The Height of a Random Binary Search Tree

BRUCE REED

McGill University, Montreal Quebec, Canada and CNRS, Paris, France

Abstract. Let H_n be the height of a random binary search tree on n nodes. We show that there exist constants $\alpha = 4.311\cdots$ and $\beta = 1.953\cdots$ such that $\mathbf{E}(H_n) = \alpha ln\, n - \beta ln\, ln\, n + O(1)$, We also show that $\mathbf{Var}(H_n) = O(1)$.

Categories and Subject Descriptors: E.1 [**Data Structures**]: *trees*; G.2 [**Discrete Mathematics**]; G.3 [**Probability and Statistics**]

General Terms: Algorithms, Theory

Additional Key Words and Phrases: Binary search tree, height, probabilistic analysis, random tree, asymptotics, second moment method

1. *The Results*

A binary search tree is a binary tree to each node of which we have associated a key; these keys are drawn from some totally ordered set and the key at v cannot be larger than the key at its right child nor smaller than the key at its left child. Given a binary search tree T and a new key k, we insert k into T by traversing the tree starting at the root and inserting k into the first empty position at which we arrive. We traverse the tree by moving to the left child of the current node if k is smaller than the current key and moving to the right child otherwise. Given some permutation of a set of keys, we construct a binary search tree from this permutation by inserting them in the given order into an initially empty tree.

The height H_n of a random binary search tree T_n on n nodes, constructed in this manner starting from a random equiprobable permutation of $1, \ldots, n$, is known to be close to $\alpha ln\, n$ where $\alpha = 4.31107\cdots$ is the unique solution on $[2, \infty)$ of the equation $\alpha ln\,((2e)/\alpha) = 1$ (here and elsewhere, ln is the natural logarithm and log is the base 2 logarithm). First, Pittel [1984] showed that $H_n/ln\, n \to \gamma$ almost surely

---

A preliminary version of this article without proofs can be found in *Proceedings of the Symposium on Theory of Computing (STOC 2000)*, ACM, New York, 2000, pp. 479–483.

This research was carried out with the support of a NATO collaboration grant and the NSERC operating grants of Luc Devroye and David Avis.

Author's address: School of Computer Science, McGill University, 3480 University, Montreal, Que., Canada, E-mail: breed@cs.mcgill.ca.

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 $\mathbf{E}(H_n) \sim \alpha ln\ n$. 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 and Reed [1995] proved that $\mathbf{Var}(H_n) = O((ln\ ln\ n)^2)$, but this does not quite confirm Robson's findings. It is the purpose of this note to prove that for $\beta = \frac{3}{2ln\,(\alpha/2)}$, we have:

THEOREM 1. $\mathbf{E}(H_n) = \alpha ln\ n - \beta ln\ ln\ n + O(1)$ *and* $\mathbf{Var}(H_n) = O(1)$.

Remark 1. By our definition of α, we obtain $\beta = 3\alpha/(2\alpha - 2)$; however, as we shall see, the definition given above sheds more light on the reason β takes this value.

For more information on random binary search trees, one may consult Pyke [1965], Knuth [1973a, 1973b], Aho et al. [1975, 1983], Mahmoud and Pittel [1994], Devroye [1987, 1990], Mahmoud [1992], and Pittel [1994].

Remark 2. After I announced these results, Drmota developed an alternative proof of the fact that $\mathbf{Var}(H_n) = O(1)$ using completely different techniques. As our two proofs illuminate different aspects of the problem, we have decided to have the results published in the same journal [Drmota 2003].

2. *A Model*

If we construct a binary search tree from a permutation of $1, \ldots, n$ and i is the first key in the permutation then: i appears at the root of the tree, the tree rooted at the left child of i contains the keys $1, \ldots, i-1$ and its shape depends only on the order in which these keys appear in the permutation, and the tree rooted at the right child of i contains the keys $i+1, \ldots, n$ and its shape depends only on the order in which these keys appear in the permutation.

From this observation, one deduces that H_n is also the number of levels of recursion required when Vanilla Quicksort (i.e., the version of Quicksort in which the first element in the permutation is chosen as the pivot) is applied to a random permutation of $1, \ldots, n$.

Our observation also allows us to construct T_n from the top down. To ease our exposition, we think of T_n as a labeling of a subtree of T_∞, the complete infinite binary tree.

We expose the key associated with each node t of T_n. To underscore the relationship with Quicksort, we refer to the key at t as the *pivot* at t. Suppose then that we have exposed the pivots for some of the nodes forming a subtree of T_∞, rooted at the root of T_∞. Suppose further that for some node t of T_∞, all of the ancestors of t are in T_n and we have chosen their pivots. Then, these choices determine the set of keys K_t which will appear at the (possibly empty) subtree of T_n rooted at t, but will have no effect on the order in which we expect the keys in K_t to appear. Indeed, each permutation of K_t is equally likely. Thus, each of the keys in K_t will be equally likely to be the pivot. We let n_t be the number of keys in this set and specify the pivot at t by choosing a uniform element i_t of $1, \ldots, n_t$ and using the i_tth largest element of K_t as the pivot at t. We actually chose a uniform element u_t of $[0\,..\,1]$ and set $i_t = \lceil n_t u_t \rceil$. Thus, the subtree rooted at the left child of t has

$\lfloor n_t u_t \rfloor$ nodes and the subtree rooted at the right child of t has $\lfloor n_t(1-u_t) \rfloor$ nodes (note that $n_t u_t$ is an integer with probability 0, so we ignore this possibility). Now we can choose the u_t simultaneously and independently, as follows.

Each node x of T_∞ has a right son $r(x)$ and a left son $l(x)$. We consider a random labeled 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. If the labels on the edges of the path from the root to a node y of R_∞ are $L_1, \ldots, L_i$, then we define

$$h_n(y) = \lfloor \cdots \lfloor \lfloor nL_1 \rfloor L_2 \rfloor \cdots L_i \rfloor.$$

Fact 1. As discussed above, 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 $h_n(y) \geq 1$. More strongly, $h_n(y)$ is n_y. (See, e.g., Devroye [1986].)

We let T_n' be the subtree of T obtained by taking a copy R of R_∞ and letting T_n' consist of those nodes of R with $f(y) \geq 1/n$. The following easy fact, immediate via induction from the triviality $a - 1 \leq \lfloor a \rfloor \leq a$, shows that T_n' contains T_n:

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 suggest that H_n should be close to the height H_n' of T_n'. We will first discuss H_n' and then strengthen our results to show that they also apply to H_n. Thus, to begin, we will prove:

THEOREM 2. $\mathbf{E}(H_n') = \alpha \ln n - \beta \ln \ln n + O(1)$ *and* $\mathbf{Var}(H_n') = O(1)$.

In doing so, we use $X_{n,h}$ to denote the (random) set of nodes of T_n' at depth h. We use n_t' to denote $nf(t)$ which by Facts 1 and 2 is close to n_t. Finally, we give a name to our estimate of $\mathbf{E}(H_n)$.

Definition 1. We let $h' = h_n'$ be $\alpha \ln n - \beta \ln \ln n$.

3. *The Crux of the Matter*

We are interested in computing for various n and h, $\mathbf{P}(|X_{n,h}| > 0)$, as the probability that T_n' has height h is simply: $\mathbf{P}(|X_{n,h}| > 0) - \mathbf{P}(|X_{n,h+1}| > 0)$. Since $\mathbf{E}(|X_{n,h}|)$ is much easier to compute, we consider it first. Note that since there are 2^h nodes of T_∞ at depth h, we have $\mathbf{E}(|X_{n,h}|) = 2^h \mathbf{P}(f(y) \geq 1/n)$ where y is a node of R_∞ at depth h.

Not surprisingly, the function f is well understood (for cognoscenti: we use that for each x in R_∞, $-\ln U(x)$ is an exponential random variable with mean 1. Thus, if y is at depth h in T_∞, then $-\ln f(y)$ is distributed as the sum of h i.i.d. exponential random variables with mean one, that is, it is gamma distributed with parameter h). Classical results on this function allow us to compute the value of $\mathbf{E}(|X_{nh}|)$.

Definition 2. We let $h^* = h_n^*$ be the smallest h for which $\mathbf{E}(|X_{n,h}|) \leq 1$, that is, for which $\mathbf{P}(f(y) \geq \frac{1}{n}) \leq 2^{-h}$.

We recall that α is the solution in $[2, \infty)$ of $\alpha ln(2e/\alpha) = 1$.

LEMMA 1. *$h^* = \alpha ln\ n - 1/2 ln\ (\alpha/2)\ ln\ ln\ n + O(1)$, and for some constant c_1, $\mathbf{E}(|X_{n,h^*}|) > c_1$. Further, there exist positive constants c_2 and c_3 such that (a) for $i = o(\sqrt{ln\ n})$, $c_2(2/\alpha)^i \le \mathbf{E}(|X_{n,h^*+i}|) \le c_3(2/\alpha)^i$, and (b) for all positive i: $\mathbf{E}(|X_{n,h^*+i}|) \le c_3 2^{-i}$. Finally, we have (c) for n sufficiently large and $h = h_n^* + o(\sqrt{ln\ n})$, $\mathbf{P}(n'_x < 2 | x \in X_{n,h}) \ge 1/2$.*

Remark 3. Note that we can assume that n is arbitrarily large here by increasing the constants so we do not need to state this explicitly. This same remark holds for many other lemmas as well.

Remark 4. Note that $h_n^* = h'_n + \log_{\alpha/2} ln\, n + O(1)$ which, since h_n^* is $\alpha \ln(n)(1 + o(1))$, is equivalent to: $h_n^* = h'_n + \log_{\alpha/2} h_n^* + O(1)$. Thus, by Lemma 1(a), $\mathbf{E}(|X_{n,h'_n}|)$ is between $d_1 h^*$ and $d_2 h^*$ for two constants d_1 and d_2.

PROOF. See Devroye [1986, p. 492] for a proof that the claimed value of h^* is correct. The same page contains a proof of (b). The rest of the lemma can be proved using similar straightforward computations. (The key is to note that the density function for a gamma distribution with parameter h is $d(x) = x^{h-1} \frac{\exp(-x)}{(h-1)!}$. To obtain $\mathbf{P}(y \in X_{n,h})$, we integrate this function from 0 to $ln\ n$. To obtain $\mathbf{P}(n'_y \ge q2)$, we integrate it from 0 to $ln\ n - ln\ 2$. When x is below $ln\ n$ and h is near h_n^*, decreasing x by t decreases $d(x)$ by a multiplicative factor of at least $(1 + o(1)) \exp(t(\alpha - 1))$. In particular, $d(x - ln\ 2) < \frac{1}{4} d(x)$ from which (c) follows easily. Further, this shows that the main contribution to the integral for $\mathbf{P}(y \in X_{n,h})$ comes from x near $ln\ n$ and suggests that to compute the ratio of the integral for h to that for $h - 1$, we need really only consider the ratio between the corresponding density functions at this point. This latter ratio is $ln(n)/h$, which is about α^{-1}. Since the size of the corresponding tree levels differ by a factor of 2, (a) seems believable. We omit a formal proof). □

If for each h near h^*, $|X_{n,h}|$ were highly concentrated around its expected value then this would imply that $\mathbf{E}(H'_n) = h^* + O(1)$, and $\mathbf{Var}(H'_n) = O(1)$. However, for each such h, $|X_{n,h}|$ is not concentrated around its expected value, and, more strongly, $\mathbf{Var}(|X_{n,h}|)$ is high, as we explain below. (This is why the expected value of H'_n is near h'_n rather than h_n^*.)

To begin, we note that, if every u_i were $1/2$, then H'_n would be $\lceil log\ n \rceil$. It is the existence of biased pivots (i.e., pivots which split the remaining set of keys unevenly) that forces the height to rise. In particular, a node x at depth h in T_∞ will only be in T'_n if the pivots along the path P_x of T_∞ from the root to x are sufficiently biased towards 1. How will this affect the probability that other nodes of T_∞ at depth h are in T'_n?

The answer to this question depends on where these biased pivots occur. To illustrate this point, we consider three unlikely but illuminating examples. For simplicity, we assume $n = 2^k$, that is, $log\ n$ is an integer.

