W$  jelölés nyilván a fehér színre utal, viszont eddig ezt a jelölést használtuk),  $B_n$  pedig a kék csúcsok számát.

Vizsgáljuk meg, mi történik ha az egyik, vagy másik típusú csúcsot választjuk beszúrható helynek. Ha egy piros csúcsot választunk, akkor ő, illetve testvér csúcsa nem lesznek a továbbiakban piros csúcsok, a testvér csúcsa kék csúccsá változik, illetve ne

feledjük, hogy keletkezik 2 új piros csúcs a választott csúcs alatt. Ha kék csúcsot választunk, akkor ő megszűnik kék lenni, illetve alatta keletkezik két piros csúcs. Így tehát fel tudjuk írni az urnamodellel reprezentációs mátrixát:

$$\mathbf{A} = \begin{pmatrix} 0 & 1 \\ 2 & -1 \end{pmatrix}.$$

Ezután pedig a 2.9. Tételt használva fel tudjuk írni a piros golyók számának határeloszlását:

$$\frac{W_n - \frac{2}{3}n}{\sqrt{n}} \xrightarrow{d} \mathcal{N}\left(0, \frac{8}{45}\right).$$

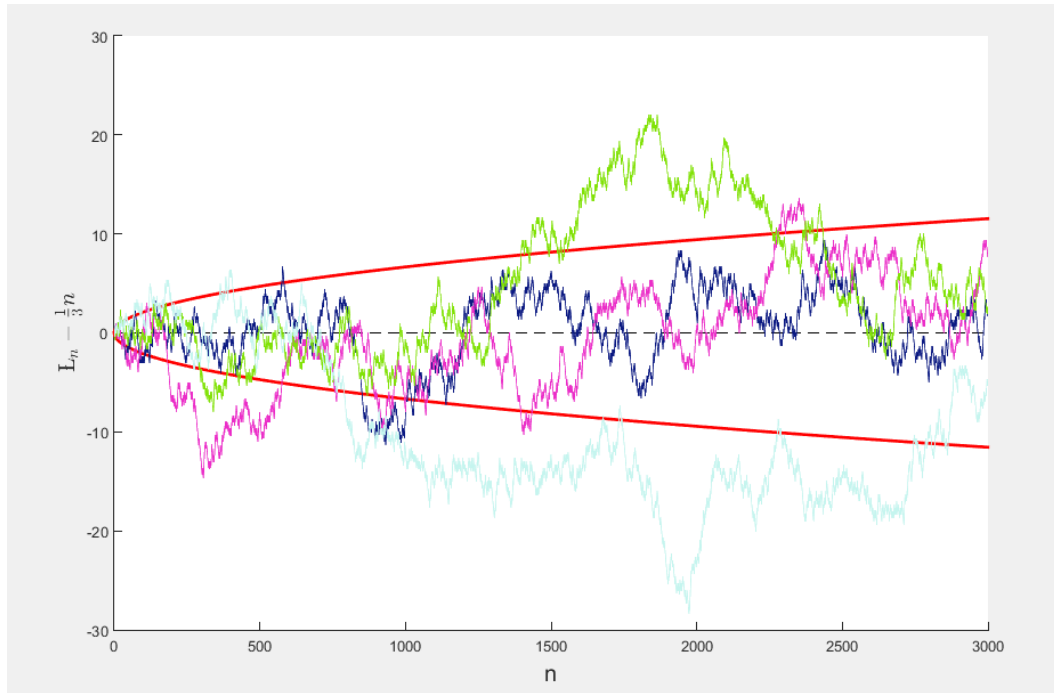
Mivel tudjuk, hogy ha  $k$  piros csúcs van a fában, akkor pontosan  $k/2$  levele van a fának, így  $W_n = 2L_n$ , amiből következik az állítás. ■

Szerettem volna meggyőződni a kapott formuláról, így szimuláltam véletlen növény bináris fákat. A kapott eredmény szemléletesebb bemutatása érdekében a tételt a következő formában használtam:

$$L_n - \frac{1}{3}n \xrightarrow{d} \mathcal{N}\left(0, \frac{2}{45}n\right),$$

mely ekvivalens alakja a kapott formulának.

A 3. ábrán (forráskód: 4.3) is megfigyelhetjük, hogy milyen kis szórású az eloszlása a levelek számának. 3000 behelyezett elem után mind a 4 kísérlet esetében kevesebb, mint 10-zel tért el a várható értéktől a levelek száma, azaz 990 és 1010 között volt.

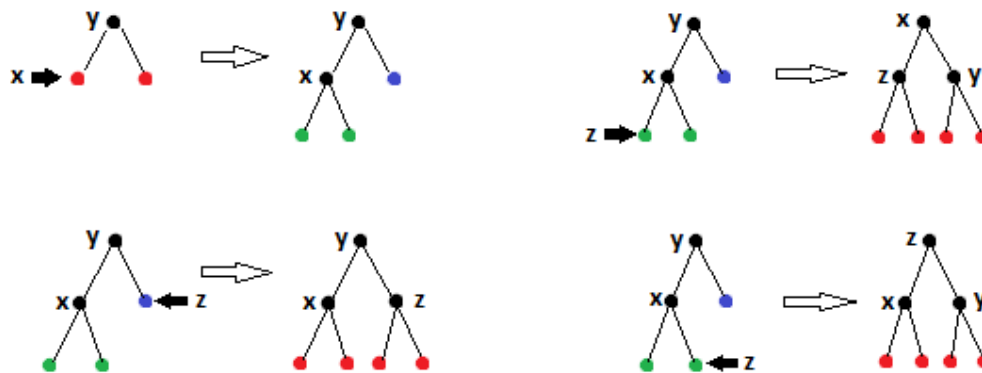


3. ábra. Véletlen növe bináris fa négy szimulációja 3000 behelyezett elemig. Az y tengely megmutatja, hogy mennyivel tér el a levelek száma az összes elem harmadától. A piros görbe az elméleti szórást ábrázolja.

## 4.2. Perem-kiegyensúlyozott fa

A perem-kiegyensúlyozott fa nagyon hasonlít a jobban elterjedt AVL fához. Az AVL fában minden csúcsra igaz, hogy a bal oldali illetve a jobb oldali leszármazott részfa magassága maximum 1-gyel tér el egymástól, más szóval, a csúcs kiegyensúlyozott. Minden beillesztés után leellenőrizzük, hogy valamelyik csúcs kiegyensúlyozatlanná vált-e, és ha igen, akkor elvégzünk egy egyszerű átalakítást, mely után újra kiegyensúlyozottá válik a fa. A perem-kiegyensúlyozott fa esetében csak akkor hajtunk végre cserét, ha a levél közvetlen környezetében alakult ki a nem kívánt tulajdonság, azaz hogyha a beillesztett csúcs szülőjének szülője válik kiegyensúlyozatlanná. Az urnamodell segítségével normális határeloszlást kaphatunk a cserék számára.

Ebben az esetben 3 különböző típusa lesz a beszúrandó helyeknek. Az első típusú helyek azok, ahol a csúcsnak, ami alá be szeretnénk illeszteni, van testvére, azaz a szülőjének 2 gyereke van. Ezek a csúcsok a kiegyensúlyozott állapotot jelentik. Piros színnel vannak jelölve a 4. ábrán. Ha kiválasztunk egy első típusú (tehát piros) csúcsot és oda illesztünk egy elemet, akkor a kiválasztott hely testvére kettes típusúvá válik, míg az új csúcs alatti két beszúrandó hely hármas típusú helyé válik. A kettes típusú helyek kék színűek, míg a hármas típusúak zöldek az ábrákon.



