On the Height of Random m-ary Search Trees

Luc Devroye
School of Computer Science, McGill University, Montreal, Canada H3A 2A7

ABSTRACT

A random m-ary seach tree is constructed from a random permutation of $1, \ldots, n$. A law of large numbers is obtained for the height H_n of these trees by applying the theory of branching random walks. In particular, it is shown that $H_n/\log n \to \gamma$ in probability as $n \to \infty$, where $\gamma = \gamma(m)$ is a constant depending upon m only. Interestingly, as $m \to \infty$, $\gamma(m)$ is asymptotic to $1/\log m$, the coefficient of $\log n$ in the asymptotic expression for the height of the complete m-ary search tree. This proves that for large m, random m-ary search trees behave virtually like complete m-ary trees.

Key Words: analysis of algorithms, asymptotic behavior, binary search tree, branching process, height of tree

1. INTRODUCTION

In Devroye [5, 6], we applied the rich theory of branching processes in the analysis of the height of random binary search trees that are constructed by repeated insertions of elements of a random permutation. The purpose of this note is to extend those results to the case of m-ary search trees.

For the random binary search tree, under the standard random permutation model for the data, a variety of techniques have been used in the analysis of various quantities. For example, the result that the expected depth of the nth node in an n-node tree is asymptotic to $2 \log (n)$ can be found in most textbooks on data structures and algorithms [1, 2, 14]. The limit law of the depth of this node, and various other properties were obtained in Lynch [16], Knuth [5], Sedgewick [24], Pittel [21], Kemp [12], and Devroye [7]. The height H_n of a (random) tree with n elements is the number of nonempty levels minus one.

Random Structures and Algorithms, Vol. 1, No. 2 (1990)
 CCC 1042-9832/90/020191-13$04.00

Flajolet and Odlyzko [8] studied H_n under other models of randomization., For the binary search tree, Robson [23] and Pittel [21] provided the first analyses of H_n. The following result will be generalized in this note:

Theorem 1. [1, 2]. *Let H_n be the height of a random binary search tree. Then*

$$\frac{H_n}{c \log(n)} \to 1 \quad \textit{in probability}$$

and

$$E(H_n) \sim c \log(n)$$

where $c = 4.31107\ldots$ is the solution of $c \log\left(\frac{2e}{c}\right) = 1$; $c \geq 2$.

Note that "in probability" is to be taken in the standard probability theoretical sense. In combinatorial jargon, this corresponds to "almost all." Thus, the first statement of Theorem 1 can be rephrased as follows: for all $\epsilon > 0$, almost all permutations of $1, \ldots, n$ yield trees of height H_n between $(c - \epsilon)\log(n)$ and $(c + \epsilon)\log(n)$ as $n \to \infty$.

The m-ary search tree is often attributed to Muntz and Uzgalis [20]. It was introduced because of external memory problems; nodes should be thought of as pages that reside in external memory (m is thus in the hundreds). Each node holds up to $m - 1$ elements, and is childless when it is not fully occupied. Within a node, the elements are usually kept in an array or a linked list in sorted order. The advantages are clear: the number of node accesses is way down when m is large, and with binary search within each node, the number of comparisons typically improves as well over its binary counterpart when we search for an element.

Several results are known about L_n (the distance between the root and the $n+1$-st element) and C_n (the number of comparisons made when that $n+1$-st element is inserted). Mahmoud and Pittel [18] have shown that L_n is in probability asymptotic to $c \log n$ with $c = 1/(h_m - 1)$, where $h_m \stackrel{\Delta}{=} \sum_{i=1}^{m} (1/i)$. (Note that we reserve the notation H_m for the height of a tree.) Under a different model of randomization, L_n was also analyzed by Halton [9]. The properties of C_n are related to those of L_n, but depend upon the kind of list search that is used in a node when that node is traversed. Mahmoud [17] showed that the average internal path length is asymptotic to $c\, n \log n$. But the jewel on the crown is the joint asymptotic distribution of (L_n, C_n), obtained by Mahmoud and Pittel [18]. The random permutation model can be made into an infinite incremental model, when we think of the inserted elements as a sequence of i.i.d. uniform $[0, 1]$ random variables. Under this model, Mahmoud and Pittel [18] proved that almost surely, $L_n/\log n \in [d - \epsilon, c + \epsilon]$ for some positive constants c, d, and any fixed $\epsilon > 0$. Since $H_n \geq L_{n-1}$, this implies that $\liminf H_n/\log n \geq c$ almost surely as well. In the present article, we identify this constant c, and show that $H_n/\log n \to c$ in probability:

TABLE I

m	$\gamma(m)$	$1/\log m$	$1/(h_m - 1)$
2	4.3110	1.442695	2.000000
3	2.4699	0.910239	1.200000
4	1.8387	0.721348	0.923077
5	1.5139	0.621335	0.779221
6	1.3133	0.558111	0.689655
7	1.1760	0.513898	0.627803
8	1.0753	0.480898	0.582121
9	0.9979	0.455120	0.546756
10	0.9362	0.434294	0.518412
20	0.6521	0.333808	0.384950
30	0.5473	0.294014	0.333891
40	0.4894	0.271085	0.305014
50	0.4515	0.255622	0.285779
60	0.4242	0.244239	0.271749
70	0.4034	0.235377	0.260903
80	0.3867	0.228205	0.252176
90	0.3731	0.222232	0.244944
100	0.3615	0.217147	0.238813

Theorem 2 (The main result). *Let $\gamma > 1/(h_m - 1)$ be defined as*

$$\gamma \stackrel{\Delta}{=} \inf\left\{ c > 1/(h_m - 1) : t + c \log(m!) - c \sum_{i=1}^{m-1} \log(t+i) < 0 \right\},$$

where $t > 0$ is the unique solution of the equation

$$\frac{1}{c} = \sum_{i=1}^{m-1} \frac{1}{t+i}.$$

Then $H_n/\log n \to \gamma$ in probability as $n \to \infty$. Furthermore, for all $p > 0$,

$$\frac{EH_n^p}{\log^p n} \sim \gamma^p \text{ as } n \to \infty.$$

Finally, $\lim_{m\to\infty} \gamma(m) \log m = 1$.

In Table I, some values for c are provided. For comparison, we also show $1/\log m$, the coefficient of $\log n$ in the asymptotic height of the best possible m-ary tree, and $1/(h_m - 1)$, the coefficient of $\log n$ in the asymptotic expression of EL_n, the depth of the $n+1$-st inserted element.

2. SPACINGS OF UNIFORM VARIABLES

We begin with an important distributional observation regarding the sizes of the m subtrees of a node. Consider a random m-ary tree with $n \geq m - 1$. Then the root node has subtrees of sizes $N_1, \ldots, N_m$. We have the following property:

Lemma 1. *The random variables $N_1, \ldots, N_m$ are identically distributed. Furthermore, N_1 is stochastically smaller than Un and stochastically larger than $U(n-m+2)-1$, where U is distributed as the minimum of $m-1$ i.i.d. uniform $[0,1]$ random variables. In fact, there exists an embedding in a rich enough probability space such that $N_1 \le Un$.*

Proof of Lemma 1. The first statement of the Lemma is rather obvious. In view of it, we can write N instead of N_1. For integer i,

$$P(N \ge i) = \frac{\binom{n-1}{m-1}}{\binom{n}{m-1}} = \frac{(n-i)\cdots(n-m-i+2)}{n(n-1)\cdots(n-m+2)}.$$

This is at most equal to $(\frac{n-i}{n})^{m-1} = P(Un \ge i)$ and at least equal to $(\frac{n-m-i+2}{n-m+2})^{m-1} = P(U(n-m+2) \ge i)$. To prove stochastic majorization, we consider a real number $x \in [0, n-m+1]$. Then

$$P(N \ge x) = P(N \ge \lceil x \rceil) \le P(Un \ge \lceil x \rceil) \le P(Un \ge x),$$

and

$$P(N \ge x) = P(N \ge \lceil x \rceil) \ge P(U(n-m+2) \ge \lceil x \rceil) \ge P(U(n-m+2) \ge x+1),$$

which was to be shown. The last claim of the lemma follows from a simple argument. Consider the interval $[0, n]$ and its n unit length subintervals. Consider m i.i.d. uniform random variables on this interval, and in particular the leftmost of these values, V. Next, remove all the points that fell into an already occupied bin (subinterval), and replace these by new independent points. Repeat this process until all bins have at most one point. Note that N_1 is equal to i if and only if the first i bins are empty and bin $i+1$ is not empty. But clearly, $N_1 \le V$, which is distributed as nU. ∎

