Ann. Inst. Statist. Math.
Vol. 55, No. 4, 885–900 (2003)

ONE-SIDED VARIATIONS ON BINARY SEARCH TREES

HOSAM M. MAHMOUD

Department of Statistics, The George Washington University, Washington, D.C. 20052, U.S.A.,
e-mail: hosam@gwu.edu

(Received July 8, 2002; revised February 24, 2003)

Abstract. We investigate incomplete one-sided variants of binary search trees. The (normed) size of each variant is studied, and convergence to a Gaussian law is proved in each case by asymptotically solving recurrences. These variations are also discussed within the scope of the contraction method with degenerate limit equations. In an incomplete tree the size determines most other parameters of interest, such as the height and the internal path length.

Key words and phrases: Random tree, limit distribution, recurrence equation, contraction, degenerate limit.

1. Introduction

The binary search tree is a popular structure for data storage and for the analysis of algorithms. For instance, it is in common use as a data structure (see Mahmoud (1992)), and it underlies Quicksort, which is one of the most popular sorting algorithms (see Knuth (1998) and Mahmoud (2000)).

The study of pruned variants of full trees (incomplete or one-sided trees) has recently been a popular subject. For instance, Prodinger (1993) analyzes various parameters of the incomplete trie, a one-sided version of a random digital tree, and Fill *et al.* (1996) follow up with a proof for the non-existence of limit distributions for the height of the incomplete trie, owing to the presence of oscillations. Itoh and Mahmoud (2003) study several incomplete variants of interval trees. There is also revived interest in algorithms to cut down trees as in the recent probabilistic analysis in Chassaing and Marchand (2002), an area of research begun in the work of Meir and Moon (1970). The pruning alluded to is one way of cutting the binary search tree down to a single branch. The present study is also related to the notion of ascendants and descendants in search trees, which was taken up in Martínez *et al.* (1998).

In this study we investigate a few one-sided variations of binary search trees. For each incomplete variant we study the distribution of the size of that variant. In each variation we shall show Gaussian tendency of the size when appropriately normed. The study parallels the continuous analog in interval trees, with similar results, but a discrete methodology instead. The results of this paper, as well as those in other sources (for example Sibuya and Itoh (1987) and Drmota (2002)) indicate that there might be a deep method of embedding binary search trees into interval trees. The two classes of random trees have rather disparate stopping rules. This interesting subject is noted, but not pursued in this paper.

2. One-sided binary search trees

A binary tree is a structure of nodes each with no children, one left child, one right child, or two children (one left and one right). Many combinatorial algorithms, such as sorting, are represented and analyzed by labeled binary trees endowed with a *search property*. Several models of randomness are often used on binary trees. In the uniform model all binary trees are equally likely: This is the model prevalent in formal language studies, compilers, computer algebra, etc. (see Kemp (1984)).

In many practical situations trees grow from a permutation of an ordered set. If $\Pi_n = (\pi_1, \ldots, \pi_n)$ is the permutation, the binary search tree grows as follows. The element π_1 is placed in the root. If $\pi_2 < \pi_1$, the element π_2 goes to the left subtree, where it is placed in a node adjoined to the root. Otherwise, π_2 goes into the right subtree, where it is placed in a node adjoined to the root. Likewise, subsequent elements of the permutation are guided to the left or right subtree, according as whether they are not or are at least as large as π_1, and the element is then recursively inserted in the subtree.

The only aspect of the permutation Π_n that pertains to the construction of the tree is the relative ranking of its elements. Thus, Π_n can be assimilated by a permutation of $\{1, \ldots, n\}$. The *random permutation model* is the probability model often assumed for data structures and combinatorial algorithms, such as sorting. In the random permutation probability model, the tree is assumed to be built from a random permutation of $\{1, \ldots, n\}$, where all $n!$ permutations are equally likely. Under this model binary search trees are not equally likely. Generally, the model favors short bushy trees to tall linear shapes. The natural balance of binary search trees under the random permutation model is an attractive property for fast search applications (see Mahmoud (1992)). The random permutation model is fairly general, as it covers, for example, the ranks of data taken from *any* continuous distribution. The random permutation probability model is assumed in the sequel. Under the random permutation model ties occur with probability zero.

We shall interchangeably use the terms incomplete trees and one-sided trees. Different variants pursuing various pruning polices will be investigated. We shall discuss five incomplete variations in this paper:

- Left preference.
- Min preference.
- Max preference.
- Proportionate preference.
- Uniform (or no) preference.

The left preference search tree develops only the left subtree. The min (max) one-sided variant always derives a one-sided tree from what would be the subtree of the smaller (larger) size in the full binary search tree. This is a way of speeding up (slowing down) the incomplete tree construction. In proportionate preference, one of the two sides is chosen with probability proportionate to the size of the subtree. The uniform preference policy simply chooses one of the two subtrees with equal probability, regardless of their size in the binary tree. The rest of the paper is organized in sections. Each of the following sections is dedicated to one particular one-sided variation. In each of these sections, a more precise statement of the construction algorithm is given.

The size is random in each variation and we shall determine its limiting distribution, when appropriately normed. We shall show that in each variation the (normed) size

exhibits Gaussian tendency. The size in one-sided variants determines almost every property of interest. For example, while in a binary search tree knowledge of the size is not sufficient to determine the height, in an incomplete tree we have a simple connection: the size is the height plus 1.