4. ábra. Csúcs beillesztése a perem-kiegyensúlyozott bináris keresőfába.

Látható az ábrán, hogy cserére a zöld csúcsok választásakor kerül sor.

**4.2. Tétel. (Panholzer és Prodinger, 1998)** Jelölje  $Z_n$  a cserék számát egy perem-  
kiegyensúlyozott fában  $n$  beszúrást követően. Ekkor

$$\frac{Z_n}{n} \xrightarrow{p} \frac{2}{7}.$$

**Bizonyítás.** Írjuk fel az ábrának megfelelően a reprezentációs mátrixát a modellnek. A sorok rendre egy piros, egy kék, illetve egy zöld kiválasztott csúcs utáni változásokat mutatják:

$$\mathbf{A} = \begin{pmatrix} -2 & 1 & 2 \\ 4 & -1 & -2 \\ 4 & -1 & -2 \end{pmatrix}.$$

Mivel minden beillesztéskor véletlenül választunk egy színes golyót, és tudjuk, hogy akkor történik csere, ha hármastípusú (zöld színű) golyót választunk, ezért a zöld színű golyók száma az urnában megmutatja aszimptotikusan, hogy hány csere történt az urnában. A bizonyításhoz a 3.4. Tételt szeretnénk segítségül hívni. Ehhez be kell látni, hogy az urnarendszer kiterjesztett. Látható, hogy a sorösszeg állandó, egyenlő 1-gyel. A mátrix sajátértékeit tekintve,  $\lambda_1 = 1, \lambda_2 = 0, \lambda_3 = -6$  adódik, tehát igaz, hogy  $\lambda_2 < \lambda_1/2$ , vagyis a rendszer megfelel a kiterjesztett urnamodellrendszer feltételeinek. Az egyetlen dolog amit ki kell számolni, az a  $\lambda_1 = 1$  sajátértékhez tartozó normált sajátvektor. Kis számolás után adódik, hogy baloldali sajátvektor  $v = \left(\frac{4}{7}, \frac{1}{7}, \frac{2}{7}\right)$ .

Ebből pedig következik a 3.4. Tételt használva, hogy

$$\frac{Z_n}{n} \xrightarrow{p} v_3 = \frac{2}{7}.$$

■

**4.3. Megjegyzés.** Megjegyezném, hogy ennél a tételnél több is igaz, normális határeloszlást lehet nyerni, ám a szórás kiszámítása igen bonyolult számolást igényel:

$$\frac{Z_n - \frac{2}{7}n}{\sqrt{n}} \xrightarrow{d} \mathcal{N}\left(0, \frac{66}{637}\right).$$

### 4.3. m-áris fa

Egy keresőfában lévő még hatékonyabb keresés érdekében vezették be a bináris fa általánosítását, az  $m$ -áris fát. A lényege, hogy egy csúcsban a bináris fával ellentétben nem 1, hanem  $m - 1$  elemet tárolunk, és minden csúcsnak  $m$  gyerekcsúcsa lehet. A kezdetben üres fából indulva tehát az első  $m - 1$  elem az első ún. gyökércsúcsba kerül. Az  $m - 1$  elemet sorbarendezve  $m$  intervallumot kapunk. (Feltesszük, hogy nincs két azonos elem.) Az  $m$ . elem behelyezése egy új csúcsba történik, abba, amelyik csúcs azt az intervallumot reprezentálja, amelyikbe ő esik. Így növekszik a fa. Ha az adott csúcs megtelik, akkor a továbbiakban a gyerekcsúcsaiba kerülnek az elemek a szülőcsúcsban található elemek által meghatározott intervallumoknak megfelelően.

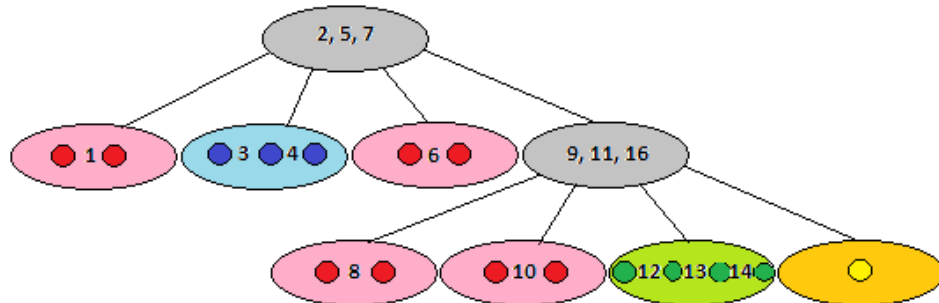


5. ábra. A  $(7, 5, 2, 16, 3, 1, 11, 9, 12, 4, 14, 6, 8, 10)$  elemek behelyezése utáni állapot, majd a  $(13)$  beszúrása az  $m$ -áris fába, ahol  $m = 4$ .

Észrevehető, hogy míg a bináris fa esetében  $n$  elem behelyezését követően tudjuk, hogy  $n$  csúcsa lesz a fának, ebben az általánosított esetben nem tudjuk, hogy hány csúcsa lesz a fának. Az urnamodell segítségével az  $m$ -áris fa csúcsai számának alakulását tudjuk meghatározni.

Az alapfeltevésünk továbbra is az a fa növekedésével kapcsolatban, hogy  $n$  behelyezett elem esetén gondolhatunk rá úgy, mintha az  $(1, 2, \dots, n)$  sorozat egy véletlen

permutációját vennénk, és aszerint helyeznénk be az elemeket. Ezt az urnás reprezentálással úgy érhetjük el, hogy az összes lehetséges kiválasztható intervallumot egyenlő eséllyel választjuk ki. A színek aszerint lesznek kiosztva, hogy az adott csúcsban hány elem van, és így hány intervallumot határoznak azok meg.



6. ábra. A  $(7, 5, 2, 16, 3, 1, 11, 9, 12, 4, 14, 6, 8, 10, 13)$  elemek behelyezése utáni állapot  $(m = 4)$ , az urnamodell reprezentálása színes golyókkal.

A 6. ábrán az  $m = 4$  esetben az  $(1, 2, \dots, 16)$  számok egy véletlen permutációja szerinti beszurása utáni állapot látható. Ha egy csúcs teli van, és van legalább egy nem üres gyereke, akkor azt a csúcsot belső csúcsnak nevezzük. Az ábrán ezt szürke színnel jelöltem. Az urna reprezentálásához 4 színre van szükségünk. A sárga csúcs egy üres csúcsra utal, ez az első szín. Ez nem valódi csúcsa a fának, tehát amikor a fa csúcsszámáról beszélünk, ezt nem számoljuk bele. A csúcsban lévő sárga színű golyó reprezentálja a  $(>16)$  intervallumot. A második szín a piros. Egy csúcs akkor piros, ha 1 elemet helyeztünk bele eddig. Nyilvánvaló, hogy egy piros csúcsban két piros golyó lesz, hiszen két intervallumot határoz meg a benne lévő elem. Hasonló mondható a kék és zöld szín esetében is. Tehát 1 lépésben a színes golyók közül választunk egyet. Gondoljuk meg, mi történik ekkor az egyes esetekben.

