ON THE EXPECTED HEIGHT OF FRINGE-BALANCED TREES

Luc Devroye
School of Computer Science
McGill University

ABSTRACT. We study the effect of a well-known balancing heuristic on the expected height of a random binary search tree. After insertion of an element, if any node on the insertion path has a subtree of size precisely $2t+1$ for a fixed integer t, then the subtree rooted at that node is destroyed and replaced by a new subtree in which the median of the $2t+1$ elements is the new root. If H_n denotes the height of the resulting random tree, we show that $H_n/\log n \to c(t)$ in probability for some function $c(t)$. In particular, $c(0) = 4.31107\ldots$ (the ordinary binary search tree), $c(1) = 3.192570\ldots$, $c(3) = 2.555539\ldots$, $c(10) = 2.049289\ldots$ and $c(100) = 1.623695\ldots$.

KEYWORDS AND PHRASES. Binary search tree, data structures, expected time analysis, height of a tree, balanced tree, fringe analysis, fringe reorganization.

CR CATEGORIES: 3.74, 5.25, 5.5.

Research of the author was sponsored by NSERC Grant A3456 and by FCAR Grant 90-ER-0291.
Address: School of Computer Science, McGill University, 3480 University Street, Montreal, Canada H3A 2A7.

1. Introduction.

Consider an ordinary binary search constructed by standard consecutive insertions of values $X_1, \ldots, X_n$. It is well-known that when the input forms a random permutation of $\{1, \ldots, n\}$ (or equivalently, when the input sequence is independent and identically distributed) that the height H_n of the tree satisfies the following convergence property:

$$\frac{H_n}{\log n} \to 4.31107\ldots \text{ a.s.}$$

(Robson, 1979; Devroye, 1986). The purpose of this note is to investigate what happens to H_n when we apply a very simple heuristic during the insertion process.

Bell (1965) and Walker and Wood (1976) introduced the following heuristic applied to the fringe of the tree: fix an integer t. After insertion of an element in the tree, verify whether one of the nodes on the insertion path is the root of a subtree of size precisely $2t+1$. If so, we perform the operation MEDIAN-ROOT on that subtree. This operation consists of finding the median of the $2t+1$ elements in the subtree, and making it the root. This can be done in a number of ways, via a splay operation, via a tree splitting method, or by complete reorganization of the subtree. As t is typically small—$t = 1$ is the most frequently studied case—, the reorganization of the subtree is less important.

Another way of looking at this, following Poblete and Munro (1985), is to consider internal nodes and external nodes, where internal nodes hold one data point and external nodes are bags of capacity $2t$. Insertion proceeds as usual. As soon as an external node overflows (i.e., when it would grow to size $2t+1$), its bag is split about the median, leaving two new external nodes (bags) of size t each, and an internal node holding the median. After the insertion process is completed, we may wish to expand the bags into balanced trees. In fact, what we do with the bags does not matter: if H'_n is the maximal distance between an internal node and the root, and H_n is the height of the tree, i.e, the maximal distance between any expanded external node and the root, it is easy to see that

$$H'_n < H_n \leq H'_n + 2t \ .$$

Asymptotically, this is unimportant.

Using the branching process method of proof (Devroye, 1986, 1987, 1990; see also Mahmoud, 1992) we obtain the almost sure limit value of $H_n / \log n$ for all t. For another possible proof method, see Pittel (1992). The improvement in H_n is important for small values of t. Not surprisingly, for every $\epsilon > 0$, we can find a t such that

$$\lim_{t \to \infty} \mathbf{P}\{H_n > (1+\epsilon) \log_2 n\} = 0 \ .$$

Unfortunately, the value of t increases very rapidly as $\epsilon \downarrow 0$.