We use the following notation. Within each section S_n will refer to the size of the particular one-sided flavor discussed in that section, and $\phi_n(t)$ will denote the moment generating function. For brevity, the notation S_n and $\phi_n(t)$ will be reused in a different way in each section. The symbols $\stackrel{\mathcal{D}}{=}$, $\stackrel{\mathcal{D}}{\to}$, and $\stackrel{P}{\to}$ stand respectively for equality in distribution, convergence in distribution, and convergence in probability, and $\mathcal{N}(\mu, \sigma^2)$ will represent a normally distributed random variable with mean μ and variance σ^2. Because of the discrete nature of the incomplete search trees, some distributions that will appear in the study are conveniently given in terms of the **mod** notation of programming languages. The notation $n \bmod k$ will assume the integer value of the remainder in the integer division of n by k. The indicator $\mathbf{1}_{\mathcal{E}}$ is a function that assumes the value 1 if event $\mathcal{E}$ occurs, otherwise the indicator is 0.

3. Left preference incomplete search trees

In this variation only the left side is developed. A left preference incomplete tree T_n arising from a random permutation of $\{1, \ldots, n\}$ is the binary search tree with each right subtree replaced by an empty tree. The algorithm for this construction operates as follows. The first element of the permutation, say R_n, is placed in the root of the tree. Each subsequent element of the permutation exceeding R_n is thrown out (resulting in an empty right subtree), and each subsequent member not exceeding R_n is taken into the left subtree, where recursively the tree T_{R_n-1} is attached on the left side as the only subtree of the root. Note that if the permutation Π_n is stripped of the elements $\{R_n, R_n+1, \ldots, n\}$, the order of the remaining elements (all less than R_n) forms a random permutation on $\{1, \ldots, R_n - 1\}$. The construction in T_{R_n-1} continues recursively until T_0, an empty subtree, is to be attached on the left.

The size S_n of the tree satisfies the recurrence

$$S_n = 1 + S_{R_n-1}, \qquad \text{for} \quad n \geq 1, \tag{3.1}$$

with the boundary condition $S_0 = 0$. Let $\phi_n(t)$ be the moment generating function $\boldsymbol{E}[e^{S_n t}]$. Under the random permutation model, the root label R_n is uniformly distributed on $\{1, \ldots, n\}$. Then for $n \geq 1$,

$$\begin{aligned}
\phi_n(t) &= \boldsymbol{E}[e^{(1+S_{R_n-1})t}] \\
&= e^t \sum_{r=1}^{n} \boldsymbol{E}[e^{S_{R_n-1}t} \mid R_n = r] P\{R_n = r\} \\
&= \frac{e^t}{n} \sum_{r=1}^{n} \phi_{r-1}(t).
\end{aligned}$$

Difference a version of the last telescoping recurrence, with $n-1$ replacing n, from the version with n to obtain

$$n\phi_n(t) - (n-1)\phi_{n-1}(t) = e^t \phi_{n-1}(t).$$

This recurrence can be iterated to produce the exact distribution:

$$\begin{aligned}\phi_n(t) &= \frac{n+e^t-1}{n}\phi_{n-1}(t)\\ &= \frac{(n+e^t-1)(n+e^t-2)}{n(n-1)}\phi_{n-2}(t)\\ &\vdots\\ &= \frac{(n+e^t-1)(n+e^t-2)\dots e^t}{n!}\phi_0(t).\end{aligned}$$

Using the boundary condition $S_0 = 0$, we can represent the solution in a Gamma function form suitable for asymptotics.

PROPOSITION 3.1. *Let S_n be the size of a random left preference incomplete search tree grown from a random permutation of $\{1,\dots,n\}$. The exact moment generating function of S_n is given by*

$$\phi_n(t) = \frac{\Gamma(n+e^t)}{\Gamma(e^t)\Gamma(n+1)}.$$

The exact (and consequently asymptotic) first two moments of S_n can immediately be computed from the exact distribution of Proposition 3.1 by taking derivatives with respect to t (at $t = 0$). The computation of the variance is a bit lengthy, but remains straightforward. The computation of the mean recovers an old result of Arora and Dent (1969); the variance can be found in Martínez *et al.* (1998). These papers considered the size of the left preference incomplete search tree in the context of the length of the so-called left arm of a random binary search tree (the depth of the node containing 1). For completeness we present these results. The result is compact when written in terms of the first and second degree harmonic numbers $H_n = \sum_{j=1}^n 1/j$, and $H_n^{(2)} = \sum_{j=1}^n 1/j^2$.

COROLLARY 3.1.

$$\begin{aligned}\boldsymbol{E}[S_n] &= H_n \sim \ln n,\\ \mathbf{Var}[S_n] &= H_n - H_n^{(2)} \sim \ln n.\end{aligned}$$

The limit distribution for an appropriately normed S_n can be found from Gamma function asymptotics and local expansions. Let $t = v/\sqrt{\ln n}$, for fixed v, and eventually let $n \to \infty$. By the Stirling approximation of the Gamma function we can rewrite the exact result in Proposition 3.1 in the form

$$\begin{aligned}\boldsymbol{E}\left[\exp\left\{S_n\frac{v}{\sqrt{\ln n}}\right\}\right] &= \frac{n^{\exp\{v/\sqrt{\ln n}\}-1}}{\Gamma(\exp\{v/\sqrt{\ln n}\})}\left(1+O\left(\frac{1}{n}\right)\right)\\ &= \exp\left(\left[\left(1+\frac{v}{\sqrt{\ln n}}+\frac{v^2}{2!\ln n}+O((\ln n)^{-3/2})\right)-1\right]\ln n\right)\\ &\quad\cdot\left(1+O\left(\frac{1}{n}\right)\right).\end{aligned}$$

As $n \to \infty$,

$$\boldsymbol{E}\left[\exp\left\{\frac{S_n - \ln n}{\sqrt{\ln n}} v\right\}\right] \to e^{v^2/2}.$$

The right-hand side is the moment generating function of $\mathcal{N}(0,1)$, and we have convergence in distribution by Lévy's continuity theorem.