We now return to a tree associated with the m-ary tree, following a construction from Devroye [5]. Indeed, Lemma 1 suggests that the sizes of the subtrees of the root, when divided by $n-m+1$, are about distributed like the m spacings induced on $[0,1]$ by a sample of $m-1$ i.i.d. uniform $[0,1]$ random variables. These spacings will be denoted by $(S_1, \ldots, S_m)$. A collection of random variables $X_1, \ldots, X_m$ is said to be negatively associated (NA) if for every pair of disjoint subsets A, B of $\{1, \ldots, n\}$,

$$Cov(f_1(X_i; i \in A), f_2(X_j; j \in B)) \le 0$$

whenever f_1, f_2 are increasing [11]. For such random variables, and any sequence of numbers $x_1, \ldots, x_m$

$$P\left(\bigcap_{i=1}^{m} [X_i \geq x_i]\right) \leq \prod_{i=1}^{m} P(X_i \geq x_i)\,,$$

$$P\left(\bigcap_{i=1}^{m} [X_i \leq x_i]\right) \leq \prod_{i=1}^{m} P(X_i \leq x_i)$$

[18, property P3]

Lemma 2. *Let $N_1, \ldots, N_m$ be the sizes of the subtrees of the root of a random m-ary tree, and let $S_1, \ldots, S_m$ be the spacings induced on $[0,1]$ by a sample of $m-1$ i.i.d. uniform $[0,1]$ random variables. Then, the following inequalities are valid for any positive numbers $x_1, \ldots, x_m$:*

$$P(N_1 \geq x_1, \ldots, N_m \geq x_m) \leq P(nS_1 \geq x_1, \ldots, nS_m \geq x_m)\,,$$

$$P(N_1 \geq x_1, \ldots, N_m \geq x_m) \geq P(nS_1 - (m-1) \geq x_1, \ldots, nS_m - (m-1) \geq x_m)\,.$$

Furthermore, $N_1, \ldots, N_m$ are negatively associated.

Proof of Lemma 2. We begin with the following fact (proved in Mahmoud and Pittel [18] and again in Mahmoud [17]): if $i_1, \ldots, i_k$ are nonnegative integers, then

$$P(N_1 = i_1, \ldots, N_m = i_m) = 1 \Big/ \binom{n}{m-1}\,,$$

when $\Sigma_{k=1}^{m} i_k = n - (m-1)$, and the probability is zero otherwise. This immediately implies that for any sequence of nonnegative integers $(i_1, \ldots, i_m)$,

$$\begin{aligned} P(N_1 \geq i_1, \ldots, N_m \geq i_m) &= P\left(N_1 \geq 0, N_2 \geq 0, \ldots, N_{m-1} \geq 0, N_m \geq \sum i_k\right) \\ &= P\left(N_m \geq \sum i_k\right) \\ &= \frac{\binom{n - \sum i_k}{m-1}}{\binom{n}{m-1}} \leq \left(1 - \frac{1}{n}\sum i_k\right)_+^{m-1}\,, \end{aligned}$$

where Σ denotes $\Sigma_{k=1}^{m}$, and $u_+ = \max(u, 0)$. For any sequence of nonnegative numbers $(x_1, \ldots, x_m)$, we have

$$\begin{aligned} P(N_1 \geq x_1, \ldots, N_m \geq x_m) &= P(N_1 \geq \lceil x_1 \rceil, \ldots, N_m \geq \lceil x_m \rceil) \\ &\leq \left(1 - \frac{1}{n}\sum \lceil x_k \rceil\right)_+^{m-1} \leq \left(1 - \frac{1}{n}\sum x_k\right)_+^{m-1} \\ &= P(nS_1 \geq x_1, \ldots, nS_m \geq x_m)\,. \end{aligned}$$

Here we used a well-known fact about the distribution of uniform spacings [22].