Example 1. The first $h - k$ labels on P_x are 1, the remaining k labels are $\frac{1}{2}$.

Example 2. The first k labels on P_x are $1/2$, the remaining $h - k$ labels are 1.

Example 3. All the labels on P_x are the same: $1/2^{k/h}$.

Note that, in Example 1, the $(h-k)$th node z on P_x satisfies $f(z)=1$. Its child z', which is not on P_x, satisfies $f(z')=1/2$. Thus, the random subtree of T'_n rooted at z' given the labels on P_x is simply a copy of $T'_{n/2}$ generated independently of our choices on P_x. So, the expected number of nodes at depth h in T'_n given our choices of the labels on P_x is at least the expected number of nodes at depth $k-1$ in $T'_{n/2}$. Since $k-1=log\ \frac{n}{2} \ll \alpha ln\ \frac{n}{2}$, we expect this value to be quite large, in fact it is of the order of n^{ϵ} for some positive ϵ. So, the existence of a node x as in Example 1 implies that we expect there to be many other nodes of T'_n at depth h.

In discussing Examples 2 and 3, we consider for each node z of $P_x - x$, the child z' of z not on P_x. As in the first example, the subtree of T'_n rooted at z' is an independently generated copy of $T_{nf(z')}$.

In Example 2, if z is at depth at least k, then $f(z')=0$ so T'_n does not intersect the subtree of T_∞ rooted at z'. If z is at height $i<k$, then $f(z')=1/2^{i+1}$. We can therefore calculate the expected number of nodes of T'_n under z' at depth h in T'_n given the labels of P_x, it is simply the expected size of $X_{n/2^{i+1},h-i}$ in the copy. We do not do so, we simply note that the labels on $P_{z'}$ are not biased towards 1 and so this value tends to zero as i grows. It turns out that the sum of these values is actually less than 1. So the existence of a node as in Example 2 only increases the expected number of nodes of T'_n at depth h slightly.

Before discussing Example 3, we introduce some notation. We enumerate the vertices of P_x as $z_0,\ldots,z_h$ in the order they appear when P_x is traversed starting at the root. For i with $0 \le i \le h-1$, we let z'_i be the child of z_i which is not on P_x, and let l_i be the label of the edge of P_x entering z_i.

In analyzing Example 3, we again use the fact that

$$\mathbf{E}(|X_{n,h}| : x \in T'_n \textit{ and labels on } P_x) = 1 + \sum_{i=0}^{h-1} \mathbf{E}\left(\left|X_{nf(z'_i),h-i}\right|\right).$$

We do not compute this bound explicitly. We claim only that for i between 0 and $3h/4$ the summand is above $1/10$ and more strongly that given a node as in Example 3 we expect about h nodes of T'_n at depth h. This claim could be proven using Lemma 1. The intuition behind it is that, for each i, the labels on $P_{z'_i}$ are exactly as biased as they need to be on a path from the root to a vertex of T'_n at depth h. Thus, we expect about one node of T'_n at depth h under each z'_i (there are some technical complications when i is near h which don't really matter).

Our examples all demonstrate that to estimate $\mathbf{Var}(|X_{n,h}|)$ we need to determine, for a typical x in $X_{n,h}$, whether the pivots along P_x which are biased towards 1 occur near the beginning or near the end of P_x. It turns out that, as in Example 3, our analysis relies on determining (i) how many z_i on P_x satisfy: $f(z_i)$ is larger than it would be if all the labels on P_x were identical, and (ii) by how much does $f(z_i)$ exceed this "normal" value for such z_i. Precise answers to these questions will allow us to prove Theorem 2.

4. *The Approach*

To answer the questions posed above, we adopt the approach taken in Devroye and Reed [1995]. In that paper, the authors considered those special nodes for which $f(z_i)$ is never above its normal value. They showed that for h near h^*, the expected

number of such special nodes is at least $|X_{n,h}|/h$. Combining this result with an analysis of the variance of the number of such nodes, along the lines suggested in the analysis of Example 3 in the preceding section, allowed them to prove that there is an $\epsilon > 0$ such that if $\mathbf{E}(|X_{n,h}|)$ exceeds h^2 then the probability that there is such a special node at height h exceeds ϵ. This, along with some routine analysis using Lemma 1, allowed them to prove that the expected value of H'_n is within $O(\ln \ln n)$ of h^*.

In this section, we explain their approach and discuss how we will modify it to prove Theorem 2.

We note that this section is intended to provide intuition, and we will not use any of the statements that we present except for Lemma 2, whose proof is given in full.

To begin, they considered a fixed node x at depth h in T_∞ and generated the labels on P_x by:

(i) first choosing a set S of h random reals uniformly and independently from $[0\,..\,1]$.

(ii) then choosing uniformly a random permutation $l_1, \ldots, l_h$ of S and letting l_i be the label on the ith edge of P_x.

We note that the event $x \in T'_n$ is determined in (i), whilst the difference between $f(z_i)$ and its normal value is determined in (ii). It is this independence which allowed Devroye and Reed to obtain their results by a straightforward analysis of the random permutation generated in (ii).

Remark 5. It is important here that we are considering T'_n and not T_n, as the effects of the roundoffs in the definition of the n_t vary over the permutations, and so the first step alone does not determine if x is in T_n.

For each i between 1 and h we let a'_i be the natural logarithm of $1/l_i$. We note that $x \in T'_n \iff \sum_{i=1}^{h} a'_i \le \ln n$. We normalize by setting $a_i = a_i{}' - (\sum_{i=1}^{h} a'_i)/h$, so $\sum_{i=1}^{h} a_i = 0$. We let PS_i be the partial sum $\sum_{j=1}^{i} a_j$. Then, $f(z_i)$ is above its normal value precisely if $PS_i < 0$. We say x is *good* if $x \in T'_n$ and every $PS_i \ge 0$.

The key to Devroye and Reed's proof is:

LEMMA 2. *Let $X = \{x_1, \ldots, x_h\}$ be a set of real numbers whose sum is non-negative. Let $x_{\pi(1)}, \ldots, x_{\pi(h)}$ be a uniformly random permutation of X, then*

$$\mathbf{P}\left(\nexists j \text{ such that } \sum_{i=1}^{j} x_{\pi(i)} < 0\right) \ge \frac{1}{h}.$$

Furthermore, if X sums to zero, but none of its proper subsets do, then this probability is exactly $1/h$.

Letting $X'_{n,h}$ be the set of good nodes at depth h, and summing the bound of Lemma 2 over all those choices of S in (i) for which $x \in T'_n$ yields:

$$\mathbf{E}(|X'_{n,h}|) \ge \frac{\mathbf{E}(|X_{n,h}|)}{h} \tag{1}$$

as claimed in the first paragraph of this section.

Since Lemma 2 is also used in this article, we reproduce its short proof.

PROOF OF LEMMA 2. We can choose the permutation by first exposing a uniformly random cyclic permutation $c_1, \ldots, c_h$ of X and then choosing a uniformly random starting point $j+1$. That is, we set $x_{\pi(i)} = c_{j+i}$ where addition is taken modulo h.

Since X sums to a nonnegative number, if j minimizes $\sum_{i=1}^{j} c_i$ then all the partial sums for the chosen permutation are nonnegative. Since there is at least one j obtaining this minimum, the probability that all the partial sums are nonnegative is at least $1/h$ as claimed.

Furthermore, if X sums to 0, then all the partial sums are nonnegative precisely if j minimizes $\sum_{i=1}^{j} c_i$, If none of the proper subsets of X sum to 0 then there is exactly one j achieving this minimum. So the inequality does indeed become an equation in this case. □

When considering Example 3 in the last section, we presented the intuition behind the second result that Devroye and Reed required:

$$\exists \epsilon > 0 \textit{ such that if } h \textit{ satisfies } \mathbf{E}(|X'_{n,h}| > h), \textit{ then } \mathbf{P}(X'_{n,h} \neq \emptyset) \geq \epsilon. \qquad (2)$$

Now, (1) and (2), together with the estimates on $\mathbf{E}(|X_{n,h}|)$ given in Lemma 1 imply that for some $h^- = h^* - 2log_{\alpha/2} h^* + 0(1)$ we have:

$$\exists \epsilon > 0 \textit{ such that } \mathbf{P}(X'_{n,h^-} \neq \emptyset) > \epsilon. \qquad (3)$$

By analyzing the 2^i subtrees of T'_n rooted at those nodes of T_∞ at depth i independently (a standard simple trick), Devroye and Reed [1995] showed that (3) implies:

$$\exists c > 0 \textit{ and } d < 1, \textit{ such that } \forall i \textit{ with } i = o(\sqrt{h^*}), \mathbf{P}(H'_n < h^- - i) \leq cd^i. \qquad (4)$$

This yields

$$\mathbf{E}(H'_n) > h^- + O(1). \qquad (5)$$

Thus, Devroye and Reed have shown that $\mathbf{E}(H'_n)$ lies between $h^* + O(1)$ and $h^- + O(1)$. That is, by our remark after Lemma 1, between $h' - \log_{\frac{\alpha}{2}} h^* + O(1)$ and $h' + \log_{\frac{\alpha}{2}} h^* + O(1)$.

To prove Theorem 2, we need to tighten both the upper and lower bound. To tighten the upper bound, we replace the bound

$$\mathbf{P}(X_{n,h'+i} \neq \emptyset) \leq \mathbf{E}(|X_{n,h'+i}|) = O(h2^{-i})$$

obtained from Lemma 1 (we use part (b) of the lemma for i between 1 and $h^* - h'$ and part (c) for larger i), by the bound

$$\mathbf{P}(X_{n,h'+i} \neq \emptyset) = O\left(2^{\frac{-i}{2}}\right). \qquad (6)$$

To tighten the lower bound, we show that knowing that x is a typical good node only increases the expected number of good nodes by a constant, rather than by $O(h)$ which was the bound used by Devroye and Reed to prove (2).

We close this section by presenting some intuition as to how we prove these results.

To provide an intuitive as to why (6) holds, we consider pairing each permutation Π of a given S with its reverse Π_r. We claim that either

(i) $f(z_i)$ is at least its "normal" value when the labels on P_x are given by Π, or
(ii) $f(z_{h-i})$ is at least its "normal" value when the labels on P_x are given by Π_r.

To see this, note that $f(z_i)$ is above its "normal" value if and only if

$$\prod_{j=1}^{i} l_j \geq \left(\prod_{j=1}^{h} l_j\right)^{\frac{i}{h}}.$$

Our claim implies that on average at least $h/2$ of the $f(z_i)$ are above their normal value. As in Example 3, we can deduce that the expected number of nodes in $X_{n,h}$ given that x is in $X_{n,h}$ is at least ch for some constant $c > 0$. This fact suggests (but does not imply) that

$$\mathbf{P}(X_{n,h'+i} \neq \emptyset) \leq \mathbf{E}\left(\frac{|X_{n,h'+i}|}{h'}\right) = O(2^{-i}).$$

We cannot quite prove this, but do prove (6), which is good enough for our purposes.

In order to tighten our lower bound, we need to show that for a typical good node, most of the $f(z_i)$ are well below their normal value and hence the expected number of good nodes at depth h' given the existence of a typical good node is $O(1)$. That is, we need to show that for a typical good node, the overwhelming majority of the partial sums for our random permutation of S are well above zero.

To do so, we need to refine the analysis used in the proof of Lemma 2. To begin, we remark that the event that for the S chosen in (i), there is some proper subset of S which sums to 0 has probability zero. Thus, we can apply the strong version of Lemma 2 which holds for sets none of whose proper subsets sum to zero, to obtain:

$$\textit{the probability } x \textit{ is good given that it is in } X_{n,h} \textit{ is exactly } 1/h. \tag{7}$$

Actually, for very general sets S which sum to 0, it can be shown that the probability that every partial sum PS_i of a random permutation of S is nonnegative is $O(1/h)$ and more strongly that for small a the probability every PS_i exceeds $-a$ is also near $1/h$.

