Dependence between Path Lengths and Size in Random Trees

(joint with H.-H. Chern, H.-K. Hwang and R. Neininger)

Michael Fuchs

Institute of Applied Mathematics
National Chiao Tung University

Hsinchu, Taiwan

Kraków, July 4, 2016

Random m-ary Search Trees

Proposed by Muntz and Uzgalis in 1971.

Random m-ary Search Trees

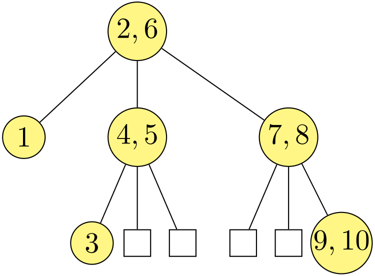
Proposed by Muntz and Uzgalis in 1971.

Input: 6, 2, 4, 8, 7, 1, 5, 3, 10, 9

Random m-ary Search Trees

Proposed by Muntz and Uzgalis in 1971.

Input: 6, 2, 4, 8, 7, 1, 5, 3, 10, 9

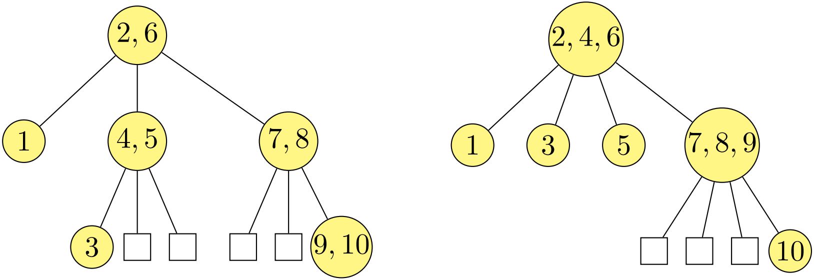


$m = 3$

Random m-ary Search Trees

Proposed by Muntz and Uzgalis in 1971.

Input: 6, 2, 4, 8, 7, 1, 5, 3, 10, 9

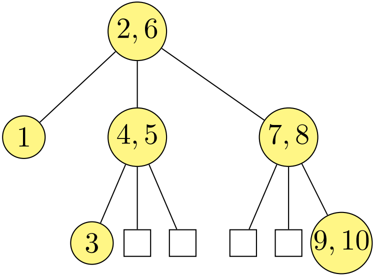


$m = 3$ $\qquad\qquad$ $m = 4$

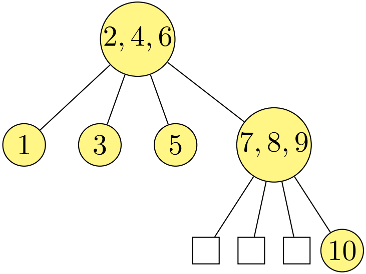
Random m-ary Search Trees

Proposed by Muntz and Uzgalis in 1971.

Input: 6, 2, 4, 8, 7, 1, 5, 3, 10, 9



$m = 3$



$m = 4$

If permutations are equally likely $\longrightarrow$ **random m-ary search trees**

Size, KPL, and NPL

- **Size** (or Storage Requirement)

  Number of nodes holding keys. Only random if $m \geq 3$.

  $S_n =$ size of a random m-ary search tree built from n keys.

Size, KPL, and NPL

- **Size** (or Storage Requirement)

  Number of nodes holding keys. Only random if $m \geq 3$.

  $S_n =$ size of a random m-ary search tree built from n keys.

- **Key Path Length** (KPL)

  Sum of all key-distances to the root.

  $K_n =$ KPL of a random m-ary search tree built from n keys.

Size, KPL, and NPL

- **Size** (or Storage Requirement)

  Number of nodes holding keys. Only random if $m \geq 3$.

  $S_n =$ size of a random m-ary search tree built from n keys.

- **Key Path Length** (KPL)

  Sum of all key-distances to the root.

  $K_n =$ KPL of a random m-ary search tree built from n keys.

- **Node Path Length** (NPL)

  Sum of all node-distances to the root.

  $N_n =$ NPL of a random m-ary search tree built from n keys.

Size: Mean

Knuth (1973):

$$\mathbb{E}(S_n) \sim \phi n,$$

where

$$\phi := \frac{1}{2(H_m - 1)}$$

and H_m are the Harmonic numbers.

Size: Mean

Knuth (1973):

$$\mathbb{E}(S_n) \sim \phi n,$$

where

$$\phi := \frac{1}{2(H_m - 1)}$$

and H_m are the Harmonic numbers.

Mahmoud and Pittel (1989):

$$\mathbb{E}(S_n) = \phi(n+1) - \frac{1}{m-1} + \mathcal{O}(n^{\alpha-1}),$$

where α is the real part of the second largest zero of

$$\Lambda(z) = z(z+1)\cdots(z+m-2) - m!.$$

Size: Phase Change for Variance

Mahmoud and Pittel (1989):

$$\mathrm{Var}(S_n) \sim \begin{cases} C_S n, & \text{if } m \le 26; \\ F_1(\beta \log n) n^{2\alpha-2}, & \text{if } m \ge 27, \end{cases}$$

where $\lambda = \alpha + i\beta$ is the second largest zero of $\Lambda(z)$.