Following the arguments of Lemma 1, we obtain also

$$\begin{aligned} &P(N_1 \geq x_1, \ldots, N_m \geq x_m) \\ &\quad \geq \left(1 - \frac{1}{n-m+2} \sum \lceil x_k \rceil\right)_+^{m-1} \geq \left(1 - \frac{1}{n-m+2} \sum (1+x_k)\right)_+^{m-1} \\ &\quad = P((n-m+2)S_1 \geq 1 + x_1, \ldots, (n-m+2)S_m \geq 1 + x_m) \\ &\quad \geq P(nS_1 - (m-1) \geq x_1, \ldots, nS_m - (m-1) \geq x_m), \end{aligned}$$

which was to be proved. Next, consider $Y_1, \ldots, Y_m$, i.i.d. uniform random variables on $\{0, 1, \ldots, n\}$. Then given $\Sigma_{i=1}^m Y_i = n - m + 1$, we note that $(Y_1, \ldots, Y_m)$ is distributed as $(N_1, \ldots, N_m)$ because uniformity is not lost by conditioning on subsets of the space. Note also that for every increasing function f and subset A of $\{1, \ldots, m\}$ of size $k < m$,

$$E\left(f(Y_i; i \in A) \mid \sum_{i \in A} Y_i = s\right)$$

is $\uparrow$ in s. Thus, by Theorem 2.6 of Joag-Dev and Proschan [11], the conditional distribution of $Y_1, \ldots, Y_m$ given $\Sigma_{i=1}^m Y_i = s$ is NA for any value of s. ∎

3. AN ASSOCIATED TREE OF RANDOM VARIABLES

We associate with the random m-ary tree a complete m-ary tree T_k with k full levels of edges. We use the symbol p rather freely to denote a path from root to leaf (clearly, there are m^k such paths). With each collection of m edges emanating from an internal node, we associate an independent copy of $(S_1, \ldots, S_m)$, the spacings induced on $[0, 1]$ by $m - 1$ i.i.d. uniform $[0, 1]$ random variables. The number of independent replicas of this is equal to the number of internal nodes. To simplify the notation, we say that edge i has random variable X_i associated with it. Then we define

$$V_k = \begin{cases} 1, & k = 0, \\ \max\limits_p \prod\limits_{i \in p} X_i, & k > 0. \end{cases}$$

For $n \geq 1$ and $k \geq 0$ integers, in the binary case, it is true [5] that $P(H_n \geq k) \leq P(V_k \geq 1/n)$ and that $P(H_n \geq k) \geq P(V_k \geq (k+1)/n)$. This connection allows one to focus on the associated tree only. Unfortunately, these inequalities (or similar ones with $1 + k$ replaced by $1 + k(m-1)$) are much more difficult to establish for $m > 2$. Instead, we will bypass them, and use a more direct proof, which nevertheless exploits the connection with the associated tree.

There is a second type of tree, which is defined as the previous one, except that each instance of $(S_1, \ldots, S_m)$ is replaced by m i.i.d. copies of S_1, denoted by $(S_1(1), \ldots, S_1(m))$. The random variable V_k is formally replaced by Z_k. The connection between Z_k and H_n is dealt with in Lemma 4.

4. AN UPPER BOUND

Lemma 3. *Let $S_1(1), \ldots, S_1(k)$ be i.i.d. random variables, each distributed as S_1 (defined in Lemma 2), where $k \geq 1$ is an integer. We have*

$$P(H_n \geq k) \leq m^k P\left(n \prod_{i=1}^{k} S_1(i) \geq 1\right).$$

When $k = \lceil c \log n \rceil$ for some constant $c > \gamma > 1/(h_m - 1)$, this is further bounded by $n^{-\eta}$, where $0 < \eta \uparrow \infty$ as $c \uparrow \infty$. In particular, we can conclude that

$$\limsup_{n \to \infty} P(H_n / \log n \geq \gamma + \epsilon) = 0$$

for all $\epsilon > 0$. Furthermore, for all $p > 0$,

$$\limsup_{n \to \infty} \frac{EH_n^p}{\log^p n} \leq \gamma^p$$

Proof of Lemma 3. Let $N_1, \ldots, N_m$ be the sizes of the subtrees of the root, where the N_is are as defined in Lemma 2. Thus,

$$P(H_n \geq k) \leq \sum_{i=1}^{m} P(H_{N_i} \geq k-1) = mP(H_{N_i} \geq k-1).$$

