Analytical Approach to Some Problems Involving Order Statistics

W. Szpankowski*
Department of Computer Science
Purdue University
W. Lafayette, IN 47907

June 16, 1995

*Part of this work was done jointly with P. Jacquet, P. Kirschenhofer, G. Louchard, H. Mahmoud, and H. Prodinger.

Outline of the Talk

1. General Approach to Order Statistics:
   (a) Inclusion-Exclusion Formula
   (b) Refreshment: Limiting Distributions for I.I.D. Case: Probabilistic Method
2. Analytical Approach to Limiting Distribution of Order Statistics
3. Applications:
   (a) I.I.D. Case (Sum of Geometric RV)
   (b) Dependent Case (Depth in a trie)
   (c) Dependent Case: Open Problems (e.g., height in a trie)
4. Order Statistics by a Functional Equation Approach:
   (a) Depth in a Digital Search Tree
   (b) Leader Election Algorithm
   (c) A Probabilistic Counting Algorithm

General Approach to Order Statistics

ORDER STATISTICS for a sequence $X_1, X_2, \ldots, X_n$ is

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)} \ .$$

Throughout the talk we assume that $X_1, X_2, \ldots, X_n$ are DISCRETE random variables.

Let $X_1, X_2, \ldots, X_n$ be a sequence of EXCHANGEABLE random variables.

DEFINITION. A sequence $X_1, \ldots, X_n$ is called EXCHANGEABLE if for any $1 \leq i_1 \leq \cdots \leq i_r \leq n$ we have

$$\Pr\{X_{i_1} < x_{i_1}, \ldots, X_{i_r} < x_{i_r}\} = \Pr\{X_1 < x_1, \ldots, X_r < x_r\} \ .$$

We will use also the following notation:

$$\begin{aligned} M_n &= \max\{X_1, \ldots, X_n\} = X_{(n)} \\ M_{r,n} &= X_{(r)} \ . \end{aligned}$$

General Approach – Continuation

The following is a consequence of the INCLUSION-EXCLUSION PRINCIPLE:

$$\begin{aligned}\Pr\{M_n > k\} &= \Pr\left\{\bigcup_{r=1}^{n}\{X_r > k\}\right\} \\ &= \sum_{r=1}^{n}(-1)^{r+1}\binom{n}{r}\Pr\{X_1 > k, \ldots, X_r > k\} \ .\end{aligned}$$

In general:

$$\begin{aligned}\Pr \quad & \{X_{(n-k+1)} \le x\} = \{\text{At most k-1 of } X_j \ge x \text{ occur}\} \\ = \quad & \sum_{t=0}^{k-1}\sum_{r=0}^{n-t}(-1)^r\binom{r+t}{t}\binom{n}{r+t}\Pr\{X_1 \ge x, \ldots, X_{r+t} \ge x\} \ .\end{aligned}$$

CONCLUSION: One has to deal with alternating sums !

Limiting Distribution – Probabilistic Method for I.I.D. Case

Assume for a moment that $X_1, \ldots, X_n$ are CONTINUOUS RV. Let $F(x) = \Pr\{X_i < x\}$ and $H_n(x) = \Pr\{M_n < x\}$.

Let exist sequences a_n and $b_n > 0$ such that

$$\lim_{n\to\infty} H_n(a_n + b_n x) = H(x) \ .$$

Then, $H(x)$ is called LIMITING DISTRIBUTION of M_n

Depending on the tail behevior of $F(x)$ (more precisely: behavior of $\lim_{t\to\infty} \frac{1-F(tx)}{1-F(t)}$ or $\lim_{t\to\infty} \frac{1-F(t+x)}{1-F(t)}$) there are THREE possible limiting distributions (i.e., extreme distributions), that is:

$$\begin{aligned}
H_{1,\gamma} &= \exp\left(-x^{-\gamma}\right) & x > 0 \ , \\
H_{2,\gamma} &= \exp\left(-(-x)^{\gamma}\right) & x < 0 \ , \\
H_{3,0} &= \exp\left(-e^{-x}\right) & -\infty < x < \infty
\end{aligned}$$

Often a_n can be computed $a_n = \inf\{x : \ 1 - F(x) \le 1/n\}$.

Limiting Distribution – Probabilistic Method for DISCRETE I.I.D. Case

We are back again to $\textsc{Discrete}$ i.i.d. RV. Let $\Pr\{X = k\} = p_k$. Anderson in 1970 noticed that:

If

$$\frac{p_k}{\sum_{j=k}^{\infty} p_j}$$

does not converge to 0 or ∞ as $k \to \infty$, then there are no such sequences a_n and b_n that

$$\frac{M_n - a_n}{b_n}$$

converges to a non-degenerate distribution.

Limiting Distribution – Example

EXAMPLE: Geometric Distribution.

Assume that $\Pr\{X = k\} = p^k(1-p)$ for $k \geq 0$. Then, limiting distribution does not exist, but asymptotically for large n

$$\Pr\{M_n < \log_{1/p} n + m\} \sim \exp\left(-p^{m - \{\log_{1/p} n + m\}}\right)$$

where $\{t\} = t - \lfloor t \rfloor$. Indeed, we have

$$\begin{aligned} \Pr\{M_n < k\} &= (F(k))^n = (1 - (1 - F(k))^n \\ &= \exp(n \log(1 - (1 - F(k))) \\ &\sim \exp(-n(1 - F(k)) \end{aligned}$$

Observe that $1 - F(k) = p^k$. Thus, for $k = \lfloor \log_{1/p} n + m \rfloor$ we have

$$\begin{aligned} \Pr\{M_n < \lfloor \log_{1/p} n + m \rfloor\} &= \Pr\{M_n < \log_{1/p} n + m\} \\ &\sim \exp\left(np^{\lfloor \log_{1/p} n + m \rfloor}\right) \\ &= \exp\left(-p^{m - \{\log_{1/p} n + m\}}\right) . \end{aligned}$$