Látható, hogy ha kiválasztjuk például az egyik piros színű golyót, akkor ő, illetve a párja eltűnnek, és 3 db kék színű golyó keletkezik. Általánosan, ha  $i$ -edik színű golyót választunk, akkor  $i$  db golyó eltűnik az  $i$ -edik színből és  $i + 1$  golyó keletkezik az  $i + 1$ -edik színből. Ha pedig  $m$ -edik színű golyót választunk, akkor a csúcs belső csúcscsá változik, és  $m$  db üres (az ábrán sárga) csúcs keletkezik.

Ezek alapján felírhatjuk az  $m$ -áris fa reprezentációs mátrixát:



$$\mathbf{A} = \begin{pmatrix} -1 & 2 & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & -2 & 3 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & -3 & 4 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & & \ddots & & & \vdots & \vdots \\ 0 & 0 & & \dots & & & -(m-2) & m-1 \\ m & 0 & & \dots & & & 0 & -(m-1) \end{pmatrix}$$

Vizsgáljuk meg a mátrixot, hogy kiterjesztett urnamodell rendszer-e. Hasonlóan, mint a perem-kiegyensúlyozott fa esetében, a 3.2. definícióban található (i) – (iii) pontok könnyen láthatóan teljesülnek, az egyetlen kérdés a sajátértékekkel kapcsolatos. Számítsuk ki a mátrix sajátértékeit. Írjuk fel a karakterisztikus polinomot az első oszlop szerint kifejtve:

$$\begin{aligned} & (-1 - \lambda)(-2 - \lambda) \dots (-(m-1) - \lambda) \\ & + (-1)^m \cdot m \cdot 2 \cdot 3 \cdot \dots \cdot (m-1) = 0. \end{aligned}$$

Picit átrendezve adódik, hogy

$$(\lambda + 1)(\lambda + 2) \dots (\lambda + m - 1) = m!.$$

Indexeljük a sajátértékeket szokás szerint a valós részük nagysága alapján:

$$\Re \lambda_1 \geq \Re \lambda_2 \geq \dots \geq \Re \lambda_m.$$

**4.4. Állítás.** *Megfigyelve a karakterisztikus polinomot a következő tulajdonságokat kapjuk:*

- (i)  $\lambda_1 = 1$ .
- (ii) Ha  $m$  páratlan, akkor  $-m - 1$  is valós sajátérték ( $= \lambda_m$ ).
- (iii) A többi sajátérték mindegyike komplex, nemnulla képzetes résszel.
- (iv) Ha  $m \leq 26$ , akkor  $\Re \lambda_2 < \frac{1}{2}$ , és ha  $m > 26$ , akkor  $\Re \lambda_2 > \frac{1}{2}$ .

**Bizonyítás.**

- (i) Az nyilvánvaló, hogy  $\lambda = 1$  sajátérték. Ellenőriznünk kell, hogy nincs ennél nagyobb valós részű sajátérték. Tegyük fel, hogy  $\lambda'$  sajátérték és  $\Re \lambda' > 1$ . Ekkor

$$|(\lambda' + 1)(\lambda' + 2) \dots (\lambda' + m - 1)| \geq \Re(\lambda' + 1)\Re(\lambda' + 2) \dots \Re(\lambda' + m - 1) > m!,$$

tehát nem lehet 1-nél nagyobb valós részű sajátérték.

(ii) Behelyettesítve az egyenletbe adódik.

(iii) Meg kell vizsgálnunk, lehet-e másik valós sajátértéke az egyenletnek. Vegyük  $\lambda$  nagysága szerint a lehetséges eseteket:

- $\lambda < -m - 1$ :

$$|(\lambda + 1)(\lambda + 2) \dots (\lambda + m - 1)| > m!$$

- $-i - 1 < \lambda \leq -i$ ,  $1 \leq i \leq m$ :

$$|(\lambda + 1)(\lambda + 2) \dots (\lambda + m - 1)| < i(i - 1) \cdot \dots \cdot 2 \cdot 1 \cdot 1 \cdot 2 \cdot \dots \cdot (m - i - 1) \leq m!$$

- $\lambda > -1$ :

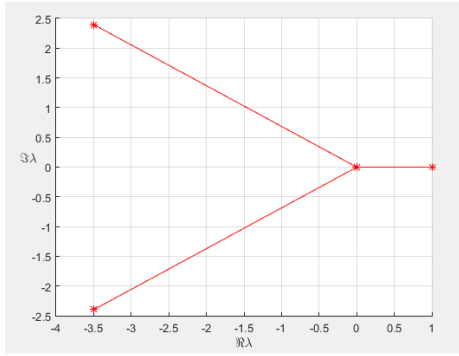
$$(\lambda + 1)(\lambda + 2) \dots (\lambda + m - 1) > m!$$

(iv) E tulajdonság bizonyítása bonyolult számolásokat igényel, melyek megtalálhatók Mahmoud és Smythe [12] cikkében.

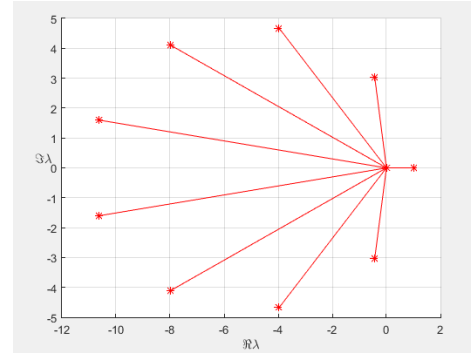
■

Ugyan az utolsó pontot nem bizonyítottuk, számítógép segítségével magunk is meggyőződhetünk az eredményről. Így tettem magam is. A 7. ábra (forráskód: 4.3) a sajátértékek elhelyezkedését mutatja különböző  $m$  értékek esetén.

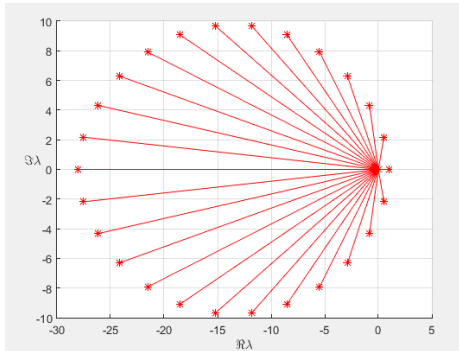
Az  $m = 27$  eset azért van megmutatva, mert akkor nem teljesül először, hogy  $\Re \lambda_2 < \frac{1}{2}$ , ez látható is a 7d. ábrán.