Size: Phase Change for Variance

Mahmoud and Pittel (1989):

$$\mathrm{Var}(S_n) \sim \begin{cases} C_S n, & \text{if } m \le 26; \\ F_1(\beta \log n) n^{2\alpha-2}, & \text{if } m \ge 27, \end{cases}$$

where $\lambda = \alpha + i\beta$ is the second largest zero of $\Lambda(z)$.

Here, $F_1(z)$ is the periodic function

$$\begin{aligned} F_1(z) = 2\frac{|A|^2}{|\Gamma(\lambda)|^2}&\left(-1 + \frac{m!(m-1)|\Gamma(\lambda)|^2}{\Gamma(2\alpha+m-2) - m!\Gamma(2\alpha-1)}\right) \\ &+ 2\Re\left(\frac{A^2 e^{2iz}}{\Gamma(\lambda)^2}\left(-1 + \frac{m!(m-1)\Gamma(\lambda)^2}{\Gamma(2\lambda+m-2) - m!\Gamma(2\lambda-1)}\right)\right) \end{aligned}$$

with $A = 1/(\lambda(\lambda-1)\sum_{0\le j\le m-2}\frac{1}{j+\lambda})$.

Size: Phase Change for Limit Law

Theorem (Mahmoud & Pittel (1989); Lew & Mahmoud (1994))

For $3 \leq m \leq 26$,

$$\frac{S_n - \mathbb{E}(S_n)}{\sqrt{\mathrm{Var}(S_n)}} \xrightarrow{d} N(0,1),$$

where $N(0,1)$ *is the standard normal distribution.*

Size: Phase Change for Limit Law

Theorem (Mahmoud & Pittel (1989); Lew & Mahmoud (1994))

For $3 \leq m \leq 26$*,*

$$\frac{S_n - \mathbb{E}(S_n)}{\sqrt{\mathrm{Var}(S_n)}} \xrightarrow{d} N(0,1),$$

where $N(0,1)$ *is the standard normal distribution.*

Theorem (Chern & Hwang (2001))

For $m \geq 27$*,*

$$\frac{S_n - \mathbb{E}(S_n)}{\sqrt{\mathrm{Var}(S_n)}}$$

does not converge to a fixed limit law.

KPL: Moments

Mahmoud (1986):

$$\mathbb{E}(K_n) = 2\phi n \log n + c_1 n + o(n),$$

where c_1 is an explicitly computable constant.

KPL: Moments

Mahmoud (1986):

$$\mathbb{E}(K_n) = 2\phi n \log n + c_1 n + o(n),$$

where c_1 is an explicitly computable constant.

Mahmoud (1992):

$$\mathrm{Var}(K_n) \sim C_K n^2,$$

where

$$C_K = 4\phi^2 \left(\frac{(m+1)H_m^{(2)} - 2}{m-1} - \frac{\pi^2}{6} \right)$$

with $H_m^{(2)} = \sum_{1 \le j \le m} 1/j^2$.

KPL: Moments

Mahmoud (1986):

$$\mathbb{E}(K_n) = 2\phi n \log n + c_1 n + o(n),$$

where c_1 is an explicitly computable constant.

Mahmoud (1992):

$$\mathrm{Var}(K_n) \sim C_K n^2,$$

where

$$C_K = 4\phi^2 \left(\frac{(m+1)H_m^{(2)} - 2}{m-1} - \frac{\pi^2}{6} \right)$$

with $H_m^{(2)} = \sum_{1 \le j \le m} 1/j^2$.

So, no phase change here for the variance!

KPL: Limit Law

Theorem (Neininger & Rüschendorf (1999))

We have,

$$\frac{K_n - \mathbb{E}(K_n)}{n} \xrightarrow{d} K,$$

where K is the unique solution of

$$X \stackrel{d}{=} \sum_{1\le r\le m} V_r X^{(r)} + 2\phi \sum_{1\le r\le m} V_r \log V_r$$

with $X^{(r)}$ an independent copy of X and

$$V_r = U_{(r)} - U_{(r-1)},$$

where $U_{(r)}$ is the r-th order statistic of m i.i.d. uniform RVs.

Node Path Length (NPL)

N_n = sum of all node-distances in an m-search tree built from n keys.

Node Path Length (NPL)

N_n = sum of all node-distances in an m-search tree built from n keys.

Broutin and Holmgren (2012):

$$\mathbb{E}(N_n) = 2\phi^2 n \log n + c_2 n + o(n),$$

where c_2 is an explicitly computable constant.

Node Path Length (NPL)

N_n = sum of all node-distances in an m-search tree built from n keys.

Broutin and Holmgren (2012):

$$\mathbb{E}(N_n) = 2\phi^2 n \log n + c_2 n + o(n),$$

where c_2 is an explicitly computable constant.

