ON THE VARIANCE OF THE HEIGHT OF RANDOM BINARY SEARCH TREES

Luc Devroye
and
Bruce Reed
School of Computer Science
McGill University

ABSTRACT. Let H_n be the height of a random binary search tree on n nodes. We show that there exists a constant $\alpha = 4.31107\ldots$ such that $\mathbf{P}\{|H_n - \alpha \log n| > \beta \log\log n\} \to 0$, where $\beta > 15\alpha/\ln 2 = 93.2933\ldots$. The proof uses the second moment method, and does not rely on properties of branching processes. We also show that $\mathrm{Var}\{H_n\} = O((\log\log n)^2)$.

KEYWORDS AND PHRASES. Binary search tree. Probabilistic analysis. Random tree. Asymptotics. Height. Second moment method.

Introduction.

The height H_n of a random binary search tree on n nodes, constructed in the usual manner starting from a random equiprobable permutation of $1, \ldots, n$, is known to be close to $\alpha \log n$ where $\alpha = 4.31107\ldots$ is the unique solution on $[2, \infty)$ of the equation $\alpha \log((2e)/\alpha) = 1$. First, Pittel (1984) showed that $H_n / \log n \to \gamma$ almost surely as $n \to \infty$ for some positive constant γ. This constant was known not to exceed α (Robson, 1979), and it was shown in Devroye (1986) that $\gamma = \alpha$, as a consequence of the fact that $\mathbf{E}H_n \sim \alpha \log n$. Robson (1982) has found that H_n does not vary much from experiment to experiment, and seems to have a fixed range of width not depending upon n. Devroye (1987) proved that $H_n - \alpha \log n = O(\sqrt{\log n \log \log n})$ in probability, but this does not quite confirm Robson's findings. It is the purpose of this note to prove

THEOREM.

$$\mathbf{E}H_n = \alpha \log n + O(\log \log n)$$

and

$$\mathrm{Var}\,\{H_n\} = O((\log \log n)^2)\ .$$

While this is a major step forward, we still do not know whether $\mathrm{Var}\,\{H_n\} = O(1)$. For more information on random binary search trees, one may consult Knuth (1973), Aho, Hopcroft and Ullman (1975, 1983), Mahmoud and Pittel (1984), Devroye (1990), Mahmoud (1992) and Pittel (1992).

Finally, we note that the present paper contains the first proof of the asymptotic properties of H_n that is not based upon the theory of branching processes or branching random walks. We merely employ a well-known representation of random binary search trees due to Devroye (1986), and combine it with the second moment method which has found so many other applications in the theory of random graphs (see, e.g., Palmer, 1985).

Notation and definitions.

Let T_∞ be the complete infinite binary tree. Each node x has a right son $r(x)$ and a left son $l(x)$. We consider a random labelled tree R_∞ obtained from T_∞ by choosing a uniform $[0,1]$ random variable $U(x)$ for each node x of T_∞ and labelling the edge $(x, r(x))$ by $U(x)$ and the edge $(x, l(x))$ by $1 - U(x)$. The label of edge a is denoted $L(a)$. We let R_k be the random tree consisting of the first k edge levels of R_∞.

For each node y of R_∞, we let $f(y)$ be the product of the labels of the edges on the unique path from the root to y. We remark that for each $x \in R_\infty$, $-\log U(x)$ is an exponential random variable with mean 1. If the labels on the path from the root to a node y of R_∞ are $U_1, \ldots, U_i$, then we define

$$h_n(y) = \lfloor \cdots \lfloor \lfloor nU_1 \rfloor U_2 \rfloor \cdots U_i \rfloor \ .$$

Also, $-\log f(y)$ is distributed as the sum of i i.i.d. exponential random variables with mean one, i.e., it is gamma distributed with parameter i.

FACT 1. *It is well known that we can construct a random binary search tree T_n on n nodes by taking a copy R of R_∞ and letting T_n consist of those nodes y of R with $h_n(y) \geq 1$. (See, e.g., Devroye, 1986.)*

FACT 2. *Let y be a node of R_∞ at depth i (i.e., at edge-distance i from the root). Then*

$$nf(y) - i \leq h_n(y) \leq nf(y) \ .$$

Facts 1 and 2 basically allow us to obtain refined information regarding H_n merely by studying R_∞. The inequality in Fact 2 introduces a certain looseness; in fact, it will limit the accuracy of the results on H_n to be $O(\log\log n)$.