THEOREM 3.1. *Let S_n be the size of a left preference incomplete search tree grown from the random permutation $\{1,\ldots,n\}$. As $n \to \infty$,*

$$\frac{S_n - \ln n}{\sqrt{\ln n}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1).$$

The asymptotic distribution of Theorem 3.1 as well as the exact and asymptotic mean and variance of Corollary 3.1 can also be obtained from the theory of records in the manner discussed in Devroye (1988) in the context of recursive trees.

4. Min incomplete search tree

The min strategy is meant to speed up the construction of the incomplete search tree. In this strategy only the subtree with minimal size in the binary search tree is recursively developed. If the two sides are of equal size, we can arbitrarily choose either one. We shall stick to a left choice in breaking ties, but all other flavors are mathematically equivalent from the point of view of parameters such as the probability distribution of the size of the incomplete tree.

A min incomplete search tree T_n is grown from a permutation of $\{1,\ldots,n\}$ by first allocating a root for R_n, the first element in the permutation. Let $Z_n = \min\{R_n - 1, n - R_n\}$. If $Z_n = R_n - 1$, the right subtree remains empty, and the left grows a tree T_{Z_n} recursively on the permutation $\{1,\ldots,R_n - 1\}$. Else, $Z_n = n - R_n$, and the left subtree remains empty; the right grows a tree T_{Z_n} recursively on the permutation $\{R_n + 1,\ldots,n\}$. The size S_n of the min tree satisfies the recurrence

$$S_n = 1 + S_{Z_n}, \quad \text{for} \quad n \geq 1, \tag{4.1}$$

with the boundary condition $S_0 = 0$. Let $\phi_n(t)$ be the moment generating function of S_n. We shall develop asymptotic relations for $\phi_n(t)$ by conditioning on Z_n. The variable Z_n "almost" has a uniform distribution. That is, its distribution on $\{0,\ldots,\lfloor \frac{1}{2}(n-1)\rfloor\}$ is uniform except for a possible perturbation at the upper end of the distribution. More precisely, $F_{Z_n}(z)$, the distribution function of Z_n, is given by

$$F_{Z_n}(z) = 1 - P\{Z_n > z\} = 1 - P\{R_n - 1 > z, n - R_n > z\}.$$

In the range $z = 0,1,\ldots,\lfloor \frac{1}{2}(n-1)\rfloor - 1$ this yields

$$\begin{aligned} F_{Z_n}(z) &= 1 - P\{z+1 < R_n < n - z\} \\ &= 1 - P\{z+2 \leq R_n \leq n - z - 1\} \\ &= 1 - \frac{n - 2z - 2}{n} \\ &= \frac{2}{n}(z+1). \end{aligned}$$

We thus have

$$P\{Z_n = z\} = F_{Z_n}(z) - F_{Z_n}(z-1) = \frac{2}{n}, \quad \text{if} \quad z \in \left\{1, \ldots, \left\lfloor \frac{1}{2}(n-1) \right\rfloor - 1\right\}.$$

This expression is also valid at the uppermost point $\lfloor \frac{1}{2}(n-1) \rfloor$, if n is even. However if n is odd, the range $z+2$ to $n-z-1$ is empty at $z = \lfloor \frac{1}{2}(n-1) \rfloor$. Instead, we have $F_{Z_n}(\lfloor \frac{1}{2}(n-1) \rfloor) = 1$, and it follows that

$$P\left\{Z_n = \left\lfloor \frac{1}{2}(n-1) \right\rfloor\right\} = \frac{1}{n}.$$

Using the **mod** notation we have

$$P\{Z_n = z\} = \frac{2}{n} - \frac{1}{n}(n \bmod 2)\mathbf{1}_{\{z = \lfloor \frac{1}{2}(n-1) \rfloor\}},$$

in the entire range $n = \{0, \ldots, \lfloor \frac{1}{2}(n-1) \rfloor\}$. We can now formulate a recurrence by conditioning on Z_n:

$$\begin{aligned}
\phi_n(t) &= \boldsymbol{E}[e^{(1+S_{Z_n})t}] \\
&= e^t \sum_{z=0}^{\lfloor \frac{1}{2}(n-1) \rfloor} \boldsymbol{E}[e^{S_{Z_n} t} \mid Z_n = z] P\{Z_n = z\} \\
&= \frac{e^t}{n} \sum_{z=0}^{\lfloor \frac{1}{2}(n-1) \rfloor} \phi_z(t)(2 - (n \bmod 2)\mathbf{1}_{\{z = \lfloor \frac{1}{2}(n-1) \rfloor\}}) \\
&= \frac{2e^t}{n} \sum_{z=0}^{\lfloor \frac{1}{2}(n-1) \rfloor} \phi_z(t) - \frac{e^t}{n}(n \bmod 2)\phi_{\lfloor \frac{1}{2}(n-1) \rfloor}(t).
\end{aligned}$$

We can get rid of the telescoping sum by differencing. If n is even, this yields

$$n\phi_n(t) - (n-1)\phi_{n-1}(t) = e^t \phi_{\frac{1}{2}(n-2)}(t),$$

and if n is odd, the differencing yields

$$n\phi_n(t) - (n-1)\phi_{n-1}(t) = e^t \phi_{\frac{1}{2}(n-1)}(t).$$

For all parities of n, we get the recurrence

$$n\phi_n(t) - (n-1)\phi_{n-1}(t) = e^t \phi_{\lfloor \frac{1}{2}(n-1) \rfloor}(t). \tag{4.2}$$