Now, consider such a set S and corresponding permutation Π, and ask what is the probability of the event E that $PS_{\frac{h}{2}}$ is 0 and all the partial sums for Π are nonnegative? If $PS_{\frac{h}{2}} = 0$, then our permutation Π breaks S into two subsets S_1 and S_2 both of which have $\frac{h}{2}$ elements and sum to 0. Further as all the partial sums for Π are nonnegative so are the partial sums for the permutation Π_1 of S_1 consisting of the first half of Π and the permutation Π_2 of S_2 consisting of the second half of Π. Clearly, we can first generate our choice for (S_1, S_2) and then generate Π_1 and Π_2 independently. If S is sufficiently well-behaved then we can show that the probability a random permutation of each S_i stays above 0 is also $O(\frac{1}{h})$. So, the foregoing discussion implies that the probability of E is $O((\frac{2}{h})^2)$.

Since, by Lemma 2, the probability that all the PS_i for Π are nonnegative is at least $\frac{1}{h}$, we see that the probability that $PS_{\frac{h}{2}}$ is 0 given all the PS_i are nonnegative is $O(\frac{1}{h})$. A similar analysis shows that $PS_{\frac{h}{2}}$ is unlikely to be near 0, as is PS_i for any i sufficiently far from 0 and h. Using a more delicate analysis with the same flavor, we can deal with the extremities of the permutation, and obtain the desired tightening of the lower bound.

5. *Some Independence Properties of Conditioned Exponentials*

The results we prove about random permutations hold for a large class of distributions. However, the proof of this general result is rather gruesome. A proof sketch of this general result will eventually be presented elsewhere. We present here an elegant proof which relies on the independence properties of the distribution that arises in the study of random binary search trees. These properties and the resultant proof simplification were pointed out to me by Luc Devroye.

So let $E_1, \ldots, E_n$ be independently chosen exponential random variables with mean 1. Let σ_n be their sum. Let $D_n = (d_1, \ldots, d_n)$ where $d_i = \frac{E_i}{\sigma_n}$. For $1 \leq i \leq n$, let $T_i = \sum_{j=1}^{i} d_j$. We need the following

PROPERTY 1: *σ_n is independent of D_n.*

PROOF. See Chapter 8 of Shorack and Wellner [1986]. □

PROPERTY 2: *D_n is distributed like the spacings on $[0\,..\,1]$ determined by $n-1$ independent uniform variables, and hence every permutation of a given multi-set is equally likely to be D_n.*

PROOF. See Chapter 8 of Shorack and Wellner [1986]. □

PROPERTY 3. *For $1 \leq i < j \leq n$, if we condition on $T_i = a$ and $T_j = b$, then the set $(d_{i+1}/(b-a), \ldots, d_j/(b-a))$ is distributed like D_{j-i}.*

PROOF. Specify the choices of $S_1 = \sum_{l=1}^{i} E_l$, $S_2 = \sum_{l=i+1}^{j} E_l$ and $S_3 = \sum_{i=l+1}^{n} E_l$. Then, our conditioning is that $S_1 = a\sigma_n$, $S_2 = (b-a)\sigma_n$, and $S_3 = (1-b)\sigma_n$. The result follows from the second of these equations and Property 1. □

PROPERTY 4. *For any integer a between 1 and n and real b between 0 and 1, the probability that T_a exceeds b is precisely the probability that if we pick a subset B of a set A of $n-1$ elements by putting each element of A into B independently with probability b, we obtain: $|B| < a$.*

PROOF. Follows immediately from Property 2, we just specify which of the uniforms are less than b. □

PROPERTY 5. *The probability that there exists an i such that $E_i > 6 \ln n$ is less than $\frac{1}{n^5}$.*

PROOF. The probability that E_i exceeds $6 \ln n$ is $\frac{1}{n^6}$. Since there are n choices for i, the result follows. □

6. *The Key Ideas*

We let $y_i = d_i - \frac{1}{n}$, that is, to obtain the y_i we normalize the d_i so that they sum to 0. We let $U_i = \sum_{j=1}^{i} y_j = T_i - \frac{i}{n}$. We say that D_n exceeds 0, written D_n *exc* 0 if for all i we have $U_i \geq 0$. For a nonnegative integer a, we say that D_n exceeds $-a$, written D_n *exc* $-a$ if for all i we have $U_i > \frac{-a}{n}$. By Lemma 2, $\mathbf{P}(D_n \text{ exc } 0) = \frac{1}{n}$. The crux of the proof is the following result:

LEMMA 3. *For any positive integer a, $\mathbf{P}(D_n \text{ exc } -a) = O(\frac{a^4}{n})$.*

PROOF. We can assume that a is at least 3, as if the statement holds for 3, it also holds for 1 and 2. We can also assume that n is large and $a < n^{\frac{1}{4}}$. Consider D_{n+2a^2}. Let $Ev(b,c)$ be the event that $U_{a^2} = \frac{b}{n+2a^2}$, and $U_{n+a^2} = \frac{c}{n+2a^2}$.

$$\begin{aligned} &\textit{If } b, c > 0, \textit{ then given } Ev(b,c);\\ &\quad \mathbf{P}((U_i > 0 \forall i \le a^2) \cap (U_i > 0 \forall i \ge n + a^2)) \ge a^{-4}. \end{aligned} \tag{8}$$

PROOF OF (8). Choose as unordered sets the values of $\{y_1, \ldots, y_{a^2}\}$ and $\{y_{n+a^2}, \ldots, y_{n+2a^2}\}$. Then, specify the values of these y_i by choosing a uniformly random permutation of each of these sets. By Lemma 2, since $b > 0$, the permutation of the first set yields $U_i > 0 \forall i \le a^2$ with probability at least $\frac{1}{a^2}$. Similarly, since $c > 0$, the permutation of the second set yields $U_i > 0 \forall i \ge n + a^2$ with probability at least $\frac{1}{a^2}$. Furthermore, these choices are independent. The result follows. □

We now consider the probability that U_i is above 0 for all i between a^2 and $n+a^2$ given $Ev(b,c)$. We focus on b, c with $n > b > c > a$.

Note that, for i between a^2 and $n + a^2$, we have:

$$U_i = U_{a^2} + \sum_{j=a^2+1}^{i} d_j - \frac{i - a^2}{n + 2a^2}.$$

$$= \frac{b}{n+2a^2} + \frac{n+(c-b)}{n+2a^2}\left(\left(\sum_{j=a^2+1}^{i} d_j \frac{n+2a^2}{n+(c-b)}\right) - \frac{i-a^2}{n}\right) + \frac{c-b}{n+2a^2}\frac{i-a^2}{n}$$

$$= \frac{b}{n+2a^2} + \frac{n+(c-b)}{n+2a^2}\left(\left(\sum_{j=a^2+1}^{i} \frac{d_j}{T_{n+a^2} - T_{a^2}}\right) - \frac{i-a^2}{n}\right) + \frac{c-b}{n+2a^2}\frac{i-a^2}{n}$$

$$= \frac{b}{n+2a^2}\frac{n+a^2-i}{n} + \frac{c}{n+2a^2}\frac{i-a^2}{n} + \frac{n+(c-b)}{n+2a^2}\left(\left(\sum_{j=a^2+1}^{i} \frac{d_j}{T_{n+a^2} - T_{a^2}}\right) - \frac{i-a^2}{n}\right).$$

Since $b, c > a$ and $a^2 < i \le n + a^2$ this yields:

$$U_i \ge \frac{a}{n+2a^2} + \frac{n+(c-b)}{n+2a^2}\left(\left(\sum_{j=a^2+1}^{i} \frac{d_j}{T_{n+a^2} - T_{a^2}}\right) - \frac{i-a^2}{n}\right).$$

Since $a > 0$ and $n + c - b > 0$, if U_i is not positive, then $(\sum_{j=a^2+1}^{i} \frac{d_j}{T_{n+a^2} - T_{a^2}}) - \frac{i-a^2}{n}$ is negative. As $c - b < 0$, this implies:

$$U_i > min\left(0, \frac{a}{n+2a^2} + \frac{n}{n+2a^2}\left(\left(\sum_{j=a^2+1}^{i} \frac{d_j}{T_{n+a^2} - T_{a^2}}\right) - \frac{i-a^2}{n}\right)\right).$$

Thus, if

$$\left(\sum_{j=a^2+1}^{i} \frac{d_j}{T_{n+a^2} - T_{a^2}}\right) - \frac{i-a^2}{n} \geq \frac{-a}{n},$$

then $U_i > 0$. So, by Property 3, the probability that $U_i > 0$ for all i between a^2 and $n + a^2$ given $Ev(b, c)$ is at least $\mathbf{P}(D_n \; exc \; -a)$.

Since this probability is clearly independent of our choices for the permutations of the y_i with $i \leq a^2$ or $i \geq n + a^2$, we have:

$$For\ n > b > c > a: \quad \mathbf{P}(D_{n+2a^2} \; exc \; 0 | Ev(b, c)) \geq \frac{1}{a^4}\mathbf{P}(D_n \; exc \; -a). \tag{9}$$

We now show that the probability that $Ev(b, c)$ holds for some choice of b and c with $n > b > c > a$ is $\Omega(1)$. Since $\mathbf{P}(D_{n+2a^2} \; exc \; 0) = \frac{1}{n+2a^2}$ this completes the proof.

More precisely, we show:

$$\mathbf{P}((2a < U_{a^2} < 3a) \cap (a < U_{n+a^2} < 2a)) = \Omega(1). \tag{10}$$

PROOF OF (10). This probability can be computed using the correspondence discussed in the proof of Property 4. We provide the details but suggest the reader bypass them.

Consider a set X of $n + 2a^2 - 1$ independent uniform elements of $[0 \,..\, 1]$. Let A_1 be the event that less than a^2 elements of X are in the interval $[0..\frac{a^2+2a}{n+2a^2}]$. Let A_2 be the event that less than a^2 elements of X are in the interval $[\frac{n+a^2+2a}{n+2a^2} \,..\, 1]$. Let A_3 be the event that more than a^2 elements of X are in the interval $[0 \,..\, \frac{a^2+3a}{n+2a^2})$. Let A_4 be the event that more than a^2 elements of x are in the interval $(\frac{n+a^2+a}{n+2a^2} \,..\, 1]$. Then, by Property 2,

$$\mathbf{P}((2a < U_{a^2} < 3a) \cap (a < U_{n+a^2} < 2a)) = \mathbf{P}(A_1 \cap A_2 \cap A_3 \cap A_4).$$

So, it is enough to show that the right hand side of this inequality is $\Omega(1)$.

To do so, we let A_5 be the event that between $\frac{a}{2}$ and $\frac{3a}{2}$ of the elements of X are in the interval $(\frac{a^2+2a}{n+2a^2} \,..\, \frac{a^2+3a}{n+2a^2})$. Similarly, we let A_6 be the event that between $\frac{a}{2}$ and $\frac{3a}{2}$ of the elements of X are in the interval $(\frac{n+a^2+a}{n+2a^2} \,..\, \frac{n+a^2+2a}{n+2a^2})$.

We let A_1' be the event that less than a^2 but more than $a^2 - \frac{a}{2}$ of the elements of X are in the interval $[0 \,..\, \frac{a^2+2a}{n+2a^2}]$. Similarly, we let A_2' be the event that less than a^2 but more than $a^2 - \frac{a}{2}$ of the elements of X are in the interval $[\frac{n+a^2+2a}{n+2a^2} \,..\, 1]$.

Clearly, $A_5 \cap A_6 \cap A_1' \cap A_2'$ is a subset of $A_1 \cap A_2 \cap A_3 \cap A_4$. So, it is enough to show:

$$\mathbf{P}(A_5 \cap A_6 \cap A_1' \cap A_2') = \Omega(1).$$

The number of reals in X which are in an interval of length l is distributed like the sum of $n + 2a^2 - 1$ independent binomials that are 1 with probability l.

Thus, the expected number of elements of X in the interval $(\frac{a^2+2a}{n+2a^2} \,..\, \frac{a^2+3a}{n+2a^2})$ is $a - \frac{a}{n+2a^2}$ which is between a and $a + \frac{1}{2}$. Applying the Chernoff bound, we see that $\mathbf{P}(A_5) = \Omega(1)$.