We have,

$$\begin{cases} S_n \stackrel{d}{=} S_{I_1}^{(1)} + \cdots + S_{I_m}^{(m)} + 1, \\ N_n \stackrel{d}{=} N_{I_1}^{(1)} + \cdots + N_{I_m}^{(m)} + S_{I_1}^{(1)} + \cdots + S_{I_m}^{(m)}. \end{cases}$$

Node Path Length (NPL)

N_n = sum of all node-distances in an m-search tree built from n keys.

Broutin and Holmgren (2012):

$$\mathbb{E}(N_n) = 2\phi^2 n \log n + c_2 n + o(n),$$

where c_2 is an explicitly computable constant.

We have,

$$\begin{cases} S_n \stackrel{d}{=} S_{I_1}^{(1)} + \cdots + S_{I_m}^{(m)} + 1, \\ N_n \stackrel{d}{=} N_{I_1}^{(1)} + \cdots + N_{I_m}^{(m)} + S_{I_1}^{(1)} + \cdots + S_{I_m}^{(m)}. \end{cases}$$

So, one expects a strong positive dependence between S_n and N_n!

Size and NPL: Correlation (i)

Theorem (Chern, F., Hwang, Neininger)

We have,

$$\mathrm{Cov}(S_n, N_n) \sim \begin{cases} C_R n \log n, & \text{if } 3 \le m \le 13; \\ \phi F_2(\beta \log n) n^{\alpha}, & \text{if } m \ge 14, \end{cases},$$

where C_R is a constant and $F_2(z)$ is a periodic function. Moreover,

$$\mathrm{Var}(N_n) \sim \phi^2 C_K n^2.$$

Size and NPL: Correlation (i)

Theorem (Chern, F., Hwang, Neininger)

We have,

$$\mathrm{Cov}(S_n, N_n) \sim \begin{cases} C_R n \log n, & \text{if } 3 \le m \le 13; \\ \phi F_2(\beta \log n) n^{\alpha}, & \text{if } m \ge 14, \end{cases},$$

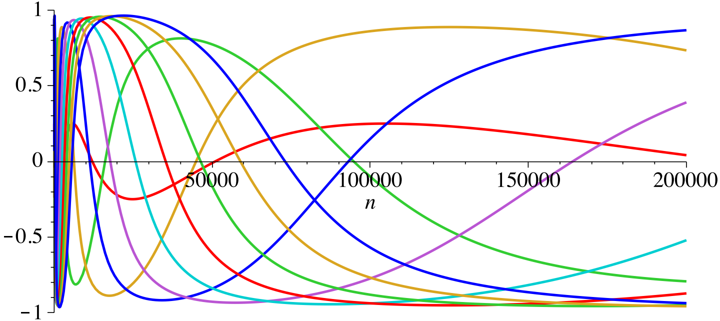
where C_R is a constant and $F_2(z)$ is a periodic function. Moreover,

$$\mathrm{Var}(N_n) \sim \phi^2 C_K n^2.$$

Thus (!),

$$\rho(S_n, N_n) \begin{cases} \longrightarrow 0, & \text{if } 3 \le m \le 26; \\ \sim \frac{F_2(\beta \log n)}{\sqrt{C_K F_1(\beta \log n)}}, & \text{if } m \ge 27. \end{cases}$$

Size and NPL: Correlation (ii)



Periodic function of $\rho(S_n, N_n)$ for $m = 27, 54, \ldots, 270$.

Pearson's Correlation Coefficient

Pearson: for RVs X and Y

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}.$$

Measures linear dependence between X and Y!

Pearson's Correlation Coefficient

Pearson: for RVs X and Y

$$\rho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}.$$

Measures linear dependence between X and Y!

Refined correlation measures:

Distance correlation, Brownian covariance, mutual information, total correlation, dual total correlation, etc.

Pearson's Correlation Coefficient

Pearson: for RVs X and Y

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}.$$

Measures linear dependence between X and Y!

Refined correlation measures:

Distance correlation, Brownian covariance, mutual information, total correlation, dual total correlation, etc.

Question: Can our counterintuitive result for $\rho(S_n, N_n)$ be ascribed to the weakness of Pearson's correlation coefficient?

Pearson's Correlation Coefficient

Pearson: for RVs X and Y

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}.$$

Measures linear dependence between X and Y!

Refined correlation measures:

Distance correlation, Brownian covariance, mutual information, total correlation, dual total correlation, etc.

Question: Can our counterintuitive result for $\rho(S_n, N_n)$ be ascribed to the weakness of Pearson's correlation coefficient? NO!

Size and NPL: Limit Law for $3 \leq m \leq 26$

Theorem (Chern, F., Hwang, Neininger)

Consider

$$Q_n = (S_n, N_n).$$

Then,

$$\mathrm{Cov}(Q_n)^{-1/2}\Big(Q_n - \mathbb{E}(Q_n)\Big) \xrightarrow{d} \Big(N, C_K^{-1/2}K\Big),$$

where N has a standard normal distribution.