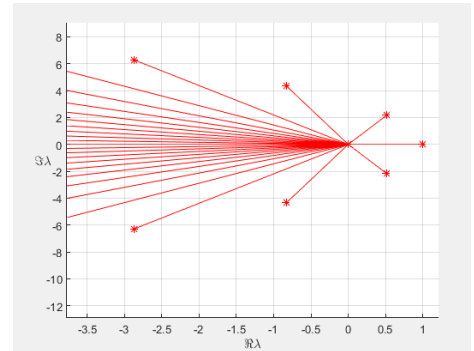
(a)  $m = 4$



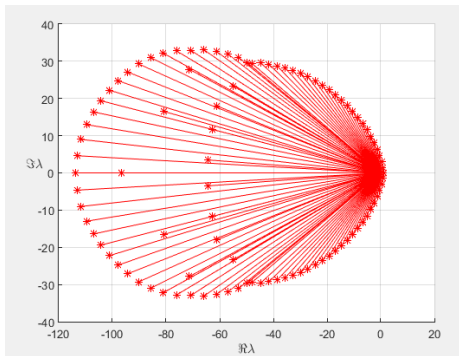
(b)  $m = 10$



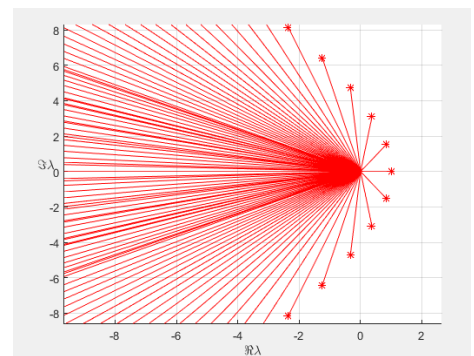
(c)  $m = 27$ , összes sajátérték



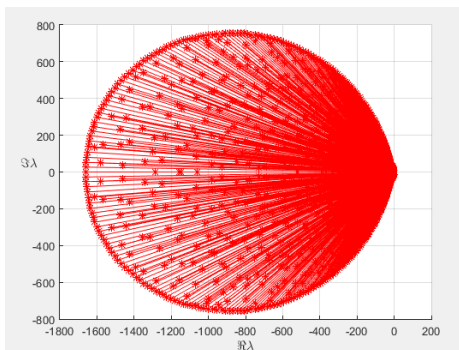
(d)  $m = 27$ , a  $\lambda_1 = 1$ -hez közeli rész



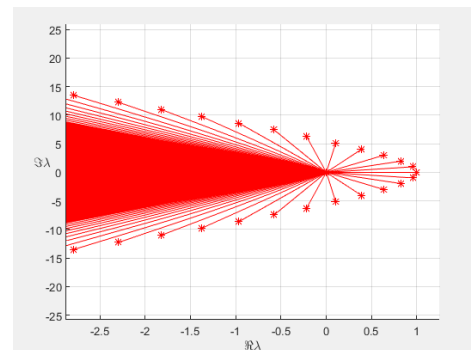
(e)  $m = 100$ , összes sajátérték



(f)  $m = 100$ , a  $\lambda_1 = 1$ -hez közeli rész



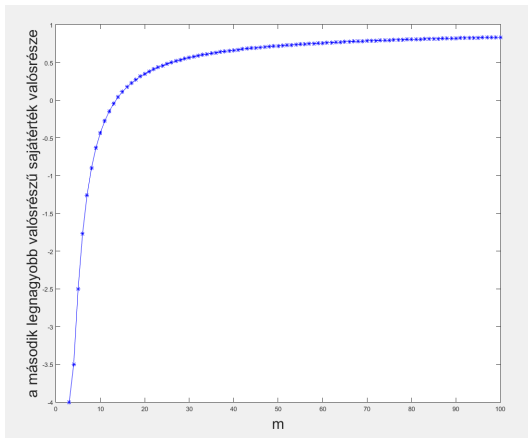
(g)  $m = 1000$ , összes sajátérték



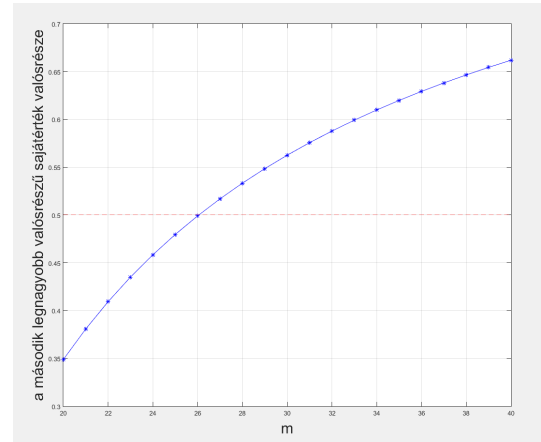
(h)  $m = 1000$ ,  $\lambda_1 = 1$ -hez közeli rész

7. ábra. Az  $m$ -áris fa reprezentációs mátrix sajátértékeinek elhelyezkedése a komplex számsíkon, különféle  $m$  értékek esetén.

Az összes sajátérték elhelyezkedésének bemutatása után vizsgáljuk meg jobban  $\lambda_2$  értékét  $m$  függvényében. A 8. ábra ezt hivatott bemutatni.



(a)  $3 \leq m \leq 100$



(b)  $20 \leq m \leq 40$

8. ábra. A második legnagyobb valós részű sajátérték  $m$  függvényében.

A sajátértékek alapos vizsgálata után a következő feladat megfelelő sajátvektort találunk  $\lambda_1 = 1$ -hez.

**4.5. Állítás.** Legyen az  $A \in \mathbb{R}^{(m-1) \times (m-1)}$  mátrix az  $m$ -áris fa reprezentációs mátrixa. Ekkor a  $\lambda_1 = 1$  sajátértékhez tartozó normált bal oldali sajátvektor  $v = (v_1, \dots, v_{m-1})$ , melyre

$$v_i = \frac{1}{(i+1)(H_m - 1)},$$

ahol  $H_n$  az  $n$ -edik harmonikus számot jelöli, azaz

$$H_n = 1 + \frac{1}{2} + \frac{1}{3} + \dots + \frac{1}{n}.$$

**Bizonyítás.** Baloldali sajátértéket keresve írjuk fel a kapott egyenletrendszert:

$$\begin{aligned} -v_1 + mv_{m-1} &= v_1 \\ 2v_1 - 2v_2 &= v_2 \\ 3v_2 - 3v_3 &= v_3 \\ &\vdots \\ (m-1)v_{m-2} - (m-1)v_{m-1} &= v_{m-1}. \end{aligned}$$

Az első egyenletből adódik, hogy  $\frac{v_1}{v_{m-1}} = \frac{m}{2}$ . Általános  $2 \leq i \leq m-2$ -re pedig az egyenletrendszer 2., 3., ...,  $i$ . egyenletét egymás után használva adódik, hogy

$$\frac{v_1}{v_i} = \frac{3}{2} \cdot \frac{4}{3} \cdot \dots \cdot \frac{i+1}{i} = \frac{i+1}{2}.$$

A kapott eredményt szemléletesebben felírva:

$$v_1 : v_2 : \dots : v_{m-1} = \frac{2}{2} : \frac{2}{3} : \frac{2}{4} : \dots : \frac{2}{m}.$$

Tekintve, hogy normált sajátvektort keresünk,

$$v_1 + \frac{2}{3}v_1 + \frac{2}{4}v_1 + \dots + \frac{2}{m}v_1 = 1,$$

amit átrendezve:

$$v_1 = \frac{1}{1 + 2\left(\frac{1}{3} + \frac{1}{4} + \dots + \frac{1}{m}\right)} = \frac{1}{2\left(\frac{1}{2} + \frac{1}{3} + \dots + \frac{1}{m}\right)} = \frac{1}{2(H_m - 1)}.$$

Felhasználva a kiszámolt  $(v_1 : v_i)$  arányokat adódik az állítás a sajátvektor többi elemére is. ■

A mátrix tüzetes vizsgálata után rátérhetünk a minket érdeklő tételre:

**4.6. Tétel. (Chern és Hwang, 2001)** *Jelölje  $S_n$  a csúcsok számát egy véletlenül nö-  
vő  $m$ -áris keresőfában. Ha  $3 \leq m \leq 26$ , akkor*

$$\frac{S_n}{n} \xrightarrow{p} \frac{1}{2(H_m - 1)}.$$

**Bizonyítás.** A reprezentációs mátrix előbbieken vizsgált tulajdonságai alapján akkor beszélhetünk kiterjesztett urnamodellről, ha  $3 \leq m \leq 26$ . Jelölje  $C_n^{(i)}$  az  $i$ . színű golyók számát  $n$  húzás után (tehát az olyan csúcsok számát a fában, amelyekben  $i - 1$  elem van). Ekkor alkalmazhatjuk a 3.4. Tételt, mely szerint

$$\frac{C_n^{(i)}}{n} \xrightarrow{p} \lambda_1 v_i = \frac{1}{(i + 1)(H_m - 1)}.$$

Mi a fa csúcsszámát keressük. Vezessünk be néhány jelölést:

- $B_n$ : Belső csúcsok száma  $n$  húzás után. Egy csúcs belsőnek számít, ha  $m - 1$  elemet tartalmaz, illetve van legalább egy nemüres gyerekcúcsa.
- $L_n$ : A levelek száma  $n$  húzás után. Levélnek számít egy csúcs, ha nemüres és  $m - 1$ -nél kevesebb elemet tartalmaz, vagy ha  $m - 1$ -et tartalmaz, akkor még nincsen leszármazott csúcs alatta.
- $\widetilde{L}_n$ : A levelek száma  $n$  húzás után, beleértve az üres leveleket is (amik tehát még nem tartalmaznak elemet, ezek felelnek meg az 1. színű golyóknak).

Három különböző módon is felírhatjuk az  $\widetilde{L}_n$  mennyiséget:

$$\begin{aligned}\widetilde{L}_n &= L_n + C_n^{(1)}, \\ \widetilde{L}_n &= (m-1)B_n + 1,\end{aligned}$$

illetve

$$\widetilde{L}_n = \sum_{i=1}^{m-1} \frac{C_n^{(i)}}{i}.$$

A második azért igaz, mert az üres fából indulva kezdetben 1 levele van a fának, majd minden egyes belső csúcs megjelenésekor  $m$  további keletkezik, viszont ami belső csúcscsá vált, az korábban levél volt.

A harmadik azért igaz, mert ha egy csúcs  $i-1$  elemet tartalmaz ( $1 \leq i \leq m$ ), akkor ez az urnamodellben pontosan annak felel meg, hogy  $i$  db golyó van az  $i$ . színből.

Mi a csúcsok  $S_n$  számát keressük, melyre igaz, hogy

$$S_n = B_n + L_n.$$

A kapott összefüggésekből a következő adódik:

$$\begin{aligned}S_n = B_n + L_n &= \frac{\widetilde{L}_n - 1}{m-1} + \widetilde{L}_n - C_n^{(1)} = \frac{m}{m-1} \widetilde{L}_n - \frac{1}{m-1} - C_n^{(1)} \\ &= \frac{m}{m-1} \sum_{i=1}^{m-1} \frac{C_n^{(i)}}{i} - C_n^{(1)} - \frac{1}{m-1}.\end{aligned}$$

Felhasználva, hogy ismerjük  $C_n^{(i)}/n$  határértékét, a következőt kapjuk:

$$\begin{aligned}\frac{S_n}{n} &\xrightarrow{p} \frac{m}{m-1} \sum_{i=1}^{m-1} \frac{1}{i(i+1)(H_m-1)} - \frac{1}{2(H_m-1)} - 0 \\ &= \frac{1}{H_m-1} \cdot \frac{m}{m-1} \sum_{i=1}^{m-1} \frac{1}{i(i+1)} - \frac{1}{2(H_m-1)} \\ &= \frac{1}{H_m-1} \cdot \frac{m}{m-1} \left( \frac{1}{1 \cdot 2} + \frac{1}{2 \cdot 3} + \frac{1}{3 \cdot 4} + \cdots + \frac{1}{(m-1)m} \right) - \frac{1}{2(H_m-1)} \\ &= \frac{1}{H_m-1} \cdot \frac{m}{m-1} \left( \left(1 - \frac{1}{2}\right) + \left(\frac{1}{2} - \frac{1}{3}\right) + \cdots + \left(\frac{1}{m-1} - \frac{1}{m}\right) \right) - \frac{1}{2(H_m-1)} \\ &= \frac{1}{H_m-1} \cdot \frac{m}{m-1} \left(1 - \frac{1}{m}\right) - \frac{1}{2(H_m-1)} \\ &= \frac{1}{2(H_m-1)}.\end{aligned}$$

■

**4.7. Megjegyzés.** Hasonlóan a perem-kiegyensúlyozott esethez, itt is kaphatunk normális határeloszlást. A szórásnégyzet kiszámítása további hosszú számolásokat igényel.

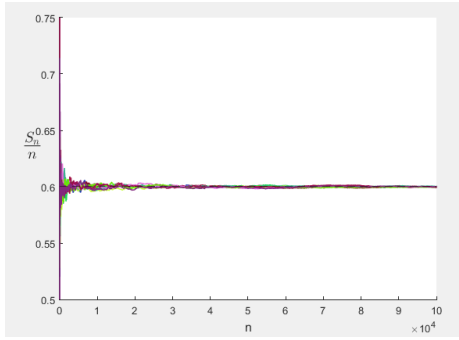
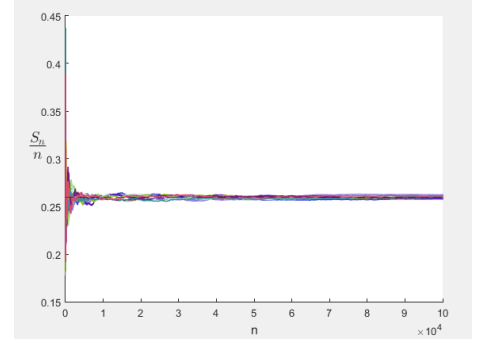
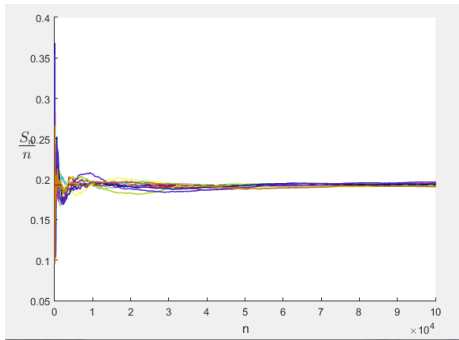
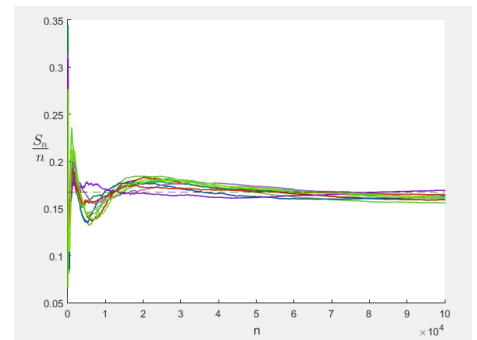
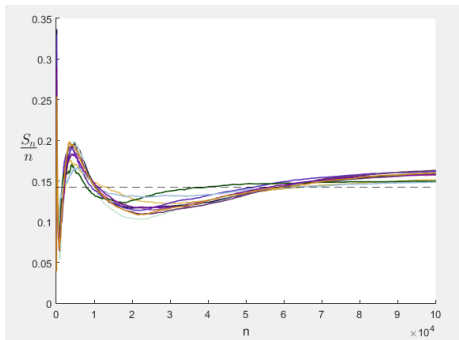
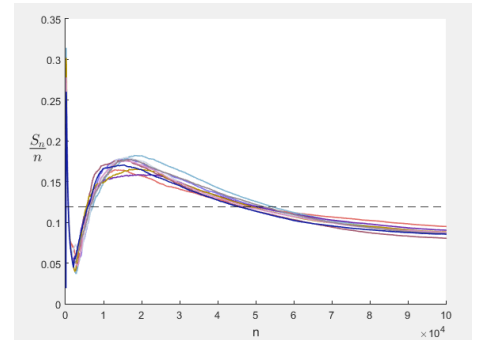
**4.8. Tétel. (Chern és Hwang, 2001)** *Jelölje  $S_n$  a csúcsok számát egy véletlenül nö-  
vő  $m$ -áris keresőfában. Ha  $3 \leq m \leq 26$ , akkor*

$$\left( \frac{S_n}{n} - \frac{1}{2(H_m - 1)} \right) \sqrt{n} \xrightarrow{d} \mathcal{N}(0, \sigma_m^2),$$