Lemmas regarding the gamma distribution.

The sum S_n of n i.i.d. exponential random variables with mean 1 is gamma (n) distributed. Its density is given by

$$g(t) = \frac{t^{n-1}e^{-t}}{(n-1)!} \ , \ t > 0 \ .$$

LEMMA 1. *Let* $\{t_n\}$ *be a sequence of numbers such that* $t_n \sim cn$ *as* $n \to \infty$ *for some* $c \in (0,1)$. *Then*

$$\mathbf{P}\{S_n < t_n\} \sim \frac{1}{1-c}\frac{e^{-t_n}(t_n)^n}{n!} \; .$$

PROOF. By integration by parts,

$$\begin{aligned}
\mathbf{P}\{S_n < t_n\} &= \int_0^{t_n} \frac{t^{n-1}e^{-t}}{(n-1)!}\,dt \\
&= e^{-t_n}\left(\frac{t_n^n}{n!} + \frac{t_n^{n+1}}{(n+1)!} + \frac{t_n^{n+2}}{(n+2)!} + \cdots\right) \\
&\sim \frac{1}{1-c}\frac{e^{-t_n}(t_n)^n}{n!} \; . \square
\end{aligned}$$

LEMMA 2. *Let* $t \in (0,1)$ *be a fixed constant. Then*

$$\frac{e^{-tn}(tn)^n}{n!} \le \mathbf{P}\{S_n < tn\} \le \frac{1}{1-t}\frac{e^{-tn}(tn)^n}{n!} \; . \square$$

PROOF. The lower bound follows directly by integration by parts as in the proof of Lemma 1. For the upper bound, note that

$$\begin{aligned}
\mathbf{P}\{S_n < tn\} &\le e^{-tn}\left(\frac{(tn)^n}{n!} + \frac{(tn)^{n+1}}{(n+1)!} + \frac{(tn)^{n+2}}{(n+2)!} + \cdots\right) \\
&\le \frac{e^{-tn}(tn)^n}{n!}\left(1 + \frac{tn}{n+1} + \left(\frac{tn}{n+1}\right)^2 + \cdots\right) \\
&\le \frac{e^{-tn}(tn)^n}{n!}\left(\frac{1}{1-t}\right) \; . \square
\end{aligned}$$

LEMMA 3.

$$A \le \sqrt{n}2^n\mathbf{P}\{S_n < n/\alpha\} \le B \; ,$$

where $A = e^{-1/12}/\sqrt{2\pi}$ *and* $B = \alpha/((\alpha-1)\sqrt{2\pi})$.

PROOF. From Lemma 2,

$$\frac{e^{-n/\alpha}(n/\alpha)^n}{n!} \le \mathbf{P}\{S_n < n/\alpha\} \le \frac{1}{1-1/\alpha}\frac{e^{-n/\alpha}(n/\alpha)^n}{n!} \; .$$

Use the fact that $n! = (n/e)^n\sqrt{2\pi n}e^{\theta/(12n)}$ for some $\theta \in (0,1)$ and the definition of α. $\square$

LEMMA 4. *There exists a universal constant C such that*

$$\mathbf{P}\{S_n \geq Cn\} \leq 2^{-2n} .$$

$C = 5$ will do.

PROOF. Take $C > 1$. By Chernoff's exponential bounding method (Chernoff, 1952), for $t > 0$,

$$\mathbf{P}\{S_n \geq Cn\} \leq \mathbf{E}e^{tS_n}e^{-tCn} = (1-t)^{-n}e^{-tCn} = \left(Ce^{1-C}\right)^n ,$$

where we took $1 - t = 1/C$. For C large enough (e.g., $C \geq 5$), this is less than 4^{-n}. □

LEMMA 5. *Let $E_1, E_2, \ldots, E_n$ be i.i.d. random variables with a density, and let a be a fixed constant. Then*

$$\mathbf{P}\{E_1 < a, E_1 + E_2 < 2a, \ldots, E_1 + \cdots + E_n < na \mid E_1 + \cdots + E_n < na\} \geq \frac{1}{n} .$$