The asymptotics of the solution in the left preference incomplete tree come from Stirling approximation to gamma function. This suggests an asymptotic solution for the present recurrence. For fixed t, we can try the asymptotic form

$$\phi_n(t) = c(t) n^{g(t)} \left(1 + \frac{K_t}{n} + O\left(\frac{1}{n^2}\right)\right), \quad \text{as} \quad n \to \infty,$$

for functions $c(t)$, $g(t)$, and K_t that depend only on t. The constant in O must also depend on t. Whatever they are, $g(t)$, K_t and the O constant must all approach 0, as

$t \to 0$, and $c(t) \to 1$, as $t \to 0$, to meet the requirement $\phi_n(t) \to 1$, as $t \to 0$, for all $n \geq 1$. Indeed, this can be a unique asymptotic solution if

$$c(t)n^{g(t)+1}\left(1+\frac{K_t}{n}+O\left(\frac{1}{n^2}\right)\right)-c(t)(n-1)^{g(t)+1}\left[1+\frac{K_t}{n-1}+O\left(\frac{1}{(n-1)^2}\right)\right]$$
$$=e^t c(t)\left\lfloor\frac{1}{2}(n-1)\right\rfloor^{g(t)}\left(1+\frac{K_t}{\left\lfloor\frac{1}{2}(n-1)\right\rfloor}+O\left(\frac{1}{n^2}\right)\right).$$

In expanded form, this relation is

$$n^{g(t)+1}+K_t n^{g(t)}+O(n^{g(t)-1})-[n^{g(t)+1}-(g(t)+1)n^{g(t)}+O(n^{g(t)-1})]$$
$$\times\left(1+\frac{K_t}{n}+O\left(\frac{1}{n^2}\right)\right)$$
$$=e^t\left[\left(\frac{n}{2}\right)^{g(t)}+O(n^{g(t)-1})\right]\left(1+\frac{2K_t}{n}+O\left(\frac{1}{n^2}\right)\right).$$

When simplified, this reads

$$(g(t)+1)n^{g(t)}+O(n^{g(t)-1})=e^t\left(\frac{n^{g(t)}}{2^{g(t)}}+O(n^{g(t)-1})\right).$$

This is possible if $g(t)$ is the solution to the equation

$$2^{g(t)}(g(t)+1)=e^t.$$

This implicit equation is similar but not identical to that which appeared in Itoh and Mahmoud's (2003) investigation on min incomplete interval trees. The solution to this equation can be expressed in terms of the unique principal branch of Lambert's function, which satisfies:

$$W(x)e^{W(x)}=x.$$

Lambert's function is closely related to the tree function $T(x)$, which appears often in the enumeration of classes of trees, in fact $T(x)=-W(-x)$. In our case,

$$g(t)=\frac{1}{\ln 2}W((2\ln 2)e^t)-1.$$

The asymptotic distribution is developed from local expansions.

LEMMA 4.1. *As* $t \to 0$,

$$g(t)=g_1 t+g_2\frac{t^2}{2}+O(t^3),$$

where

$$g_1=\frac{1}{1+\ln 2},$$
$$g_2=\frac{1}{(1+\ln 2)^3}.$$

PROOF. See Itoh and Mahmoud (2003). □

Let t approach 0 at an appropriate rate, by taking $t = v/\sqrt{\ln n}$, for fixed v, as $n \to \infty$. So,

$$\boldsymbol{E}\left[\exp\left\{S_n \frac{v}{\sqrt{\ln n}}\right\}\right] \sim c\left(\frac{v}{\sqrt{\ln n}}\right) e^{g(v/\sqrt{\ln n})\ln n}.$$

Using the second-order local expansion of $g(t)$ as in Lemma 4.1, we have

$$\boldsymbol{E}\left[\exp\left\{S_n \frac{v}{\sqrt{\ln n}}\right\}\right] \sim c\left(\frac{v}{\sqrt{\ln n}}\right) \exp\left(g_1 v \frac{\ln n}{\sqrt{\ln n}} + \frac{g_2 v^2}{2} + O\left(\frac{1}{\sqrt{\ln n}}\right)\right).$$

It follows that, as $n \to \infty$,

$$\boldsymbol{E}\left[\exp\left\{\frac{S_n - g_1 \ln n}{\sqrt{\ln n}} v\right\}\right] \to e^{g_2 v^2/2}.$$

This convergence can also be discussed rigorously within the scope of the emerging contraction method (see the Appendix).

THEOREM 4.1. *Let S_n be the size of a min incomplete search tree grown from a random permutation of $\{1, \ldots, n\}$. As $n \to \infty$,*

$$\frac{S_n - \dfrac{1}{1+\ln 2}\ln n}{\sqrt{\ln n}} \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{1}{(1+\ln 2)^3}\right).$$

COROLLARY 4.1.

$$\boldsymbol{E}[S_n] \sim \frac{1}{1+\ln 2}\ln n \approx 0.590616109 \ln n,$$
$$\mathbf{Var}[S_n] = \frac{1}{(1+\ln 2)^3}\ln n \approx 0.2060230748 \ln n.$$

Note that the average size of the min incomplete search tree is smaller than that of the left preference incomplete search tree, as a result of acceleration.

5. Max incomplete search trees

The max variant reverses the accelerating policy followed in the min tree. The max variant slows down the development. To obtain the max tree from the binary tree, at each node the smaller of the two sides is replaced by an empty subtree, the larger is recursively further grown.