2. The expected depth of a node.

The depth D_n of the last node when the fringe heuristic is used has been studied by the theory of Markov processes or urn models in a series of papers, notably by Poblete and Munro (1985), Aldous, Flannery and Palacios (1988). See also Gonnet and Baeza-Yates (1991, p. 109). Poblete and Munro (1985) showed that

$$\lim_{n\to\infty} \frac{\mathbf{E}D_n}{\log n} = d(t)$$

and

$$\frac{D_n}{\log n} \to d(t) \text{ i.p. } ,$$

where

$$\frac{1}{d(t)} = \sum_{i=t+1}^{2t+1} \frac{1}{1+i} .$$

It is a simple exercise to show that $d(t) \to 1/\log 2$ as $t \uparrow \infty$. The values of H_n we will obtain below are larger than these. A comparison of the limits will be provided.

3. The main result.

THEOREM 1. *A random binary search tree constructed with the aid of the fringe heuristic with parameter t has the following property:*

$$\frac{H_n}{\log n} \to c(t) \text{ i.p. } ,$$

where $c(t)$ is the unique solution greater than $d(t)$ of the equation

$$\lambda(c) - c \sum_{i=t+1}^{2t+1} \log\left(1 + \frac{\lambda(c)}{i}\right) + c \log 2 = 0 , \tag{1}$$

and $\lambda(c)$ is defined by the equation

$$\frac{1}{c} = \sum_{i=t+1}^{2t+1} \frac{1}{\lambda + i} .$$

A table of values for $d(t)$ and $c(t)$ is given below.

t	$c(t)$	$d(t)$
0	4.311070	2.000000
1	3.192570	1.714286
2	2.779633	1.621622
3	2.555539	1.575985
4	2.411554	1.548863
5	2.309726	1.530900
6	2.233133	1.518130
7	2.172976	1.508587
8	2.124195	1.501186
9	2.083648	1.495279
10	2.049289	1.490455
20	1.863726	1.467601
30	1.782617	1.459539
40	1.734851	1.455420
50	1.702554	1.452920
60	1.678898	1.451241
70	1.660617	1.450035
80	1.645976	1.449128
90	1.633883	1.448420
100	1.623695	1.447853

Remark 1. The existence of a solution. The constant $\lambda(c)$ is well-defined whenever

$$\frac{1}{c} \leq \sum_{i=t+1}^{2t+1} \frac{1}{1+i} \stackrel{\text{def}}{=} \frac{1}{d} \ .$$

This follows from the fact that the left-hand-side of (1) is a concave function of c, and that at $c = d$ (note that $\lambda(c) = 1$), the left-hand-side of (1) is positive (and takes the value 1). The derivative of (1) is

$$\lambda'(c) - \sum_{i=t+1}^{2t+1} \log\left(1 + \frac{\lambda(c)}{i}\right) - c\lambda'(c) \sum_{i=t+1}^{2t+1} \frac{1}{\lambda(c)+i} + \log 2 = -\sum_{i=t+1}^{2t+1} \log\left(1 + \frac{\lambda(c)}{i}\right) + \log 2 \ ,$$

and the second derivative of (1) is

$$-\frac{1}{\partial c/\partial\lambda} \sum_{i=t+1}^{2t+1} \frac{1}{\lambda(c)+i} = -\frac{1}{c\partial c/\partial\lambda} < 0 \ .$$

Thus, the left-hand side of (1) is a concave function in c.

4. A property of spacings.

As in Devroye (1985, 1986), we establish a crucial link between random binary search trees and trees of random variables. The root of a random binary search tree splits the $n-1$ remaining elements in two sets, one for each subtree, where the size of the left subtree is distributed as $\lfloor nU \rfloor$, and U is a uniform $[0,1]$ random variable. In the model that we are considering, the split is not according to a uniform $[0,1]$ random variable. Rather, the value U should be replaced by V, the median of $2t+1$ i.i.d. uniform $[0,1]$ random variables. Such a random variable will be called a median-of-uniform (MOU) random variable, and we will denote all such random variables by V or V_i. The sizes of the subtrees are jointly distributed as $(N, n-1-N)$, where $nV - t - 1 \leq N \leq nV + t$, and $\leq$ denotes stochastic ordering.

