PROBABILISTIC BEHAVIOR OF ASYMMETRIC LC-TRIES

Luc Devroye
School of Computer Science
McGill University
3450 University Street
Montreal H3A 2K6
Canada
luc@cs.mcgill.ca

Wojcieh Szpankowski
Department of Computer Sciences
Purdue University
250 N. University Street
West Lafayette, Indiana, 47907-2066
USA
spa@cs.purdue.edu

September 4, 2003

ABSTRACT. LC tries were introduced by Andersson and Nilsson in 1993. They are compacted versions of tries in which, from the top down, maximal height complete subtrees are level compressed. We show that when the input consists of n independent strings with independent Bernoulli (p) bits, $p \neq 1/2$, then the expected depth of a typical node is in probability asymptotic to

$$\frac{\log \log n}{\log \left(1 - \frac{\mathcal{H}}{\mathcal{H}_{-\infty}}\right)},$$

where $\mathcal{H} = -p \log p - (1-p) \log(1-p)$ is the Shannon entropy of the source, and $\mathcal{H}_{-\infty} = \log(1/\min(p, 1-p))$. The height is in probability asymptotic to

$$\frac{\log n}{\mathcal{H}_2}$$

where $\mathcal{H}_2 = \log \left(1/(p^2 + (1-p)^2)\right)$.

KEYWORDS AND PHRASES. Trie, probabilistic analysis, law of large numbers, LC trie, height of a tree.

CR CATEGORIES: 3.74, 5.25, 5.5.

1991 MATHEMATICS SUBJECT CLASSIFICATIONS: 60D05, 68U05.

Corresponding authors' address: Luc Devroye, School of Computer Science, McGill University, 3480 University Street, Montreal, Canada H3A 2K6.

The first author's research was sponsored by NSERC Grant A3456 and FCAR Grant 90-ER-0291. The second author was supported by NSF Grant CCR-0208709 and NIH grant R01 GM068959-01.

Introduction

Consider the complete infinite binary tree B in which we associate with each node $u \in B$ the bit string $b_1, \ldots, b_k$ that describes the path from the root to u. The number of zeros (left edges) is denoted by $L(u)$, and the number of ones by $R(u)$. Define the function

$$\mu(u) = np^{R(u)}(1-p)^{L(u)}.$$

Note that this is nothing but n times the probability of reaching u if we flip independent biased coins, and travel from the root down, going left with probability $1-p$ and right with probability p. This simple labeled binary tree will be essential in understanding our analysis of LC tries.

Tries are efficient data structures that were initially developed and analyzed by Fredkin (1960) and Knuth (1973), and intended to store n strings $X_1, \ldots, X_n$. The binary expansion of X_i gives rise to an infinite binary string $(X_{i1}, X_{i2}, \ldots)$ which in turn defines an infinite path in a binary tree in the following manner: from the root, follow the X_{i1}-st child, then its X_{i2}-nd child, and so forth. The collection of nodes and edges visited by the union of the n paths is the infinite trie $T_{n,\infty}$. It is of course embedded in our complete infinite binary tree B. In this paper, we will assume throughout that $X_1, \ldots, X_n$ are i.i.d. (independent and identically distributed), and that $(X_{i1}, X_{i2}, \ldots)$ are independent Bernoulli (p) random bits, a "1" occurring with probability $p \in (0, 1/2]$. The probability that the path for X_1 visits u is

$$p^{R(u)}(1-p)^{L(u)}.$$

The expected number of paths that visit u is $\mu(u)$. The actual number of paths that visit u is a binomial $(n, \mu(u)/n)$ random variable, which we shall denote by $N(u)$.

The trie T_n is the subtree of $T_{n,\infty}$ that consists of all nodes u with $N(u) > 0$ and $N(v) > 1$, where v is the parent of u. It is easy to verify that if the input strings are all different, this tree is finite and has n leaves. Also, for $u \in T_n$, $N(u)$ is the number of leaves in the subtree rooted at u. Let us introduce the random variables $Z(u)$, where $Z(u)$ is the largest integer such that the complete tree of height $Z(u)$ and rooted at u is embedded in T_n. This means that all $2^{Z(u)}$ nodes at distance $Z(u)$ from u exist in T_n, but not all $2^{Z(u)+1}$ nodes at distance $Z(u)+1$.

The LC trie is a further compactification of T_n, in which the following operation is repeated recursively from the root down: denote the $2^{Z(r)}$ depth $Z(r)$ descendants of the root r of the trie T_n by u_i, and let the subtrees of the u_i's be T_i', $1 \le i \le 2^{Z(r)}$. In the LC trie, create a root node that corresponds to r, and give this root $2^{Z(r)}$ children, each corresponding to a u_i. Apply the level compression process recursively to each T_i'. The resulting (usually non-binary) tree is called the LC trie. Note that in LC tries, the number of children of each node is a power of 2. For compact and simple array implementations, we refer to the work of Andersson and Nilsson. The idea of level compression was proposed by Andersson and Nilsson (1993). LC tries are first defined there, and an early average case analysis may be found in that paper and in Andersson and Nilsson (1994). LC tries were suggested by Andersson and Nilsson (1995) for string searching, as improvements of suffix trees. Nilsson and Karlsson (1998, 1999) noted their usefulness for fast address look-up for internet routers and IP address look-up. Experimental comparisons can be found in Iivonen, Nilsson and Tikkanen (1999) and Nilsson and Tikkanen (1998).

