ELSEVIER

Information Processing Letters 65 (1998) 195–199

Information Processing Letters

On the richness of the collection of subtrees in random binary search trees

Luc Devroye [1]

School of Computer Science, McGill University, 3480 University Street, Montreal, Quebec, Canada H3A 2A7

Received 27 June 1997
Communicated by F. Dehne

Abstract

The purpose of this paper is to settle two conjectures by Flajolet, Gourdon and Martinez (1996). We confirm that in a random binary tree on n nodes, the expected number of different subtrees grows indeed as $\Theta(n/\log n)$. Secondly, if K is the largest integer such that all possible shapes of subtrees of cardinality less than or equal to K occur in a random binary search tree, then we show that $K \sim \log n/\log\log n$ in probability. © 1998 Published by Elsevier Science B.V.

Keywords: Probabilistic analysis; Random binary search trees; Random permutation; Subtrees; Computational complexity

1. Introduction

The catalyst for this paper is the work of Flajolet, Gourdon and Martinez [4]: if N is the number of different (shapes of) subtrees in a random binary search tree on n nodes (which are constructed by insertion of a uniform random permutation of n numbers), then these authors showed that

$$\mathrm{E}\{N\} \leqslant \frac{(4+\mathrm{o}(1))n}{\log_2 n}.$$

They conjectured that this is indeed the right order of growth. Without attempting to obtain the best constant, we show the following.

Theorem 1.

$$\mathrm{E}\{N\} = \Theta\left(\frac{n}{\log n}\right).$$

The size of N matters to those who use compression methods for storing or transmitting the shapes of binary search trees. The richness of the collection of subtrees may also be measured in a different manner. Let K be the largest integer such that all possible (shapes of) subtrees of size K or less occur as subtrees. Based upon similar properties for strings shown by Flajolet, Kirschenhofer and Tichy [5], Flajolet, Gourdon and Martinez [4] conjecture that K should be close to $\log_4 n$. We settle this conjecture by showing the following.

Theorem 2.

$$\frac{K}{(\log n)/(\log\log n)} \to 1 \quad \textit{in probability.}$$

[1] Email: luc@cs.mcgill.ca. Research of the author was sponsored by NSERC Grant A3456 and by FCAR Grant 90-ER-0291.

0020-0190/98/$19.00
PII S0020-0190(97)00206-8

2. Proof of Theorem 1

For Theorem 1, it is good to recall some properties of paged trees. Let t denote a binary tree, and let $|t|$ be its size. Let N_t be the number of subtrees of a random binary search tree on n nodes whose shape is identical to t. The b-index of a tree is a tree that retains only the nodes of size $> b$. The size of this b-index is denoted by

$$B = \sum_{t:\,|t|>b} |N_t|.$$

The idea is that subtrees of size $\leqslant b$ can be trimmed away and stored in pages of capacity b in peripheral storage. The b-index resides in main storage. We need two results about B.

Lemma 3 (Knuth [7, p. 122]). *For $n > b \geqslant 2$,*

$$\mathrm{E}\{B\} = \frac{2(n+1)}{b+2} - 1.$$

Lemma 4 (Flajolet, Gourdon and Martinez [4]). *For $b \geqslant 2$,*

$$\mathrm{Var}\{B\} = \frac{2(b-1)b(b+1)(n+1)}{3(b+2)^2}.$$

In the last paper, the authors also obtain a Gaussian limit law for B. Both results and the limit law can also be obtained from the general results of Devroye [2].

We briefly recall the proof of the upper bound for Theorem 1, as given by Flajolet, Gourdon and Martinez. The number of binary trees on k nodes is

$$C_k = \frac{1}{k+1}\binom{2k}{k}.$$

Define the threshold

$$b = \lfloor (1-\varepsilon)\log_4 n \rfloor.$$

Then we have

$$N \leqslant \sum_{i=0}^{b} C_i + \sum_{t:\,|t|>b} |N_t|.$$

We know that as $k \to \infty$,

$$C_k \sim \frac{4^k}{\sqrt{\pi}k^{3/2}},$$

so that for n large enough, if $\varepsilon < 1/2$,

$$\sum_{i=0}^{b} C_i = \mathrm{O}\Big(\frac{4^b}{b^{3/2}}\Big) = \mathrm{O}\Big(\frac{n^{1-\varepsilon}}{(\log_4 n)^{3/2}}\Big)$$