Given that A_5 holds, the expected number of elements of X in the interval $(\frac{n+a^2+a}{n+2a^2} \,..\, \frac{n+a^2+2a}{n+2a^2})$ is distributed like the sum of $n + 2a^2 - d - 1$ independent

binomials which are 1 with probability $\frac{a}{n+2a^2-a}$ for some d between $\frac{a}{2}$ and $\frac{3a}{2}$. Thus, the conditional expected size of this subset of X is $a+O(\frac{a^2}{n+2a^2})$ which is $a+O(1)$. Applying the Chernoff bound, we see that $\mathbf{P}(A_6|A_5)=\Omega(1)$, and hence $\mathbf{P}(A_5\cap A_6)=\Omega(1)$.

Given that $A_5\cap A_6$ holds, the expected number of elements of X in the interval $[0\,..\,\frac{a^2+2a}{n+2a^2}]$ is distributed like the sum of $n+2a^2-d-1$ independent binomials which are 1 with probability $\frac{a^2+2a}{n+2a^2-2a}$ for some d between a and $3a$. Thus, the conditional expected size of this subset of X is $a^2+2a+O(\frac{a^3}{n+2a^2})$ which, since $a<n^{\frac{1}{4}}$, is $a^2+2a+O(1)$. Since the standard deviation for the sum of independent binomials is the square root of this expected value, we see that $\mathbf{P}(A_1'|A_6\cap A_5)=\Omega(1)$, and hence $\mathbf{P}(A_1'\cap A_6\cap A_5)=\Omega(1)$.

Given that $A_1'\cap A_5\cap A_6$ holds, the expected number of elements of X in the interval $[\frac{n+a^2+2a}{n+2a^2}\,..\,1]$ is distributed like the sum of $n+a^2-d-1$ independent binomials which are 1 with probability $\frac{a^2-2a}{n+a^2-4a}$ (note that this probability is positive since a is assumed to be at least 3), for some d between $\frac{a}{2}$ and $3a$. Thus, the conditional expected size of this subset of X is $a^2-2a+O(\frac{a^2}{n+a^2-4a})$ which is $a^2-2a+O(1)$. Since the standard deviation for the sum of independent binomials is the square root of this expected value, we see that $\mathbf{P}(A_2'|A_1'\cap A_6\cap A_5)=\Omega(1)$, and so (10) holds. □

Using the previous lemma, we can easily show that if D_n exceeds 0 then it is well above 0 most of the time. Specifically, we have;

LEMMA 4. *For every integer $a>0$ letting R_a be the event*

$$\exists k,\ a^{40}<k\le\frac{n}{2}\ \textit{s.t.}\ U_k<\frac{30(\log k)^7}{n}\ \textit{or}\ U_{n-k}<\frac{30(\log k)^7}{n}$$

we have;

$$\mathbf{P}((D_n\ exc\ -a)\cap R_a)=O\left(\frac{1}{na^2}\right).$$

Remark 6. We need $k>a^{40}$ so that $a^8=O(k^{1/4})$ and $a^2=o(k^{1/10})$, two inequalities used in the proof. The right-hand sides of the bounds on U_k and U_{n-k} are chosen because this is the bound required in later proofs. Actually the result holds for any right-hand side which is $o(k^{1/280})$ with the same proof. Here and below, the O and o notation is with respect to a, k and n. that is, the constants involved are independent of all three parameters.

PROOF. We can assume n is large. Fix an a as in the statement of the lemma. We can assume a is bigger than any fixed constant as for bounded a the result follows directly from Lemma 3. In particular, we can assume that for any $k>a^{40}$, $30(\log k)^7<k^{\frac{1}{40}}<\frac{k}{2}$. Choose an integer k between a^{40} and $\frac{n}{2}$. For any real b with $-a<b<30(\log k)^7$, we let $Ev(b)$ be the event that $U_k=\frac{b}{n}$. and $Ev'(b)$ be the event that $U_{n-k}=\frac{b}{n}$.

The key to the proof is to apply Property 3 and Lemma 3 to show:

$$\begin{aligned}\mathbf{P}(D_n\ exc\ -a|Ev(b))\ &\le\ \mathbf{P}(D_k\ exc\ -2a-2|b|)\\ &\times\ \mathbf{P}(D_{n-k}\ exc\ -2a-2|b|)=O\left(\frac{(a+|b|)^8}{nk}\right).\end{aligned}\quad(11)$$

and

$$\mathbf{P}(D_n \; exc \; -a | Ev'(b)) = O\left(\frac{(a+|b|)^8}{nk}\right). \tag{12}$$

We do so below.

Using Property 4, it is easy to show

$$\textit{the probability that } U_k \textit{ lies in any interval of size } \frac{1}{n} \textit{ is } O\left(\frac{1}{\sqrt{k}}\right). \tag{13}$$

For completeness, we also prove this result below.

It follows from (13) that the probability that $Ev(b)$ holds for some b between $-a$ and $30(\log\ k)^7$ is $O(\frac{a+(\log\ k)^7}{\sqrt{k}})$. By symmetry (reversing the permutation) the probability that $Ev'(b)$ holds for some b between $-a$ and $30(\log\ k)^7$ is $O(\frac{a+(\log\ k)^7}{\sqrt{k}})$.

Applying (11) and (12), we obtain that the probability that for this specific value of k we have

$$(D_n \; exc \; -a) \cap (U_k \; < \; \frac{30(\log\ k)^7}{n} \; or \; U_{n-k} \; < \; \frac{30(\log\ k)^7}{n}) \text{ is } O\left(\frac{a+(\log\ k)^7}{\sqrt{k}} \frac{(a+(\log\ k)^7)^8}{nk}\right).$$

Since $k \geq a^{40}$ and $30(\log k)^7 < k^{\frac{1}{40}}$, this is $o(n^{-1}k^{\frac{-5}{4}}) = o(n^{-1}k^{\frac{-9}{8}}a^{-2})$. So, summing over all values of k yields the desired result. It remains only to prove (11), (12), and (13).

PROOF OF (11). We want to bound the conditional probability that D_n exceeds $-a$ given that $Ev(b)$ holds.

We consider first the conditional probability that $U_i \geq -a$ for all i between 1 and k. By Property 3, we can generate these U_i by taking an independent D'_k distributed like D_k with corresponding T'_j and U'_j and setting $T_i = T'_i(\frac{k}{n} + \frac{b}{n})$. Now suppose $T'_i = (\frac{i}{k} - \frac{x}{k})$ for some $x > 0$. Then, $T_i = \frac{i}{n} - \frac{x}{n} - \frac{xb}{kn} + \frac{ib}{kn}$.

If b is nonnegative, then since $i \leq k$, T_i is at most $\frac{i}{n} - \frac{x}{n} + \frac{b}{n}$ and so U_i is at most $\frac{b-x}{n}$. If b is negative, then since $k > 2a > 2|b|$, T_i is at most $\frac{i}{n} - \frac{x}{n} + \frac{x}{2n}$ and so U_i is at most $\frac{-x}{2n}$.

Thus, in either case, if x is more than $2a + |b|$ then U_i is less than $\frac{-a}{n}$. So, if D_n exceeds $-a$ then D'_k must exceed $2a + |b|$ and hence also $2a + 2|b|$.

We consider next the conditional probability that $U_i \geq -a$ for all i between $k+1$ and n. By Property 3, we can generate these U_i by taking a D'_{n-k} independently of D'_k distributed like D_{n-k} with corresponding T'_j and U'_j and setting $T_{k+j} = \frac{k}{n} + \frac{b}{n} + T'_j(\frac{n-k}{n} - \frac{b}{n})$. Now suppose $T'_j = (\frac{j}{n-k} - \frac{x}{n-k})$ for some $x > 0$. Then, $T_{k+j} = \frac{k+j}{n} - \frac{x}{n} + \frac{b}{n} + \frac{xb}{(n-k)n} - \frac{jb}{(n-k)n}$.

If b is nonnegative, then since $n - k > \frac{n}{2} > 60(\log\ n)^7 > 2b$, T_{j+k} is at most $\frac{k+j}{n} - \frac{x}{n} + \frac{b}{n} + \frac{x}{2n}$ and so U_{j+k} is at most $\frac{b}{n} - \frac{x}{2n}$. If b is negative, then since $n - k > j$, T_{k+j} is at most $\frac{k+j}{n} - \frac{x}{n} + \frac{b}{n} - \frac{b}{n}$ and so U_i is at most $\frac{-x}{n}$.

Thus, in either case, if x is more than $2a + 2|b|$ then U_i is less than $\frac{-a}{n}$. So, if D_n exceeds $-a$ then D'_{n-k} must exceed $2a + 2|b|$.

We have obtained the desired result by considering $i \leq k$ and $i \geq k$ separately. □

PROOF OF (12). We want to bound the conditional probability that D_n exceeds $-a$ given that $Ev'(b)$ holds. The proof is very similar to that of (11).

We consider first the conditional probability that $U_i \geq -a$ for all i between 1 and $n-k$. By Property 3, we can generate these U_i by taking an independent D'_{n-k} distributed like D_{n-k} with corresponding T'_j and U'_j and setting $T_i = T'_i(\frac{n-k}{n} + \frac{b}{n})$. Now suppose $T'_i = (\frac{i}{n-k} - \frac{x}{n-k})$ for some $x > 0$. Then, $T_i = \frac{i}{n} - \frac{x}{n} - \frac{xb}{(n-k)n} + \frac{ib}{(n-k)n}$.

If b is nonnegative, then since $i \leq n-k$, T_i is at most $\frac{i}{n} - \frac{x}{n} + \frac{b}{n}$ and so U_i is at most $\frac{b-x}{n}$. If b is negative, then since $n-k > 2a > 2|b|$, T_i is at most $\frac{i}{n} - \frac{x}{n} + \frac{x}{2n}$ and so U_i is at most $\frac{-x}{2n}$.

Thus, in either case, if x is more than $2a + |b|$, then U_i is less than $\frac{-a}{n}$. So, if D_n exceeds $-a$, then D'_{n-k} must exceed $2a + |b|$ and hence also $2a + 2|b|$.

We consider next the conditional probability that $U_i \geq -a$ for all i between $n-k+1$ and n. By Property 3, we can generate these U_i by taking a D'_k independently of D'_{n-k} distributed like D_k with corresponding T'_j and U'_j and setting $T_{k+j} = \frac{n-k}{n} + \frac{b}{n} + T'_j(\frac{k}{n} - \frac{b}{n})$. Now suppose $T'_j = (\frac{j}{k} - \frac{x}{k})$ for some $x > 0$. Then, $T_{n-k+j} = \frac{n-k+j}{n} - \frac{x}{n} + \frac{b}{n} + \frac{xb}{kn} - \frac{jb}{kn}$.

If b is nonnegative, then since $\frac{k}{2} > 30(\log k)^7 > b$, T_{n-k+j} is at most $\frac{n-k+j}{n} - \frac{x}{n} + \frac{b}{n} + \frac{x}{2n}$ and so U_{j+k} is at most $\frac{b}{n} - \frac{x}{2n}$. If b is negative, then since $j \leq k$, T_{n-k+j} is at most $\frac{n-k+j}{n} - \frac{x}{n} + \frac{b}{n} - \frac{b}{n}$ and so U_i is at most $\frac{-x}{2n}$.

Thus, in either case, if x is more than $2a + 2|b|$, then U_i is less than $\frac{-a}{n}$. So, if D_n exceeds $-a$ then D'_{n-k} must exceed $2a + 2|b|$.

We have again obtained the desired result by considering $i \leq k$ and $i \geq k$ separately. □

PROOF OF (13). Let X be a set of $n-1$ independent uniform elements of $[0\,..\,1]$. Consider any real a and let $a' = \frac{k}{n} + a$. Then, the probability that U_k lies in the interval $[a\,..\,a + \frac{1}{n}]$ is the probability of the event E that the kth largest element of X lies in the interval $I = [a'\,..\,a' + \frac{1}{n}]$.

The number r of elements of X in I is the sum of $n-1$ independent binomial variables each of which is 1 with probability at most $\frac{1}{n}$ (we have at most instead of equal to here because a' may be negative or exceed $\frac{n-1}{n}$). Thus, the probability that $r = j$ is at most $\frac{1}{j!}$. So, we have that $\mathbf{P}(r > \log k) = O(\frac{1}{k})$.