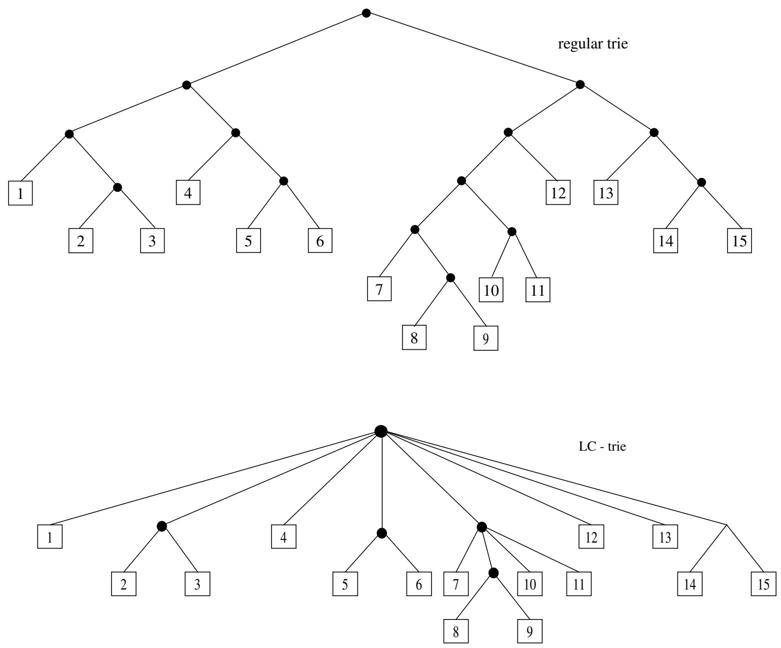

FIGURE 1. A standard binary trie and the corresponding LC trie.

The quantities of interest to us in a trie or LC trie are D_n, the depth of the n-th string (which is thus also the depth of a typical string, as all tries considered here are permutation-invariant), and H_n, the height of the LC trie.

Andersson and Nilsson (1993) showed that for $p \neq 1/2$, $p \in (0, 1)$, $\mathbb{E}\{D_n\} = \Theta(\log \log n)$. Devroye (2001) showed that for $p = 1/2$, $D_n \sim \log^* n$ in probability, and provides various tail bounds. We note below that for asymmetric tries ($p \neq 1/2$), D_n is of the order of $\log \log n$ in probability. This is due to the fact that in the symmetric case, $Z(r)$ is close to $\log_2 n - O(\log \log n)$, and thus, the $2^{Z(r)}$ subtrees of the root after compaction are of order $\log n$. After k iterations of compaction, the subtrees drop to about the k-times iterated logarithm of n. In the asymmetric case, $Z(r)$ is of the order of a constant times $\log n$, with the $2^{Z(r)}$ subtrees having varying sizes. Those that matter, i.e., those that have the bulk of the strings visiting them, are of size about n^β for some $\beta \in (0, 1)$ depending upon p only. After k iterations, the typical subtree sizes drop to about n^{β^k}, and this is of constant order when k is roughly $\log \log n$. The idea, then, is to get a good grip on the exponent β.

It is helpful to introduce the Rényi entropy $\mathcal{H}_\alpha$ of order α (see Szpankowski (2001, p. 157)):

$$\mathcal{H}_\alpha = \frac{\log\left(p^\alpha + (1-p)^\alpha\right)}{1-\alpha}$$

where $\mathcal{H}_\alpha$ decreases in α. Important values are

$$\begin{aligned}
\mathcal{H}_{-\infty} &= -\log(\min(p, 1-p));\\
\mathcal{H}_0 &= \log 2;\\
\mathcal{H}_1 &= \mathcal{H} = -p \log p - (1-p)\log(1-p) \text{(the Shannon entropy)};\\
\mathcal{H}_2 &= -\log\left(p^2 + (1-p)^2\right);
\end{aligned}$$

$$\mathcal{H}_\infty = -\log(\max(p, 1-p)).$$

In this paper, we obtain the following law of large numbers:

THEOREM 1. *For a random LC trie with* $p \in (0,1)$, $p \neq 1/2$,

$$D_n - \frac{\log\log n}{\log\left(1 - \frac{\mathcal{H}}{\mathcal{H}_{-\infty}}\right)} = O(\log\log\log n)$$

in probability.

The height of the random LC trie was analyzed by Devroye (2002) for $p = 1/2$:

$$\frac{H_n}{\log_2 n} \to 1$$

in probability. This result will be extended here to the asymmetric Bernoulli model. We note that the height is much larger than the depth because of a simple probabilistic phenomenon, the likelihood of a long common prefix among two strings. In fact, it is very likely that there are two strings that agree in their $(2\log n)/\mathcal{H}_2$ first bits. Half of these bits get level compressed to about $O(\log\log n)$, but the second half "sticks out", and is not compressed at all, leading to a height of roughly $(\log n)/\mathcal{H}_2$.

THEOREM 2. *For a random LC trie with* $p \in (0,1)$,

$$H_n - \frac{\log n}{\mathcal{H}_2} = O(\log\log n)$$

in probability.

THE MAIN PARAMETERS FOR RANDOM TRIES. The asymptotic behavior of tries under the uniform model is well-known. The height is studied by Régnier (1981), Mendelson (1982), Flajolet and Steyaert (1982), Flajolet (1983), Devroye (1984), Pittel (1985, 1986), and Szpankowski (1988, 1991). For the depth of a node, see, e.g., Pittel (1986), Jacquet and Régnier (1986), Flajolet and Sedgewick (1986), Kirschenhofer and Prodinger (1986), and Szpankowski (1988). For example, it is known that

$$H_n \sim \frac{2\log n}{\mathcal{H}_2} \quad \text{in probability.}$$

This is asymtotically exactly twice the value for the random LC trie, regardless of p. The limit law of H_n was obtained in Devroye (1984). For other models, we refer to Devroye (1982, 1984, 1990), Régnier (1988), Szpankowski (1988, 2001) and Pittel (1985).