uniformly over all such ε. Take

$$\varepsilon = \frac{\log\log n}{\log n},$$

so that the upper bound becomes

$$\mathrm{O}\Big(\frac{n}{(\log n)^{5/2}}\Big).$$

From Lemma 3,

$$\mathrm{E}\Big\{\sum_{t:\,|t|>b} |N_t|\Big\} = \frac{2(n+1)}{b+2} - 1.$$

As $b \sim \log_4 n$, we have $\mathrm{E}\{N\} = \mathrm{O}(n/\log n)$.

For a lower bound, we argue not very differently. Let A denote the event that among subtrees of size $> b$, some duplicates occur, where

$$b = \lceil (4+\varepsilon)\log_3 n \rceil$$

and $\varepsilon > 0$ is arbitrary. Let $p_{k,t}$ denote the probability that a random binary search tree on k nodes is identical to a given tree t. This parameter has been studied by Fill [3]. We do not need any deep results on $p_{k,t}$ beyond

$$p_{k,t} = \prod_{u\in t} \frac{1}{|u|},$$

where the product is over all nodes u of t, and $|u|$ denotes the size of the subtree rooted at u. Flajolet, Gourdon and Martinez [4] provide the upper bound

$$p_{k,t} \leqslant 2^{-k/4}, \quad k \geqslant 4.$$

Fill [3] showed that

$$p_{k,t} \leqslant \mathrm{e}^{-ck+\mathrm{O}(\log^2 k)}$$

where $c = \ln(4) - \sum_{j=1}^{\infty} 2^{-j}|\ln(1-2^{-j})| \approx 0.946$. Both bounds will do, but we provide a simple non-asymptotic bound for further reference.

Lemma 5. *For all $k \geqslant 0$,*

$$\sup_{t:\,|t|=k} p_{k,t} \leqslant 3^{-(k-1)/2}.$$

Proof. We proceed by induction on k and show that

$$\sup_{t:\,|t|=k} p_{k,t} \leqslant c2^{-Ck}$$

for some constants c and C. Clearly, for $k = 0$ and $k = 1$, the formula is valid provided that $c \geqslant 1$ and $C \leqslant \log_2 c$. For $k = |t| = 2$, $p_{2,t} = 1/2$, so $c4^{-C} \geqslant 1/2$. Assuming that $k \geqslant 3$, $|t| = k$, and that the left and right subtrees of the root are l and r with $|l| + |r| = k - 1$, we have

$$\begin{aligned} p_{k,t} &\leqslant \frac{p_{|l|,l}\,p_{|r|,r}}{k} \leqslant \frac{c^2 2^{-C(|l|+|r|)}}{k} = \frac{c^2 2^C 2^{-Ck}}{k} \\ &\leqslant c2^{-Ck} \end{aligned}$$

provided that $k \geqslant c2^C$. All the inequalities for c and C can be simultaneously satisfied if we pick $c = \sqrt{3}$ and $C = \log_4 3$. Thus,

$$\sup_{t:\,|t|=k} p_{k,t} \leqslant 3^{-(k-1)/2}. \qquad \square$$

We conclude that the probability that two subtrees both have sizes $> b$ and have identical shapes is

$$\mathrm{P}\{A\} \leqslant n^2 3^{-b/2} \to 0$$

by choice of b. Here we used the union bound and the fact that if two nodes are not in an ancestor/descendant relationship, then conditional on the sizes of the subtrees being m and n, the subtrees themselves are independent random binary search trees of sizes m and n respectively. Therefore, if A^c denotes the complement of A,

$$\begin{aligned} \mathrm{E}\{N\} &\geqslant \mathrm{E}\{B I_{A^c}\} \\ &= \mathrm{E}\{B\} - \mathrm{E}\{B I_A\} \\ &\geqslant \mathrm{E}\{B\} - \sqrt{\mathrm{E}\{B^2\}\mathrm{P}\{A\}} \\ &\qquad \text{(by the Cauchy–Schwarz inequality)} \\ &= \frac{2(n+1)}{b+1} - 1 - o\left(\sqrt{(\mathrm{E}^2\{B\} + \mathrm{Var}\{B\})}\right) \\ &\sim \frac{2n}{(4+\varepsilon)\log_3 n} \end{aligned}$$

by Lemmas 3 and 4 and our choice of b. This concludes the proof of Theorem 1. $\square$