If we condition on $r = j$, then the probability of E is simply the probability that between $k-j$ and k of the remaining $n-j$ elements of X are less than a'. This is clearly at most j times the maximum over all s betwen $k-j$ and k of $\mathbf{P}(E(s, a'))$ where $E(s, a')$ is the event that exactly s of the remaining $n-j$ elements of X are less than a'. Now, $\mathbf{P}(E(s, a'))$ is the probability that the sum of $n-j$ independent identically distributed 0-1 binomials each of which is 1 with probability a' sums to s. Since $s \leq k < \frac{n}{2}$, standard results on the binomial distribution yield: $\mathbf{P}(E(s, a')) = O(\frac{1}{\sqrt{s}})$. So we have for all $j \leq \log k$:

$$\mathbf{P}(E|r = j) \leq O\left(\frac{j}{\sqrt{k}}\right).$$

Since $\mathbf{P}(r = j) \leq \frac{1}{j!}$, we see that

$$\mathbf{P}(E \cap (r \leq \log k)) \leq O\left(\sum_{j=1}^{k} \frac{1}{(j-1)!\sqrt{k}}\right) = O\left(\frac{1}{\sqrt{k}}\right).$$

The result follows. □

We have now obtained enough information about the shape of our random permutations to allow us to prove Theorems 1 and 2. We do so in the next two sections.

7. *The Proof of Theorem* 2

We now show that H'_n is concentrated around $h'_n = \lceil \alpha ln\ n - \beta ln\ ln\ n \rceil$ via three lemmas proven below.

LEMMA 5. $\mathbf{P}(X_{n,h'_n} \neq \emptyset) = \Omega(1)$.

LEMMA 6. *There exist (universal) constants $C_1 > 2$ and $d > 1$ such that, for $i > 0$ with $i = o(\sqrt{ln\ n})$, we have:*

$$\mathbf{P}(X_{n,h'_n - i} = \emptyset) < C_1 d^{-i}.$$

LEMMA 7. *There is a (universal) constant $C_2 > 2$ such that, for $i > 0$, we have:*

$$\mathbf{P}(X_{n,h'_n + i} \neq \emptyset) < C_2 2^{-i/2}.$$

Now, Lemma 6 implies that for some constant C (in fact, $C = 16$ will do)

$$\mathbf{E}(H'_n) > h'_n - C \log_d C_1 + o(1).$$

Similarly, Lemma 7 implies that for a constant C (in fact, $C = 16$ will do)

$$\mathbf{E}(H'_n) < h'_n + C \log C_2 + o(1).$$

Thus, $\mathbf{E}(H'_n) = h'_n + O(1)$. Combining Lemmas 6 and 7 then yields $\mathbf{Var}(H'_n) = O(1)$ (since $\sum \frac{i^2}{b^i}$ is O(1) for any $b > 1$).

So, it remains only to prove these three lemmas. Their proofs rely on the results of the last section. In particular, as discussed in Section 4, we apply them to the labels on the path from the root to a node x of the tree.

The crucial fact that allows us to apply the results is that we can first expose the products of labels to determine if x is in T'_n and then expose the actual values of the labels to check if x is good or determine other properties of the ordered set of labels on this path. Before presenting the proofs of this theorem we make this idea more precise.

We can assume n is large for the remainder of the section, since all three results are asymptotic.

Thus, we again consider some node x at depth h of T_∞, the path $P_x = z_o, \ldots, z_h = x$ from the root of T_∞ to x, and the nodes $z'_0, \ldots, z'_{h-1}$. Once again we let a'_i be the negative of the natural logarithm of the label on the edge into z_i and let a_i be $a'_i - \frac{\sum_{i=1}^{h} a'_i}{h}$. We let Π $(= \Pi(x))$ be the permutation of the a_i given by the labeling of P_x. We let $\sigma_h = \sum_{j=1}^{h} a_j$. We let $PS_i = \sum_{j=1}^{i} a_j$. For a positive real a, We say Π is above $-a$ (written $\Pi > -a$) if each PS_i is. We say Π is above

0 (written $\Pi > 0$) if for all i, $PS_i \geq 0$. We say x is good if Π is above 0. We say x is above $-a$ if Π is.

We can expose the a'_i by

(i) choosing σ_h, according to the appropriate probability distribution (i.e., σ_h is gamma distributed with parameter h),
(ii) choosing D_h as in Property 2 of Section 5, and
(iii) setting $a'_i = d_i \sigma_h$.

Note that $PS_i = \sigma_h U_i$. Thus, $\mathbf{P}(\Pi > -a)$ is simply $\mathbf{P}(D_h \; exc \; \frac{-ah}{\sigma_h})$ and $\mathbf{P}(x \; is \; \; good\,)$ is $\mathbf{P}(D_h \; exc \; 0)$.

We consider $h \cong \alpha ln\, n$ (as we know the height of T'_n is not far from this value) and $\sigma_h \cong ln\, n$ (since this is the breakpoint at which x is in $X_{n,h}$). So, $h/\sigma_h \cong \alpha$. In particular, we often consider choices for h and σ_h such that $h < 5\sigma_h$.

In such a situation, Lemma 3 implies:

Conditioned on a choice of σ_h, *which is less than* $5h$:

$$\mathbf{P}(\Pi > \; -a) = \mathbf{P}\left(D_h \; exc \; \frac{-ah}{\sigma_h}\right) < \mathbf{P}(D_h \; exc \; -5a) = O\left(\frac{a^4}{h}\right). \quad (14)$$

In such a situation, Lemma 4 implies:

Conditioned on a choice of σ_h, *which is less than* $5h$, *for every integer* $a > 0$:

$$\mathbf{P}\Bigg((\Pi > \; -a) \cap \Bigg(\exists k, \; a^{40} < k \leq \frac{h}{2} \; such \; that \; PS_k < 6(\log\, k)^7 \; or \; PS_{n-k} < 6(\log\, k)^7\Bigg)\Bigg) = O\left(\frac{1}{ha^2}\right). \quad (15)$$

Having translated Lemmas 3 and 4 into the language we use for discussing our random binary search trees, we are now ready to proceed with the proofs of Lemmas 5, 6, and 7.

PROOF OF LEMMA 5. We claimed in Section 4 that for a *typical* good node, we have that for i far from 0 and h, PS_i is well above 0. We note that (15) is an explicit formulation of this claim, which we have proven correct. To prove Lemma 5, we want to restrict our attention to these typical good nodes, as they are easier to handle.

Specifically, we let $h' = h'_n$ defined earlier and restrict our attention to the set $Y_{n,h'}$ of those x in $X_{n,h'}$ which satisfy $n'_x < 2$. Lemma 1(c) implies that the expected size of $Y_{n,h'}$ is at least half that of $X_{n,h'}$. Applying Lemma 2, we see that, on average, $\frac{1}{h'}|Y_{n,h'}|$ of the elements of $Y_{n,h'}$ are good. Applying (15), we see that there is a constant r such that for all a^* larger than r, we expect at least one half of the good elements of $Y_{n,h'}$ to satisfy:

$$\forall k \; with \; \frac{h'}{2} \geq k > a^*, PS_k > 6(\log\, k)^7 \; and \; PS_{h'-k} > 6(\log\, k)^7.$$

We choose a^* larger than r and also large enough to satisfy some implicit inequalities below. We let $A = A_{n,h'}$ consist of those x in $Y_{n,h'}$ which are good, and satisfy the property above. We have shown that $\mathbf{E}(|A_{n,h'}|) > \frac{1}{4h'}\mathbf{E}(|X_{n,h'}|) = \Omega(1)$. We will show that the variance of $|A|$ is bounded, thereby completing the proof of Lemma 5.

Remark 7. The constant one half above is arbitrary as is the choice of 2 as our bound on n'_x. By increasing r and the bound on n'_x, we could obtain that $A_{n,h'}$ contains all but an arbitrarily small proportion of the good nodes in $X_{n,h'}$. Thus, we can indeed think of the nodes of $A_{n,h'}$ as being typical good nodes.

The approach we take is quite simple. Our choice of $A_{n,h'}$ ensures that for k between a^* and $h' - a^*$, $n_{z'_k}$ is well below its normal value and hence the expected number of good nodes under z'_k will be small. The remaining values of k will not contribute much to the sum as there are only a constant number of them. The details follow; although somewhat tedious they are just straightforward counting.

Because we need a similar result in our proof of Theorem 1, we will actually prove a more general statement.

Suppose we have exposed a set of choices for the labels on P_x for a node x at height h' which yield

(a) $x \in X_{n,h'}$,
(b) $n'_x \leq 2$,
(c) *x is good, and*
(d) *For some constant C^*,*

$$\forall k \text{ with } \frac{h'}{2} \geq k > C^*, PS_k > 6 \log\ k \text{ and } PS_{h'-k} > 6 \log\ k.$$

Then the conditional expected value of $|X'_{n,h'}|$ is at most D for some constant which depends on C^*. (16)

Note that the event that x is in $A_{n,h'}$ is the disjoint union of a set of events each of which satisfy (a)–(d) with $C^* = a^*$ (each of these events is a choice of labels on P_x). Thus, summing (7) over this set of events yields the conditional expected value of $|X'_{n,h'}|$ given that x is in $A_{n,h'}$ is at most $D = D(a^*)$. Since $A_{n,h'}$ is a subset of $X'_{n,h'}$, this implies that the conditional expected value of $|A_{n,h'}|$ given that x is in $A_{n,h'}$ is at most D. It follows that the expected value of $|A_{n,h}|^2$ and hence the variance of $|A_{n,h'}|$ are $O(1)$, as required. So, it remains only to prove (16).

PROOF OF (16). Throughout the proof, we set $h = h'_n$. We consider a fixed x and a fixed set of choices for the labels on P_x, satisfying (a)–(d). For $1 \leq k \leq h$, we let E_k be the expected number of nodes of $X'_{n,h}$ under z'_k given this set of choices. We need to show that the $\sum_{k=0}^{h-1} E_k = O(1)$. We note first that $E_{h-j} < 2^{j-1}$ because there are only 2^{j-1} nodes of depth h under z'_{h-j}. Thus, $\sum_{j=1}^{\lceil C^* \rceil} E_{h-j} = O(1)$.

We consider next k between $\frac{h}{2}$ and $h - C^*$. We let $j = h - k$.

We note that since $n'_x < 2$ and $PS_k > 6 \log j > 6 ln\ j$ we have that:

$$n'_{z'_k} \leq n'_{z_k} \leq \frac{n}{j^6} \left(\frac{2}{n}\right)^{k/h} \leq \frac{2}{j^6} n^{j/h}.$$

Since $ln\ n = h + \beta ln\ ln\ n/\alpha + O(1)$, we obtain:

$$n'_{z'_k} = O\left(j^{-6} \exp\left(\frac{j}{\alpha}\right) \exp\left(\frac{j \beta ln\ ln\ n}{\alpha h}\right)\right).$$

If $j < \frac{h}{ln\ ln\ n}$, then this is $O(\frac{e^{j/\alpha}}{j^6})$. Otherwise, since $j < h$ and $ln\ h = ln\ \ ln\ n + O(1)$, it is $O(j^{-6}\exp(\frac{j}{\alpha})h^{\frac{\beta}{\alpha}})$. Now, $j > \frac{h}{ln\ ln\ n} > \sqrt{h}$ and $\beta < \alpha$ so $n'_{z'_k}$ is $O(\frac{e^{j/\alpha}}{j^4})$.

Thus, $n' = n'_{z'_k}$ is $O(\exp(j/\alpha - 4ln\ j))$, We now consider the copy of T_∞ rooted at z'_k. We let T' be the corresponding copy of $T'_{n'}$ and note that this corresponds to the subtree of T'_n under z'_k. We let $i^* = \alpha(\frac{j}{\alpha} - 4ln\ j) - \frac{1}{2ln\ \alpha/2} ln\ (j/\alpha - 4\ ln\ j)$ and note that by Lemma 1(c), we have that $h^*_{n'}$ is $i^* + O(1)$.