PROOF. Define $F_i = E_i - a$ for all i. Define $E_r = E_{r-n}$ when $n < r \leq 2n$. Then, by symmetry,

$$\begin{aligned}
&\mathbf{P}\{E_1 < a, E_1 + E_2 < 2a, \ldots, E_1 + \cdots + E_n < na \mid E_1 + \cdots + E_n < na\} \\
&\quad = \mathbf{P}\{F_1 < 0, F_1 + F_2 < 0, \ldots, F_1 + \cdots + F_n < 0 \mid F_1 + \cdots + F_n < 0\} \\
&\quad = \frac{1}{n}\sum_{i=1}^{n} \mathbf{P}\{F_i < 0, F_i + F_{i+1} < 0, \ldots, F_i + \cdots + F_{i+n-1} < 0 \mid F_1 + \cdots + F_n < 0\} \\
&\quad = \mathbf{P}\{F_S < 0, F_S + F_{S+1} < 0, \ldots, F_S + \cdots + F_{S+n-1} < 0 \mid F_1 + \cdots + F_n < 0\} ,
\end{aligned}$$

where S is independent of the E_i's and uniformly distributed on $\{1, \ldots, n\}$. Now, fix $E_1, \ldots, E_n$, and let $s \in \{1, \ldots, n\}$ be the (unique) value at which $\sum_{i>0, i<s} F_i$ is maximal. If $s = 1$, then $\sum_{i=1}^{j} F_i < 0$ for all $j > 0$. If $s > 1$, then, as $\sum_{i=1}^{n} F_i < 0$, we see that $\sum_{i=s}^{s+j} F_i = \sum_{i=1}^{s+j} F_i - \sum_{i=1}^{s-1} F_i < 0$ for all $j \geq 0$. Thus,

$$\begin{aligned}
&\mathbf{P}\{F_S < 0, F_S + F_{S+1} < 0, \ldots, F_S + \cdots + F_{S+n-1} < 0 \mid F_1 + \cdots + F_n < 0\} \\
&\quad \geq \mathbf{P}\{S = s\} = \frac{1}{n} . \ \square
\end{aligned}$$

Proof of the Theorem.

Lemma 6. *Consider positive integers $n > k$. Then*

$$\mathbf{P}\{H_n \geq k\} \geq \mathbf{P}\{\exists \text{ leaf } y \in R_k \ : \ f(y) \geq (k+1)/n\} \ .$$

Proof. This follows immediately from facts 1 and 2. □

Lemma 7. *There exists a constant $d > 0$ such that for sufficiently large j,*

$$\mathbf{P}\{\exists \text{ leaf } y \in R_j \ : \ f(y) \geq (j+1)/\exp(j/\alpha + d\log(j/\alpha))\} \geq 1 - \frac{1}{j^3} \ .$$

We may pick $d = \epsilon + 15/\log 2$ for any small $\epsilon > 0$.

Proof. To follow. □

Lemma 8. *Let d be the constant of Lemma 7. Then, for sufficiently large n,*

$$\mathbf{P}\{H_n \geq \alpha \log n - d\alpha \log\log n - 1\} \geq 1 - \frac{1}{(\alpha \log n)^3} \ .$$

We may choose $d = \epsilon + 15/\log 2$ for any small $\epsilon > 0$.

Proof. Follows from Lemmas 6 and 7 by setting $j = k = \lfloor \alpha \log n - d\alpha \log\log n \rfloor$. □

Lemma 9.

$$\mathbf{P}\{H_n \geq \lceil \alpha \log n + i \rceil\} \leq \left(\frac{2}{\alpha}\right)^i \ , \ i \geq 0 \ .$$

Proof. See page 492 of Devroye (1986). □

Note that the Theorem follows from Lemmas 8 and 9 without further work.

Proof of Lemma 7.