3. Proof of Theorem 2

3.1. An upper bound

Let $\varepsilon > 0$. Define $k = \lceil (1+\varepsilon)\log n/\log\log n \rceil$. Verify that $k! = n^{1+\varepsilon+o(1)}$. Denote by L_k a binary tree on k nodes consisting of a chain of left children. If the random binary search tree is constructed incrementally by standard insertions of $X_1, \ldots, X_n$, a random permutation of $1, \ldots, n$, then we let T_i be a subtree rooted at the node for X_i. The size of T_i is $|T_i|$. We have

$$\begin{aligned} \mathrm{P}\{K > k\} &\leqslant \mathrm{P}\left\{\bigcup_{i=1}^n [T_i = L_k]\right\} \\ &\leqslant \sum_{i=1}^n \mathrm{P}\{T_i = L_k\} \\ &\leqslant \sum_{i=1}^n \mathrm{P}\{T_i = L_k \mid |T_i| = k\} \\ &= \frac{n}{k!} \\ &= \frac{1}{n^{\varepsilon+o(1)}} \to 0. \end{aligned}$$

3.2. A lower bound

For an accompanying lower bound, we define $k = \lceil (1-\varepsilon)\log n/\log\log n \rceil$, and note that $k! = n^{1-\varepsilon+o(1)}$, where $\varepsilon \in (0,1)$. The random binary search tree many also be thought of as based upon an i.i.d. uniform $[0,1]$ sequence $X_1, \ldots, X_n$. Assuming n even, the partial tree based upon $X_1, \ldots, X_{n/2}$ has $n/2 + 1$ external nodes (these are at all possible positions for insertion of a new node). When the tree is completed by adding $X_{n/2+1}, \ldots, X_n$, these external nodes grow to trees labeled $T_1, \ldots, T_{n/2+1}$ (note the change in definition from the first part of the proof). Some of these trees may have size 0. We recall that there are $C_l = \frac{1}{l+1}\binom{2l}{l}$ possible shapes of binary trees on l nodes. Each of these shapes has a probability of occurrence at least equal to $1/l!$ under the random binary search tree model (this is easy to show by induction). Let us denote by T the vector of cardinalities $|T_1|, \ldots, |T_{n/2+1}|$. Note that given these cardinalities, the shapes of the T_i's are clearly independent. Thus, if $N_i = \sum_{j=1}^{n/2+1} I_{|T_j|=i}$,

$$
\begin{aligned}
&\mathrm{P}\{K < k | T\} \\
&\leqslant \sum_{i=1}^{k} \mathrm{P}\{\text{one of the } C_i \text{ binary trees is not represented by the } T_j\text{s} \mid T\} \\
&\leqslant \sum_{i=1}^{k} C_i \sup_{t:\,|t|=i} \mathrm{P}\{t \text{ is not represented by the } T_j\text{s} \mid T\} \\
&= \sum_{i=1}^{k} C_i \sup_{t:\,|t|=i} \prod_{j=1}^{n/2+1} \mathrm{P}\{T_j \neq t \mid |T_j|\} \\
&\leqslant \sum_{i=1}^{k} C_i \sup_{t:\,|t|=i} \prod_{j=1}^{n/2+1} \left(I_{|T_j| \neq i} + I_{|T_j|=i}(1 - 1/i!)\right) \\
&= \sum_{i=1}^{k} C_i \prod_{j=1}^{n/2+1} \left(1 - I_{|T_j|=i}/i!\right) \\
&\leqslant \sum_{i=1}^{k} C_k \mathrm{e}^{-\sum_{j=1}^{n/2+1} I_{|T_j|=i}/i!} \\
&= \sum_{i=1}^{k} C_k \mathrm{e}^{-N_i/i!}.
\end{aligned}
$$

Therefore,

$$
\mathrm{P}\{K < k\} \leqslant \sum_{i=1}^{k} C_k \mathrm{E}\{\mathrm{e}^{-N_i/i!}\}.
$$