> Proof. Let U denote the minimum of $2t+1$ i.i.d. uniform $[0,1]$ random variables, and let M be the minimum of $2t+1$ integers drawn without replacement from $\{1, \ldots, n\}$. A simple argument shows that $M \leq nU + 1$, where $\leq$ denotes stochastic ordering. Thus, $n - M \geq nW - 1$, where W is distributed as the maximum of $2t+1$ i.i.d. uniform $[0,1]$ random variables. Iterating this $t+1$ times, an induction argument shows that $N \geq nV - (t+1)$, where V is distributed as the median of $2t+1$ i.i.d. uniform $[0,1]$ random variables. Since $n-1-N$ is distributed as N, we see that $N \leq nV + t$ as well, where all the inequalities still refer to stochastic dominance.

Each of the subtrees can be split in a similar fashion, requiring this time two new MOU random variables. This process can be repeated at all levels and it leads to a tree in which node values are subtree sizes that are approximately obtained as (truncated) products of MOU random variables. More formally, let T_k be a complete binary tree with k full levels of edges. The total number of edges is $2^1 + 2^2 + \ldots + 2^k = 2^{k+1} - 2$. We will use the symbol p for a path from root to leaf (there are 2^k such paths in T_k). Consider all edges pairwise in level order and from left to right, and associate with each pair an independent random vector distributed as $(V, 1-V)$ where V is distributed as a MOU random variable. Take one of the 2^k leaves of T_k and let p be the path from that leaf to the root of T_k. Let $\{V_1, V_2, \ldots, V_k\}$ denote the collection of MOU random variables encountered on the path from leaf to root. The quantity

$$nV_1V_2 \cdots V_k$$

is approximately equal to the size of the subtree rooted at a leaf in the random binary search tree. A node corresponds to a real internal node in the search tree when its values is $\geq 2t+1$. Indexing the MOU random variables in an obvious manner, we define

$$Z_k = \begin{cases} 1 & (k=0) \\ \max_p \prod_{i \in p} V_i & (k>0) \end{cases} .$$

LEMMA 1. *Let $n \geq 1$, $k \geq 1$ be given integers. Then the internal node height H'_n of a random binary search tree on n nodes with fringe reorganization is related to the random variables Z_k via the following inequalities:*

$$\mathbf{P}\{H'_n \geq k\} \geq \mathbf{P}\left\{Z_k \geq \frac{1+2t+k(t+1)}{n}\right\} ,$$

$$\mathbf{P}\{H'_n \geq k+2l\} \leq \mathbf{P}\left\{Z_k \geq \frac{1}{n}\right\} + 2^l \mathbf{P}\left\{\prod_{i=1}^{l} V_i \geq \frac{l}{k-l}\right\} ,0<l<k .$$

Here $V_1, \ldots, V_k$ are i.i.d. MOU *random variables.*

PROOF. A simple induction argument shows that the size of the subtree rooted at a node with path p to the root is at least equal to

$$n \prod_{i \in p} V_i - k(t+1) .$$

If this is at least $2t+1$ for some node, then $H'_n \geq k$. On the other hand, if $H'_n \geq k+2l$, then some node at distance k from the root has a subtree of size at least equal to $2l(t+1)+1$. On the other hand, if $N(p)$ denotes the size of the subtree rooted at the node at distance k from the root, where p is the path to the root, and $V_1, \ldots, V_k$ are the MOU variables associated with the edges on this path, with V_k nearest the root, then $N(p)$ is in distribution not greater than

$$\begin{aligned} & n \prod_{i=1}^{k} V_i + t + t \sum_{j=2}^{k} \prod_{i=j}^{k} V_i \\ & \leq n \prod_{i=1}^{k} V_i + lt + (k-l)t \prod_{i=k-l+1}^{k} V_i . \end{aligned}$$