The mathematical development is mutatis mutandis similar to the min tree case, with max mirror images of min. Therefore, we shall only sketch both description and analysis. A max incomplete tree T_n grows from a permutation of $\{1, \ldots, n\}$. Firstly, a root is allocated for R_n, the first element of the permutation. Let $Z_n = \max\{R_n - 1, n - R_n\}$. A tree T_{Z_n} is grown recursively along the side corresponding to the larger of the two subtrees in the full binary tree, the other side is voided.

The size S_n of the max tree satisfies the recurrence

$$S_n = 1 + S_{Z_n}, \qquad \text{for} \quad n \geq 1,$$

with the boundary condition $S_0 = 0$. Let $\phi_n(t)$ be the moment generating function of S_n. We shall develop asymptotic relations for $\phi_n(t)$ by conditioning on Z_n. The distribution of Z_n is uniform over $\{\lceil \frac{1}{2}(n-1) \rceil, \ldots, n-1\}$, with the possible exception of a parity perturbation at the lowest point of the distribution. Similarly to developments in the min case, one finds

$$P\{Z_n = z\} = \frac{2}{n} - \frac{1}{n}(n \bmod 2)\mathbf{1}_{\{z = \lceil \frac{1}{2}(n-1) \rceil\}}.$$

For $n \geq 1$,

$$\begin{aligned}
\phi_n(t) &= \boldsymbol{E}[e^{(1+S_{Z_n})t}] \\
&= e^t \sum_{z=\lceil \frac{1}{2}(n-1) \rceil}^{n-1} \boldsymbol{E}[e^{S_{Z_n} t} \mid Z_n = z] P\{Z_n = z\} \\
&= \frac{2e^t}{n} \sum_{z=\lceil \frac{1}{2}(n-1) \rceil}^{n-1} \phi_z(t) - \frac{e^t}{n}(n \bmod 2)\phi_{\lceil \frac{1}{2}(n-1) \rceil}(t).
\end{aligned}$$

Just like in the case of min trees, going through the routine of differencing and arguing for all parities of n, one obtains the recurrence

$$n\phi_n(t) - (n-1)\phi_{n-1}(t) = 2e^t \phi_{n-1}(t) - e^t \phi_{\lceil \frac{1}{2}(n-2) \rceil}(t).$$

By arguments that parallel the min tree case, one sees that $d(t)n^{h(t)}(1 + A_t/n + O(n^{-2}))$ is an asymptotic solution for fixed t, if

$$h(t) + 1 = e^t(2 - 2^{-h(t)}), \tag{5.1}$$

and if $h(t), A_t \to 0$, and $d(t) \to 1$, as $t \to 0$, and the hidden constant in O is also a function of t tending to 0, as $t \to 0$. The function $h(t)$ that solves the functional equation must then have the local expansion

$$h(t) = h_1 t + h_2 \frac{t^2}{2} + O(t^3), \qquad \text{as} \quad t \to 0,$$

with

$$\begin{aligned}
h_1 &= \frac{1}{1 - \ln 2}, \\
h_2 &= \frac{1 - 2\ln^2 2}{(1 - \ln 2)^3}.
\end{aligned}$$

The general idea of the proof is to develop the coefficients $h_1 = h'(0)$, and $h_2 = h''(0)$ from functional equation for $h'(t)$ and $h''(t)$ that can be obtained by taking the first and second derivative of the implicit equation (5.1). The details of this proof are omitted, and the reader can refer to Itoh and Mahmoud (2003) for a more detailed discussion of

a similar result. This convergence can also be discussed rigorously within the scope of the emerging contraction method (see the Appendix).

By a development parallel to the asymptotic procedure in the min incomplete search tree (with $g(t)$ replaced by $h(t)$), we obtain the following.

THEOREM 5.1. *Let S_n be the size of a max incomplete search tree grown from a random permutation of $\{1,\dots,n\}$. As $n \to \infty$,*

$$\frac{S_n - \dfrac{1}{1-\ln 2}\ln n}{\sqrt{\ln n}} \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{1-2\ln^2 2}{(1-\ln 2)^3}\right).$$

COROLLARY 5.1.

$$\boldsymbol{E}[S_n] \sim \frac{1}{1-\ln 2}\ln n \approx 3.258891353 \ln n,$$

$$\mathbf{Var}[S_n] = \frac{1-2\ln^2 2}{(1-\ln 2)^3}\ln n \approx 1.353067446 \ln n.$$

Note that the average size of the max incomplete search tree is larger than that of the left preference incomplete search tree, which is to be expected in view of the deceleration in the algorithm.

6. Proportionate preference incomplete interval trees

The proportionate preference incomplete search tree develops one of the two subtrees with probability proportionate to its size. More precisely, the construction algorithm operates as follows. The first element R_n is placed in the root. A subtree is chosen with a particular probability to be further developed, the other remains empty. The random choice of the subtree is obtained through the mechanism of an independent random variable V_n uniform on the set $\{1,\dots,n-1\}$; if $V_n < R_n$ the left subtree is developed, the right is truncated, otherwise $V_n \geq R_n$, and the right subtree is developed, the left is truncated. The selector V_n chooses the left subtree with conditional probability $(R_n-1)/(n-1)$, or chooses the right subtree with conditional probability $(n-R_n)/(n-1)$, given R_n. A subtree is grown recursively on the random permutation belonging to the size of the selected subtree.