To bound this, it helps to condition on $X = (X_1, \ldots, X_{n/2})$. Conditional on X, the sizes $|T_j|$ are indeed multinomially distributed. As the components of a multinomial random vector are negatively associated (see [6]), we have for all $\lambda > 0$,

$$
\mathrm{E}\left\{\mathrm{e}^{-\lambda \sum_{j=1}^{n/2+1} I_{|T_j|=i}} \,\middle|\, X\right\} \leqslant \prod_{j=1}^{n/2+1} \mathrm{E}\{\mathrm{e}^{-\lambda I_{|T_j|=i}} \mid X\}.
$$

The points X_i, $1 \leqslant i \leqslant n/2$, define $n/2+1$ spacings $S_1, \ldots, S_{n/2+1}$. Given S_j, $|T_j|$ is binomial $(n/s, S_j)$. Thus, if $p_{j,i}$ is the probability that such a binomial takes the value i,

$$
\begin{aligned}
\mathrm{E}\{\mathrm{e}^{-\lambda I_{|T_j|=i}} \mid S_j\} &= p_{j,i}\mathrm{e}^{-\lambda} + (1 - p_{j,i}) \\
&\leqslant \mathrm{e}^{-p_{j,i}(1-\mathrm{e}^{-\lambda})}
\end{aligned}
$$

so that

$$
\begin{aligned}
&\mathrm{E}\left\{\mathrm{e}^{-\lambda \sum_{j=1}^{n/2+1} I_{|T_j|=i}} \,\middle|\, X\right\} \\
&\quad \leqslant \mathrm{e}^{-\sum_{j=1}^{n/2+1} p_{j,i}(1-\mathrm{e}^{-\lambda})} \stackrel{\text{def}}{=} \mathrm{e}^{-Z_i(1-\mathrm{e}^{-\lambda})},
\end{aligned}
$$

where

$$
Z_i \stackrel{\text{def}}{=} \sum_{j=1}^{n/2+1} p_{j,i}.
$$

Observe that Z_i is a function of X that is such that if one of the components of X is replaced by another value, then Z_i changes by at most 4. Therefore, by the independence of the components of X, and by McDiarmid's inequality ([9]; see also [1]), for $a > 0$,

$$
\mathrm{P}\{|Z_i - \mathrm{E}\{Z_i\}| > a\mathrm{E}\{Z_i\}\} \leqslant 2\mathrm{e}^{-a^2(\mathrm{E}\{Z_i\})^2/(n/2)4^2}.
$$

Take $a = 1/2$ and note that

$$
\begin{aligned}
&\mathrm{E}\left\{\mathrm{e}^{-\lambda \sum_{j=1}^{n/2+1} I_{|T_j|=i}} \,\middle|\, X\right\} \\
&\quad \leqslant \mathrm{e}^{-(1/2)\mathrm{E}\{Z_i\}(1-\mathrm{e}^{-\lambda})} + 2\mathrm{e}^{-(\mathrm{E}\{Z_i\})^2/32n}.
\end{aligned}
$$

Summarizing the above bounds, we have

$$
\begin{aligned}
\mathrm{P}\{K < k\} &\leqslant \sum_{i=1}^{k} C_k \mathrm{E}\{\mathrm{e}^{-N_i/i!}\} \\
&\leqslant \sum_{i=1}^{k} C_k \mathrm{e}^{-(1/2)\mathrm{E}\{Z_i\}(1-\mathrm{e}^{-1/i!})} \\
&\quad + \sum_{i=1}^{k} C_k 2\mathrm{e}^{-(\mathrm{E}\{Z_i\})^2/32n} \\
&\leqslant kC_k \sup_{1\leqslant i\leqslant k} \mathrm{e}^{-(1/2)\mathrm{E}\{Z_i\}(1-\mathrm{e}^{-1/k!})} \\
&\quad + 2kC_k \sup_{1\leqslant i\leqslant k} \mathrm{e}^{-(\mathrm{E}\{Z_i\})^2/32n}.
\end{aligned}
$$