LC tries: analysis of the depth

Assume without loss of generality that $0 < p < 1/2$. We obtain upper and lower bounds for D_n separately. In fact, it is convenient to work with D_{n+1}, as X_{n+1} is independent of T_n. The proof in this section evolves around the following idea: associate with each node u in the infinite trie the expected number $\mu(u)$ of strings that are expected to visit it. Truncating this tree to all u with $\mu(u) \geq c \log n$ yields a deterministic binary subtree. LC compression applied to this tree yields a tree in which the path of X_{n+1} is easy to trace, and in which all compression levels are explicitly known, as $\mu(u) \geq c \log n$ implies that $N(u) > 0$ with overwhelming probability. With this argument, the lower bound is somewhat easier to deal with than the upper bound.

LEMMA 1. *There exists a constant $M > 0$ such that*

$$\lim_{n\to\infty} \mathbf{P}\left\{ D_{n+1} < \frac{\log\log n - \log\log\log n - M}{\log\left(\frac{1}{1-\frac{\mathcal{H}}{\log(1/p)}}\right)} \right\} = 0.$$

PROOF. Let $u_0, u_1, \ldots$ be the nodes on the path of X_{n+1} at which compaction takes place. We will require the quantities $N(u_i)$, $\mu(u_i)$, $R(u_i)$, $L(u_i)$ and $Z(u_i)$. The index i refers throughout this proof to the path distance $\delta(u)$ of node u to the root. Thus, in particular, $i = \delta(u_i)$. By the law of large numbers, we have $R(u_i)/i \to p$ in probability and $L(u_i)/i \to 1-p$ in probability. By Hoeffding's inequality (Hoeffding, 1963), for a sequence of positive numbers a_i tending to 0 with i, and for $\epsilon > 0$, using the event

$$A = \bigcap_{i=1}^{\infty} \left[\left| \frac{R(u_i)}{i} - p \right| \leq a_i \right],$$

we have

$$\mathbf{P}\{A^c\} \leq \sum_{i=1}^{\infty} 2e^{-2ia_i^2} \leq \epsilon$$

if we take $a_i = \sqrt{\log(2i(i+1)/\epsilon)/2i}$. Thus, on A, we have for all i, if θ denotes an arbitrary number with absolute value less than or equal to $2\log(1/p)$,

$$\begin{aligned}
\log \mu(u_i) &= \log n + R(u_i)\log p + L(u_i)\log(1-p) \\
&= \log n + i[(R(u_i)/i)\log p + (L(u_i)/i)\log(1-p)] \\
&= \log n + i[p\log p + (1-p)\log(1-p)] + \theta a_i \\
&= \log n - i\mathcal{H} + \theta a_i \ .
\end{aligned}$$

Consider integers $a, t > 0$ and observe that if v is the rightmost descendant of u, a levels below u, then, for any u,

$$\mathbf{P}\{Z(u) \geq a + t | N(u)\} \leq \mathbf{P}\left\{N(v) \geq 2^t | N(u)\right\} \leq \mathbf{P}\left\{\text{binomial}(N(u), p^a) \geq 2^t | N(u)\right\}.$$

Thus, for any u on the path for X_{n+1}, at distance $i = \delta(u)$ from the root,

$$
\begin{aligned}
\mathbb{P}\{Z(u) \geq a + t | A\} &\leq \mathbb{P}\{N(u) \geq 2\mu(u) | A\} + \mathbb{P}\left\{N(v) \geq 2^t | N(u) < 2\mu(u), A\right\} \\
&\leq \mathbb{P}\{\text{binomial}(n, \mu(u)/n) \geq 2\mu(u) | A\} + \mathbb{P}\left\{\text{binomial}(\lfloor 2\mu(u) \rfloor, p^a) \geq 2^t | A\right\} \\
&\leq \mathbb{E}\{e^{-C\mu(u)} | A\} + \mathbb{E}\left\{ \binom{\lfloor 2\mu(u) \rfloor}{2^t} (p^a)^{2^t} \middle| A \right\} \\
&\leq e^{-Cne^{-i\mathcal{H}} e^{-|\theta| a_i}} + \mathbb{E}\left\{ \frac{(2\mu(u) p^a)^{2^t}}{(2^t)!} \middle| A \right\} \\
&\leq e^{-C'' ne^{-i\mathcal{H}}} + \mathbb{E}\left\{ \frac{\left(2ne^{-i\mathcal{H}} e^{|\theta| a_1} p^a\right)^{2^t}}{(2^t)!} \middle| A \right\} \\
&\leq e^{-C'' ne^{-i\mathcal{H}}} + \mathbb{E}\left\{ \frac{\left(ne^{C' - i\mathcal{H}} p^a\right)^{2^t}}{(2^t)!} \middle| A \right\}
\end{aligned}
$$

where C is a universal constant (Hoeffding, 1963), $C' = \log 2 + 2a_1 \log(1/p)$, and $C'' = Ce^{-2\log(1/p)a_1}$. We define

$$
a(u) = \left\lceil \frac{\log\left(ne^{C' - i\mathcal{H}}\right)}{\log(1/p)} \right\rceil
$$

to get the bound

$$
\mathbb{P}\{Z(u) \geq a(u) + t | A\} \leq e^{-C'' ne^{-i\mathcal{H}}} + \frac{1}{(2^t)!}.
$$

With

$$
a'(u) = \frac{\log n - i\mathcal{H} + C'}{\log(1/p)},
$$

we have

$$
\mathbb{P}\left\{Z(u) \geq a'(u) + t + 1 | A\right\} \leq e^{-C'' ne^{-i\mathcal{H}}} + \frac{1}{(2^t)!}.
$$

And from this, with $c > 1$, and

$$
B = \bigcap_{0 \leq i \leq \frac{\log n - c\log\log n}{\mathcal{H}}} \left[Z(u_i) \leq a'(u_i) + t + 1\right],
$$

we note that

$$
\mathbb{P}\{B^c | A\} \leq \sum_{i=0}^{\left\lfloor \frac{\log n - c\log\log n}{\mathcal{H}} \right\rfloor} e^{-C'' ne^{-i\mathcal{H}}} + \frac{1 + c\log n}{(2^t)!} \leq \left(1 + \frac{\log n}{\mathcal{H}}\right) e^{-C''(\log n)^c} + \frac{1 + \frac{\log n}{\mathcal{H}}}{(2^t)!}.
$$