Moreover, N and K are independent!

Size and NPL: Limit Law for $3 \leq m \leq 26$

Theorem (Chern, F., Hwang, Neininger)

Consider

$$Q_n = (S_n, N_n).$$

Then,

$$\mathrm{Cov}(Q_n)^{-1/2}\Big(Q_n - \mathbb{E}(Q_n)\Big) \xrightarrow{d} \Big(N, C_K^{-1/2}K\Big),$$

where N has a standard normal distribution.

Moreover, N and K are independent!

Thus, asymptotic independence for $3 \leq m \leq 26$ is also observed in the bivariate limit law!

Size and NPL: Limit Law for $m \geq 27$

Theorem (Chern, F., Hwang, Neininger)

Consider

$$Y_n = \left(\frac{S_n - \phi n}{n^{\alpha-1}}, \frac{N_n - \mathbb{E}(N_n)}{n} \right).$$

Then,

$$\ell_2(Y_n, (\Re(n^{i\beta}\Lambda), \phi K)) \longrightarrow 0,$$

where ℓ_2 is the minimal L_2-metric and Λ is the unique solution of

$$W \stackrel{d}{=} \sum_{1 \leq r \leq m} V_r^{\lambda-1} W^{(r)}$$

with $W^{(r)}$ independent copies of W.

Size and KPL

Same results hold for size and KPL, e.g.,

$$\rho(S_n, K_n) \begin{cases} \longrightarrow 0, & \text{if } 3 \le m \le 26; \\ \sim \rho(S_n, N_n), & \text{if } m \ge 27. \end{cases}$$

Size and KPL

Same results hold for size and KPL, e.g.,

$$\rho(S_n, K_n) \begin{cases} \longrightarrow 0, & \text{if } 3 \leq m \leq 26; \\ \sim \rho(S_n, N_n), & \text{if } m \geq 27. \end{cases}$$

Theorem (Chern, F., Hwang, Neininger)

We have $\rho(K_n, N_n) \sim 1$ *and*

$$\|N_n - \phi K_n - (\mathbb{E}(N_n - \phi K_n))\|_2 = o(n).$$

In particular,

$$\left(\frac{K_n - \mathbb{E}(K_n)}{n}, \frac{N_n - \mathbb{E}(N_n)}{n}\right) \xrightarrow{d} (K, \phi K).$$

Trie

Proposed by René de la Briandais in 1959.

Trie

Proposed by René de la Briandais in 1959.

Name from the word data retrieval (suggested by Fredkin).

Trie

Proposed by René de la Briandais in 1959.

Name from the word data retrieval (suggested by Fredkin).

Example:

$\square$

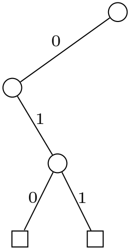
011011
010101
101110
010000
101010
001100

Trie

Proposed by René de la Briandais in 1959.

Name from the word data retrieval (suggested by Fredkin).

Example:



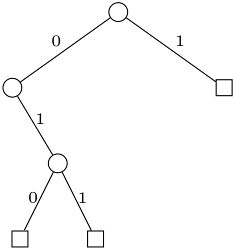
011011
010101
101110
010000
101010
001100

Trie

Proposed by René de la Briandais in 1959.

Name from the word data retrieval (suggested by Fredkin).

Example:



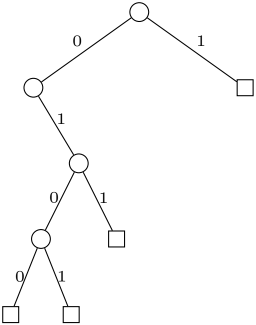
011011
010101
101110
010000
101010
001100

Trie

Proposed by René de la Briandais in 1959.

Name from the word data retrieval (suggested by Fredkin).

Example:



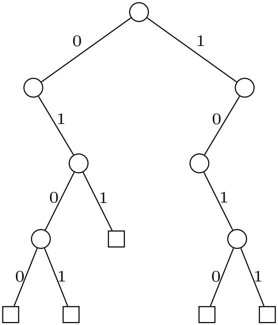
011011
010101
101110
010000
101010
001100

Trie

Proposed by René de la Briandais in 1959.

Name from the word data retrieval (suggested by Fredkin).

Example:



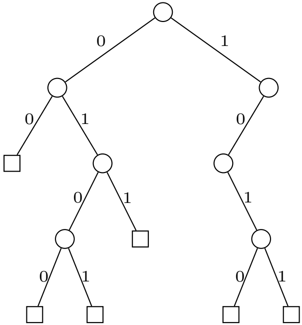
011011
010101
101110
010000
101010
001100

Trie

Proposed by René de la Briandais in 1959.

Name from the word data retrieval (suggested by Fredkin).

Example:



011011
010101
101110
010000
101010
001100

Notation and Random Model

Two types of nodes:

- Internal nodes: only used for branching;
- External nodes: nodes which hold data.