Since there exists an embedding in which H_n is an increasing function of n (use sequential insertion for example), and since $N_1 \leq nS_1(1)$ on a rich enough probability space (apply Lemma 1), we see that the upper bound can be replaced by $mP(H_{\lfloor nS_1(1) \rfloor} \geq k-1)$. We can now argue by induction on k, after noting that $P(H_n \geq 0) = 1$ if and only if $n \geq 1$, if we agree that the height of an empty tree is minus one. Thus,

$$P(H_n \geq k) \leq m^k P\left(H_{\lfloor n \prod_{i=1}^{k} S_1(i) \rfloor} \geq 0\right) = m^k P\left(n \prod_{i=1}^{k} S_1(i) \geq 1\right).$$

To bound the last expression, we use Chebyshev's (or Jensen's) inequality, with a yet-to-be-determined parameter $t > 0$:

$$P(H_n \geq k) \leq m^k P\left(n \prod_{i=1}^{k} S_1(i) \geq 1\right) \leq m^k n^t \prod_{i=1}^{k} E(S_1^t(i)) = m^k n^t E^k(S_1^t)$$

$$= m^k n^t \left(\int_0^1 P(S_1 > x^{1/t})\, dx\right)^k = m^k n^t \left(\int_0^1 t(1-y)^{m-1} y^{t-1}\, dy\right)^k$$

$$= m^k n^t \left(\frac{\Gamma(m)\Gamma(t+1)}{\Gamma(m+t)}\right)^k = n^t \left(\frac{\Gamma(m+1)}{\prod_{i=1}^{m-1}(t+i)}\right)^k. \tag{1}$$

Let t^* be the unique positive value of t that minimizes this expression. It is easy to check that it can be obtained by setting the derivative of the logarithm of the bound to zero. This yields the equation

$$\log n - k \sum_{i=1}^{m-1} \frac{1}{t+i} = 0 .$$

In our proof, we are interested in taking $k = \lceil c \log n \rceil$ for some $c > 1/(h_m - 1)$. Thus, instead of working with t^*, we **define** t as the unique solution of the equation

$$\sum_{i=1}^{m-1} \frac{1}{t+i} = \frac{1}{c} .$$

The properties of the function t are derived in Section 7. For $c = 1/(h_m - 1)$, we have $t = 1$. Furthermore, $t \uparrow \infty$ as $c \uparrow \infty$, and for $c > 1/(h_m - 1)$ (and thus $t > 1$), (1) does not exceed

$$\exp\left(\log n\left(t + c \log (m!) - c \sum_{i=1}^{m-1} \log (t+i)\right)\right) . \tag{2}$$

Recall the definition of γ in Theorem 2. Note next that for $c > \gamma$, the upper bound (2) becomes

$$e^{-\eta \log n} = n^{-\eta} ,$$

where $\eta > 0$ depends upon c only. In the next section, we will prove that as $c \uparrow \infty$, $\eta \uparrow \infty$. For fixed $p > 0$, we find $d > c$ so large that $\eta(d) > p$, and we choose $c = \gamma + \epsilon$ for arbitrary small $\epsilon > 0$. Then,

$$\begin{aligned} EH_n^p &\le (c \log n)^p + (d \log n)^p P(H_n \ge c \log n) + n^p P(H_n \ge d \log n) \\ &\le (c \log n)^p + o(1) . \end{aligned}$$

This concludes the proof of Lemma 3. ∎

5. LOWER BOUNDS

Lemma 4. *Let $n \ge 1$ and $k \ge 0$ be integers. Then*

$$P\left(Z_k \ge \frac{1 + k(m-1)}{n}\right) \le P(H_n \ge k) .$$

Proof of Lemma 4. We proceed by induction on k. For $k = 0$, the statement is obviously true for all $n \ge 1$. We assume that the inequality holds up to $k - 1$ and for all n. First we show that the bound remains valid for k and all $n \ge 1$. The bound is obviously valid for all $n \le m - 1$, so we need only consider $n > m - 1$.