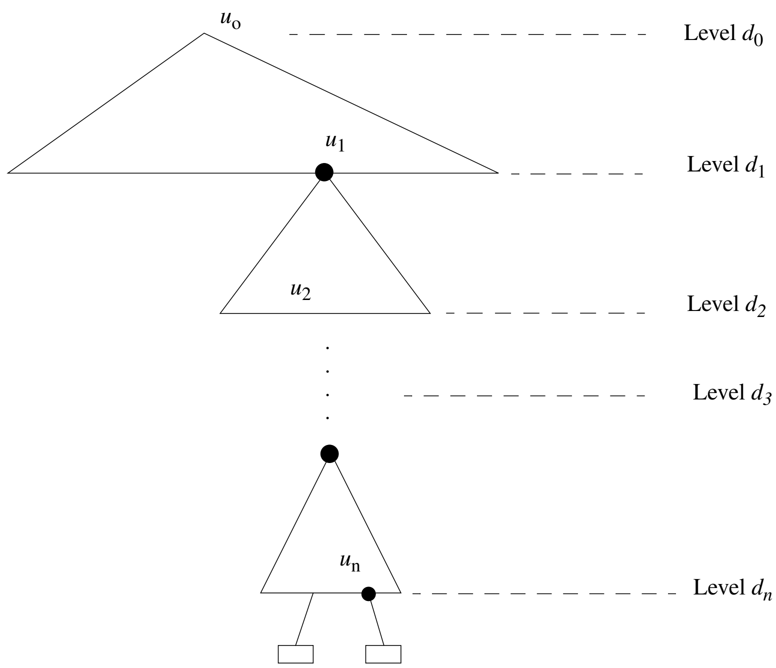

FIGURE 2. The depths of the nodes on the path of X_{n+1} at which compaction takes place are denoted by $d_0 = 0 < d_1 < d_2 < \cdots$.

Denote the depths of the nodes on the path of X_{n+1} at which compaction takes place by $d_0 = 0 < d_1 < d_2 < \cdots$. We stress that these are depths in the original trie, measured as path distances from the root. We have

$$d_{j+1} = d_j + Z(u_{d_j})$$

for all j. Assume that both events A and B occur. Then, for j such that $d_j < (\log n - c \log\log n)/\mathcal{H}$,

$$Z(u_{d_j}) < \frac{\log n - d_j \mathcal{H} + C'}{\log(1/p)} + t + 1,$$

and thus,

$$d_{j+1} \le d_j \left(1 - \frac{\mathcal{H}}{\log(1/p)}\right) + \frac{\log n}{\log(1/p)} + t',$$

where $t' = \frac{C'}{\log(1/p)} + t + 1$. It is easy to prove then by induction on j that

$$d_j \le \left[\frac{\log n}{\log(1/p)} + t'\right] \times \frac{1 - \left(1 - \frac{\mathcal{H}}{\log(1/p)}\right)^j}{\mathcal{H}/\log(1/p)} \le \frac{\log n}{\mathcal{H}} + \frac{t' \log(1/p)}{\mathcal{H}} - \frac{\log n \left(1 - \frac{\mathcal{H}}{\log(1/p)}\right)^j}{\mathcal{H}}.$$

We choose ξ such that $d_\xi \le (\log n - c \log\log n)/\mathcal{H}$ (the range for which the above induction argument is valid) and $B^* = [N(u_{d_\xi}) \ge 2]$ occurs with high probability. The inequality on d_ξ holds if we have

$$t' \log(1/p) - \log n \left(1 - \frac{\mathcal{H}}{\log(1/p)}\right)^\xi \le -c \log\log n,$$

which is satisfied if

$$\xi = \left\lfloor \frac{\log\log n - \log(c\log\log n + t'\log(1/p))}{\log\left(\frac{1}{1-\frac{\mathcal{H}}{\log(1/p)}}\right)} \right\rfloor .$$

In that case, we can conclude that $D_{n+1} \geq \xi$. We see that on A and B, with $C''' = \exp(-2\log(1/p)a_1)$,

$$\begin{aligned}
\mu(u_{d_\xi}) &\geq C''' n e^{-d_\xi \mathcal{H}} \\
&\geq C''' e^{-t'\log(1/p)} n^{(1-\mathcal{H}/\log(1/p))^\xi} \\
&\geq C''' e^{c\log\log n} \\
&= C''' (\log n)^c .
\end{aligned}$$

Thus, $N(u_{d_\xi})$ is stochastically greater than a binomial $(n, C'''(\log n)^c/n)$ random variable. [We say that X is stochastically greater than Y if for all x, $\mathbb{P}\{X \geq x\} \geq \mathbb{P}\{Y \geq x\}$.] Summarizing, we have

$$\begin{aligned}
\mathbb{P}\{D_{n+1} < \xi\} &\leq \mathbb{P}\{A^c\} + \mathbb{P}\{B^c|A\} + \mathbb{P}\{(B^*)^c|A,B\} \\
&\leq \epsilon + \left(1 + \frac{\log n}{\mathcal{H}}\right) e^{-C''(\log n)^c} + \frac{1+\frac{\log n}{\mathcal{H}}}{(2^t)!} + \mathbb{P}\left\{\text{binomial}\left(n, \frac{C'''(\log n)^c}{n}\right) < 2\right\} \\
&= \epsilon + o(1)
\end{aligned}$$

provided that $c > 1$ and $(2^t)!/\log n \to \infty$. A simple choice like $t \sim \log\log n$ will do. This implies that $t' = O(\log\log n)$, and thus, as ϵ was arbitrary, we showed that there exists a constant $M > 0$ such that

$$\lim_{n\to\infty} \mathbb{P}\left\{ D_{n+1} < \frac{\log\log n - \log\log\log n - M}{\log\left(\frac{1}{1-\frac{\mathcal{H}}{\log(1/p)}}\right)} \right\} = 0. \square$$

For the upper bound, we argue in two steps.