*valamilyen  $m$ -től függő, kiszámolható  $\sigma_m^2$  szórásnégyzetre.*

A kapott eredményt szerettem volna kipróbálni különböző  $m$  értékekre. A kísérletek eredményei a 9. ábrán (forráskód: 4.3) láthatóak.

Észrevehetjük az ábrákon, hogy ha  $m$ -et nagyobbra állítjuk, akkor nem teljesül a konvergencia, mely kisebb  $m$ -ekre teljesül is, ahogyan a tétel is állítja.

(a)  $m = 3$ (b)  $m = 10$ (c)  $m = 20$ (d)  $m = 30$ (e)  $m = 50$ (f)  $m = 100$ 

9. ábra. Véletlen növény  $m$ -áris fa csúcsainak száma osztva a behelyezett elemek számával, különböző  $m$  értékek esetén. Mindegyik esetben 10 kísérlet eredménye látható. Egy kísérlet  $n = 10^5$  fába helyezett elemig tart. A szaggatott vonal a 4.8. Tételben szereplő  $\frac{1}{2(H_m - 1)}$  mennyiséget jelzi.

Felmerülhet tehát a kérdés, vajon mi történik, ha  $m \geq 27$ ?

*Chauvin* és *Pouyanne* vizsgálták a kérdést meg részletesebben 2004-es [2], illetve 2014-es [3] cikkükben. A következő eredményt kapták:

**4.9. Tétel. (Chauvin, Pouyanne, 2004)** Legyen  $X_n = (X_n^{(1)}, \dots, X_n^{(m-1)})$ , ahol  $X_n^{(i)}$  jelölje az  $i - 1$  elemet tartalmazó csúcsok számát  $n - 1$  behelyezett elem után egy  $m$ -áris



fába. Ekkor

$$\frac{X_n - nX}{n^{\sigma_2}} - \rho \left( C \cos(\tau_2 \log n + \phi) + S \sin(\tau_2 \log n + \phi) \right) \xrightarrow{n \rightarrow \infty} 0,$$

ahol

- $\sigma_2$  és  $\tau_2$  a mátrix második legnagyobb valós részű sajátérték valós, illetve képzetes része,
- $X$  a már általunk is vizsgált vektor,  $X^{(i)} = \frac{1}{i(i+1)(H_m-1)}$ ,
- $C$  és  $S$  a második sajátértéktől függő meghatározható vektorok,
- $\rho$  és  $\phi$  valószínűségi változók.

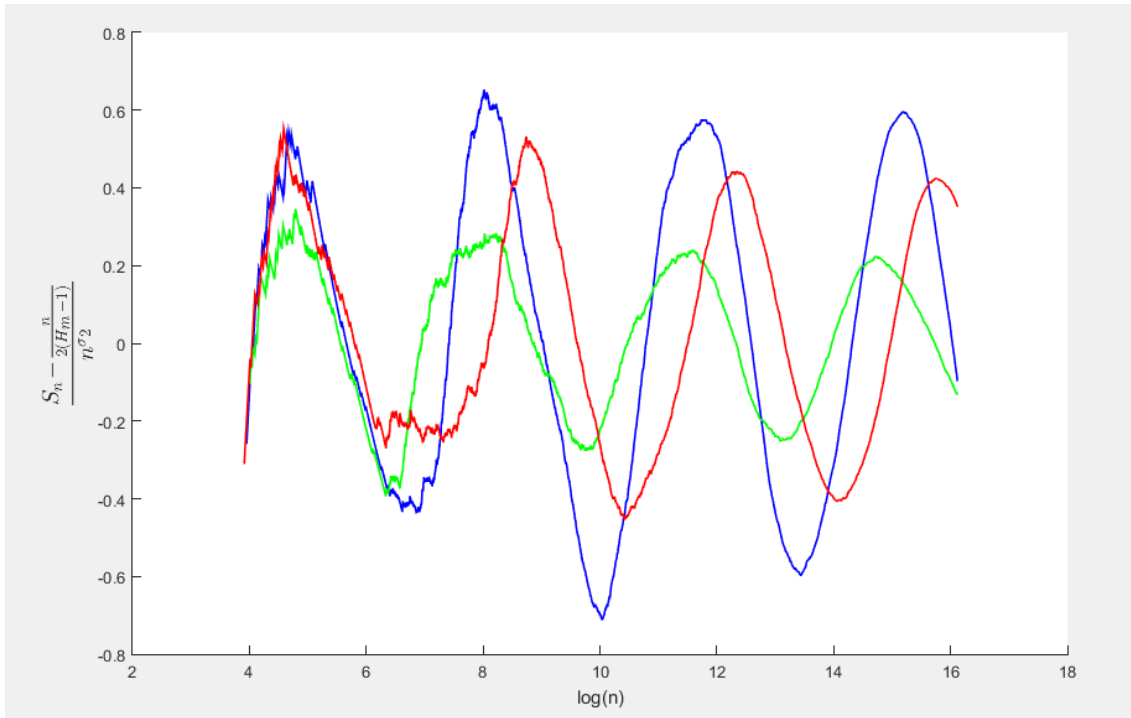
Ahogy az a tételből is látható, ők nem kifejezetten az  $m$ -áris fa csúcsainak az összsámát vizsgálták, hanem az egyes színű golyók számának eloszlását, melyet egy  $m-1$  dimenziós Markov-folyamattal modelleztek. Az eredmény részletesebb taglalása meghaladná e dolgozat kereteit, ám szeretnék néhány megjegyzést fűzni hozzá.

A tétel a mi általunk vizsgált összes csúcsszámot tekintve teljesen hasonlóan néz ki:

$$\frac{S_n - \frac{n}{2(H_m-1)}}{n^{\sigma_2}} - \rho \left( C \cos(\tau_2 \log n + \phi) + S \sin(\tau_2 \log n + \phi) \right) \xrightarrow{n \rightarrow \infty} 0$$

Megfigyelhetjük, hogy az első tag az  $m \leq 26$  esetben kapott tételre hasonlít, ott a nagyságrendet illetően  $n$  más hatványa szerepel. A kapott eloszlásban véletlen változók ( $\rho$ ,  $\phi$ ) szerepelnek. Ezt szerettem volna jobban megvizsgálni. Észrevehetjük, hogy a két szögfüggvény belsejében ugyanaz az argumentum van. Így ha többször szimulálunk egy véletlen növekvő  $m$ -áris fát ( $m > 26$ ), akkor a tétel szerint  $\log n$  függvényében a  $\frac{S_n - \frac{n}{2(H_m-1)}}{n^{\sigma_2}}$  mennyiséget ábrázolva olyan görbéket kell kapnunk, melyeknek különböző az amplitúdója, illetve máshol vannak a hullámhegyei, viszont a periódus nagyságának meg kell egyeznie. A  $\rho$  véletlen változó felel a különböző nagyságú amplitúdókért,  $\phi$  pedig a fáziskülönbségekért. El is készítettem a szimulációt hozzá, melynek eredménye a 10. ábrán (forráskód: 4.3) látható.