Analytical Approach – General Case

Define

$$M_n(z) = \sum_{k=0}^{\infty} \Pr\{M_n = k\} z^k \quad , \quad \widehat{M}_n(z) = \sum_{k=0}^{\infty} \Pr\{M_n > k\} z^k \ ,$$

$$X(z) = \sum_{k=0}^{\infty} \Pr\{X = k\} z^k \quad , \quad \widehat{X}_r(z) = \sum_{k=0}^{\infty} \Pr\{X > k\} z^k \ ,$$

and

$$\widehat{F}_r(z) = \sum_{k=0}^{\infty} \Pr\{X_1 > k, \ldots, X_r > k\} z^k \ .$$

Our previous equation implies

$$\widehat{M}_n(z) = \sum_{r=1}^{n} (-1)^{r+1} \binom{n}{r} \widehat{F}_r(z) \ .$$

Analytical Approach – I.I.D.

Assume that $X_1, \ldots, X_n$ are independently identically distributed random variables. Then:

$$\Pr\{X_1 > k, \ldots, X_r > k\} = (P\{X > k\})^r \ ,$$

and

$$\widehat{F}_r(z) = \widehat{X}^{(r)}(z) = \underbrace{\widehat{X}(z) \odot \cdots \odot \widehat{X}(z)}_{\text{r times}} \ ,$$

where

$$\widehat{X}^{(r)}(z) = \sum_{k=0}^{\infty} P^r\{X > k\} z^k \ ,$$

where $\odot$ denotes the Hadamard product, and (r) denotes r Hadamard's products.

Thus, we have

$$\widehat{M}(z) = \sum_{r=1}^{n} (-1)^{r+1} \binom{n}{r} \widehat{X}^{(r)}(z) \ .$$

Analytical Approach – Example

Let us now consider a special case, namely, let X be a sum of geometric distributions. More generally, we assume

$$\widehat{X}(z) = \frac{a}{(1-\rho z)^d}$$

where a, ρ and d are constants. We need to estimate $\widehat{X}^{\textcircled{r}}(z)$. But (assume d is an integer)

$$[z^n]\widehat{X}(z) = a\binom{n+d-1}{d-1}\rho^n .$$

Hence, with $d_1 = d-1$

$$[z^n]\widehat{X}^{\textcircled{r}}(z) = a^r \frac{n^{d_1 r}}{(d_1!)^r}\rho^{nr} + \Theta(n^{(d-1)r-1})$$

and

$$\widehat{X}^{\textcircled{r}}(z) = \frac{(d_1 r)!}{(d_1!)^r}\frac{a^r}{(1-\rho^r z)^{d_1 r+1}} + \Theta\left(\frac{1}{(1-\rho^r z)^{d_1 r}}\right) .$$

Example – Continuation

By Rice's method or Mellin-like approach we have

$$\sum_{r=1}^{n}(-1)^r\binom{n}{r}f(r) = -\frac{1}{2\pi i}\oint_{\mathcal{C}}\frac{f(s)}{s}\omega_n(s)ds$$

where

$$\omega_n(s) = \frac{1}{(1-\frac{s}{1})(1-\frac{s}{2})\cdots(1-\frac{s}{n})} \ .$$

In our case:

$$f(s) = \frac{\Gamma(d_1 s+1)a^s}{\Gamma(d)^s(1-\rho^s z)^{d_1 s+1}} \ .$$

Using

$$\omega_n(s) = n^s\Gamma(1-s)(1+O(1/n)) \ ,$$

we finally obtain

$$\begin{aligned}\widehat{M}_n(z) &= \frac{1}{1-z}+\left(1+O\left(\frac{1}{n}\right)\right)\oint_{\mathcal{C}}\frac{n^s a^s\Gamma(-s)\Gamma(d_1 s+1)}{\Gamma^s(d)(1-\rho^s z)^{d_1 s+1}} \\ &+ \text{ smaller order terms .}\end{aligned}$$

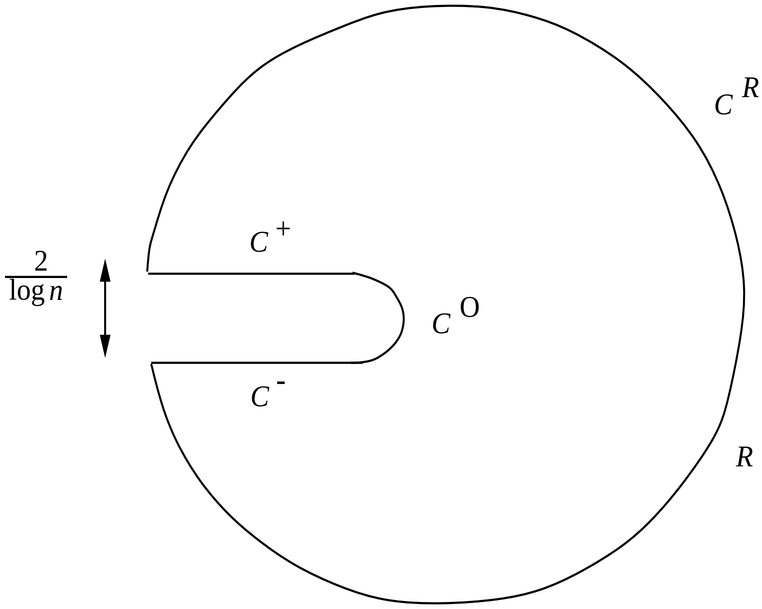
Example – Continuation

After substitution $1 - \rho^s z = u$, the integral becomes

$$I_n(z) = -\frac{(an)^{\lg(z)}}{\Gamma^{\lg z}(d)\log\rho}.$$

$$\frac{1}{2\pi i}\oint \frac{an)^{-\lg(1-u)}\Gamma(-\lg z + \lg(1-u))\Gamma(1 + d_1(\lg z - \lg(1-u))}{(1-u)u^{d_1(\lg z - \lg(1-u))+1}\Gamma^{-\lg(1-u)}(d)} d$$

where $\lg(\cdot) = \log_{1/\rho}(\cdot)$.