Lemma 2. *If $a > 0$, $b \geq 1$ and the integer n are fixed, then*

$$\sum_{i=0}^{n} e^{-a/b^i} \leq \frac{e^{-a/b^n}}{1 - e^{-a(b-1)/b^n}} .$$

Proof. Rewrite the sum as

$$e^{-a/b^n} \sum_{i=0}^{n} e^{-a(b^i-1)/b^n} \leq e^{-a/b^n} \sum_{i=0}^{n} e^{-ai(b-1)/b^n}$$

from which the inequality follows immediately. $\square$

In the first step, we establish an upper bound of $O(\log\log n)$.

LEMMA 3. *There exists a constant $K > 0$ such that*

$$\lim_{n\to\infty} \mathbf{P}\{D_{n+1} \geq K \log\log n\} = 0.$$

PROOF. We use the notation of the proof of Lemma 1. Consider integers $a, t > 0$ and observe that if v_j, $1 \leq j \leq 2^{a+1}$ are the descendants of u, $a+1$ levels below u, then, for any u,

$$\begin{aligned}
\mathbf{P}\{Z(u) \leq a|N(u)\} &\leq \sum_{j=1}^{2^{a+1}} \mathbf{P}\{N(v_j) = 0|N(u)\} \\
&\leq \sum_{k=0}^{a+1} \binom{a+1}{k} \mathbf{P}\left\{\text{binomial}(N(u), p^k(1-p)^{a+1-k}) = 0|N(u)\right\} \\
&= \sum_{k=0}^{a+1} \binom{a+1}{k} (1 - p^k(1-p)^{a+1-k})^{N(u)} \\
&\leq 2^{a+1} e^{-N(u)p^{a+1}}.
\end{aligned}$$

Thus, for any u on the path for X_{n+1}, at distance $i = \delta(u)$ from the root,

$$\begin{aligned}
\mathbf{P}\{Z(u) \leq a|A\} &\leq \mathbf{P}\{N(u) \leq \mu(u)/2|A\} + \mathbf{E}\left\{e^{a+1-N(u)p^{a+1}}|N(u) > \mu(u)/2, A\right\} \\
&\leq \mathbf{P}\{\text{binomial}(n, \mu(u)/n) \leq \mu(u)/2|A\} + \mathbf{E}\left\{e^{a+1-\mu(u)p^{a+1}/2}|A\right\} \\
&\leq \mathbf{E}\{e^{-C\mu(u)}|A\} + \mathbf{E}\left\{e^{a+1-\mu(u)p^{a+1}/2}|A\right\} \\
&\leq e^{-Cne^{-i\mathcal{H}}e^{-|\theta|a_i}} + e^{a+1-ne^{-i\mathcal{H}}e^{-|\theta|a_i}p^{a+1}/2} \\
&\leq e^{-C''ne^{-i\mathcal{H}}} + e^{a+1-C'''ne^{-i\mathcal{H}}p^{a+1}/2}
\end{aligned}$$

where C is a universal constant (different from that in Lemma 2), and C'' and C''' are defined as in Lemma 2. Define

$$a(u) = \left\lfloor \frac{\log n - i\mathcal{H} - 2\log\log n}{\log(1/p)} \right\rfloor,$$

and note that

$$\mathbf{P}\{Z(u) \leq a(u)|A\} \leq e^{-C''ne^{-i\mathcal{H}}} + e^{2+\frac{\log n}{\log(1/p)} - C'''(p^2/2)(\log n)^2}.$$

Finally, set

$$B = \bigcap_{0 \leq i \leq \frac{\log n - 3\log\log n}{\mathcal{H}}} [Z(u_i) > a(u_i)],$$

and observe that

$$\mathbf{P}\{B^c|A\} \leq \left(2 + \frac{\log n}{\mathcal{H}}\right) \times \left(e^{-C''(\log n)^3} + e^{2+\frac{\log n}{\log(1/p)} - C'''(p^2/2)(\log n)^2}\right) \to 0.$$

Denote the depths of the nodes on the path of X_{n+1} at which compaction takes place by $d_0 = 0 < d_1 < d_2 < \cdots$. We have

$$d_{j+1} = d_j + Z(u_{d_j})$$

for all j. Assume that both events A and B occur. Then, for j such that $d_j < (\log n - 3\log\log n)/\mathcal{H}$,

$$Z(u_{d_j}) \geq \frac{\log n - d_j\mathcal{H} - 2\log\log n}{\log(1/p)}$$

and thus,

$$d_{j+1} \geq d_j \left(1 - \frac{\mathcal{H}}{\log(1/p)}\right) + \frac{\log n - 2\log\log n}{\log(1/p)},$$

It is easy to prove then by induction on j that

$$d_j \geq \left[\frac{\log n - 2\log\log n}{\log(1/p)}\right] \times \frac{1-(1-\mathcal{H}/\log(1/p))^j}{\mathcal{H}/\log(1/p)} \geq \frac{\log n - 2\log\log n}{\mathcal{H}} - \frac{\log n \left(1 - \frac{\mathcal{H}}{\log(1/p)}\right)^j}{\mathcal{H}}.$$