A két véletlen változó felfogható úgy is mint egy komplex értékű valószínűségi változó, melynek hossza  $\rho$ , szöge pedig  $\phi$ . Valójában pontosan ez van az egész háttérben. A két említett cikkben természetesen részletesen megvizsgálták ezt a valószínűségi változót, illetve az egész formulát.



10. ábra. Az  $\frac{S_n - \frac{n}{2(H_m - 1)}}{n^{\sigma_2}}$  mennyiség ábrázolva  $\log n$  függvényében. Három kísérlet látható,  $n = 10000000 = 10^7$  behelyezett elemig. Megfigyelhető az amplitúdóbeli randomitás, illetve a fáziseltolódás az egyes kísérletek között.

## Irodalomjegyzék

- [1] Athreya, K. and Karlin, S. (1968). Embedding of urn schemes into continuous time Markov branching process and related limit theorems. *The Annals of Mathematical Statistics*, **39**, 1801–1817.
- [2] Chauvin, B. and Pouyanne, N. (2004).  $m$ -ary Search trees when  $m \geq 27$ : A strong asymptotics for the space requirements. *Random Structures and Algorithms*, **24**, 133–154.
- [3] Chauvin, B., Liu, Q. and Pouyanne, N. (2014). Limit distributions for multitype branching processes of  $m$ -ary search trees *Annales de l'Institut Henri Poincaré - Probabilités et Statistiques*, **50**, 628–654.
- [4] Chern, H. and Hwang, H. (2001). Phase Change in random  $m$ -ary search trees and generalizes quicksort. *Random Structures and Algorithms*, **19**, 316–358.
- [5] Chern, H., Hwang, H. and Tsai, T. (2002). An asymptotic theory for Cauchy-Euler differential equations with applications to the analysis of algorithms. *Journal of Algorithms*, **44**, 177–225.
- [6] Freedman D. (1965). Bernard Friedman's urn. *The Annals of Mathematical Statistics*, **36**, 956–970.
- [7] Gouet, R. (1989). A martingale approach to strong convergence in a generalized Pólya-Eggenberger urn model. *Statistics and Probability Letters*, **8**, 225–228.
- [8] Gouet, R. (1993). Martingale functional central limit theorems for a generalized Pólya urn. *The Annals of Probability*, **21**, 1624–1639.
- [9] Janson, S. (2004). Functional limit theorems for multitype branching processes and generalized Pólya urns. *Stochastic Processes and Applications*, **110**, 177–245.
- [10] Knuth, D. (1998). *The Art of Computer Programming*, Vol. 3: Sorting and Searching, 2nd ed. Addison-Wesley, Reading, Massachusetts.
- [11] Laforgia, A. and Natalini, P. (2013). On Some Inequalities for the Gamma Function. *Advances in Dynamical Systems and Applications*, **2**, 261–267.
- [12] Mahmoud, H. and Smythe, R. (1995). Probabilistic analysis of bucket recursive trees. *Theoretical Computer Science*, **144**, 221–249.
- [13] Mahmoud, H. (2008). *Pólya urn models*. CRC press, Boca Raton

- [14] Panholzer, A. and Prodinger, H. (1998). An analytic approach for the analysis of rotations in fringe-balanced binary search trees. *The Annals of Combinatorics*, **2**, 173–184.
- [15] Smythe R. T. (1996). Central limit theorems for urn models. *Stochastic Processes and their Applications* **65** 115–137.

## Függelék

A Függelékben szimulációkhoz használt Matlab kódok találhatóak.

A Pólya–Eggenberger urna szimulációja különböző kezdeti feltételekből.(1. ábra)

```
1 %simulate a Polya–Eggenberger Urn A=[s 0;0 s]
2 %simulate the process from (1,1) and from (1,2)
3
4 clear all;
5 n=50; %num of experiments
6 m=200; %we do the simulation while there is less than m balls
   are in the urn
7
8 % simulations when there is 1 white and 2 blue balls in the
   urn initially
9 sumX1=0; %these will count the avarage value of the
   experiments
10 sumY1=0;
11
12 for i=1:n
13     X=zeros(m,1); %X[i] tells how many white balls are in the
   urn after i-1 draw
14     Y=zeros(m,1); %Y[i] is similar to X, (blue balls)
15     x=1; %it shows the actual number of white balls
16     y=1; %it shows the actual number of blue balls
17     X(1)=1;
18     Y(1)=1;
19     while x+y<=m %it chosses a random ball
20         if rand < x/(x+y)
21             x=x+1;
22         else
23             y=y+1;
24         end
25         X(x+y-1)=x;
26         Y(x+y-1)=y;
27     end
```

```

28     sumX1=sumX1+x;
29     sumY1=sumY1+y;
30     hold on;
31     plot(X,Y, 'r-', 'LineWidth',0.5);
32     axis([0 m 0 m]);
33
34
35     % simulations when there is 1 white and 2 blue balls in the
        urn
36     % initially
37     sumX2=0;
38     sumY2=0;
39
40     X=zeros(m,1);
41     Y=zeros(m,1);
42     x=2;
43     y=1;
44     X(1)=2;
45     Y(1)=1;
46     while x+y<=m+1
47         if rand < x/(x+y)
48             x=x+1;
49         else
50             y=y+1;
51         end
52         X(x+y-2)=x;
53         Y(x+y-2)=y;
54     end
55     hold on;
56     plot(X,Y, 'b-', 'LineWidth',0.5);
57     axis([0 m 0 m]);
58     sumX2=sumX2+x;
59     sumY2=sumY2+y;
60 end
61
62 %plotting the avarageresults
63 plot([1 sumX1/n],[1 sumY1/n], 'r-', 'Linewidth',3);

```

```

64 plot([1 sumX2/n],[1 sumY2/n], 'b—', 'Linewidth',3);
65 title('Pólya–Eggenberger urna (s=1) két különböző kiindulási
        helyzetből');
66 text(m*3/4-4*m/10,m*3/4+2*m/10,'Egy kísérlet 200 betett
        golyóig tart','FontSize',12);
67 text(m*3/4-m/10,m*3/4+m/10,'50–50 kísérlet','FontSize',12);
68 text(m*3/4-m/10,m*3/4,'(1,1)–ből indulók','Color','red','
        FontSize',12);
69 text(m*3/4-m/10,m*3/4-m/10,'(2,1)–ből indulók','Color','blue',
        'FontSize',12);
70 xlabel('kék golyók száma az urnában');
71 ylabel('fehér golyók száma az urnában');

```

A bináris fa szimulációja, a levelek számának viselkedését vizsgálva. (3. ábra)

```

1 %binaryTreeSimulation.m
2 %We will simulate a random binary tree.
3
4 clear all;
5
6 n=3000; %we will add 3000 items to the tree (then we should
        get approximately ~1000 leaves
7 t=0:n-1;
8
9 %firstly, we draw the expected value, and the standard dev
        lines that we
10 %got according to the theorem
11
12 x_up=zeros(n,1);
13 x_low=zeros(n,1);
14 x_avg=zeros(n,1);
15 for i=0:n-1
16     x_up(i+1)=(2/45*i)^(1/2); %this is the standard
        deviation