Notation and Random Model

Two types of nodes:

- Internal nodes: only used for branching;
- External nodes: nodes which hold data.

Random Model:

Bits are independent Bernoulli random variables with mean p.

Notation and Random Model

Two types of nodes:

- Internal nodes: only used for branching;
- External nodes: nodes which hold data.

Random Model:

Bits are independent Bernoulli random variables with mean p.

- $p = 1/2$: *symmetric trie*;
- $p \neq 1/2$: *asymmetric trie*.

Notation and Random Model

Two types of nodes:

- Internal nodes: only used for branching;
- External nodes: nodes which hold data.

Random Model:

Bits are independent Bernoulli random variables with mean p.

- $p = 1/2$: *symmetric trie*;
- $p \neq 1/2$: *asymmetric trie*.

Question: correlation between size and path-length in random tries?

Size, EPL, and IPL

- **Size**

  Number of internal nodes.

Size, EPL, and IPL

- **Size**

  Number of internal nodes.

- **External Path Length** (EPL)

  Sum of all distances between external nodes and root.

Size, EPL, and IPL

- **Size**

  Number of internal nodes.

- **External Path Length** (EPL)

  Sum of all distances between external nodes and root.

- **Internal Path Length** (IPL)

  Sum of all distances between internal nodes and root.

Size, EPL, and IPL

- **Size**

  Number of internal nodes.

- **External Path Length** (EPL)

  Sum of all distances between external nodes and root.

- **Internal Path Length** (IPL)

  Sum of all distances between internal nodes and root.

We again use S_n, K_n, N_n and

$$\begin{array}{ccc} \text{EPL} & \leftrightsquigarrow & \text{KPL;} \\ \text{IPL} & \leftrightsquigarrow & \text{NPL.} \end{array}$$

Some Notation

We use the following notation:

Some Notation

We use the following notation:

- Entropy: $h = -p\log p - q\log q$;

Some Notation

We use the following notation:

- Entropy: $h = -p\log p - q\log q$;
- If $\log p/\log q \in \mathbb{Q}$, then

$$\frac{\log p}{\log q} = \frac{r}{\ell}, \qquad \gcd(r,\ell) = 1.$$

Some Notation

We use the following notation:

- Entropy: $h = -p\log p - q\log q$;
- If $\log p/\log q \in \mathbb{Q}$, then

$$\frac{\log p}{\log q} = \frac{r}{\ell}, \qquad \gcd(r,\ell) = 1.$$

and

$$\chi_k = \frac{2rk\pi i}{\log p}, \qquad (k \in \mathbb{Z}).$$

Some Notation

We use the following notation:

- Entropy: $h = -p\log p - q\log q$;
- If $\log p/\log q \in \mathbb{Q}$, then

$$\frac{\log p}{\log q} = \frac{r}{\ell}, \qquad \gcd(r,\ell) = 1.$$

and

$$\chi_k = \frac{2rk\pi i}{\log p}, \qquad (k \in \mathbb{Z}).$$

- For a sequence g_k:

$$\mathscr{F}[g](z) = \begin{cases} \sum_{k\in\mathbb{Z}} g_k z^{-\chi_k}, & \text{if } \log p/\log q \in \mathbb{Q}; \\ g_0, & \text{if } \log p/\log q \notin \mathbb{Q}. \end{cases}$$

Means and Variances

Shape parameters	$\frac{1}{n}$(mean) $\sim$	$\frac{1}{n}$(variance) $\sim$
Size S_n	$\mathscr{F}[\cdot](n)$	$\mathscr{F}[g^{(1)}](n)$
NPL N_n	$\frac{\mathbb{E}(S_n)}{n}\frac{\log n}{h}$	$\frac{\mathrm{Var}(S_n)}{n}\frac{(\log n)^2}{h^2}$
KPL K_n	$\frac{\log n}{h}+\mathscr{F}[\cdot](n)$	$\frac{pq\log^2\frac{p}{q}}{h^2}\cdot\frac{\log n}{h}+\mathscr{F}[g^{(3)}](n)$
Depth D_n	$\mathbb{E}(D_n)=\frac{\mathbb{E}(K_n)}{n}$	$\mathrm{Var}(D_n)=\frac{\mathrm{Var}(K_n)}{n}+O(1)$

Means and Variances

Shape parameters	$\frac{1}{n}$(mean) $\sim$	$\frac{1}{n}$(variance) $\sim$
Size S_n	$\mathscr{F}[\cdot](n)$	$\mathscr{F}[g^{(1)}](n)$
NPL N_n	$\frac{\mathbb{E}(S_n)}{n}\frac{\log n}{h}$	$\frac{\mathrm{Var}(S_n)}{n}\frac{(\log n)^2}{h^2}$
KPL K_n	$\frac{\log n}{h}+\mathscr{F}[\cdot](n)$	$\frac{pq\log^2\frac{p}{q}}{h^2}\cdot\frac{\log n}{h}+\mathscr{F}[g^{(3)}](n)$
Depth D_n	$\mathbb{E}(D_n)=\frac{\mathbb{E}(K_n)}{n}$	$\mathrm{Var}(D_n)=\frac{\mathrm{Var}(K_n)}{n}+O(1)$

Note that

$$N_n \approx S_n\frac{\log n}{h}$$

and for $p \neq q$

$$\mathrm{Var}(D_n) \sim \frac{\mathrm{Var}(K_n)}{n}.$$

Main Approaches

Main Approaches

- **Rice Method**

  Exercise 54 in Section 5.2.2 of Knuth's book. Developed into a systematic tool by Flajolet and Sedgewick.