Define the semi-depth D^*_{n+1} of X_{n+1} as the largest integer j such that

$$d_j < \frac{\log n - 3\log\log n}{\mathcal{H}}.$$

Note that necessarily, replacing j by D^*_{n+1} above, if A and B happen,

$$\log n - 3\log\log n \geq \log n - 2\log\log n - \log n \left(1 - \frac{\mathcal{H}}{\log(1/p)}\right)^{D^*_{n+1}}$$

and thus,

$$D^*_{n+1} \leq \frac{\log\log n - \log\log\log n}{\log\left(\frac{1}{1-\frac{\mathcal{H}}{\log(1/p)}}\right)}.$$

If A occurs, then

$$\mu\left(u_{d_{1+D^*_{n+1}}}\right) \leq \frac{ne^{-d_{D^*_{n+1}}\mathcal{H}}}{C'''} \leq \frac{(\log n)^3}{C'''}.$$

Thus, conditional on A, $N\left(u_{d_{1+D^*_{n+1}}}\right)$ is stochastically less than a binomial $(n, (\log n)^3/(C'''n))$ random variable, which we call R. Observe that $D_{n+1} - (1 + D^*_{n+1})$ is bounded from above by a random variable distributed as D_{R+1}, because subtrees of tries behave as tries. Note that $D_{R+1} \leq H_R$, where H_n denotes the height of the uncompacted trie with n nodes. From the results cited in the introduction, and the weak law of large numbers for the binomial, we have

$$\begin{aligned}
&\mathbb{P}\left\{D_{R+1} > \frac{3\log\log n}{\log\left(\frac{1}{p^2+(1-p)^2}\right)}\right\} \\
&\leq \mathbb{P}\{A^c\} + \mathbb{P}\left\{\text{binomial}\left(n, \frac{(\log n)^3}{C'''n}\right) > \frac{2}{C'''}(\log n)^3\right\} + \mathbb{P}\left\{H_{(2/C''')(\log n)^3} > \frac{3\log\log n}{\log\left(\frac{1}{p^2+(1-p)^2}\right)}\right\} \\
&\to 0.
\end{aligned}$$

Thus,

$$\begin{aligned}
&\mathbb{P}\left\{D_{n+1} > \frac{\log\log n - \log\log\log n}{\log\left(\frac{1}{1-\frac{\mathcal{H}}{\log(1/p)}}\right)} + 1 + \frac{3\log\log n}{\log\left(\frac{1}{p^2+(1-p)^2}\right)}\right\} \\
&\leq \mathbb{P}\{A^c\} + \mathbb{P}\{B^c|A\} + o(1) \leq \epsilon + o(1).
\end{aligned}$$

Thus, as ϵ was arbitrary, we showed that there exists a constant $K > 0$ such that

$$\lim_{n\to\infty} \mathbf{P}\{D_{n+1} \geq K \log\log n\} = 0. \quad \square$$

Finally, we obtain a more refined upper bound.

Lemma 4. *There exists a constant $M > 0$ such that*

$$\lim_{n\to\infty} \mathbf{P}\left\{ D_{n+1} \geq \frac{\log\log n + M \log\log\log n}{\log\left(\frac{1}{1-\frac{\mathcal{H}}{\log(1/p)}}\right)} \right\} = 0.$$

Proof. We use the notation of the proof of Lemma 3, and the following corollary of it:

$$\mathbf{P}\left\{ D_{R+1} \geq K \log\log\left(\frac{2(\log n)^3}{C'''}\right) \Big| A \right\} \leq o(1) + \mathbf{P}\left\{ D_{\frac{2(\log n)^3}{C'''}} \geq K \log\log\left(\frac{2(\log n)^3}{C'''}\right) \right\} = o(1).$$

From the distributional inequality

$$D_{n+1} - (1 + D^*_{n+1}) \leq D_{R+1},$$

we conclude the following:

$$\mathbf{P}\left\{ D_{n+1} > \frac{\log\log n - \log\log\log n}{\log\left(\frac{1}{1-\frac{\mathcal{H}}{\log(1/p)}}\right)} + 1 + K \log\log\left(\frac{2(\log n)^3}{C'''}\right) \right\}$$
$$\leq \mathbf{P}\{A^c\} + \mathbf{P}\{B^c | A\} + \mathbf{P}\left\{ D_{R+1} \geq K \log\log\left(\frac{2(\log n)^3}{C'''}\right) \Big| A \right\} \leq \epsilon + o(1).$$

This completes the proof of Lemma 4. $\square$

The height of the asymmetric LC trie

Consider the deterministic binary tree S_n consisting of all nodes u with $\mu(u) \geq c \log n$. We show that very likely T_n contains S_n.

Lemma 5. *Fix $c > 1$. Then*

$$\mathbf{P}\{S_n \subset T_n\} \geq 1 - \frac{2n(1 + c\log n)}{c \log n (n/e)^c} \to 1$$

as $n \to \infty$.

PROOF. Observe that $N(u)$ is binomial $(n, \mu(u)/n)$ and is thus stochastically larger than a binomial $(n, c\log n/n)$ random variable for all $u \in S_n$. For any such u, we have

$$\begin{aligned}
\mathbf{P}\{N(u) < 2\} &\le \mathbf{P}\{\text{binomial}(n, c\log n/n) < 2\} \\
&= (1 - c\log n/n)^n + n(c\log n/n)(1 - c\log n/n)^{n-1} \\
&\le (1 + c\log n)(1 - c\log n/n)^{n-1} \\
&\le (1 + c\log n)e^{-\frac{c(n-1)\log n}{n}} \\
&\le \frac{(1 + c\log n)}{(n/e)^c} .
\end{aligned}$$