The size S_n of the tree satisfies the recurrence

$$S_n = 1 + \mathbf{1}_{\{V_n \leq R_n - 1\}} S_{R_n - 1} + \mathbf{1}_{\{V_n \geq R_n\}} \tilde{S}_{n-R_n}, \qquad \text{for} \quad n \geq 1,$$

where $\tilde{S}_j$ is distributed like S_j, for each j, and S_k and $\tilde{S}_j$ are independent, for every j and k. It is important to note that S_{R_n-1} and $\tilde{S}_{n-R_n}$ are only conditionally independent, given R_n, otherwise without the knowledge of R_n they are dependent through their joint dependency on R_n. The boundary condition is $S_0 = 0$. Let $\phi_n(t)$ be the moment generating function of S_n. For $n \geq 1$,

$$\phi_n(t) = \boldsymbol{E}[e^{(1+\mathbf{1}_{\{V_n \leq R_n-1\}} S_{R_n-1} + \mathbf{1}_{\{V_n \geq n-R_n\}} \tilde{S}_{n-R_n})t}]$$

$$= e^t \left[\sum_{r=1}^{n} \sum_{v=1}^{r-1} \boldsymbol{E}[e^{S_{r-1}t}] P\{R_n = r\} P\{V_n = v\} + \sum_{r=1}^{n} \sum_{v=r}^{n-1} \boldsymbol{E}[e^{\tilde{S}_{n-r}t}] P\{R_n = r\} P\{V_n = v\} \right]$$

$$= \frac{e^t}{n(n-1)} \left[\sum_{r=1}^{n} (r-1)\phi_{r-1}(t) + \sum_{r=1}^{n} (n-r)\phi_{n-r}(t) \right].$$

Changing the second summation variable r to $n - r + 1$ renders the second summation the same as the first:

$$n(n-1)\phi_n(t) = 2e^t \sum_{r=1}^{n} (r-1)\phi_{r-1}(t).$$

Differencing a version of this recurrence with $n-1$ replacing n, from the version with n, we obtain an iteratable recurrence

$$\phi_n(t) = \frac{n + 2e^t - 2}{n} \phi_{n-1}(t).$$

Iterating all the way back to $\phi_0(t) = 1$, we have an exact representation in terms of the Gamma function:

$$\phi_n(t) = \frac{\Gamma(n + 2e^t - 1)}{\Gamma(e^t)\Gamma(n+1)}.$$

This exact distribution is similar to the exact distribution of the left preference incomplete search tree, and can be manipulated with the same Gamma function asymptotics and methods of local expansion.

THEOREM 6.1. *Let S_n be the size of a proportionate preference search tree grown from a random permutation of $\{1, \ldots, n\}$. As $n \to \infty$,*

$$\frac{S_n - 2\ln n}{\sqrt{\ln n}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 2).$$

COROLLARY 6.1.

$$\boldsymbol{E}[S_n] \sim 2\ln n,$$
$$\mathbf{Var}[S_n] \sim 2\ln n.$$

It is to be expected that the average size of the proportionate tree lies somewhere between that of the min and that of the max incomplete search tree as the recursive selection of the smaller and larger subtrees for the developed side mixes the two cases. The proportionate preference incomplete search tree is also larger on average than the left preference incomplete search tree, because larger subtrees are favored probabilistically.

7. No preference incomplete interval trees

In the uniform preference (no preference) variation equal weight is given to the two subtrees, regardless of their size. A subtree is chosen with probability 1/2. The choice of the subtree to develop is determined by a fair coin flip, say with heads we develop the left subtree and with tails we develop the right subtree. The first element of the permutation, say R_n, goes into the root. In the case of heads we grow a subtree on the left side on the remaining elements in the permutation that are less than R_n (which form a random permutation on $\{1, \ldots, R_n - 1\}$); the right subtree remains empty. In the case of tails we grow a subtree on the right side on the remaining elements in the permutation that are greater than R_n (which form a random permutation on $\{R_n + 1, \ldots, n\}$); the left subtree remains empty.

Let the events $\mathcal{H}$ and $\mathcal{T}$ stand respectively for the outcomes heads and tails. The recurrence for the size S_n is

$$S_n = 1 + \mathbf{1}_{\mathcal{H}} S_{R_n - 1} + \mathbf{1}_{\mathcal{T}} \tilde{S}_{n - R_n}; \tag{7.1}$$

the tilde stands for equality in distribution and conditional independence, as explained in Section 6.

Let $\phi_n(t)$ be the moment generating function of S_n. By conditioning on the outcome Q of the coin flip, the relation (7.1) yields

$$\begin{aligned} \phi_n(t) &= \boldsymbol{E}[e^{(1+\mathbf{1}_{\mathcal{H}} S_{R_n-1} + \mathbf{1}_{\mathcal{T}} \tilde{S}_{n-R_n})t}] \\ &= e^t (\boldsymbol{E}[e^{S_{R_n-1}t}] P\{Q = \mathcal{H}\} + \boldsymbol{E}[e^{\tilde{S}_{n-R_n}t}] P\{Q = \mathcal{T}\}) \\ &= \boldsymbol{E}[e^{(1+S_{R_n-1})t}], \end{aligned}$$

where we used the identical distribution of $n - R_n$ and $R_n - 1$, and of S_j and $\tilde{S}_j$, for all j. Hence, for $n \geq 1$,

$$S_n = 1 + S_{R_n - 1}.$$

This recurrence is the same as that of the recurrence of the left preference tree (cf. (3.1)). Consequently, the results are the same as those in that case; the no preference case grows a zig-zag path of length distributed like the leftmost arm of a binary tree.

Acknowledgements

This research was conducted while the author was visiting the Institute of Statistical Mathematics, Japan. The author gratefully acknowledges the institute's generous support. Special thanks are due to Professor Yoshiaki Itoh for his encouragement. The author thanks Dr. Ralph Neininger for a valuable consultation, and Professor Dr. Helmut Prodinger for pointing out a connection to the tree function.

Appendix

The one-sided variations discussed can be handled within the rigorous approach of the contraction method. We illustrate the approach on the min tree. However, the method is applicable to all five variants discussed.