Main Approaches

- **Rice Method**

  Exercise 54 in Section 5.2.2 of Knuth's book. Developed into a systematic tool by Flajolet and Sedgewick.

- **Two-stage Approach**

  Introduced by Jacquet and Régnier. Further developed by Jacquet and Szpankowski.

Main Approaches

- **Rice Method**

  Exercise 54 in Section 5.2.2 of Knuth's book. Developed into a systematic tool by Flajolet and Sedgewick.

- **Two-stage Approach**

  Introduced by Jacquet and Régnier. Further developed by Jacquet and Szpankowski.

- **Poisson Variance and Covariance**

  F., Hwang and Zacharovas.

Main Approaches

- **Rice Method**

  Exercise 54 in Section 5.2.2 of Knuth's book. Developed into a systematic tool by Flajolet and Sedgewick.

- **Two-stage Approach**

  Introduced by Jacquet and Régnier. Further developed by Jacquet and Szpankowski.

- **Poisson Variance and Covariance**

  F., Hwang and Zacharovas.

- **Other Approaches**

  Elementary (Schachinger), probabilistic (Devroye, Janson, etc.)

Size: Variance

Theorem (Régnier & Jacquet (1989); Kirschenhofer & Prodinger (1991); F., Hwang, Zacharovas (2014))

We have,

$$\mathrm{Var}(S_n) \sim \mathscr{F}[g^{(1)}](n)n,$$

where

$$\begin{aligned} g_k^{(1)} =& \frac{\chi_k \Gamma(-1+\chi_k)}{h}\left(1 - \frac{\chi_k+3}{2^{1+\chi_k}}\right) \\ & - \frac{1}{h^2}\sum_{j\in\mathbb{Z}} \Gamma(\chi_j+1)\Gamma(\chi_{k-j}+1) \\ & - \frac{2}{h}\sum_{j\geq 1} \frac{(-1)^j (j+1+\chi_k)\Gamma(j+\chi_k)\left(p^{j+1}+q^{j+1}\right)}{(j-1)!(j+1)(1-p^{j+1}-q^{j+1})}. \end{aligned}$$

Size and NPL: Variances and Covariance

Theorem (F., Hwang, Zacharovas (2014))

We have,

$$\mathrm{Var}(S_n) \sim \mathscr{F}[g^{(1)}](n)n,$$

and

$$\mathrm{Cov}(S_n, N_n) \sim \mathscr{F}[g^{(1)}](n)\frac{n\log n}{h}$$

and

$$\mathrm{Var}(N_n) \sim \mathscr{F}[g^{(1)}](n)\frac{n\log^2 n}{h^2}.$$

Size and NPL: Variances and Covariance

Theorem (F., Hwang, Zacharovas (2014))

We have,

$$\mathrm{Var}(S_n) \sim \mathscr{F}[g^{(1)}](n)n,$$

and

$$\mathrm{Cov}(S_n, N_n) \sim \mathscr{F}[g^{(1)}](n)\frac{n\log n}{h}$$

and

$$\mathrm{Var}(N_n) \sim \mathscr{F}[g^{(1)}](n)\frac{n\log^2 n}{h^2}.$$

Corollary

We have,

$$\rho(S_n, N_n) \longrightarrow 1.$$

Size and NPL: Limit Law

Theorem (F. & Lee (2015))

We have,

$$\left(\frac{S_n - \mathbb{E}(S_n)}{\sqrt{\mathrm{Var}(S_n)}}, \frac{N_n - \mathbb{E}(N_n)}{\sqrt{\mathrm{Var}(N_n)}}\right) \xrightarrow{d} \mathcal{N}(0, E_2),$$

where E_2 is the 2×2 unit matrix

$$E_2 = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}.$$

Size and NPL: Limit Law

Theorem (F. & Lee (2015))

We have,

$$\left(\frac{S_n - \mathbb{E}(S_n)}{\sqrt{\mathrm{Var}(S_n)}}, \frac{N_n - \mathbb{E}(N_n)}{\sqrt{\mathrm{Var}(N_n)}}\right) \xrightarrow{d} \mathcal{N}(0, E_2),$$

where E_2 is the 2×2 unit matrix

$$E_2 = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}.$$

NPL was also investigated by Nguyen-The (2004) in his PhD-thesis, but his result is incorrect.