Now note that $|S_n| \le 2n/(c\log n)$ and conclude by the union bound. □

LEMMA 6. *Assume $p < 1/2$ and $[S_n \subset T_n]$. Let $\lambda(u)$ denote the LC trie depth of a node $u \in S_n$ that is present in the LC trie after compaction (recalling that the LC trie has fewer nodes than T_n). Then*

$$\lambda(u) \le \left\lceil \frac{1}{\log(1/(1-p))} \right\rceil + \left\lceil \frac{\log\log n}{\log(1/\beta)} \right\rceil ,$$

where

$$\beta = 1 - \frac{\log(1/(1-p))}{\log(1/p)}.$$

PROOF. Note that a node u has $Z(u) \ge d$ if the nearest descendant outside S_n is at distance d. Thus, we must have

$$\mu(u)p^d < c\log n$$

and therefore

$$Z(u) \ge \frac{\log\left(\frac{\mu(u)}{c\log n}\right)}{\log(1/p)}.$$

Consider any sequence of nodes in S_n with u_0 the root, and u_{i+1} a $Z(u_i)$-level descendant of u_i. Then

$$\mu(u_{i+1}) \le \mu(u_i)(1-p)^{Z(u_i)} \le \mu(u_i)(1-p)^{\frac{\log\left(\frac{\mu(u)}{c\log n}\right)}{\log(1/p)}} \le \alpha(\mu(u_i))^\beta$$

where $\alpha = (c\log n)^{1-\beta}$, and we note that $\beta \in (0,1)$. By repeating this, we have

$$\mu(u_i) \le n^{\beta^i}\alpha^{1+\beta+\cdots+\beta^{i-1}} \le \alpha^{1/(1-\beta)}n^{\beta^i} = c\log n \times n^{\beta^i}.$$

For

$$i = \left\lceil \frac{\log\log n}{\log(1/\beta)} \right\rceil ,$$

the upper bound is not more than $ce\log n$. Let k be the smallest integer such that $e(1-p)^k < 1$. All k-level descandants of such u_i have μ-value less than $(1-p)^k ce\log n < c\log n$ and fall thus outside S_n. The Lemma has been proved. □

Theorem 2 is implied by Lemmas 7 and 8 below.

Lemma 7. *Let $\omega_n \uparrow \infty$ arbitrarily slowly. Then*

$$\lim_{n\to\infty} \mathbf{P}\left\{ H_n < \frac{\log n - \omega_n}{\log\left(\frac{1}{p^2+(1-p)^2}\right)} \right\} = 0.$$

Proof. Assume without loss of generality that $p \le 1/2$. Let U be the collection of all nodes u with $\mu(u) \in [1, 1/p]$ such that no child v of u has $\mu(v) \in [1, 1/p]$. Let N be the number of nodes $u \in U$ with $N(u) = 2$. Note that no two nodes in U are ancestors of each other, so that clearly, $|U| \ge np$. Furthermore, the vector $V \stackrel{\text{def}}{=} (N(u) : u \in U)$ is multinomially distributed with probabilities all between $1/n$ and $1/(np)$. Therefore, each individual $N(u)$ is asymtotically stochastically bounded by a Poisson (1) random variable from below and a Poisson $(1/p)$ random varable from above. It is easy to show that there exists a constant $c > 0$ such that $\mathbf{P}\{N \ge cn\} \to 1$. If we condition on the vector V, then the subtrees rooted at $u \in U$ are independent. Consider a particular subtree with $N(u) = 2$. Let $W(u)$ denote the number of consecutive bits, starting at u, in which the two strings in the subtree of u agree. Observe that

$$\mathbf{P}\{W(u) \ge L\} = (p^2 + (1-p)^2)^L.$$

Therefore,

$$\begin{aligned}
\mathbf{P}\{H_n < L\} &\le \mathbf{P}\left\{\cap_{u\in U:N(u)=2}[W(u) < L]\right\} \\
&= \mathbf{E}\left\{ \prod_{u\in U:N(u)=2} \mathbf{P}\{W(u) < L | N(u) = 2\} \right\} \\
&= \mathbf{E}\left\{ \prod_{u\in U:N(u)=2} \left(1 - (p^2 + (1-p)^2)^L\right) \right\} \\
&= \mathbf{E}\left\{ \left(1 - (p^2 + (1-p)^2)^L\right)^N \right\} \\
&\le \mathbf{E}\left\{ e^{-N(p^2+(1-p)^2)^L} \right\} \\
&\le \mathbf{P}\{N < cn\} + e^{-cn(p^2+(1-p)^2)^L} \\
&\to 0
\end{aligned}$$

if we set $L = (\log n - \omega_n)/\log(1/(p^2 + (1-p)^2))$, where $\omega_n \to \infty$ arbitrarily slowly. □

Lemma 8. *Let $\omega_n \uparrow \infty$ arbitrarily slowly. Then*

$$\lim_{n\to\infty} \mathbf{P}\left\{ H_n > \frac{\log n + 2\log\log n + \omega_n}{\log\left(\frac{1}{p^2+(1-p)^2}\right)} \right\} = 0.$$

Proof. Assume without loss of generality that $p \le 1/2$. Let U be he collection of all nodes u with $\mu(u) \in [2\log n, (2/p)\log n]$ such that no child v of u has $\mu(v) \in [2\log n, (2/p)\log n]$. Note that no two nodes in U are ancestors of each other, so that clearly, $|U| \le n$. Furthermore, the vector $V \stackrel{\text{def}}{=} (N(u) : u \in U)$ is multinomially distributed with probabilities all between $(2/n)\log n$ and $(2/(np))\log n$. If we condition on the vector V, then the subtrees rooted at $u \in U$ are independent. Consider a particular subtree rooted a u. Let $H(u)$ denote the height of the subtree (with LC compaction) rooted at u. Observe that

$$\mathbf{P}\{H(u) \geq L | N(u)\} \leq \binom{N(u)}{2} (p^2 + (1-p)^2)^L.$$

Therefore, if $D(u)$ denotes the LC-depth of node u, using $H_n \leq \max_{u \in U}(D(u) + H(u))$, for $C > 0$,

$$\begin{aligned}
\mathbf{P}\{H_n \geq L + C \log\log n\} &\leq \mathbf{P}\left\{\cup_{u \in U}[H(u) \geq L]\right\} + \mathbf{P}\left\{\max_{u \in U} D(u) \geq C \log\log n\right\} \\
&\leq n \max_{u \in U} \mathbf{E}\{(N(u))^2\}(p^2 + (1-p)^2)^L + \mathbf{P}\left\{\max_{u \in U} D(u) \geq C \log\log n\right\} \\
&\leq n \frac{\log n}{p} \left(1 + \frac{\log n}{p}\right) (p^2 + (1-p)^2)^L + \mathbf{P}\left\{\max_{u \in U} D(u) \geq C \log\log n\right\}.
\end{aligned}$$