Manipulating our expression for i^*, we obtain

$$h^*_{n'} = j - \left(4\alpha - \frac{1}{2\ ln\ \alpha/2}\right) ln\ j + O(1) < j - 6ln\ j + O(1).$$

Applying Lemma 1(b) to T', we obtain that E_k is $O(2^{h^*_{n'}-j}) = O(1/j^3)$. Summing over all k yields $\sum_{k=(h/2)}^{h-\lceil C^*\rceil} E_k = O(1)$.

A similar computation shows that $\sum_{k=\lceil C^*\rceil}^{h/2} E_k = O(1)$. However, we need to be a bit more careful. So we fix some k in this range. We show that $E_k = o(\frac{1}{k^2})$ from which the desired result follows. We know, as in the previous case, that for some $b' > 6\ \log\ k > 6\ ln\ k + 1$ we have:

$$n' = n'_{z'_k} \le n'_{z_k} \le \frac{2}{\exp(b')} n^{(h-k)/h}.$$

Thus, we can write n' as $\exp(-b)n^{(b-k)/n}$ for some $b > 6 \log\ k - 1$. Set $j = h - k$. Then $j = (ln\ n' + b)\frac{h}{\ln\ n}$. If $b > 6\ ln\ j$, then we can proceed as in the case $k > \frac{h}{2}$ to obtain the desired result. So, we can assume that $b < 6\ ln\ j < 6\ ln\ h$. Hence, $j - \alpha ln\ n' = o(\sqrt{ln(n)})$.

We now consider the copy of T_∞ rooted at z'_k. We let T' be the corresponding copy of $T'_{n'}$ and X' be the nodes of T' at depth j. Note first that E_k is at most the expected number of nodes of X' *which are above* -1-b; for the other nodes of X' will not correspond to good nodes of $X_{n,h}$. Our bound on $j - \alpha \ln\ n'$ implies that we can apply (14) to T' to obtain that E_k is at most $O(b^4/h|\mathbf{E}(|X'|))$. We claim that $\mathbf{E}(|X'|) = o(h/b^4k^3)$. Thus, $E_k = o(1/k^3)$, a stronger result than we needed.

To prove our claim, we need to calculate $h^*_{n'}$ and then apply Lemma 1, as in the previous case.

Mimicking the proof above with j^{-6} replaced by $\exp(-b)$, we obtain:

$$n' = O\left(\exp(-b)\exp\left(\frac{j}{\alpha}\right)h^{\beta/\alpha}\right).$$

Setting

$$i^* = \alpha ln\ n' - \frac{1}{2ln\ (\alpha/2)} ln\ ln\ n'.$$

and applying Lemma 1, we have that $h^*_{n'} = i^* + O(1)$. Manipulating our expression for i^*, we obtain:

$$h^*_{n'} \le j + \beta ln\ h - \alpha b - \frac{1}{2ln\ \frac{\alpha}{2}}\ ln\ \left(\frac{j + \beta ln\ h}{\alpha} - b\right) + O(1).$$

Since $j > \frac{h}{2}$ and $b = o(j)$, this yields:

$$h^*_{n'} \leq j + \left(\beta - \frac{1}{2\ ln\ \alpha/2}\right) ln\ h - \alpha b + O(1) = j + \frac{1}{ln\ \alpha/2} ln\ h - \alpha b + O(1).$$

Our bound on $j - \alpha \ln n'$ thus allows us to apply Lemma 1(a) to show that we expect $O((\frac{\alpha}{2})^{(1/ln\frac{\alpha}{2})ln\ h - \alpha b})$ nodes at height j in T'. That is, $\mathbf{E}(|X'|) = O(h\frac{2}{\alpha}^{\alpha b}) = o(h2^{-\alpha b})$. Since $b > 6\log k - 1$, we have $b > 3\log k + 4\log b$, so $\mathbf{E}(|X'|)$ is $O(\frac{h}{b^4k^3})$, as claimed.

Finally, we consider k in the range 1 to C^*. We need to show that $\sum_{k=1}^{\lceil C^* \rceil} E_k = O(1)$. So we fix some k in this range. We show that $E_k = O(1)$ from which the desired result follows. We know that $n'_{z'_k} \leq n'_{z_k}$; since x is good and $n'_x \leq 2$, this yields:

$$n' = n'_{z'_k} \leq 2n^{\frac{h-k}{h}}.$$

So, for some b between -1 and $ln\ n$, we can write

$$n' = e^{-b} n^{\frac{h-k}{h}}.$$

Remark 8. Because k is so close to 0, we have that $n^{(h-k)/h}$ is between n and n/F for some constant F which depends on C^* and that $j - h = O(1)$. It may help the reader to be aware of these facts in the sequel.

We now consider the copy of T_∞ rooted at z'_k. We let T' be the corresponding copy of $T'_{n'}$ and X' be the nodes of T' at depth $j = h - k$.

Now, E_k is at most the expected number of nodes of X' which are above $-(b+4)$; for the other nodes of X' will not correspond to good nodes of $X_{n,h}$.

If $b > ln\ h + 1$, then $n' < \frac{n}{h}$ and so $h^*_{n'} < h^*_{n/h} = h^*_{n'} - \alpha ln\ h + O(1) < h - 2ln\ h + O(1) < j - 2ln\ (h) + 0(1)$. So, applying Lemma 1(a), we obtain $\mathbf{E}(|X'|) = O(2^{2lnh}) = O(\frac{1}{h})$, a stronger result then we need.

Thus, we can assume that $b < ln\ h + 1$. This implies that $j - h^*_{n'} = o(\sqrt{ln\ n'})$ and hence $j < 4.5\ ln\ n'$. Applying (14), we see that the E_k is at most $O(\frac{(b+4)^4}{h}|\mathbf{E}(|X'|))$.

Now, since $n' \leq \frac{n}{e^b}$ and $j = h'_n - O(1)$, Lemma 1(b) implies that $\mathbf{E}(|X'|)$ is $O(h\frac{2}{\alpha}^{-b})$. Thus, $E_k = O(\frac{(b+4)^4}{2^b})$, which is $O(1)$ as claimed.

This completes the proof of (16) and Lemma 5. □

PROOF OF LEMMA 6. Expose the tree $R_{\lceil \frac{i}{100} \rceil}$ consisting of the first $\lceil \frac{i}{100} \rceil$ levels of R_∞. Let B be the set of $2^{\lceil \frac{i}{100} \rceil}$ leaves of $R_{\lceil \frac{i}{100} \rceil}$. Let B' be the subset $\{x | n'_x > n\exp(-\frac{i}{5})\}$ of B. Since $-ln\frac{n'_x}{n}$ is gamma distributed with parameter $\lceil \frac{i}{100} \rceil$, its expected value is $\lceil \frac{i}{100} \rceil$ and the probability it exceeds $\frac{i}{5}$ is $o(2^{\frac{-i}{100}})$ (once again, this is easy to prove by considering the density function). So, $\mathbf{E}(|B - B'|) = o(1)$, and $\mathbf{P}(|B - B'| > 2^{\frac{i}{100}-1}) \leq 2^{1-\frac{i}{100}}$. Thus,

$$\mathbf{P}\left(|B'| < 2^{\frac{i}{100}-1}\right) \leq 2^{1-\frac{i}{100}}. \tag{17}$$

Now, for each $x \in B'$, we let A_x be the event that there is a node of the subtree of T'_n rooted at x with depth $h' - i$. By Lemma 5, we see that for some fixed $\epsilon > 0$, we have $\mathbf{P}(A_x) \geq \epsilon$. Furthermore, having exposed the labels on $R_{\lceil \frac{i}{100} \rceil}$, the events

A_x and A_y are independent for distinct x and y. Thus,

$$\mathbf{P}(H'_n < h' - i \mid |B'| = l) \leq (1-\epsilon)^l.$$

Combining this result with (17) yields the lemma. □

PROOF$_i$ OF LEMMA 7. We let $h = h'_n + i$. We will show that $\mathbf{P}(X_{n,h} \neq \emptyset) = O(\frac{2^{\frac{i}{2}}}{h}\mathbf{E}(|X_{n,h}|) + \frac{1}{2^{\frac{i}{2}}})$. By Lemma 1 (part (a) for $i < h^*_n - h'_n$, part (b) for larger i), this implies the desired result. Clearly, Lemma 1(b) permits us to assume $i \leq 2 \log h$.

We set $d = \lceil 4 \log h \rceil$, we let X_d be those nodes in $X_{n,h}$ which are *not* above $-d$ and Y_d be those which are above $-d$. We consider X_d and Y_d separately.

If X_d is not empty, then there is a node t of depth $l < h$ such that $n'_t \geq n^{(h-l)/h}h^4$. The number of such nodes is distributed like the number of nodes of $T_{\frac{n^{l/h}}{h^4}}$ at height l. Since $h > \alpha ln(n) - 2ln\ ln\ n$, we have that $\frac{n^{l/h}}{h^4} = O(\frac{e^{l/\alpha}}{h^3})$. So, applying Lemma 1, we obtain that the expected number of such nodes at height l is $O(\frac{1}{h^3})$. Since there are fewer than h choices for l, it follows that $\mathbf{P}(X_d \neq \emptyset) = O(\frac{1}{h^2})$. Since $i \leq 2 \log h$, this yields

$$\mathbf{P}(X_d \neq \emptyset) = O\left(\frac{1}{2^{i/2}}\right).$$

So, it remains to show

$$\mathbf{P}(Y_d \neq \emptyset) = O\left(\frac{2^{i/2}}{h}\mathbf{E}(|X_{n,h}|) + \frac{1}{2^{i/2}}\right).$$

Now,

$$\mathbf{P}(Y_d \neq \emptyset) \leq \mathbf{E}(|Y_d|) \leq O\left(\frac{d^4}{h}\mathbf{E}(|X_{n,h}|)\right),$$

where we obtain the last inequality by applying (14) (our bound on i implies we have $\sigma h < 5h$; this plus our bound on d allows us to apply (14), (15), and Lemma 1(a) here and below). So, we can assume $i \leq 10 \log d$ or we are done.

Now, we set $b = \lfloor 2^{\frac{i}{10}} \rfloor$, and let X_b be the set of nodes in $X_{n,h}$ which are above $-b$. We also consider separately for all integers a between b and d, the set X_a consisting of those elements of $X_{n,h}$, which are above $-a$ but not above $1-a$. Applying (14) as before, we obtain: $\mathbf{P}(X_b \neq \emptyset) = O(\frac{2^{\frac{i}{2}}}{h}\mathbf{E}(|X_{n,h}|))$. We shall show that for each a between b and d, $\mathbf{P}(X_a \neq \emptyset) = O(\frac{1}{h(a+1)^2}\mathbf{E}(|X_{n,h}|) + \frac{1}{2^{\frac{i}{2}}(a+1)^2})$. The desired result follows.

So, fix an a between b and d. By (15), we need only prove that $\mathbf{P}(Y_a \neq \emptyset) = O(\frac{1}{h(a+1)^2}\mathbf{E}(|X_{n,h}|) + \frac{1}{2^{\frac{i}{2}}(a+1)^2})$ where Y_a consists of those elements x of X_a satisfying:

$$\nexists k \ \textit{such that}\ (a-1)^{40} \leq k \leq h - (a-1)^{40}\ \textit{with}\ PS_k < 0.$$

Thus, in particular, for $c = (a-1)^{40}$, the minimum PS_k is achieved on some k with $k < c$ or $k > h - c$. For each k with $k < c$ or $k > h - c$, we let y_k be the random number of elements of Y_a for which PS_i is minimized at $i = k$.

Step 1: $\mathbf{P}(\exists k < c \text{ such that } y_k \neq 0) \leq O(\frac{1}{2^{\frac{i}{2}}(a+1)^2})$.

It is enough to show that, for each k with $k < c$, we have:

$$\mathbf{P}(y_k \neq 0) = O\left(\frac{1}{2^{\frac{i}{2}}(a+1)^{42}}\right).$$

For each node t at height $k < c$, we let W_t be the event that there is a descendant of t in Y_a whose partial sum is minimized at PS_k. We note that if W_t holds then $n'_t > \exp(a-1)n^{(h-k)/h}$, that is $f(t) \geq \exp(a-1)n^{-k/h}$. Thus, setting $n_0 = \exp(1-a)n^{k/h}$, we have:

$$\mathbf{P}(y_k \neq 0) = \mathbf{P}(X_{n_0,k} \neq \emptyset) \leq \mathbf{E}(|X_{n_0,k}|).$$

Thus, we need only show that this expected value is at most $O(\frac{1}{2^{\frac{i}{2}}(a+1)^{42}})$.