Size and NPL: Limit Law

Theorem (F. & Lee (2015))

We have,

$$\left(\frac{S_n - \mathbb{E}(S_n)}{\sqrt{\mathrm{Var}(S_n)}}, \frac{N_n - \mathbb{E}(N_n)}{\sqrt{\mathrm{Var}(N_n)}}\right) \xrightarrow{d} \mathcal{N}(0, E_2),$$

where E_2 is the 2×2 unit matrix

$$E_2 = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}.$$

NPL was also investigated by Nguyen-The (2004) in his PhD-thesis, but his result is incorrect.

Question: same result for size and KPL?

Size and KPL: Covariance

Theorem (F. & Hwang)

We have,

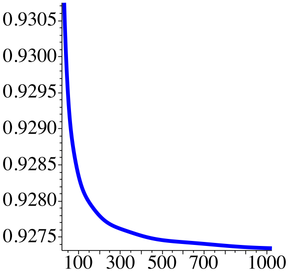
$$\mathrm{Cov}(S_n, K_n) \sim \mathscr{F}[g^{(2)}](n)n,$$

where

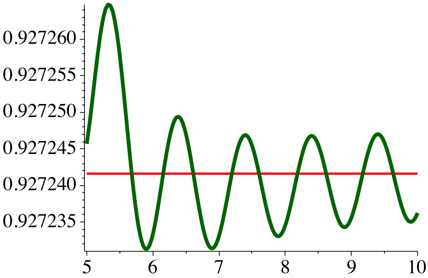
$$\begin{aligned} g_k^{(2)} &= \frac{\Gamma(\chi_k)}{h}\left(1 - \frac{\chi_k + 2}{2^{\chi_k+1}}\right) \\ &\quad - \frac{1}{h^2}\sum_{j\in\mathbb{Z}\setminus\{0\}} \Gamma(\chi_{k-j}+1)(\chi_j - 1)\Gamma(\chi_j) \\ &\quad - \frac{\Gamma(\chi_k+1)}{h^2}\left(\gamma + 1 + \psi(\chi_k+1) - \frac{p\log^2 p + q\log^2 q}{2h}\right) \\ &\quad + \frac{1}{h}\sum_{j\geq 2}\frac{(-1)^j(2j^2 - 2j + 1 + (2j-1)\chi_k)\Gamma(j-1\chi_k)(p^j + q^j)}{j!(1-p^j-q^j)}. \end{aligned}$$

Correlation Coefficient

$p = q = 1/2$:



$\rho(S_n, K_n)$



$$\frac{\mathrm{Cov}(S_n, K_n)}{\sqrt{\mathrm{Var}(S_n)(\mathrm{Var}(K_n)+1.046)}}$$

Size and KPL: Correlation Coefficient

Theorem (F. & Hwang)

We have,

$$\rho(S_n, K_n) \sim \begin{cases} 0, & \textit{if } p \neq q; \\ F(n), & \textit{if } p = q, \end{cases}$$

where

$$F(n) = \frac{\mathscr{F}[g^{(2)}](n)}{\sqrt{\mathscr{F}[g^{(1)}](n)\mathscr{F}[g^{(3)}](n)}}$$

is a periodic function with

$$\textit{average value} = 0.927\cdots \ \textit{and amplitude} \leq 1.5 \times 10^{-5}.$$

Size and KPL: Correlation Coefficient

Theorem (F. & Hwang)

We have,

$$\rho(S_n, K_n) \sim \begin{cases} 0, & \textit{if } p \neq q; \\ F(n), & \textit{if } p = q, \end{cases}$$

where

$$F(n) = \frac{\mathscr{F}[g^{(2)}](n)}{\sqrt{\mathscr{F}[g^{(1)}](n)\mathscr{F}[g^{(3)}](n)}}$$

is a periodic function with

$$\textit{average value} = 0.927\cdots \quad \textit{and amplitude} \leq 1.5 \times 10^{-5}.$$

Question: is now this behavior due to the weakness of Pearson's correlation coefficient?

Size and KPL: Correlation Coefficient

Theorem (F. & Hwang)

We have,

$$\rho(S_n, K_n) \sim \begin{cases} 0, & \textit{if } p \neq q; \\ F(n), & \textit{if } p = q, \end{cases}$$

where

$$F(n) = \frac{\mathscr{F}[g^{(2)}](n)}{\sqrt{\mathscr{F}[g^{(1)}](n)\mathscr{F}[g^{(3)}](n)}}$$

is a periodic function with

$$\textit{average value} = 0.927\cdots \quad \textit{and amplitude} \leq 1.5 \times 10^{-5}.$$

Question: is now this behavior due to the weakness of Pearson's correlation coefficient? Again NO!

Size and KPL: Limit Laws

Theorem (F. & Hwang)

- $p \neq q$:

$$\left(\frac{S_n - \mathbb{E}(S_n)}{\sqrt{\mathrm{Var}(S_n)}}, \frac{K_n - \mathbb{E}(K_n)}{\sqrt{\mathrm{Var}(K_n)}}\right) \xrightarrow{d} \mathcal{N}(0, I_2),$$

where I_2 is the 2×2 identity matrix.