The first term tends to 0 with n if we choose $L = (\log n + 2\log\log n + \omega_n)/\log(1/(p^2 + (1-p)^2))$, where $\omega_n \to \infty$ arbirarily slowly. If $U \subset S_n \subset T_n$, then $\max_{u \in U} D(u) \leq C \log\log n$ for some constant C, by Lemma 6. Thus, the last term in the upper bound is further bounded by

$$\mathbf{P}\{S_n \not\subset T_n\}$$

Taking $c = 3/2$ in the definition of S_n insures that $U \subset S_n$ and that $\mathbf{P}\{S_n \not\subset T_n\} \to 0$ (Lemma 5). $\square$

References.

A. Andersson and S. Nilsson, "Improved behaviour of tries by adaptive branching," *Information Processing Letters*, vol. 46, pp. 295–300, 1993.

A. Andersson and S. Nilsson, "Faster searching in tries and quadtrees – an analysis of level compression," in: *Proceedings of the Second Annual European Symposium on Algorithms*, pp. 82–93, 1994.

A. Andersson and N. Nilsson, "Efficient implementation of suffix trees," *Software Practice and Experience*, vol. 25(2), pp. 129–141, 1995 .

L. Devroye, "A note on the average depth of tries," *Computing*, vol. 28, pp. 367–371, 1982.

L. Devroye, "A probabilistic analysis of the height of tries and of the complexity of triesort," *Acta Informatica*, vol. 21, pp. 229–237, 1984.

L. Devroye, "An analysis of random LC tries," *Random Structures and Algorithms*, vol. 15, pp. 359–375, 2001.

P. Flajolet and J. M. Steyaert, "A branching process arising in dynamic hashing, trie searching and polynomial factorization," in: *Lecture Notes in Computer Science*, vol. 140, pp. 239–251, Springer-Verlag, New York, 1982.

P. Flajolet, "On the performance evaluation of extendible hashing and trie search," *Acta Informatica*, vol. 20, pp. 345–369, 1983.

P. Flajolet and R. Sedgewick, "Digital search trees revisited," *SIAM Journal on Computing*, vol. 15, pp. 748–767, 1986.

E. H. Fredkin, "Trie memory," *Communications of the ACM*, vol. 3, pp. 490–500, 1960.

W. Hoeffding, "Probability inequalities for sums of bounded random variables," *Journal of the American Statistical Association*, vol. 58, pp. 13–30, 1963.

J.-P. Iivonen, S. Nilsson, and M. Tikkanen, "An experimental study of compression methods for functional tries," in: *Workshop on Algorithmic Aspects of Advanced Programming Languages (WAAAPL'99)*, 1999.

P. Jacquet and M. Régnier, "Trie partitioning process: limiting distributions," in: *Lecture Notes in Computer Science*, vol. 214, pp. 196–210, 1986.

P. Kirschenhofer and H. Prodinger, "Some further results on digital trees," in: *Lecture Notes in Computer Science*, vol. 226, pp. 177–185, Springer-Verlag, Berlin, 1986.

P. Kirschenhofer, H. Prodinger, and W. Szpankowski, "On the balance property of PATRICIA tries: external path length viewpoint," *Theoretical Computer Science*, vol. 68, pp. 1–17, 1989.

P. Kirschenhofer, H. Prodinger, and W. Szpankowski, "On the variance of the external path length in a symmetric digital trie," *Discrete Applied Mathematics*, vol. 25, pp. 129–143, 1989.

D. E. Knuth, *The Art of Computer Programming, Vol. 3 : Sorting and Searching*, Addison-Wesley, Reading, Mass., 1973.

H. Mendelson, "Analysis of extendible hashing," *IEEE Transactions on Software Engineering*, vol. 8, pp. 611–619, 1982.

A. Nilsson and G. Karlsson, "Fast address look-up for Internet routers," in: *Proceedings IFIP 4th International Conference on Broadband Communications*, pp. 11–22, 1998.

S. Nilsson and M. Tikkanen, "An experimental study of compression methods for dynamic tries," *Submitted to Algorithmica*, 1998.

S. Nilsson and G. Karlsson, "IP-address lookup using LC-tries," *IEEE Journal on Selected Areas in Communications*, vol. 17(6), pp. 1083–1092, 1999.

B. Pittel, "Asymptotical growth of a class of random trees," *Annals of Probability*, vol. 13, pp. 414–427, 1985.

B. Pittel, "Path in a random digital tree: limiting distributions," *Advances in Applied Probability*, vol. 18, pp. 139–155, 1986.

M. Régnier, "On the average height of trees in digital searching and dynamic hashing," *Information Processing Letters*, vol. 13, pp. 64–66, 1981.

M. Régnier, "Trie hashing analysis," in: *Proceedings of the Fourth International Conference on Data Engineering*, pp. 377–381, IEEE, Los Angeles, 1988.

W. Szpankowski, "Some results on V-ary asymmetric tries," *Journal of Algorithms*, vol. 9, pp. 224–244, 1988.

W. Szpankowski, "On the height of digital trees and related problems," *Algorithmica*, vol. 6, pp. 256–277, 1991.

W. Szpankowski, *Average Case Analysis of Algorithms on Sequences*, Springer-Verlag, New York, 2001.