The last product is shared by precisely 2^{k-l} paths. Therefore, by Bonferroni's inequality,

$$\begin{aligned}\mathbf{P}\{H_n' \geq k+2l\} &\leq \mathbf{P}\left\{\max_p N(p) \geq 2l(t+1)+1\right\} \\ &\leq \mathbf{P}\left\{Z_k \geq \frac{1}{n}\right\} + 2^l \mathbf{P}\left\{\prod_{i=1}^{l} V_i \geq \frac{l}{k-l}\right\} . \square\end{aligned}$$

Next, we need a simple result about the MOU distribution related to spacings of uniform random variables (see Pyke, 1965).

LEMMA 2. *A* MOU *random variable V is distributed as*

$$\prod_{i=t+1}^{2t+1} U_i^{1/i} ,$$

where the U_i's are i.i.d. uniform $[0,1]$.

PROOF. Note that the maximum of $2t+1$ i.i.d. uniform $[0,1]$ random variables is distributed as $U_{2t+1}^{1/(2t+1)}$. Apply this rule recursively to obtain the given property. $\square$

Finally, we recall a theorem regarding trees of random variables from the theory of branching random walks (Hammersley (1974), Kingman (1973), Biggins (1976, 1977)): note that

$$\log Z_k = \begin{cases} 0 & (k=0) \\ \max_p \sum_{i\in p} \log V_i & (k>0) \end{cases} .$$

The maximum is over all 2^k paths in the tree of length 2^k. Then, we have

$$\frac{\log Z_k}{k} \to \gamma \text{ a.s.}$$

as $k \to \infty$, where

$$\gamma = \inf\left\{x > -\sum_{j=t+1}^{2t+1} \frac{1}{j} : 2M(x) < 1\right\}$$

and

$$M(x) = \inf_{\lambda>0} \mathbf{E}\left\{V^\lambda e^{-\lambda x}\right\} .$$

In our case, by Lemma 2, we see that

$$M(x) = \inf_{\lambda>0} \prod_{j=t+1}^{2t+1} \frac{1}{1+\lambda/j} e^{-\lambda x} .$$

The minimum of the latter expression is obtained when λ is the solution of the equation

$$x = -\sum_{j=t+1}^{2t+1} \frac{1}{j+\lambda} .$$

This only has a positive solution when

$$0 > x > -\sum_{j=t+1}^{2t+1} \frac{1}{j} .$$

The function $M(x)$ is decreasing in x. Over the range of interest here, it decreases from 1 to 0. Therefore, γ is well-defined.

Another way of viewing this is from the exponential inequality angle. Indeed, for $\lambda > 0$,

$$\begin{aligned}
\mathbf{P}\{\log Z_k > kx\} &\leq 2^k \mathbf{P}\left\{\log \prod_{i=1}^{k} V_i > kx\right\} \\
&= 2^k \mathbf{P}\left\{\sum_{i=1}^{k} \log V_i > kx\right\} \\
&\leq 2^k \mathbf{E}\left\{\prod_{i=1}^{k} e^{\lambda \log V_i} e^{-\lambda kx}\right\} \\
&= \left(2\mathbf{E}\left\{V_1^\lambda e^{-\lambda x}\right\}\right)^k \\
&= (2M(x))^k
\end{aligned}$$

if we choose λ as outlined above. The upper bound tends to zero as $k \to \infty$ when $2M(x) < 1$.