But $\widehat{M}^+(z) = \widehat{M}^-(z) = O(e^{-\sqrt{\log n}})$, $\widehat{M}^R(z) = O(R^{-n})$.

Example – Continuation

Thus, the main contribution comes from $\mathcal{C}^o$, and after some tedious analysis we obtain

$$\widehat{M}(z) = \frac{1}{1-z} - \left(1 + O\left(\frac{\log\log n}{\log n}\right)\right) \frac{z^{a_n}\Gamma(-\mathrm{lg} z)}{\log \rho^{-1}}$$

where

$$a_n = \mathrm{lg} n + d_1 \mathrm{lglg} n - \mathrm{lg}\Gamma(d) \ .$$

Now, we apply Cauchy to get for any integer k

$$\begin{aligned}
\Pr\{M_n \le a_n + k\} &= \Pr\{M_n \le \lfloor a_n \rfloor + k\} \\
&= \frac{1}{2\pi i} \oint \frac{z^{a_n}\Gamma(-\mathrm{lg} z)}{\log \rho^{-1} z^{\lfloor a_n \rfloor + k + 1}} dz \\
&= \frac{1}{2\pi i} \oint \frac{z^{\{a_n\}}\Gamma(-\mathrm{lg} z)}{\log \rho^{-1} z^{k+1}} dz
\end{aligned}$$

where $\{a_n\} = a_n - \lfloor a_n \rfloor$.

Apply now Cauchy's residue theorem (i.e., for $z = -\rho^m$, $m = 0, 1, \ldots$) to obtain the asymptotic distribution presented in the next Theorem.

Example – Final Result

THEOREM. *Let* $a_n = \mathrm{lg} n + d_1 \mathrm{lg}\mathrm{lg} n - \mathrm{lg}\Gamma(d)$*. Then, for any integer* k *we have*

$$\Pr\{M_n \le a_n + k\} = \frac{1}{\log \rho^{-1}} e^{-\rho^{k+1+\{a_n\}}} (1 + O(\frac{\log\log n}{\log n})) ,$$

that is, the limiting distribution does not exists due to the fact that $\{a_n\}$ *is dense in* $[0, 1]$ *but not uniformly dense.*

Dependent RV: Depth in a Trie

We now show how our previous analysis almost without a change can be used to derive an asymptotic distribution of a dependent sequence or random variables.

We consider several examples related to digital trees. Thus, a brief reminder of digital trees as shown below.

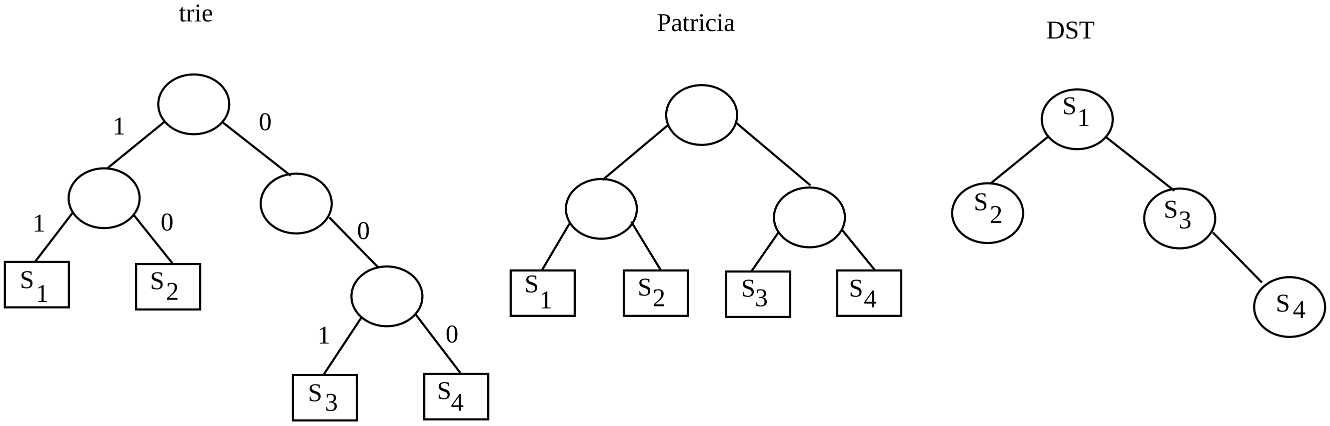


Figure 1: A trie, Patricia trie and a digital search tree (DST) built from the following four strings $S_1 = 11100\ldots$, $S_2 = 10111\ldots$, $S_3 = 00110\ldots$, and $S_4 = 00001\ldots$.

Depth in a Trie

Let $S_1, \ldots, S_n$ be n strings of "0" and "1" symbols occurring with probability p and $q = 1 - p$, respectively. Build a trie from these n sequences, and let D_n be the depth of a randomly selected external node. Define:
C_{ij} – the length of the longest prefix common to the i and the j strings S_i and S_j.
Then:

$$D_n \stackrel{d}{=} D_n(1) = \max\{C_{12}, C_{13}, \ldots, C_{1,n}\} \ .$$

where $\stackrel{d}{=}$ means "equal in distribution".