The method was introduced by Rösler (1991). Several extensions were contributed by Rachev and Rüschendorf (1995). Recently general contraction theorems and multivariate extensions were added by Rösler (2001), and Neininger (2002). Rösler and Rüschendorf (2001) survey the area.

The general philosophy of the approach is to start from a recurrence on a random variable, say X_n. Under appropriate norming, the recurrence carries over to X_n^*, the normalized random variable. The norming is usually affected by asymptotic centering (subtracting off an asymptotic equivalent of the mean) and asymptotic scaling by an asymptotic equivalent of the standard deviation. The rest of the argument is devoted to demonstrating that the distributional equation for X_n^* converges in the limit to a limiting distributional equation on a limit random variable X. Usually, the limiting distribution is the unique fixed-point solution to the limiting equation. The convergence itself is argued by showing that the distance in some metric space, such as Wasserstein's or Zolotarev's, between the laws of X_n and X approaches 0.

The method is challenged in some tree applications by the appearance of a degenerate limit equation of the form

$$X \stackrel{\mathcal{D}}{=} X, \tag{A.1}$$

which is of no help in characterizing the limit distribution, because of course any random variable with arbitrary distribution satisfies such a degenerate equation. The depth of a randomly selected node in a random binary search tree provides such an instance (see Mahmoud and Neininger (2003)).

Only very recently has this issue been finessed in Neininger and Rüschendorf (2002), where it is shown that under some mild conditions on growth rates, one still gets normal limit laws as the *unique* solution of a degenerate equation like (A.1).

Recall the definitions $g_1 = 1/(1+\ln 2)$, and $g_2 = g_1^3$ (cf. Lemma 4.1). In our case, we can start from the recurrence (4.1) and progress to normalize:

$$S_n^* := \frac{S_n - g_1 \ln n}{\sqrt{g_2 \ln n}} = \frac{S_{Z_n} - g_1 \ln Z_n}{\sqrt{g_2 \ln Z_n}} \times \frac{\sqrt{\ln Z_n}}{\sqrt{\ln n}} + A_n(Z_n),$$

where

$$A_n(u) := \frac{1}{\sqrt{g_2 \ln n}} + \frac{g_1 \ln u - g_1 \ln n}{\sqrt{g_2 \ln n}}.$$

Hence

$$S_n^* = S_{Z_n}^* \frac{\sqrt{\ln Z_n}}{\sqrt{\ln n}} + A_n(Z_n).$$

But then, as $n \to \infty$,

$$\frac{1}{\sqrt{g_2 \ln n}} \to 0,$$

and

$$\frac{Z_n}{n} \stackrel{P}{\to} \frac{1}{2} U,$$

with U being a standard Uniform$(0,1)$ random variable, so that

$$\frac{\ln Z_n - \ln n}{\sqrt{\ln n}} \stackrel{P}{\to} 0.$$

Thus, $A_n(Z_n) \xrightarrow{P} 0$, and $\sqrt{\ln Z_n}/\sqrt{\ln n} \xrightarrow{P} 1$. If $S_n^* \xrightarrow{\mathcal{D}} S$, where S is some limiting random variable one would then get the degenerate equation

$$S \stackrel{\mathcal{D}}{=} S.$$

This can be shown rigorously in either of the afore-mentioned metric spaces. But that would not help. One can then resort to the finesse of Neininger and Rüschendorf (2002).

For the reader's convenience, we repeat their main result here. Suppose Y_n is a sequence of random variables that satisfies the recurrence

$$Y_n \stackrel{\mathcal{D}}{=} Y_{I_n} + b_n, \qquad \text{for all} \quad n \geq n_0 \geq 1, \tag{A.2}$$

and $(I_n, b_n), Y_k$ are independent, with b_n random, and $I_n \in \{0, \ldots, n\}$, with $P\{I_n = n\} < 1$, for $n \geq n_0$. Denote $\boldsymbol{E}[Y_n]$ by μ_n, and $\mathbf{Var}[Y_n]$ by σ_n. Let $\|X\|_p$ denote the L_p norm of a random variable X.

THEOREM A.1. (Neininger and Rüschendorf (2002)) *Let Y_n be a sequence of random variables satisfying* (A.2) *with $\|Y_n\|_3 < \infty$, for all $n \geq 0$, and*

$$\limsup_{n\to\infty} \boldsymbol{E}\left[\ln\left(\frac{I_n \vee 1}{n}\right)\right] < 0, \tag{A.3}$$

$$\sup_{n\geq 1} \left\|\ln\left(\frac{I_n \vee 1}{n}\right)\right\|_3 < \infty. \tag{A.4}$$

Further more, assume that for real numbers α, λ, κ, with $\alpha > 0$, and $0 \leq \lambda < 2\alpha$, the mean and variance of Y_n satisfy

$$\|b_n - \mu_n + \mu_{I_n}\|_3 = O(\ln^{\kappa} n), \tag{A.5}$$

$$\sigma_n^2 = C \ln^{2\alpha} n + O(\ln^{\lambda} n), \tag{A.6}$$

for some positive constant C. If

$$\beta := \left(\frac{3}{2}(1 \wedge [2\alpha - [(2\kappa) \vee \lambda]])\right) \wedge (\alpha - \kappa + [1 \wedge (2\alpha - \lambda)]) > 1,$$

then

$$\frac{Y_n - \boldsymbol{E}[Y_n]}{\sqrt{C}\ln^{\alpha} n} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1),$$

and we have the rate of convergence $O(\ln^{\beta-1} n)$ in Zolotarev's metric space (see Zolotarev (1976)).