Let us introduce a random collection of uniform spacings, $(S_1, \ldots, S_m)$, independent of all the other random variables that we shall mention. Let $N_1, \ldots, N_m$ be the sizes of the subtrees of the root, i.e., $\Sigma N_i = n - (m-1)$ if $n \ge m - 1$, and each $N_i = 0$ if $n < m - 1$. Let $Z_{k-1}(1), \ldots, Z_{k-1}(m)$ be m i.i.d. copies of Z_{k-1}, defined just before Lemma 3. These random variables will be associated with the subtrees of the root. If $H_{N_1}, \ldots, H_{N_m}$ are the heights of these subtrees, then we have,

$$P(H_n \geq k) = P\left(\bigcup_{i=1}^{m} [H_{N_i} \geq k-1]\right)$$

provided that we arbitrarily define the height of the empty tree to be minus one. Thus, by our induction hypothesis, and various independencies,

$$\begin{aligned}
&P\left(\bigcup_{i=1}^{m} [H_{N_i} \geq k-1] \mid N_1, \ldots, N_m\right) \\
&\quad = 1 - P\left(\bigcap_{i=1}^{m} [H_{N_i} < k-1] \mid N_1, \ldots, N_m\right) \\
&\quad = 1 - \prod_{i=1}^{m} P(H_{N_i} < k-1 \mid N_i) \\
&\quad \geq 1 - \sum_{i=1}^{m} P(N_i Z_{k-1}(i) < 1 + (k-1)(m-1) \mid N_i) \\
&\quad = 1 - P(\max_i N_i Z_{k-1}(i) < 1 + (k-1)(m-1) \mid N_1, \ldots, N_m) .
\end{aligned}$$

Thus, unconditioning, and using Lemma 2,

$$P(H_n \geq k) \geq P(\max_i N_i Z_{k-1}(i) \geq 1 + (k-1)(m-1)) .$$

Now, we condition on the collection of $Z_{k-1}(i)$s, $1 \leq i \leq m$, and denote the conditional probability by P^*. Let $N_1^*, \ldots, N_m^*$ be i.i.d. random variables distributed as N_1. Let $S_1(1), \ldots, S_1(m)$ be i.i.d. random variables distributed as S_1. By Lemmas 2 and 1, we have

$$\begin{aligned}
&P^*(\max_i N_i Z_{k-1}(i) \geq 1 + (k-1)(m-1)) \\
&\quad = 1 - P^*\left(\bigcap_i [N_i Z_{k-1}(i) \leq 1 + (k-1)(m-1)]\right) \\
&\quad \geq 1 - P^*\left(\bigcap_i [N_i^* Z_{k-1}(i) \leq 1 + (k-1)(m-1)]\right) \\
&\quad \geq 1 - P^*\left(\bigcap_i [S_1(i)(n-m+2) Z_{k-1}(i) \leq Z_{k-1}(i) + 1 + (k-1)(m-1)]\right) \\
&\quad \geq P^*\left(\bigcup_i [nS_1(i) Z_{k-1}(i) \geq (m-1) + 1 + (k-1)(m-1)]\right) \\
&\quad = P^*(\max_i nS_1(i) Z_{k-1}(i) \geq 1 + k(m-1)) ,
\end{aligned}$$

which, if we un-condition again, and recall the definition of Z_k, yields

$$P(H_n \geq k) \geq P(nZ_k \geq 1 + k(m-1)) . \qquad \blacksquare$$

6. BRANCHING RANDOM WALKS

We recall the following result from the theory of branching random walks, as applied to random variables

$$\bar{X}_k = \max_p \sum_{i \in p} X_i ,$$

where the X_is are i.i.d. random variables associated with all the edges of a complete m-ary tree having precisely m^k leaves (i.e., k full levels of edges); by p we denote a path from the root to one of the leaves of this tree. It is assumed that X_1 is such that

$$\rho(t) = mE(e^{tX_1}) < \infty$$

for some $t > 0$.

Lemma 5 [20–23]. *$\bar{X}_k/k \to \Theta$ almost surely as $k \to \infty$, where*

$$\Theta \stackrel{\Delta}{=} \sup \{\theta : \mu(\theta) > 1\} ,$$

and

$$\mu(\theta) \stackrel{\Delta}{=} \inf \{\rho(t)e^{-t\theta} : t > 0\} .$$

The proof of Theorem 2 follows directly from Lemma 3 and Lemma 6 below.

Lemma 6. *Let γ be as in Theorem 2. Then, for any $\epsilon > 0$,*

$$\lim_{n \to \infty} P(H_n \geq (\gamma - \epsilon) \log n) = 1 .$$

Also, for all $p > 0$, $\lim\inf_{n \to \infty} EH_n^p / \log^p n \geq \gamma^p$.