Certainly, C_{ij} are DEPENDENT. But, still the following holds

$$\begin{aligned}
\Pr\{D_n \geq k\} &= \Pr\{\bigcup_{r=2}^{n} [C_{1,r} \geq k]\} \\
&= n^{-1} \sum_{r=2}^{n} (-1)^r \binom{n}{r} r \Pr\{C_{12} \geq k, \ldots, C_{1,r} \geq k\}
\end{aligned}$$

Depth in a Trie – Continuation

But

$$\Pr\{C_{12} \ge k, \ldots, C_{1,r} \ge k\} = (p^r + q^r)^k \ .$$

Let now $\widehat{G}_n(z) = \sum_{k=0}^{\infty} \Pr\{D_n \ge k\} z^k$. Then:

$$\widehat{G}_n(z) = n^{-1} \sum_{r=2}^{n} (-1)^r \binom{n}{r} \frac{r}{1 - z(p^r + q^r)} \ .$$

Using now the Rice or Mellin-Like method we immediately obtain:

$$\widehat{G}_n(z) = (1+O(1/n)) \frac{1}{2\pi i} \int_{-3/2-i\infty}^{-3/2+i\infty} \frac{n^{-s-1}\Gamma(s+1)}{1 - z(p^{-s} + q^{-s})} ds \ .$$

Depth in a Trie – Symmetric Case

Let us first consider the symmetric case, i.e., $p = q = 1/2$. Then

$$\begin{aligned}
\widehat{G}_n(z) &= \int_{-3/2-i\infty}^{-3/2+i\infty} \frac{\Gamma(s+1)n^{-s-1}}{1 - z2^{-s-1}} ds \\
&= \frac{1}{1-z} - \frac{n^{\log_2 z}\Gamma(\log_2 z)}{\log 2} .
\end{aligned}$$

By Cauchy's formula we have for an integer k

$$\begin{aligned}
\Pr\{D_n &\leq \log_2 n + k\} = \Pr\{D_n \leq \lfloor \log_2 n \rfloor + k\} \\
&= \frac{1}{2\pi i \log 2} \oint \frac{z^{\log_2 n}\Gamma(\log_2 z)}{z^{\lfloor \log_2 n \rfloor + k + 1}} dz \\
&= \frac{1}{\log 2} \exp\left(-2^{k+1+\{\log_2 n\}}\right) ,
\end{aligned}$$

thus limiting distribution does not exists.

Depth in a Trie – Asymmetric Case

Now we consider the asymmetric case, that is, $p \neq q$, and prove that D_n appropriately normalized is normally distributed. Actually, we prove that for a complex θ the following (Goncharev's) condition holds

$$\lim_{n\to\infty} e^{-\theta\mu n/\sigma n} G_n(e^{\theta/\sigma n}) = e^{\frac{1}{2}\theta^2}$$

where

$$\begin{aligned}
\mu_n &= ED_n \sim \frac{1}{h}\log n + O(1) \\
\sigma_n^2 &= \mathrm{Var}\ D_n = c_2 \log n + O(1)
\end{aligned}$$

where $h = -p\log p - q\log q$ and $c_2 = \frac{h_2 - h^2}{h^3}$ with $h_2 = p\log^2 p + q\log^2 q$.

Depth in a Trie – Asymmetric Case

Define

$$\tau = \frac{\theta}{\sigma_n} = O\left(\frac{\theta}{\sqrt{\log n}}\right) \to 0 \ .$$

Then:

$$1 - G_n(e^\tau) = -\frac{1-e^\tau}{n} \int_{(-3/2)} \frac{n^{-s}\Gamma(s+1)}{1 - e^\tau(p^{-s}+q^{-s})} ds \ .$$

Let $s_k(\tau)$ $(k = 0, \pm 1, \ldots)$ be solutions of

$$e^{-\tau} = p^{-s} + q^{-s} \ .$$

In particular,

$$s_0(\tau) = 1 - \frac{\tau}{h} + \frac{1}{2} c_2 \tau^2 + O(\tau^3) \ .$$

Applying Cauchy's residues theorem, after some tedious algebra, we have

$$\lim_{n\to\infty} e^{-\theta \mu_n / \sigma_n} G_n(e^{\theta/\sigma_n}) = e^{\frac{1}{2}\theta^2} \left(1 + O\left(\frac{1}{\sqrt{\log n}}\right)\right) \ .$$

Depth in a Trie – Markovian Case

Similar approach was used by Jacquet and Régnier (1986) but they worked on bivariate generating functions $G(z,w) = \sum_{n\geq 0} G_n(z)\frac{w^n}{n!}$ which required DEPOISSONIZATION.

We avoided depoissonization, and can extend the analysis to Markovian models. Indeed, in this case, we assume that the next symbol depends on previous one, and the dependency is described by the matrix P. Let also $P_{[r]}$ denotes the r-th Schur product of P (i.e., elementwise product), and π ($\pi_{[r]}$) be the stationary distribution (Schur product). The unit vector is $\psi = (1,1,\ldots,1)$. Finally, we write $< \pi|P|\psi >$ to denote $< \pi, P\psi >$ where $< x, y >$ is the inner product of vectors x and y.

Then, the generating function of the depth D_n in the Markovian model becomes

$$\widehat{G}_n(z) = n^{-1}\sum_{r=2}^{n}(-1)^n\binom{n}{r} r < \pi_{[r]}|(I - P_{[r]})^{-1}|\psi_{[r]} > \quad .$$

Rest is "easy" as in the BERNOULLI case. Details can be found in Jacquet & Szpankowski, TRANS. INFORMATION THEORY, 37, 1991.

Height in a Trie – Open Problem

The HEIGHT is defined as the maximum depth, and can be computed as

$$H_n = \max_{1 \le i < j \le n} \{C_{ij}\} \ .$$

But, C_{ij} are not EXCHANGEABLE since for example

$$\begin{aligned}
\Pr\{C_{12} \ge k, C_{13} \ge k\} &= (p^3 + q^3)^k \ , \\
\Pr\{C_{12} \ge k, C_{45} \ge k\} &= (p^2 + q^2)^k (p^2 + q^2)^k \ .
\end{aligned}$$

We need an extension of Rice's method. May be the following one suffices

$$\sum_{k=m}^{\infty} (-1)^k f_k = \frac{1}{2\pi i} \int_{m-1/2-i\infty}^{m-1/2+i\infty} \frac{\pi}{\sin(\pi s)} f(s) ds \ ?$$

Order Statistics by Functional Equations Approach

For some problems involving order statistics another approach based on functional equations is more convenient. But it is rather restricted to a case by case study.