```

```

17     x_low(i+1)=-((2/45*i)^(1/2));
18     x_avg(i+1)=0;
19 end
20 hold on;
21 plot(t,x_up,'r-',t,x_low,'r-','LineWidth',2);
22 plot(t,x_avg,'k-');
23
24 %simulations
25
26 T=4; %the number of simulations
27 for i=1:T
28     b=0;
29     w=2;
30     y=[1];
31     while b+w<n+1
32         if rand < b/(b+w)
33             w=w+2;
34             b=b-1;
35         else
36             b=b+1;
37         end
38         y=[y w/2 - 1/3*(b+w)]; %L_n-n/3
39     end
40     plot(t,y,'color',rand(1,3)); %plotting the result
41 end
42 xlabel('n','FontSize',15);
43 ylabel({'$\L_n-\frac{1}{3}n$'},'Interpreter','latex','FontSize',15)

```

Az  $m$ -áris fa reprezentációs mátrixának a sajátértékeit kirajzoló függvény. (7. ábra)

```
1 %m_aryEigValues.m
```

```
2
```

```
3 %We will draw the eigenvalues of the transition matrix of the
```



```

    m-ary tree
4
5 function m_aryEigValues(m)
6 A=zeros(m-1);
7 %construct the matrix
8 for i=1:m-2
9     A(m-1,m-1)=-m+1;
10    A(i , i)=-i ;
11    A(i , i+1)=i+1;
12    end
13 A(m-1,1)=m;
14 t=eig(A); %t returns with the eigenvalues of the matrix
15 hold on;
16 grid on;
17 for i=1:length(t)
18     x=[0,real(t(i))];
19     y=[0,imag(t(i))];
20     plot(x,y,'r-*');
21 end
22 xlabel({'$\Re \lambda$'}, 'Interpreter', 'latex');
23 ylabel({'$\Im \lambda$'}, 'Interpreter', 'latex', 'rot',0);

```

A második legnagyobb valós részű sajátértéket visszaadó függvény.

```

1 %getSecondBiggestEigValue.m
2 %it will return the eigenvalue of the transition matrix of an
   m-ary tree
3 %which has the second biggest real part
4
5 function [re ,im]=getSecondBiggestEigValue(m)
6 A=zeros(m-1);
7     for i=1:m-2
8         A(i , i)=-i ;
9         A(i , i+1)=i+1;

```

```

10     end
11     A(m-1,1)=m;
12     A(m-1,m-1)=-m+1;
13     lambda=eig(A);
14     biggestRealPart=-100000;
15     secondBiggestRealPart=-100000;
16     ind1=-1;
17     ind2=-1;
18     for i=1:length(lambda)
19         if(real(lambda(i))>biggestRealPart)
20             secondBiggestRealPart=biggestRealPart;
21             biggestRealPart=real(lambda(i));
22             ind2=ind1;
23             ind1=i;
24         elseif(real(lambda(i))>secondBiggestRealPart)
25             secondBiggestRealPart=real(lambda(i));
26             ind2=i;
27         end
28     end
29     re =real(lambda(ind2))
30     im =imag(lambda(ind2))
31 end

```

Az  $m$ -áris fa szimulációja, a 4.8. Tételben kapott eredmény szerint ábrázolva. (9. ábra)

```

1 %size_of_m_ary.m
2 %simulation the size (numbers of nodes) of an m-ary search tree
3
4 function size_of_m_ary(m,n)
5 %n: one simulation ends when there are n items in the tree
6 K=10; % the number of simulations
7 for k=1:K
8     colours=zeros(m-1,1);

```

```

9     colours(1)=1;
10    S=zeros(1,n);
11    internalNodes=0;
12    for i=1:n
13        %we know that i potencial places are
14        %we have to choose one place random
15        c=randi(i);
16        j=1;
17        while c>0
18            c=c-colours(j);
19            if c>0
20                j=j+1;
21            end
22        end
23        %the chosen ball's colour is the j-th
24        if j==m-1
25            colours(m-1)=colours(m-1)-(m-1);
26            colours(1)=colours(1)+m;
27            internalNodes=internalNodes+1;
28        else
29            colours(j)=colours(j)-j;
30            colours(j+1)=colours(j+1)+j+1;
31        end
32        for j=2:(m-1)
33            S(i)=S(i)+(colours(j)/j);
34        end
35        S(i)=S(i)+internalNodes;
36        S(i)=S(i)/i;
37    end
38    start=min(100,m);
39    t=start:n;
40    hold on;
41    plot(t,S(start:n),'-', 'color',rand(1,3),'LineWidth',1.2);
42 end
43 xlabel('n','FontSize',12);
44 ylabel({'$\frac{S_n}{n}$'},'Interpreter','latex','FontSize',
    ,20,'rot',0);

```

```

45
46 %now we count the theoretical limit
47 H=0;
48 for i=2:m
49     H=H+1/i ;
50 end
51 H=2*H;
52 H=1/H;
53 plot ( t ,H*ones (n-m+1,1) , 'k—' );
54 end

```

Az  $m$ -áris fa szimulációja a 4.9. Tételben található eredmény szerint ábrázolva. (10. ábra)

```

1 %size_of_m_ary2.m
2
3 %simulation the size of an m-ary search tree , when m>26
4
5 function size_of_m_ary2(m,n)
6 K=3;%the number of simulations
7 %firstly we count the harmonic number that we will use
8 H=0;
9 for i=2:m
10     H=H+1/i ;
11 end
12 H=2*H;
13 H=1/H;
14 %H is the reciprocal of 2 * (the m-th Harmonic number-1)
15 [re ,im]=getSecondBiggestEigValue(m)%we will use only 're'
16 %this is the real part of the second biggest eigenvalue
17
18 %simulations
19 for k=1:K
20     colours=zeros (m-1,1);
21     colours (1)=1;
22     S=zeros (1,n);
23     internalNodes=0;

```

```

24     for i=1:n
25         %we know that i potencial places are
26         %we have to choose one place random
27         c=randi(i);
28         j=1;
29         while c>0
30             c=c-colours(j);
31             if c>0
32                 j=j+1;
33             end
34         end
35         %the chosen ball's colour is the j-th
36         if j==m-1           %if the chosen ball's colour is
            the last
37             colours(m-1)=colours(m-1)-(m-1);
38             colours(1)=colours(1)+m;
39             internalNodes=internalNodes+1;
40         else               % if not the last colour
41             colours(j)=colours(j)-j;
42             colours(j+1)=colours(j+1)+j+1;
43         end
44         for j=2:(m-1)
45             S(i)=S(i)+(colours(j)/j);
46         end
47         S(i)=S(i)+internalNodes; %we are adding the internal
            nodes
48         S(i)=(S(i)-i*H)/(i^re);
49     end
50     t=m:n;
51     hold on;
52     plot(log(t),S(m:n),'-', 'color',[mod(k,3)==0,mod(k+1,3)
            ==0,mod(k+2,3)==0], 'LineWidth',1.2);
53 end
54 xlabel('log(n)');
55 ylabel({'$\frac{S_n-\frac{n}{2(H_{m-1})}}{n^{\sigma_2}}$'},
        'Interpreter','latex','FontSize',20);
56 end

```