5. Proof of Theorem 1.

From Lemma 1 and the last result of the previous section, we see that

$$\mathbf{P}\left\{Z_k \geq \frac{1}{n}\right\} = \mathbf{P}\left\{\frac{\log Z_k}{k} \geq \frac{-\log n}{k}\right\} \to 0$$

when we choose $k \sim c \log n$ and $-1/c > \gamma$. Also, as each V_i (notation of Lemma 1) is stochastically smaller than $(1+U)/2$, where U is uniform $[0,1]$, we see that $\mathbf{E}V_i^\lambda \leq \mathbf{E}(1+U)^\lambda/2^\lambda \leq 2/(1+\lambda)$. Thus, for any $\lambda > 0$, by Jensen's inequality, and $0 < l < k$,

$$2^l \mathbf{P}\left\{\prod_{i=1}^{l} V_i \geq \frac{l}{k-l}\right\} \leq 2^l \frac{\mathbf{E}^l V_i^\lambda (k-l)^\lambda}{l^\lambda} \leq \left(\frac{4}{(1+\lambda)}\right)^l \left(\frac{k-l}{l}\right)^\lambda \to 0$$

if we choose $l = \lfloor \sqrt{k} \rfloor$, $k \sim c \log n$ (as above), and $\lambda = 4$. By Lemma 1, we thus conclude that $\mathbf{P}\{H'_n \geq k + 2\sqrt{k}\} \to 0$. Furthermore,

$$\begin{aligned}\mathbf{P}\{H'_n \geq k\} &\geq \mathbf{P}\left\{\frac{\log Z_k}{k} \geq \frac{\log(2t+1+k(t+1)) - \log n}{k}\right\} \\ &\to 1\end{aligned}$$

when we choose $k \sim c \log n$ and $-1/c < \gamma$. Putting both statements together, we obtain

$$\frac{H'_n}{\log n} \to -\frac{1}{\gamma} \text{ i.p. } .$$

Set $-1/\gamma \equiv c(t)$. This leads precisely to the definition of λ and to equation (1) (which is equivalent to $2M(-1/c) = 1$) in Theorem 1. $\square$

As a by-product of the inequality

$$\mathbf{P}\{Z_k \geq 1/n\} \leq \left(2M\left(\frac{-\log n}{k}\right)\right)^k ,$$

we see that $\mathbf{E}H'^q_n / \log^q n \to (c(t))^q$ for all $q > 0$. Using arguments as in Pittel (1984), we may also deduce strong convergence of $H'_n / \log n$ to $c(t)$.

4. References.

D. Aldous, B. Flannery, and J. L. Palacios, "Two applications of urn processes: the fringe analysis of search trees and the simulation of quasi-stationary distributions of Markov chains," *Probability in the Engineering and Information Sciences*, vol. 2, pp. 293–307, 1988.

C. J. Bell, "An investigation into the principles of the classification and analysis of data of an automatic digital computer," Doctoral Dissertation, Leeds University, 1965.

J. D. Biggins, "The first and last-birth problems for a multitype age-dependent branching process," *Advances in Applied Probability*, vol. 8, pp. 446–459, 1976.

J. D. Biggins, "Chernoff's theorem in the branching random walk," *Journal of Applied Probability*, vol. 14, pp. 630–636, 1977.

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.

G. H. Gonnet and R. Baeza-Yates, *Handbook of Algorithms and Data Structures*, Addison-Wesley, Workingham, England, 1991.

J. M. Hammersley, "Postulates for subadditive processes," *Annals of Probability*, vol. 2, pp. 652–680, 1974.

J. F. C. Kingman, "Subadditive ergodic theory," *Annals of Probability*, vol. 1, pp. 883–909, 1973.

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

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.

P. V. Poblete and J. I. Munro, "The analysis of a fringe heuristic for binary search trees," *Journal of Algorithms*, vol. 6, pp. 336–350, 1985.

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. S. Vitter and P. Flajolet, "Average-case analysis of algorithms and data structures," in: *Handbook of Theoretical Computer Science, Volume A: Algorithms and Complexity*, ed. J. van Leeuwen, pp. 431–524, MIT Press, Amsterdam, 1990.

A. Walker and D. Wood, "Locally balanced binary trees," *Computer Journal*, vol. 19, pp. 322–325, 1976.