Size and KPL: Limit Laws

Theorem (F. & Hwang)

- $p \neq q$:

$$\left(\frac{S_n - \mathbb{E}(S_n)}{\sqrt{\mathrm{Var}(S_n)}}, \frac{K_n - \mathbb{E}(K_n)}{\sqrt{\mathrm{Var}(K_n)}}\right) \xrightarrow{d} \mathcal{N}(0, I_2),$$

where I_2 is the 2×2 identity matrix.

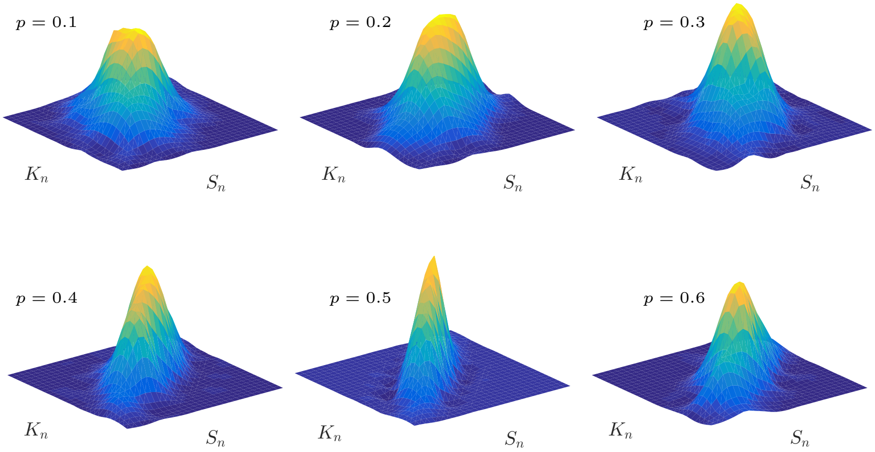
- $p = q$:

$$\Sigma_n^{-1/2}\Big(S_n - \mathbb{E}(S_n), K_n - \mathbb{E}(K_n)\Big) \xrightarrow{d} \mathcal{N}_2(0, I_2),$$

where Σ_n is the (asymptotic) covariance matrix:

$$\Sigma_n := n \begin{pmatrix} \mathscr{F}[g^{(1)}](n) & \mathscr{F}[g^{(2)}](n) \\ \mathscr{F}[g^{(2)}](n) & \mathscr{F}[g^{(3)}](n) \end{pmatrix}.$$

Joint Distribution of S_n and K_n



Summary

trees	$\rho(S_n, K_n)$	$\rho(S_n, N_n)$
tries	$\begin{cases} p \neq q :\to 0 \\ p = q : \text{periodic} \end{cases}$	~ 1
m-ary search trees	$\begin{cases} 3 \leq m \leq 26 :\to 0 \\ m \geq 27 : \text{periodic} \end{cases}$	

Summary

trees	$\rho(S_n, K_n)$	$\rho(S_n, N_n)$
tries	$\begin{cases} p \neq q :\to 0 \\ p = q : \text{periodic} \end{cases}$	~ 1
m-ary search trees	$\begin{cases} 3 \leq m \leq 26 :\to 0 \\ m \geq 27 : \text{periodic} \end{cases}$	

- Similar results for fringe-balanced binary search trees and quadtrees:

  H.-H. Chern, M. Fuchs, H.-K. Hwang, R. Neininger. Dependence and phase changes in random m-ary search trees, arxiv:1501.05135.

Summary

trees	$\rho(S_n, K_n)$	$\rho(S_n, N_n)$
tries	$\begin{cases} p \neq q :\to 0 \\ p = q : \text{periodic} \end{cases}$	~ 1
m-ary search trees	$\begin{cases} 3 \leq m \leq 26 :\to 0 \\ m \geq 27 : \text{periodic} \end{cases}$	

- Similar results for fringe-balanced binary search trees and quadtrees:

  H.-H. Chern, M. Fuchs, H.-K. Hwang, R. Neininger. Dependence and phase changes in random m-ary search trees, arxiv:1501.05135.

- Similar results for m-ary tries, m-ary PATRICIA tries and bucket digital search trees.

Summary

trees	$\rho(S_n, K_n)$	$\rho(S_n, N_n)$
tries	$\begin{cases} p \neq q :\to 0 \\ p = q : \text{periodic} \end{cases}$	~ 1
m-ary search trees	$\begin{cases} 3 \leq m \leq 26 :\to 0 \\ m \geq 27 : \text{periodic} \end{cases}$	

- Similar results for fringe-balanced binary search trees and quadtrees:

  H.-H. Chern, M. Fuchs, H.-K. Hwang, R. Neininger. Dependence and phase changes in random m-ary search trees, arxiv:1501.05135.

- Similar results for m-ary tries, m-ary PATRICIA tries and bucket digital search trees.

- Better explanation of our results?