Now, $h^*_{n_o} = \alpha\ ln\ n_0 - \frac{1}{2}\log_{\frac{\alpha}{2}}\ ln\ n_0 + O(1)$. Further, $ln\ n_0 = \frac{kln\ n}{h} - a + O(1)$ and since $i \leq 2\log h$, $h = \alpha\ ln\ n + O(ln\ ln\ n)$. Thus, $ln\ n_0 = \frac{k}{\alpha} + O(\frac{k\ ln\ ln\ n}{h}) - a + O(1)$. Since $k < a^{40} < (5\log h)^{40}$, we have: $ln\ n_0 = \frac{k}{\alpha} - a + O(1)$. Thus, $h^*_{n_o} = k - \alpha a - \frac{1}{2}\log_{\alpha/2}(k - a\alpha) + O(1)$. Applying Lemma 1(b), we see that $\mathbf{E}(|X_{n_0,k}|) = O(2^{-\alpha a + 1/2\log_{\frac{\alpha}{2}}(k-a\alpha)})$. Since $k < a^{40}$, $\frac{1}{2}ln_{\frac{\alpha}{2}}(k - a\alpha) = o(a)$. Thus, as $a \geq b \geq \lceil 2^{\frac{i}{10}} \rceil$, $\mathbf{E}(|X_{n_0,k}|) = O(\frac{1}{2^{\frac{i}{2}}(a+1)^{42}})$, as required. □

Step 2. $\mathbf{P}(\exists k > h - c$ such that $y_k \neq 0) = O(\frac{\mathbf{E}(|X_{n,h}|)}{h(a+1)^2})$.

It is enough to show that for each k with $k > h - c$, we have: $\mathbf{P}(y_k \neq 0) = O(\frac{\mathbf{E}(|X_{n,h}|)}{h(a+1)^{42}})$. For each node t at depth $k > h - c$, we let W_t be the event that there is at least one descendant of t in Y_a whose partial sum is minimized at PS_k. We note that if W_t holds then $n'_t > \exp(a-1)\,n^{\frac{h-k}{h}}$. That is, $f(t) \geq \exp(a-1)\,n^{\frac{-k}{h}}$. Furthermore, given that $f(t)$ is this high, if we consider the normalized labels on the path from the root to t we see that if W_t holds then corresponding permutation must be above $-a-1$, as otherwise we contradict our choice of k.

By (14), the probability the latter event holds (given the first) is $O(\frac{(a+1)^4}{k})$. Thus, setting $n_0 = n^{\frac{k}{h}}\exp(1-a)$ we see that $\mathbf{P}(y_k \neq 0) = O(\mathbf{E}(|X_{n_0,k}|)\frac{(a+1)^4}{k})$. Now, $k > h - c > h - a^{40}$. So, since $a \leq d = \lceil 4\ \log\ h \rceil$, we have that $k - h = o(\sqrt{lnn})$. In particular, $\frac{k}{h} = O(1)$. Thus, to complete this step, we need only show:

$$\mathbf{E}(|X_{n_0,k}|) = O(\mathbf{E}(|X_{n,h}|)(a+1)^{-46}).$$

Now, $h - h^*_n = O(\log h) = o(\sqrt{\log n})$. Thus, by Lemma 1(a), $\mathbf{E}(|X_{n,h}|) \geq c_2(\frac{2}{\alpha})^{h-h^*_n}$, for some constant c_2. Furthermore, we claim that $h^*_n - h^*_{n_0} = o(\sqrt{\log n})$ and so $k - h^*_{n_0}$ is also $o(\sqrt{\log n})$. Given this claim, by Lemma 1(a), $\mathbf{E}(|X_{n_0,k}|) \leq c_3(\frac{2}{\alpha})^{k-h^*_{n_0}}$, for some constant c_3. Thus,

$$\mathbf{E}(|X_{n_0,k}|) = O\left(\mathbf{E}(X_{n,h})\left(\frac{2}{\alpha}\right)^{k-h+h^*_n-h^*_{n_0}}\right).$$

We show that $k - h + h^*_n - h^*_{n_0} = \alpha a + O(1)$. Since $\alpha a > 46\log(a+1) + O(1)$ this completes our proof of this step. Further, since $a = o(\sqrt{ln\ n})$ by definition, and we proved above that $h - k = o(\sqrt{ln\ n})$, this result yields $h^*_n - h^*_{n_0} = o(\sqrt{ln\ n})$ as claimed above.

Thus, it only remains to show: $k - h + h_n^* - h_{n_0}^* = \alpha a + O(1)$. To begin, we note that

$$h_n^* = \alpha ln\, n - \log_{\alpha/2} ln\, n + O(1)$$

and

$$h_{n_0}^* = \alpha ln\, n \left(1 - \frac{h-k}{h}\right) - \alpha a - \log_{\frac{\alpha}{2}}\ ln\, n_0 + O(1).$$

We proved above that $h - k = o(\sqrt{ln\, n})$. By definition $a = O(\sqrt{ln\, n})$. Taking these two facts together we obtain that $|ln\, n_0 - ln\, n\} = o(ln\, n)$. This yields:

$$h_n^* - h_{n_0}^* = \alpha ln\, n \left(\frac{h-k}{h}\right) + \alpha a + O(1).$$

Manipulating our formula for $h_n^* - h_{n_0}^*$ further by substituting in $h = \alpha ln\, n + o(\sqrt{ln\, n})$ gives:

$$h_n^* - h_{n_0}^* = h - k + o(\sqrt{ln\, n}) \left(\frac{h-k}{h}\right) + \alpha a + O(1).$$

Again applying $h - k = o(\sqrt{ln\, n})$ we obtain:

$$h_n^* - h_{n_0}^* = h - k + \alpha a + O(1).$$

This is the statement we needed to prove.

8. *A Stronger Result*

We turn now to the height of T_n. Since T_n' is at least as tall as T_n we have as an immediate corollary of Lemma 7 that

LEMMA 8. *There is a (universal) constant C_2 such that for $i > 0$ with $i = o(log\, n)$ we have:*

$$\mathbf{P}(H_n \geq h_n' + i) < C_2 2^{\frac{-i}{2}}.$$

We show below that

LEMMA 9. *There is a (universal) constant C_3 such that the probability that T_n contains a node at height $h_n' - C_3$ is $\Omega(1)$.*

With this result in hand, as shown below, it is straightforward to mimic the proof of Lemma 6 using Lemma 9 in the proof of Lemma 5 to obtain:

LEMMA 10. *There exist (universal) constants C_1, C_3 and $d < 1$ such that for $i > 0$ with $i = o(\sqrt{\log n})$ we have:*

$$\mathbf{P}(H_n < h_n' - C_3 - i) < C_1 d^i.$$

Theorem 1 follows immediately from these results, in the same way that Theorem 2 followed from Lemmas 5, 6, and 7.

So, it remains only to prove Lemma 9 and Lemma 10.

PROOF OF LEMMA 9. In comparing T_n to T_n', we need to consider the difference between n_z and n_z' which arises as a result of rounding. We denote this difference

by Δ_z. We note that if z has w as a parent and L_{wz} is the label of the edge zw then Δ_z is the sum of $L_{wz}\Delta_w$ and $n_w L_{wz} - \lfloor n_w L_{wz} \rfloor$. This observation is the key to dealing with the difference between T_n and T'_n.

We begin the proof with a simple example illustrating how we can use this observation. Note first that $n_w L_{zw} - \lfloor n_w L_{wz} \rfloor$. is at most 1. Thus, if Δ_w is at most 2 and L_{wz} is less than a half, then Δ_z is less than $\frac{1}{2}(2) + 1 = 2$. Applying this fact inductively, we see that if every label on P_x is less than $\frac{1}{2}$ then $n_z - n'_z \leq 2$ for every z on P_x. Now, if we also have that x is in T'_n then $n'_x \geq 1$ and this implies that the grandfather w of x satisfies $n'_w \geq 4$ because of our upper bound on the label values on P_x. Thus, $n_w \geq 4 - \Delta_w \geq 2$ and w is in T_n.

More generally, we can show that, if there are enough *small* labels spread along P_x, then the round-off errors will not accumulate and so if x is in T'_n there is an ancestor at distance $O(1)$ from x which is in T_n.

To complete the proof, we need to show that the probability we have such a node x at height h'_n in T'_n is $\Omega(1)$. To do so, we essentially apply the proof of Lemma 5, showing that the requirements we need to impose on P_x so as to bound the roundoff errors are so loose that the proof still goes through.

We now turn to the details. We set $h = h'_n$. In the last section, we considered the subset $A_{n,h}$ consisting of those good x in $X_{n,h}$ with $n'_x < 2$ such that the corresponding normed permutation is nonnegative and satisfies for a universal constant C,

$$\forall\ k \text{ with } C \leq k \leq \frac{h}{2} : (PS_k > (6\log k)^7) \cap (PS_{n-k} > 6(\log k)^7).$$

We showed that the probability that $A_{n,h}$ is nonempty is $\Omega(1)$. We now consider the set $B_{n,h}$ consisting of those $x \in X_{n,h}$ such that $n'_x < 2$ and the corresponding normed permutation satisfies for some universal constant d,

(i) $\forall\ k \text{ with } 2^d \leq k \leq \frac{h}{2} : (PS_k > 6\log k) \cap (PS_{n-k} > 6\log k)$

(ii) x *is good*,

(iii) $\forall\ j \text{ with } d \leq j \leq \lfloor \log \frac{h}{2} \rfloor, \quad \prod_{i=h-2^j}^{h-2^{j-1}-1} l_i < \frac{1}{4}$

We will show that, for some constant D:

$$\text{the probability that } B_{n,h} \text{ is nonempty is at least } \frac{1}{D}. \tag{18}$$

Using this fact, we will prove the following restatement of Lemma 9:

$$\begin{gathered}\text{There is a constant } C_3 \text{ such that the probability that} \\ H_n \text{ is at least } h - C_3 \text{ is at least } \frac{1}{D}.\end{gathered} \tag{19}$$

PROOF OF (19). We can assume h is large as otherwise by increasing C_3 we can make $h - C_3$ negative. We actually prove the stronger statement that if $x \in B_{n,h}$ then the $h - C_3$th node on the path P_x from the root to x is in T_n.

Recall that the path P_x from the root to x is enumerated as $z_0, \ldots, z_h$. We let C_3 be 2^{2d+4}. Since $n'_{z_{h-2^d}} > 1$, it follows immediately from (iii) that $n'_{h-C_3} \geq 4^{d+4} = 2^{2d+8}$. We will show that in addition, $\Delta_{z_{h-C_3}} < 2^{2d+6}$ thereby proving that z_{h-C_3} is in T_n. To do so, we must consider how large the difference Δ_{z_i} between n_{z_i} and n'_{z_i} can become for the z_i on P_x as the rounding errors accumulate.

For each integer j between d and $j^* = \lfloor \log \frac{h}{2} \rfloor$, we claim that $\Delta_{z_{h-2^j}} < 2^{j+2}$. This claim applied to $j = 2d + 4$ implies the desired result.

We prove our claim by backwards induction on j. Note first that $\Delta_{z_{i+1}} - \Delta_{z_i} \leq 1$. Thus, for any $i \leq h$, $\Delta_{z_i} < h$. Hence, the result holds for j^*. Now, we assume that the result holds for $j+1$ for some j between d and j^* and prove it for j. The new rounding errors introduced in the pivoting between $z_{h-2^{j+1}}$ and z_{h-2^j} are at most 2^j. The effect of the old rounding errors is multiplied by $\frac{1}{4}$ by (iii). Thus, the new rounding error is at most $\frac{2^{j+3}}{4} + 2^j < 2^{j+2}$. □

So, it remains to prove (18).