Note that this route does not require at all checking first that the contraction method gives a degenerate equation of the form $Y \stackrel{\mathcal{D}}{=} Y$. The theorem stands by itself, and gives a way out if the limit equation is degenerate.

In our incomplete tree investigation, $b_n \equiv 1$, and $I_n = Z_n$. All the conditions of Theorem A.1 are satisfied: Obviously, the third moment (in fact all the moments) of $S_n \leq \frac{1}{2}(n+1)$ exist, and from the distribution of Z_n we can quickly verify conditions

(A.3) and (A.6). Indeed, recurrence equations for the moments can be readily found from the recurrence (4.2). For instance, by taking derivatives of (4.2), we get

$$n\mu_n - (n-1)\mu_{n-1} = 1 + \mu_{\lfloor \frac{1}{2}(n-1) \rfloor}.$$

It follows from an easy induction that

$$g_1 \ln n \le \mu_n \le 1 + g_1 \ln n.$$

One can now verify (A.5) for large n, by conditioning on Z_n, as follows. We have

$$\boldsymbol{E}[|b_n - \mu_n + \mu_{Z_n}|^3] \le \frac{2g_1}{n} \ln n + \frac{2}{n} \sum_{k=1}^{\lfloor \frac{1}{2}(n-1) \rfloor} (g_1 \ln n - g_1 \ln k)^3 = O(1);$$

such a computation can be carried out by comparing sums to bounding integrals. We can take $\kappa = 0$.

Likewise for (A.6), one finds by induction

$$\sigma_n = g_2 \ln n + O(1).$$

Thus, $\alpha = \frac{1}{2}$, and $\lambda = 0$. By conditioning on Z_n, we can verify that

$$\begin{aligned} \boldsymbol{E}\left[\ln\left(\frac{Z_n \vee 1}{n}\right)\right] &= \frac{2}{n} \ln \frac{1}{n} + \left[\frac{2}{n} \sum_{k=1}^{\lfloor \frac{1}{2}(n-1) \rfloor - 1} \ln\left(\frac{k}{n}\right)\right] \\ &\quad + \frac{1}{n}(2 - n \bmod 2) \ln\left(\frac{\left\lfloor \frac{1}{2}(n-1) \right\rfloor}{n}\right) \\ &\to -1 - \ln 2. \end{aligned}$$

Likewise,

$$\sup_{n \ge 1} \left\| \ln\left(\frac{Z_n \vee 1}{n}\right) \right\|_3 = 6 + 6 \ln 2 + 3 \ln^2 2 + \ln^3 2.$$

All the conditions for Theorem A.1 hold, with $\beta = \frac{3}{2}$; Theorem 4.1 follows, with $O(\ln^{-1/2} n)$ rate of convergence (in Zolotarev's metric space).

References

Arora, S. and Dent, W. (1969). Randomized binary search technique, *Communications of the ACM*, **12**, 77–80.

Chassaing, P. and Marchand, R. (2002). Cutting a random tree (and UNION-FIND algorithms) (manuscript).

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

Drmota, M. (2002). The random bisection problem and the distribution of the height of binary search trees (manuscript).

Fill, J., Mahmoud, H. and Szpankowski, W. (1996). On the distribution for the duration of a randomized leader election algorithm, *Ann. Appl. Probab.*, **6**, 1260–1283.

Itoh, Y. and Mahmoud, H. (2003). One-sided variations on interval trees, *J. Appl. Probab.*, **40**, 1–17.

Kemp, R. (1984). *Fundamentals of the Average Case Analysis of Particular Algorithms*, Wiley-Teubner, Stuttgart.

Knuth, D. (1998). *The Art of Computer Programming, Vol. III: Searching and Sorting*, Addison-Wesley, Reading, Massachusetts.

Mahmoud, H. (1992). *Evolution of Random Search Trees*, Wiley, New York.

Mahmoud, H. (2000). *Sorting: A Distribution Theory*, Wiley, New York.

Mahmoud, H. and Neininger, R. (2003). Distribution of distances in random binary search trees, *Ann. Appl. Probab.*, **13**, 253–276.

Martínez, C., Panholzer, A. and Prodinger, H. (1998). Descendants and ascendants in random search trees, *Electronic Journal of Combinatorics*, **5**, R20; 29 pages + Appendix (10 pages).

Meir, A. and Moon, J. (1970). Cutting down random trees, *J. Austral. Math. Soc.*, **11**, 313–324.

Neininger, R. (2002). On a multivariate contraction method for random recursive structures with applications to Quicksort, *Random Structures Algorithms*, **19**, 498–524.

Neininger, R. and Rüschendorf, L. (2002). On the contraction method with degenerate limit equation (manuscript).

Prodinger, H. (1993). How to select a loser, *Discrete Math.*, **120**, 149–159.

Rachev, S. and Rüschendorf, L. (1995). Probability metrics and recursive algorithms, *Adv. in Appl. Probab.*, **27**, 770–799.

Rösler, U. (1991). A limit theorem for "Quicksort", *RAIRO Inform. Théor. Appl.*, **25**, 85–100.

Rösler, U. (2001). On the analysis of stochastic divide and conquer algorithms, *Algorithmica*, **29**, 238–261.

Rösler, U. and Rüschendorf, L. (2001). The contraction method for recursive algorithms, *Algorithmica*, **29**, 3–33.

Sibuya, M. and Itoh, Y. (1987). Random sequential bisection and its associated binary tree, *Ann. Inst. Statist. Math.*, **39**, 69–84.

Zolotarev, V. M. (1976). Approximation of the distributions of sums of independent random variables with values in infinite-dimensional spaces, *Theory Probab. Appl.*, **21**, 721–737.