We discuss below three problems, namely: depth in a digital search tree, electing a leader, and a probabilistic counting algorithm.

Depth in a Digital Search Tree:

$$D_m(i) = \max\{\min\{C_{i1}, D_m(1)\}, \ldots, \min\{C_{i,i-1}, D_m(i-1)\}\} .$$

Electing a Leader: n people throw coins to find a winner that the last one that throws, say. a head is a winner). How long it takes?

$$H_n = 1+\max\{j \geq 0 : L_j > 1\} = 1+\max_{1\leq j<K}\{C_j\} = 1+C_{K-1}.$$

Probabilistic Counting. The following quantity is seeked that estimates the cardinality of a set:

$$R_{N,d} = \min\{k : \ bitmap(i) < d+1 \ 0 \leq i < k \ bitmap(i) = d+1\}$$

Depth in a Digital Search Tree

Let $\overline{B}_m(k) := EB_m(k)$ be the average number of internal nodes at level k in a digital tree built over m independent strings (we rather use m instead of n to distinguish it from the Lempel-Ziv model discuss below).

$$\Pr\{D_m = k\} = \frac{\overline{B}_m(k)}{m} \ .$$

Let $B_m(u) = \sum_{k=0}^{\infty} \overline{B}_m(k) u^k$. Then,

$$B_{m+1}(u) = 1 + u \sum_{j=0}^{m} \binom{m}{j} p^j q^{m-j} (B_j(u) + B_{m-j}(u))$$

Using a general solution (cf. Szpankowski, TCS 1991) we have

$$B_m(u) = m - (1-u) \sum_{k=2}^{m} (-1)^k \binom{m}{k} Q_{k-2}(u)$$

where

$$Q_k(u) = \prod_{j=2}^{k+1} (1 - up^j - uq^j) \quad , \quad Q_0(u) = 1 \ .$$

Depth in a Digital Search Tree – Symmetric Case

We have $ED_m = \log_2 m + O(1)$ and $\mathrm{Var}\ D_m = O(1)$. But:

$$B_m(u) = m - (1-u)\sum_{k=2}^{m}(-1)^k\binom{m}{k}Q_{k-2}(u)$$

and

$$[u^n]Q_{k-2}(u) = -\sum_{l=0}^{n}\frac{R_{n-l}}{Q_l 2^{l(k-1)}}\ .$$

where $R_i = (-1)^{i+1}\frac{2^{-i(i+1)/2}}{Q_i}$.

THEOREM. Symmetric Case.
Let $Q_k = \prod_{j=1}^{k}(1-2^{-j})$, and define $\psi(m) = \log_2 m - \lfloor \log_2 m \rfloor$. Then, for the symmetric Bernoulli we obtain for any integer K

$$\lim_{m\to\infty} |\Pr\{D_m \le \log_2 m + K\} -$$

$$2^{K-\psi(m)}\left(1+\frac{1}{2Q_\infty}\sum_{i=0}^{\infty}(-1)^{i+1}\frac{2^{-i(i+1)/2}}{Q_i}e^{-2^{-(K-\psi(m)-1-i)}}\right)\Bigg| = 0\ .$$

The function $\psi(m)$ is dense in $[0,1]$ but not uniformly dense, thus the limiting distribution of D_m does not exist.

Digital Search Tree – Asymmetric Case

In this case $\mathrm{Var}\ D_m = (h_2 - h^2)/h^3 \log m + O(1)$. Thus, one can expect the central limit theorem to hold. Indeed, we have

$$D_m(u) = 1 - \frac{1-u}{m} \sum_{k=2}^{m} (-1)^k \binom{m}{k} Q_{k-2}(u)$$

and our goal is to prove that Goncharov's condition holds:

$$\lim_{m\to\infty} e^{-\vartheta\mu_m/\sigma_m} D_m(e^{\vartheta/\sigma_m}) = e^{\vartheta^2/2}$$

where ϑ is a complex number such that $\Re(\vartheta) > 0$.

As in the case of depth in a trie we apply Mellin-like approach. Define

$$Q(u,s) = \frac{P(u,0)}{P(u,s)} = \frac{Q_\infty(u)}{P(u,s)}$$

where $P(u,s) = \prod_{j=2}^{\infty}(1 - up^{s+j} - uq^{s+j})$. Then:

$$1 - D_m(u) = \frac{1-u}{m2\pi i} \int_{-3/2-i\infty}^{-3/2+i\infty} \Gamma(s) m^{-s} Q(u, -s-2) ds \ .$$

Digital Search Tree – Asymmetric Case

Finally, as before, after applying residue theorem, tedious algebra, etc., we prove that

$$e^{-\vartheta\mu m/\sigma m} D_m(e^{\vartheta/\sigma m}) = e^{\vartheta^2/2}(1+O(vm^{-\beta v^2})+O(1/m)) \to e^{\vartheta^2/2} \,.$$

THEOREM: Asymmetric Case
We have

$$\frac{D_m - ED_m}{\sqrt{\text{Var } D_m}} \to N(0,1)$$

and the moments of D_m converge to the appropriate moments of the normal distribution. More generally, for any complex ϑ such that $\Re(\vartheta) > 0$

$$e^{-\vartheta c_1 \log m} E(e^{\vartheta D_m}) = e^{c_2 \frac{\vartheta^2}{2} \log m} \left(1 + O(\frac{1}{\sqrt{\log m}})\right)$$

where $c_1 = 1/h$ and $c_2 = (h_2 - h^2)/h^3$.

In addition, there exist positive constants A and $\alpha < 1$ such that

$$\Pr\left\{\left|\frac{D_m - c_1 \log m}{\sqrt{c_2 \log m}}\right| > k\right\} \le A\alpha^k$$

uniformly in k for large m.