PROOF OF (18). The proof follows the lines of that of Lemma 5. We let C be as in the definition of $A_{n,h}$ in that lemma, and choose a constant d with $d > \log C$ which is large enough to satisfy some implicit inequalities below. We recall that $j^* = \dagger \lfloor \log \frac{h}{2} \rfloor$. We first show that the expected number of elements in $B_{n,h}$ is $\Omega(1)$. To complete the proof, we need to show that for any x at depth h, and any choices for the labels on P_x which show that x is in $B_{n,h}$, the expected size of $B_{n,h}$ is $O(1)$. To do so, we simply apply (16) with $C^* = C$. So, we need only show that the expected number of elements in $B_{n,h}$ is $\Omega(1)$.

To prove this bound on the expectation, we show that the expected size of $B_{n,h}$ is not much smaller than that of $A_{n,h}$. To do so, we consider a random element x of $Y_{n,h}$ and consider generating a random permutation Π of the multiset of labels S on P_x via a 2-step process. We first generate a random permutation Π_1 of S uniformly. We then swap some of the elements of Π to obtain a new uniform permutation Π_2. We specify this swapping procedure precisely in a moment. We let $\mathcal{B}$ be the set of permutations of S for which x is in $B_{n,h}$ and let $\mathcal{A}$ be the set of permutations of S for which x is in $A_{n,h}$. We show that the probability that Π_2 is in $\mathcal{B}$ given that Π_1 is in $\mathcal{A}$ is $\Omega(1)$. This yields the desired result, as it shows that $\mathbf{E}(|B_{n,h}|) = \Omega(1)\mathbf{E}(|A_{n,h}|)$ and we proved in the preceding section that $\mathbf{E}(|A_{n,h}|) = \Omega(1)$.

Our swapping procedure has the following form:

(1) we pick a uniformly random set of $\binom{j^*+1}{2} - \binom{d+1}{2}$ integers $k_{\binom{d+1}{2}+1}$, $k_{\binom{d+1}{2}+2}, \ldots, k_{\binom{j^*+1}{2}}$ between $\frac{h}{4}$ and $\frac{h}{2} - 1$,

(2) we swap the values in $\{a_{k_{\binom{d+1}{2}+1}}, \ldots, a_{k_{\binom{j^*+1}{2}}}\}$ with the values

$$\{a_{h-2^{d+1}+1}, a_{h-2^{d+1}+2}, a_{h-2^{d+1}+d+1}, \ldots, a_{h-2^j+1}, \ldots, a_{h-2^j+j}, \ldots, a_{h-2^{j^*}+j^*}\}.$$

where, for $d+1 \leq j \leq j^*$ and $1 \leq r \leq j$, a_{h-2^j+r} is swapped with $a_{k_{\binom{j}{2}+r}}$.

How do these swaps affect the values of the PS_k??

For $k < \frac{h}{4}$ and $k > h - 2^d$, they have no effect. For k between $\frac{h}{4}$ and $h - 2^d$, let S_k^1 be the set of labels that appeared at or before the kth position and have now been swapped so they appear after it and let S_k^2 be the set of those labels which appeared after the kth position and have now been swapped so they appear at or before it. Here we are thinking of the labels as the normalized negative logarithms. We have that the decrease in PS_k is

$$\sum_{x \in S_k^1} x - \sum_{x \in S_k^2} x.$$

Now, we consider k with $\frac{h}{4} \leq k < \frac{h}{2}$. Note that $|S_k^1| + |S_k^2| \leq 2\binom{j^*+1}{2} \leq (\log\ h)^2$. Clearly, since each l_i is at most 1, each a_i is at least -1. Hence, $\sum_{x \in S_k^2} x \geq -(\log\ h)^2$. Suppose that:

$$\nexists i \text{ such that } a_{k_i} > (\log h)^3. \tag{20}$$

Then, for all such k, the decrease in PS_k is less than $((\log h)^3 + 1))(\log h)^2$, and hence PS_k still exceeds $6 \log\ k$.

For k with $\frac{h}{2} \leq k \leq h - C$, we define d_k to be $h - k$, we note that $|S_k^1|, |S_k^2| \leq (\log\ d_k)^2$. Hence, $\sum_{x \in S_k^2} x \geq -(\log\ d_k)^2$. So provided

$$\sum_{x \in S_k^1} x \leq 5(\log\ d_k)^7, \tag{21}$$

PS_k will still exceed $6\ \log\ d_k$.

In order for (iii) in the definition of $B_{n,h}$ to hold for x, it is enough to insist that

$$\textit{for } j \textit{ with } d+1 \leq j \leq j^* : \prod_{i=\binom{j}{2}+1}^{\binom{j+1}{2}} l_{k_i} \leq \frac{1}{4} \tag{22}$$

where the l_i are the $[0\,..\,1]$ uniforms corresponding to S and their indices refer to Π_1.

Thus, if (20) and (22) hold and (21) holds for all k, then $\Pi_2 \in \mathcal{B}$. We now show that the intersection of these three events occurs with probability $\Omega(1)$ thereby completing the proof.

To begin, we show that the probability that (20) fails is o(1). This probability is at most $(\log h)^2 \mathbf{P}(a_{i^*} \geq (\log h)^3)$ where i^* is a random index between $\frac{h}{4}$ and $\frac{h}{2}$. Since $\sum_{i=\frac{h}{4}}^{\frac{h}{2}} a_i \leq ln\ n$, this latter probability is $O(\frac{4ln\ n}{h(\log h)^3}) = O((\log h)^{-3})$. So the probability that (20) fails is $O((\log h)^{-1})$.

To bound the probability that (21) fails for some k, we proceed in a similar manner. To begin, we note that since $|S_k^1| \leq (\log d_k)^2$, this probability is at most $(\log d_k)^2 \mathbf{P}(a_{i^*} \geq 5(\log d_k)^5)$ where i^* is a random index between $\frac{h}{4}$ and $\frac{h}{2}$. Since $\sum_{i=\frac{h}{4}}^{\frac{h}{2}} a_i \leq ln\ n$, this latter probability is $O(\frac{4ln\ n}{h(\log d_k)^5}) = O((\log d_k)^{-5})$. So the probability that (21) fails for a particular k is $O((\log d_k)^{-3})$.

Since (21) holds provided it holds for all all values of i between $\frac{h}{2}$ and $h - 2^d$ for which we swap a_i, we obtain by induction that for j with $d \leq j \leq j^*$

$$\begin{aligned}&\mathbf{P}((21)\ \textit{holds}\ \forall\ k\ \textit{with}\ h - 2^j \leq k \leq h - 2^d) \\ &\quad = O\left(\frac{j}{j^3}\right) + \mathbf{P}((21)\ \textit{holds}\ \forall\ k\ \textit{with}\ h - 2^{j-1} \leq k \leq h - 2^d).\end{aligned}$$

So

$$\mathbf{P}((21)\ \textit{holds}\ \forall\ k\ \textit{with}\ \frac{h}{2} \leq k \leq h - 2^d) = O\left(\sum_{j=d}^{\infty} \frac{1}{j^2}\right) = O(d^{-1}).$$

The proof of (22) is straightforward. Let E be the event that $|\{i \,|\, \frac{h}{4} \leq i \leq \frac{h}{2}\}, l_i > 1 - \frac{1}{5000}\}| > \frac{h}{8}$. The uncoditioned probability that some l_i is more than $1 - \frac{1}{5000}$ is $\frac{1}{5000}$. Thus, the unconditioned probability that E holds is $o(2^h)$. Since

the probability $x \in A_{n,h}$ is $\Omega(2^h)$, it follows that for $x \in A_{n,h}$ the probability E holds is $o(1)$. Given that $x \in A_{n,h}$ and E fails to hold, the probability that fewer than 15000 of the i between $\binom{j}{2}+1$ and $\binom{j+1}{2}$ satisfy $l_i < 1 - \frac{1}{5000}$ is $O(2^{-j})$. But clearly, this is an upper bound on the conditioned probability

$$\prod_{i=\binom{j}{2}+1}^{\binom{j+1}{2}} l_{k_i} > \frac{1}{4}.$$

So, the probability that (22) fails, given that $x \in A_{n,h}$ and E fails, is $O(\sum_{j=d}^{\infty} 2^{-j}) = O(2^{-d})$. Hence, the probability that (22) holds for some $x \in A_{n,h}$ is $1 - O(2^{-d})$.

The desired result follows as we are free to make d as large as we like. □

To complete the proof of Theorem 2, this section and the article, we present:

PROOF OF LEMMA 10. Expose the tree $R_{\lceil \frac{i}{100} \rceil}$ consisting of the first $\frac{i}{100}$ levels of R_∞. Let B be the set of $2^{\lceil \frac{i}{100} \rceil}$ leaves of $R_{\lceil \frac{i}{100} \rceil}$. Let B' be the subset $\{x | n'_x > n2^{-\frac{i}{5}}\}$ of B. Now, (17) states:

$$\mathbf{P}\left(|B'| < 2^{\frac{i}{100}-1}\right) \leq 2^{1-\frac{i}{100}}.$$

Note that for each $x \in B'$, Fact 2 implies that $n_x > n2^{\frac{-i}{5}} - \lceil \frac{i}{100} \rceil$. Thus $h'_{n_x} > h' - i$. For each such x, we let A_x be the event that there is a node of the subtree of T_n rooted at x with depth $h' - C_3 - i$. By Lemma 9, we see that for some fixed $\epsilon > 0$, we have $\mathbf{P}(A_x) \geq \epsilon$. Furthermore, having exposed the labels on $R_{\lceil \frac{i}{100} \rceil}$, the events A_x and A_y are independent for distinct x and y. Thus,

$$\mathbf{P}(H'_n < h' - C_3 - i \mid |B'| = l) \leq (1-\epsilon)^l.$$

Combining this result with (17) yields the lemma. □

ACKNOWLEDGMENTS. I would like to thank Luc Devroye and Colin McDiarmid for many long discussions about random binary search trees. The proof given here is significantly shorter than the first proof of this result I obtained. This is due in large measure to Luc's input. Both Luc and Colin also read versions of the manuscript and made helpful suggestions on the presentation. I would also like to thank three referees and Mike Saks for very helpful suggestions on the presentation of the results.

REFERENCES

AHO, A. V., HOPCROFT, J. E., AND ULLMAN, J. D. 1975. *The Design and Analysis of Computer Algorithms*. Addison-Wesley, Reading, Mass.

AHO, A. V., HOPCROFT, J. E., AND ULLMAN, J. D. 1983. *Data Structures and Algorithms*, Addison-Wesley, Reading, Mass.

CHERNOFF, H. 1952. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Ann. Math. Stat. 23*, 493–507.

DEVROYE, L. 1986. A note on the height of binary search trees. *J. ACM 33*, 489–498.

DEVROYE, L. 1987. Branching processes in the analysis of the heights of trees. *Acta Inf. 24*, 277–298.

DEVROYE, L. 1990. On the height of random m-ary search trees. *Rand. Struct. Algorithms 1*, 191–203.

DEVROYE, L., AND REED, B. 1995. On the variance of the height of random binary search trees. *SIAM J. Comput. 24*, 1107–1112.

DRMOTA, M. 2003. An analytic approach to the height of binary search trees II. *J. ACM 50*, 3 (May), 333–374.

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

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

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

MAHMOUD, H., AND PITTEL, B. 1994. On the most probable shape of a search tree grown from a random permutation. *SIAM J. Algeb. Disc. Meth. 5*, 69–81.

PITTEL, B. 1984 On growing random binary trees. *J. Math. Anal. Appl. 103*, 461–480.

PITTEL, B. 1994. Note on the heights of random recursive trees and random m-ary search trees. *Rand. Struct. Algorithms 5*, 337–347.

PYKE, R. 1965. Spacings. *J. Roy. Stat. Soc., Ser., B 7*, 395–445.

ROBSON, J. M. 1979. The height of binary search trees. *Austral. Comput. J. 11*, 151–153.

ROBSON, J. M. 1982. The asymptotic behaviour of the height of binary search trees. *Austral. Comput. Sci. Commun.* p. 88.

SHORACK, G., AND WELLNER, J. 1986. *Empirical Processes with Applications to Statistics*. Wiley, New York.

RECEIVED JUNE 2000; REVISED DECEMBER 2002; ACCEPTED DECEMBER 2002