Lemma 10. *For every i, with probability at least $1-2^{-i}$, every leaf of R_i has $f(y) \geq e^{-5i}$.*

Proof. The probability that for some leaf y of R_i, we have $f(y) < e^{-5i}$ is at most 2^i times $\mathbf{P}\{S_i \geq 5i\}$, where S_i is gamma i distributed. By Lemma 4, this does not exceed $2^i/4^i = 2^{-i}$. $\square$

Lemma 11. *For sufficiently large k, with probability at least $1/k^3$, there is a leaf y of R_k with $f(y) \geq e^{-k/\alpha}$.*

Lemma 11 will be proved further on. If Lemma 11 is true, then we can proceed with the proof of Lemma 7 as follows: first note that we can obtain a copy of R_{i+k} by making each leaf of R_i a root of a copy of R_k, where all these trees are independently labelled. Define $k = \lfloor j - A\log j \rfloor$ and $i = \lceil A \log j \rceil$ so that $j = k + i$ with some constant A to be picked further on. Note first that for j large enough, if $A > \alpha$,

$$\frac{k}{\alpha} + 5i \leq \frac{j}{\alpha} + 5A\log\left(\frac{j}{\alpha}\right) - \log(j+1) \ .$$

Then,

$$\begin{aligned}
&\mathbf{P}\left\{\nexists \text{ leaf } y \in R_j \text{ with } f(y) \geq 1/\exp(j/\alpha + 5A\log(j/\alpha) - \log(j+1))\right\} \\
&\quad \leq \mathbf{P}\left\{\exists \text{ leaf } y \in R_i \text{ with } f(y) < e^{-5i}\right\} \\
&\qquad + \mathbf{P}\left\{\nexists \text{ leaf } y \in R_j \text{ with } f(y) \geq 1/\exp(j/\alpha + d\log(j/\alpha) - \log(j+1)) \right. \\
&\qquad\qquad \left. \mid \forall \text{ leaf } y \in R_i : f(y) \geq e^{-5i}\right\} \\
&\quad \leq 2^{-i} + \mathbf{P}\left\{\text{every copy of } R_k \text{ contains no leaf } y \text{ with } f(y) \geq 1/\exp(k/\alpha)\right\} \\
&\qquad \text{(by Lemma 10)} \\
&\quad \leq 2^{-i} + \left(1 - k^{-3}\right)^{2^i} \quad \text{(by Lemma 11)} \\
&\quad \leq 2^{-i} + \exp\left(-2^i k^{-3}\right) \\
&\quad \leq j^{-A\log 2} + \exp\left(-j^{A\log 2 - 3}\right) \\
&\quad \leq j^{-3}
\end{aligned}$$

for j large enough, provided that $A\log 2 > 3$. This proves Lemma 7. We note that we can pick $d = 5A$, where $A = \epsilon + \max(\alpha, 3/\log 2)$ for any small $\epsilon > 0$. $\square$

Proof of Lemma 11. Let P be a path from the root to a leaf y of R_k. The condition $f(y) \geq 1/e^{k/\alpha}$ is equivalent to

$$\sum_{e \in P} (-\log L(e)) \leq \frac{|P|}{\alpha} \ .$$

We call a leaf y special if—in addition to the above condition—it satisfies

$$\sum_{e \in P'} (-\log L(e)) \leq \frac{|P'|}{\alpha}$$

for every subpath P' of P that originates at a terminal vertex y. Such subpaths are called terminal. Let $\mathcal{S}$ be the collection of special leaves of R_k. By Lemma 5, the expected number of special leaves is at least $1/k$ times $\mathbf{P}\{S_k < k/\alpha\}$ times 2^k. By Lemma 3,

$$\mathbf{E}|\mathcal{S}| \geq \frac{e^{-1/12}}{\sqrt{2\pi} k^{3/2}} \ .$$

Next, we consider the expected number of pairs of special leaves in order to be able to apply the second moment method. We fix a leaf z of R_k and count $|\mathcal{S}|$ given that $z \in \mathcal{S}$. To this end, let w be another leaf of R_k. Let P_z and P_w denote the paths from the root of R_k to z and w respectively. Then P_z and P_w have an initial common subsequence, the join $P_z \cap P_w$. Let $e_1, e_2, \ldots, e_k$ be the edges on the path from the root to z, and define $Q_i = \{e_1, \ldots, e_i\}$. For any j, the number of leaves of R_k whose join with P_z is Q_j is 2^{k-j}. Furthermore, the probability that a leaf $w \in R_k$ is a special leaf given that $z \in \mathcal{S}$ and that $P_z \cap P_w = Q_j$ is bounded above by the probability that for the terminal path $P' \subseteq P_w - Q_j$ with $|P'| = \max(0, k-j-1)$, we have