Proof of Lemma 6. The second half of Lemma 6 follows trivially from the first half. Also, we note that $\log Z_k$ is distributed as $\bar{X}_k$ used in Lemma 5, provided that we let X_1 be a random variable that is distributed as $\log(S_1)$. The quantities used in Lemma 5 can now be calculated for this specific example. We have

$$\rho(t) = mE(e^{tX_1}) = mE(S_1^t) = \frac{\Gamma(m+1)\Gamma(t+1)}{\Gamma(m+t)} .$$

Furthermore,

$$\mu(\theta) = \inf \left\{ e^{-t\theta} \frac{\Gamma(m+1)\Gamma(t+1)}{\Gamma(m+t)} : t > 0 \right\} ,$$

and $\Theta = \sup \{\theta : \mu(\theta) > 1\}$. The minimum in the definition of $\mu(\theta)$ is reached for that value $t > 0$ for which

$$-\theta = \sum_{i=1}^{m-1} \frac{1}{t+i}.$$

A quick inspection shows that $-\theta$ is related to t exactly as $1/c$ is related to t in its definition in Theorem 2. Hence, $\gamma = -1/\Theta$.

Lemma 5 implies that $\frac{1}{k}\log Z_k \to \Theta = -\frac{1}{\gamma}$ almost surely as $k \to \infty$. Thus, for $\epsilon > 0$, with $k \stackrel{\Delta}{=} \lceil (\gamma - \epsilon)\log n \rceil$, and applying Lemmas 4 and 5, we obtain

$$\begin{aligned} P(H_n \geq (\gamma - \epsilon)\log n) &= P(H_n \geq k) \\ &\geq P\left(Z_k \geq \frac{1 + k(m-1)}{n}\right) \\ &= P\left(\frac{1}{k}\log(Z_k) \geq \frac{1}{k}\log\left(\frac{1+k(m-1)}{n}\right)\right) \to 1 \text{ as } n \to \infty \end{aligned}$$

provided that the limit of $\frac{1}{k}\log\left(\frac{1+k(m-1)}{n}\right)$ is smaller than $-1/\gamma$. We note that this limit is $-1/(\gamma - \epsilon)$, and thus we are done. ∎

7. PROPERTIES OF γ

In this section, we are concerned with the constant $\gamma = \gamma(m)$ of Theorem 2, and with some related facts needed in the proof of Lemma 3. It helps to introduce some notation. Recall the definition of the function $t = t(c)$ as the unique solution of the equation

$$\frac{1}{c} = \sum_{i=1}^{m-1} \frac{1}{t+i}.$$

Throughout, we assume that $c \geq 1/(h_m - 1)$. At this threshold value, $t = 1$, while $t \uparrow \infty$ as $c \uparrow \infty$.

We define $\Psi(u) \stackrel{\Delta}{=} \log m! - \sum_{i=1}^{m-1} \log(u+i)$ and $\delta(c) = t + c\Psi(t(c))$. Observe that $\gamma \stackrel{\Delta}{=} \inf(c > 1/(h_m - 1): \delta(c) < 0)$, and that η in the proof of Lemma 3 is nothing but $-\delta$. It is noteworthy that $1 + c\Psi'(t(c)) = 0$ by definition of t. The function Ψ decreases monotonically from 0 (at $u = 1$) to $-\infty$ (as $u \to \infty$). We verify the following statements about the function $\delta(c)$:

A. $\delta(1/(h_m - 1)) = 1$.

B. $\delta'(c) = \Psi(t(c))$ (use the fact that $1 + c\Psi'(t(c)) = 0$). Thus, δ' decreases monotonically from 0 (at $c = 1/(h_m - 1)$) to $-\infty$ (as $c \to \infty$). This implies that δ is a concave function, and in particular, that γ is indeed well-defined.

C. For all $c \geq 1/(h_m - 1)$, $\delta''(c) \leq -\frac{m}{(m-1)c^3}$. This can be seen as follows. First note that $1/(c^2 t'(c)) = \sum_{i=1}^{m-1}(t+i)^{-2} \leq \int_t^{m+t-1} x^{-2}\,dx = (m-1)/(t(m+t-1)) \leq (m-1)/m$. Thus, $\delta''(c) = \Psi'(t(c))t'(c) = -t'(c)/c \leq -m/(m-1)c^3)$.