Lempel-Ziv Model – Asymmetric Case

Consider the parsing algorithm due to Lempel and Ziv that partitions a sequence of length n into variable phrases (blocks) such that a new block is the shortest substring not seen in the past as a phrase. For example, the string 110010100010001000 is parsed into $(1)(10)(0)(101)(00)(01)(000)(100)$. This partition can be represented on a digital search tree as shown below.

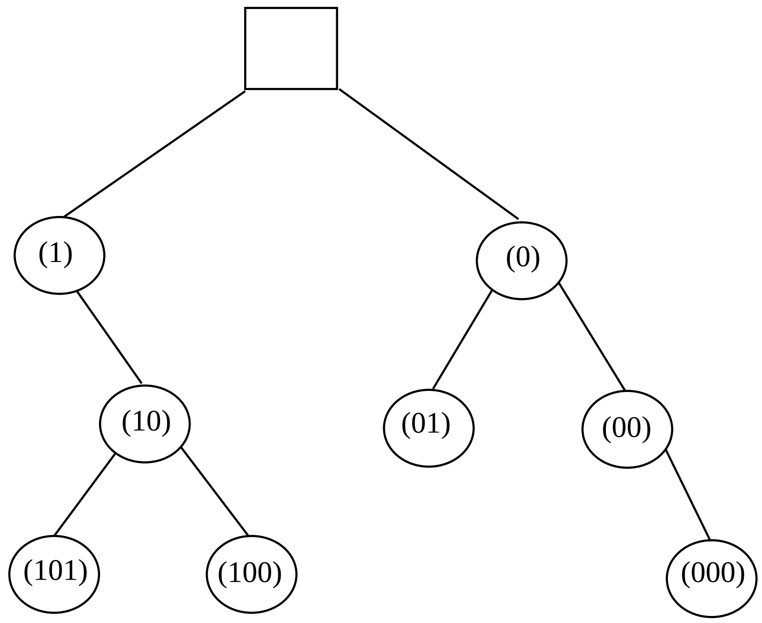


Figure 2: A digital tree representation of Ziv's parsing for the string $11001010001000100\ldots$

Lempel-Ziv Model – Asymmetric Case

Let D_n^{LZ} be the length of a randomly selected phrase when the string is of length n. Observe that:

$$\Pr\{D_n^{LZ} = k | M_n = m\} \neq \Pr\{D_m = k\}$$

where D_m is the depth in the digital tree model and M_n is the number of phrases built from a string of length n.

The following is proved in Louchard-Szpankowski, IEEE IT, 1995:

A. For large n

$$\Pr\{D_n^{LZ} = k | M_n = m\} = \left(1 + O(\sqrt{\log n / n})\right) \Pr\{D_m = k\} .$$

B. For the asymmetric alphabet when $n \to \infty$

$$Ee^{i\theta D_n^{LZ}} \sim Ee^{i\theta D_{\lfloor nh/\log n \rfloor}} ,$$

that is, the limiting distribution of D_n^{LZ} is asymptotically the same as the limiting distribution of the depth in the digital tree model with $m = \lfloor nh/\log n \rfloor$ nodes. A similar statement is also true for the symmetric model.

Leader Election Algorithm

Figure below explains how a randomized leader election algorithm works (Prodinger, DM, 1993)

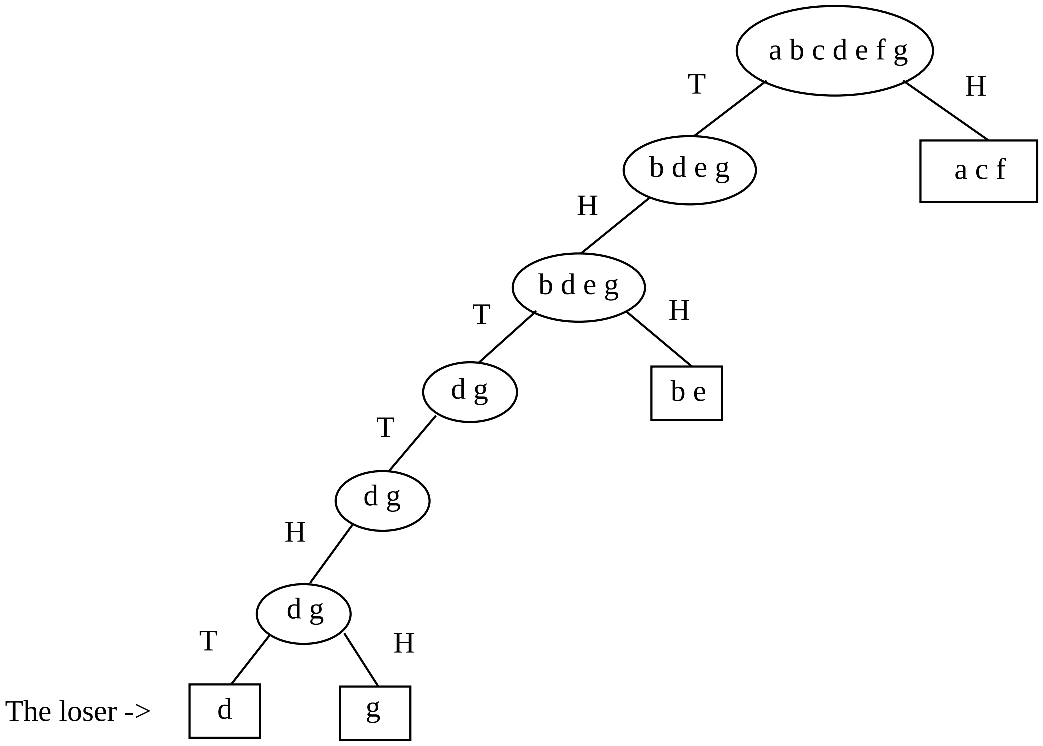


Figure 3: An incomplete trie for the elimination process starting with 7 players.