$$\sum_{e \in P'} (-\log L(e)) \leq \frac{|P'|}{\alpha} \ .$$

Note that P' contains one edge less than $P_w - Q_j$. This allows us further on to work out a conditional probability given $z \in \mathcal{S}$ without much trouble. By Lemma 3, the probability of the event mentioned above is at most

$$\frac{\alpha}{(\alpha-1)\sqrt{2\pi(k-j-1)}2^{k-j-1}} \ .$$

Thus,

$$\begin{aligned}
&\mathbf{E}\left\{|\{w \in \mathcal{S} : P_w \cap P_z = Q_j\}| \mid z \in \mathcal{S}\right\} \\
&\leq \frac{\alpha 2^{k-j}}{(\alpha-1)\sqrt{2\pi(k-j-1)}2^{k-j-1}} \\
&= \frac{2\alpha}{(\alpha-1)\sqrt{2\pi(k-j-1)}} \\
&\leq 2
\end{aligned}$$

when $k-j \geq 2$. The previous expected value is bounded by 2 when $k-j \in \{0,1\}$. Therefore,

$$\begin{aligned}
\mathbf{E}\{|\mathcal{S}| \mid z \in \mathcal{S}\} &= \sum_{j=0}^{k} \mathbf{E}\{|\{w \in \mathcal{S} : P_w \cap P_z = Q_j\}| \mid z \in \mathcal{S}\} \\
&\leq \sum_{j=0}^{k} 2 = 2k+2 \ .
\end{aligned}$$

Hence, by the second moment method,

$$\begin{aligned}
\mathbf{P}\{|\mathcal{S}| \geq 1\} &\geq \frac{\mathbf{E}|\mathcal{S}|}{1+\sup_{z \text{ leaf of } R_k} \mathbf{E}\{|\mathcal{S}||z \in \mathcal{S}\}} \\
&\geq \frac{\mathbf{E}|\mathcal{S}|}{2k+3} \\
&\geq \frac{e^{-1/12}}{\sqrt{2\pi}(2k+3)k^{3/2}} \\
&\geq \frac{1}{k^3}
\end{aligned}$$

for all k large enough. This concludes the proof of Lemma 11. $\square$

Acknowledgements.

The authors would like to thank Colin McDiarmid and an anonymous referee for helpful comments. Research of the authors was supported by NSERC grant A3456.

References.

A. V. Aho, J. E. Hopcroft, and J. D. Ullman, *The Design and Analysis of Computer Algorithms*, Addison-Wesley, Reading, Mass., 1975.

A. V. Aho, J. E. Hopcroft, and J. D. Ullman, *Data Structures and Algorithms*, Addison-Wesley, Reading, Mass., 1983.

H. Chernoff, "A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations," *Annals of Mathematical Statistics*, vol. 23, pp. 493–507, 1952.

L. Devroye, "A note on the height of binary search trees," *Journal of the ACM*, vol. 33, pp. 489–498, 1986.

L. Devroye, "Branching processes in the analysis of the heights of trees," *Acta Informatica*, vol. 24, pp. 277–298, 1987.

L. Devroye, "On the height of random m-ary search trees," *Random Structures and Algorithms*, vol. 1, pp. 191–203, 1990.

D. E. Knuth, *The Art of Computer Programming, Vol. 1: Fundamental Algorithms*, Addison-Wesley, Reading, Mass., 1973. 2nd Ed..

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

H. M. Mahmoud, *Evolution of Random Search Trees*, John Wiley, New York, 1992.

H. Mahmoud and B. Pittel, "On the most probable shape of a search tree grown from a random permutation," *SIAM Journal on Algebraic and Discrete Methods*, vol. 5, pp. 69–81, 1984.

E. M. Palmer, *Graphical Evolution*, John Wiley, New York, 1985.

B. Pittel, "On growing random binary trees," *Journal of Mathematical Analysis and Applications*, vol. 103, pp. 461–480, 1984.

B. Pittel, "Note on the heights of random recursive trees and random m-ary search trees," Technical Report, Department of Mathematics, The Ohio State University, 1992.

R. Pyke, "Spacings," *Journal of the Royal Statistical Society Series B*, vol. 7, pp. 395–445, 1965.

J. M. Robson, "The height of binary search trees," *The Australian Computer Journal*, vol. 11, pp. 151–153, 1979.

J. M. Robson, "The asymptotic behaviour of the height of binary search trees," *Australian Computer Science Communications*, p. 88, 1982.