The proof is thus finished if we can obtain a lower bound for $\mathrm{E}\{Z_i\}$. By linearity of expectation,

$$
\mathrm{E}\{Z_i\} = (n/2 + 1)\mathrm{P}\{\mathrm{binomial}(n/2, X_{(1)}) = i\},
$$

where $X_{(1)} = \min(X_1, \ldots, X_{n/2})$. The previous formula follows from the fact that all uniform spacings have identical distributions. Assume that $1 \leqslant i \leqslant k \leqslant$

$n/4$. If we consider the entire sequence $X_1, \ldots, X_n$, it should be clear that

$$\begin{aligned}
&\mathrm{P}\{\mathrm{binomial}(n/2, X_{(1)}) = i\} \\
&\quad = \frac{n/2}{n} \times \frac{n/2-1}{n-1} \times \cdots \times \frac{n/2-i+1}{n-i+1} \times \frac{n/2}{n-i} \\
&\quad \geqslant \frac{1}{2}\left(\frac{n/2-i}{n-i}\right)^i \\
&\quad = \frac{1}{2}\left(1 - \frac{n/2}{n-i}\right)^i \\
&\quad \geqslant \frac{1}{2}\mathrm{e}^{-(in/2)/(n/2-i)} \\
&\qquad\qquad (\text{use } 1-u \leqslant \mathrm{e}^{-u/(1-u)} \text{ for } u \in (0,1)) \\
&\quad = \frac{1}{2}\mathrm{e}^{-(in)/(n-2i)} \\
&\quad \geqslant \frac{1}{2}\mathrm{e}^{-2i} \\
&\quad \geqslant \frac{1}{2}\mathrm{e}^{-2k}.
\end{aligned}$$

Therefore

$$\mathrm{E}\{Z_i\} \geqslant \frac{n\mathrm{e}^{-2k}}{4}$$

and

$$\begin{aligned}
\mathrm{P}\{K < k\} &\leqslant kC_k \mathrm{e}^{-n\mathrm{e}^{-2k}(1-\mathrm{e}^{-1/k!})/8} \\
&\quad + 2kC_k \mathrm{e}^{-n\mathrm{e}^{-4k}/256}.
\end{aligned}$$

As $kC_k = \mathrm{O}(4^k) = n^{\mathrm{o}(1)}$, $\mathrm{e}^{-2k} = n^{\mathrm{o}(1)}$, $\mathrm{e}^{-4k} = n^{\mathrm{o}(1)}$ and $1 - 1/k! \sim 1/k! = n^{\varepsilon-1+\mathrm{o}(1)}$, we see that

$$\mathrm{P}\{K < k\} \leqslant n^{\mathrm{o}(1)}\mathrm{e}^{-n^{\varepsilon+\mathrm{o}(1)}} + n^{\mathrm{o}(1)}\mathrm{e}^{-n^{1+\mathrm{o}(1)}} \to 0.$$

This concludes the proof of Theorem 2. □

References

[1] K. Azuma, Weighted sums of certain dependent random variables, Tohoku Math. J. 37 (1967) 357–367.

[2] L. Devroye, Limit laws for local counters in random binary search trees, Random Structures and Algorithms 2 (1991) 303–316.

[3] J.A. Fill, On the distribution of binary search trees under the random permutation model, Random structures and Algorithms 8 (1996) 1–25.

[4] P. Flajolet, X. Gourdon, C. Martinez, Patterns in random binary search trees, Tech. Rept., INRIA Rocquencourt, 1996.

[5] P. Flajolet, P. Kirschenhofer, R.F. Tichy, Deviations from uniformity in random strings, Probability Theory and Related Fields 80 (1988) 139–150.

[6] K. Joag-Dev, F. Proschan, Negative association of random variables, with applications, Ann. of Statist. 11 (1983) 286–295.

[7] D.E. Knuth, The Art of Computer Programming, vol. 1: Fundamental Algorithms, 2nd ed., Addison-Wesley, Reading, MA, 1973.

[8] H.M. Mahmoud, Evolution of Random Search Trees, John Wiley, New York, 1992.

[9] C. McDiarmid, On the method of bounded differences. in: Surveys in Combinatorics 1989, London Mathematical Society Lecture Notes Series, vol. 141, 1989, Cambridge University Press, Cambridge, pp. 148–188.