HEIGHT H_n is the number of steps needed to find a leader (loser).

Leader Election Algorithm- Analysis

Let $G_n(z) = Ez^{H_n}$ nad $g(w,z) = \sum_{n\geq 0} G_n(z)\frac{w^n}{n!}$. The following functional equations are easy to derive

$$\begin{aligned} G_n(z) &= \frac{z}{2^n}\sum_{k=1}^{n}\binom{n}{k}G_k(z) + \frac{zG_n(z)}{2^n} \\ g(w,z) &= z(1+e^{-w/2})g\left(\frac{w}{2},z\right) \\ &+ e^{-w}\left[(1+w)(1-z) - ze^{w/2}\right] . \end{aligned}$$

Define also $h(w,z) = g(z,w)/(1-e^{-w})$. Its Mellin transform $h^*(s,z)$ is

$$h^*(s,z) = \Gamma(s)\,\zeta(s) + \frac{(1-z)\,\Gamma(s+1)\,\zeta(s+1)}{1-2^s z} .$$

Inverting the Mellin and finding the coefficient at z^k we obtain

$$\Pr\{H_{N(w)} \leq k\} = e^{-w} + \left(1-e^{-w}\right)\frac{w/2^k}{\exp(w/2^k)-1}$$

where $N(w)$ denotes the Poisson with mean w.

Leader Election Algorithm – Results

Now, we are ready for a depoissonization, that is, extracting the Bernoulli model characteristics from the Poisson model. In our case we need $\Pr\{H_n < j\}$.

In Dagstuhl I will talk about DEPOISSONIZATION LEMMA, from which we see that

$$\Pr\{H_n \leq j\} = \frac{2^{\{\log n\}-k}}{\exp\left(2^{\{\log n\}-k}\right) - 1} + O\left(n^{-1/2}\right) \ .$$

After setting $j = \lfloor \log_2 \rfloor + k$ we finally prove:

THEOREM: Uniformly over all integers k,

$$\Pr\{H_n \leq \lfloor \lg n \rfloor + k\} = \frac{2^{\{\log_2 n\}-k}}{\exp(2^{\{\log_2 n\}-k}) - 1} + O\left(\frac{1}{\sqrt{n}}\right)$$

as $n \to \infty$ and $\{\log n\} = \log n - \lfloor \log n \rfloor$.

Leader Election Algorithm- Results

$$\liminf_{n\to\infty} \Pr\{H_n \le \lfloor \log_2 n \rfloor + k\} = \frac{2^{-k+1}}{\exp(2^{-k+1}) - 1},$$
$$\limsup_{n\to\infty} \Pr\{H_n \le \lfloor \log_2 n \rfloor + k\} = \frac{2^{-k}}{\exp(2^{-k}) - 1}.$$

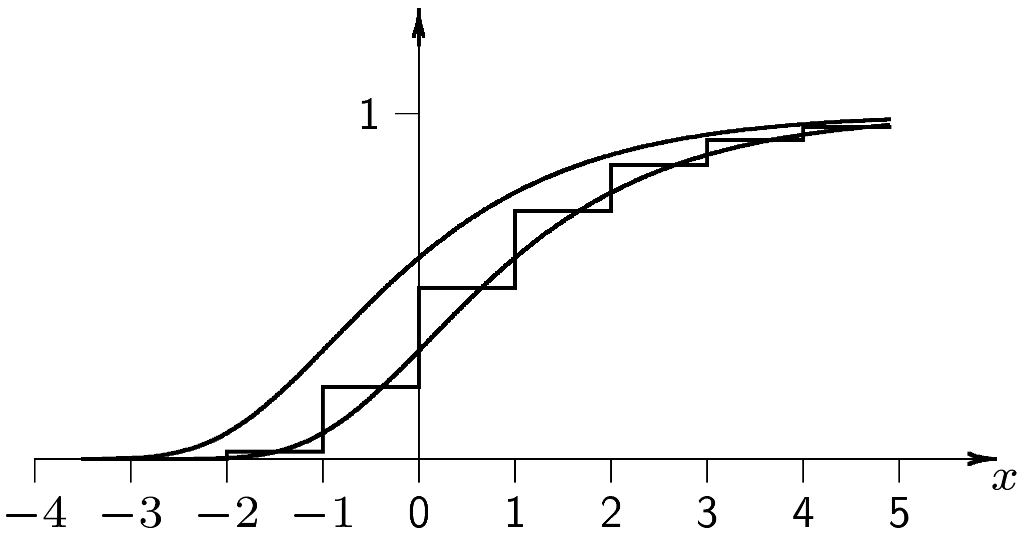


Figure 4: The distribution function of $H_{20} - \lfloor \log_2 20 \rfloor$ and the two continuous extremes.

Generalized Probabilistic Counting

In several algorithms on databases a major determinant of efficiency is the *cardinality* of the underlying set, say $\mathcal{M}$. Generalizing Flajolet & Martin idea, we suggest to do the following.