D. By Taylor's series expansion about $c = 1/(h_m - 1)$, we obtain the following inequality from observations A, B, and C:

$$\delta(c) \leq 1 - \frac{m(c - 1/(h_m - 1))^2}{2(m-1)c^3}.$$

E. If $\delta(c) \leq 0$, then we can conclude that $\gamma \leq c$. With $\epsilon > 0$ and $c = (1+\epsilon)/(h_m - 1)$, we see that $\delta(c) \leq 0$ when

$$2(1+\epsilon)^3 \leq \frac{m}{m-1}\epsilon^2(h_m - 1).$$

For any fixed ϵ, this is satisfied for all m large enough. Furthermore, it also holds for all m large enough if $\epsilon = \sqrt{(2+\xi)(m-1)/(m(h_m-1))}$ for $\xi > 0$.

F. From E, we conclude that as $m \to \infty$,

$$\gamma \leq \frac{1}{h_m - 1} + \frac{\sqrt{2} + o(1)}{(h_m - 1)^{3/2}}.$$

In particular, we can conclude that $\gamma(m) \sim 1/\log m$ as $m \to \infty$, as claimed in Theorem 2. □

ACKNOWLEDGMENTS

The author's research was sponsored by NSERC Grant No. A3456 and FCAC Grant No. EQ-1678.

REFERENCES

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

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

[3] J. D. Biggins, The first and last-birth problems for a multitype age-dependent branching process, *Adv. Appl. Probab.*, **8**, 446–459 (1976).

[4] J. D. Biggins, Chernoff's theorem in the branching random walk, *J. Appl. Probab.*, **14**, 630–636 (1977).

[5] L. Devroye, A note on the height of binary search trees, *J. ACM*, **33**, 489–498 (1986).

[6] L. Devroye, Branching processes in the analysis of the heights of trees, *Acta Inf.*, **24**, 277–298 (1987).

[7] L. Devroye, Applications of the theory of records in the study of random trees, *Acta. Inf.*, **26**, 123–130 (1988).

[8] P. Flajolet and A. Odlyzko, The average height of binary trees and other simple trees, *J. Comput. Syst. Sci.*, **25**, 171–213 (1982).

[9] J. H. Halton, The properties of random trees, Tech. Rep. 86-0214, Dept. Computer Science, University of North Carolina at Chapel Hill, 1986.

[10] J. M. Hammersley, Postulates for subadditive processes, *Ann. Probab.*, **2**, pp. 652–680 (1974).

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

[12] R. Kemp, *Fundamentals of the Average Case Analysis of Particular Algorithms*, B. G. Teubner, Stuttgart, Federal Republic of Germany, 1984.

[13] J. F. C. Kingman, Subadditive ergodic theory, *Ann. Probab.*, **1**, 883–909 (1973).

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

[15] G. Louchard, Exact and asymptotic distributions in digital and binary search trees, *Theor. Inform. Appl.*, **21**, 479–496 (1987).

[16] W. C. Lynch, More combinatorial problems on certain trees, *Comput. J.*, **7**, pp. 299–302 (1965).

[17] H. M. Mahmoud, On the average internal path length of m-ary search trees, *Acta Inf.*, **23**, 111–117 (1986).

[18] H. Mahmoud and B. Pittel, On the most probable shape of a search tree grown from a random permutation, *SIAM J. Algebraic Discrete Methods*, **5**, 69–81 (1984).

[19] H. M. Mahmoud and B. Pittel, On the joint distribution of the insertion path length and the number of comparisons in search trees, *Discrete Appl. Math.*, **20**, 243–251 (1988).

[20] R. Muntz and R. Uzgalis, in *Proc. Princeton Conf. Information Sciences and Systems*, **4**, 345–349 (1970).

[21] B. Pittel, On growing random binary trees, *J. Math. Anal. App.*, **103**, 461–480 (1984).

[22] R. Pyke, Spacings, *J. R. Stat. Soc. Ser. B*, **7**, 395–445 (1965).

[23] J. M. Robson, The height of binary search trees, *Aust. Comput. J.*, **11**, 151–153 (1979).

[24] R. Sedgewick, Mathematical analysis of combinatorial algorithms, in *Probability Theory and Computer Science* (G. Louchard and G. Latouche, Eds.), Academic, London, England 1983, pp. 123–205.

Received January 26, 1989
Revised August 6, 1989