We assume that there exists a carefully implemented hashing function $hash : \mathcal{M} \to \{0,1\}^m$. Let $y = hash(x)$, $x \in \mathcal{M}$. The procedure below constructs another string called BITMAP that is used to estimate N. As in Flajolet & Martin we define the function $\rho(y)$ for a binary string y as the leftmost position that contains value 1.

procedure $BITMAP$
for $i = 0$ **to** $m-1$ **do** $BITMAP(i) = 0$;
for $x \in \mathcal{M}$ **do**
  begin
  $y = hash(x)$;
  for $i = 0$ **to** $\rho(x)$ **do**
    $BITMAP(i) = \min\{BITMAP(i) + y(i),\ d+1\};$
  end;
end;

Observe that the resulting string BITMAP is a $d+1$-ary string. For example, for $d = 3$ we may have $BITMAP = 4444444444434100000$. As the estimate $R_{N,d}$ we take the position of the leftmost value smaller than $d+1$ (e.g., in the example above we have $R_{N,3} = 11$). More precisely we define:

$$\begin{aligned} R_{N,d} \quad &= \quad \min\{k : \; BITMAP(k) < d+1 \\ &\text{and} \qquad \text{for all } 0 \leq i < k \;\; BITMAP(i) = d+1\} \; . \end{aligned}$$

The new description goes as follows. Let us consider an empty BITMAP string, that is, with all positions filled by zeros. Assume that N objects (e.g., data, persons, etc.) can randomly insert (hit) a 1 at any position of the BITMAP, however, the probability of hitting the $j \geq 0$ position is equal to 2^{-j-1}. In terms of probabilistic counting, this means that the probability of the occurrence of the pattern like $0^j 1 \cdots$ is equal to 2^{-j-1} since 0 and 1 are equally likely (0^j denotes j consecutive zeros). Every object can hit only one time. In addition, we count the number of hits in any position of the BITMAP, but we count the number of hits only up to some value $d+1$, where d is a given parameter. In other words, the BITMAP is a $d+1$-ary string. Clearly, the BITMAP will contain a lot of $d+1$'s at the beginning of the string. This is due to the fact that the probability of a hit decreases exponentially fast with the increase of the position in the BITMAP.

Probabilistic Counting – Analysis

Let $F_N(u) = Eu^{R_{N,d}}$. Then

$$F_N(u) = \sum_{k=0}^{d} \binom{N}{k} 2^{-N} + u \sum_{k=0}^{N-d-1} \binom{N}{k} 2^{-N} F_k(u) \ .$$

We also introduce the Poisson transform of $F_N(z)$ as

$$\widehat{G}(z,u) = e^{-z} \sum_{N=0}^{\infty} F_N(u) \frac{z^N}{N!}$$

which satisfies

$$\widehat{G}(z,u) = u f_d(z/2) \widehat{G}(z/2,u) + (u-1)(f_d(z/2) - 1) \ .$$

where

$$f_d(z) = 1 - e_d(z) e^{-z} \qquad \text{and} \qquad e_d(z) = 1 + \frac{z^1}{1!} + \cdots + \frac{z^d}{d!} \ .$$

Probabilistic Counting – Analysis

From the last equation we have:

$$\frac{1-\widehat{G}(z,u)}{1-u} = \sum_{k=0}^{\infty} u^k \frac{\varphi(z2^{-k-1})}{\varphi(z)} \ .$$

where $\varphi(z) = \prod_{j=0}^{\infty} f_d(z2^j)$. Thus,

$$\begin{aligned} \widehat{h}_k(z) &= [u^k]\left(\frac{\widehat{G}(z,u)}{1-u}\right) = e^{-z}\sum_{N=0}^{\infty} \Pr\{R_{N,d} > k\}\frac{z^N}{N!} \\ &= \frac{\varphi(z2^{-k-1})}{\varphi(z)} \ . \end{aligned}$$

After DEPOISSONIZATION we obtain

$$\Pr\{R_{N,d} > k\} = \frac{\varphi(N2^{-k-1})}{\varphi(N)} + O\left(\frac{1}{\sqrt{N}}\right) \ .$$

Probabilistic Counting – Results

THEOREM. Let t be a real number.

$$\Pr\{R_{N,d} \leq \log_2 N + t - 1\} = 1 - \varphi\left(2^{-t-\{\log_2 N + t\}}\right) + O(N^{-1/2}) \; ,$$

that is,

$$\begin{aligned}
\liminf_{N\to\infty} \Pr\{R_{N,d} \leq \log_2 N + t - 1\} &= 1 - \varphi\left(2^{-t-1}\right) \\
\limsup_{N\to\infty} \Pr\{R_{N,d} \leq \log_2 N + t - 1\} &= 1 - \varphi\left(2^{-t}\right)
\end{aligned}$$

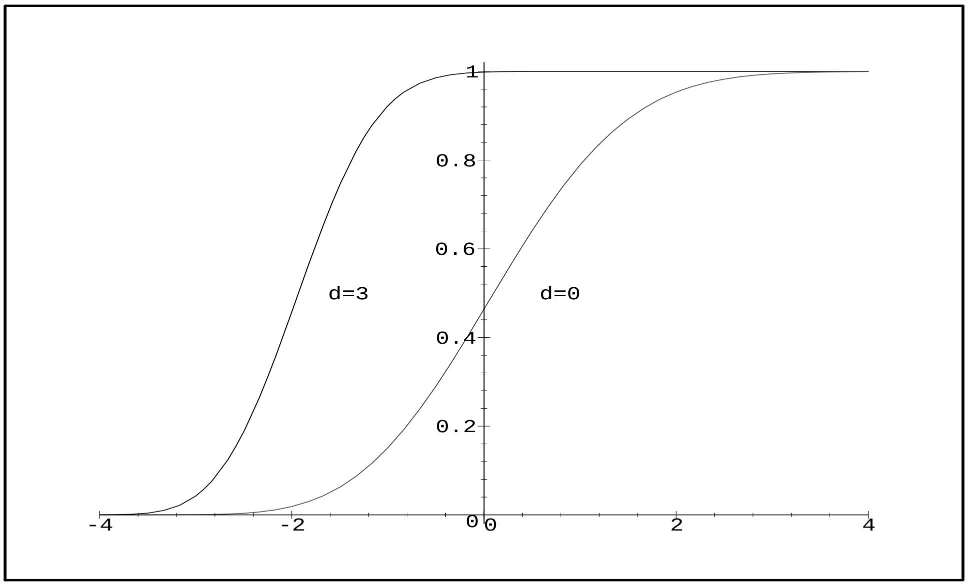


Figure 5: The upper “envelope” for the asymptotic distribution $1 - \varphi(2^{-t})$ for $d = 0$ and $d = 3$.