Distances and Finger Search in Random Binary Search Trees

Luc Devroye[1]
School of Computer Science
McGill University
3480 University Street
Montreal, H3A 2K6
Canada

Ralph Neininger[2]
Department of Mathematics
J.W. Goethe University
Robert-Mayer-Str. 10
60325 Frankfurt a.M.
Germany

November 8, 2003

Abstract

For the random binary search tree with n nodes inserted the number of ancestors of the elements with ranks k and ℓ, $1 \le k < \ell \le n$, as well as the path distance between these elements in the tree are considered. For both quantities central limit theorems for appropriately rescaled versions are derived. For the path distance the condition $\ell - k \to \infty$ as $n \to \infty$ is required. We obtain tail bounds and the order of higher moments for the path distance. The path distance measures the complexity of finger search in the tree.

AMS subject classifications. Primary: 60D05; secondary: 68U05.

Key words. Random binary search tree, finger search, path distance, limit law, analysis of algorithms.

Abbreviated title. Finger search in random binary search trees.

1 Introduction and results

In this paper we analyze the asymptotic behavior of the path distance between nodes in random binary search trees. The path distance between two nodes is the

[1]Research supported by NSERC grant A3450.
[2]Research supported by DFG grant Ne 828/1-2 (Emmy Noether Programme).

number of nodes on the shortest path connecting them in the tree. This quantity is motivated by the cost of a finger search in the tree. The finger search operation in a search tree takes as input a pointer to a node u, the current node, and either the key value of another node v or an incremental rank value Δ. The objective is to find v quickly. In the latter case, the rank of v differs from the rank of u by Δ. Finger search trees are search trees in which the finger operation takes time $O(1 + \ln \Delta)$. Various strategies are known for this. For example, Brown and Tarjan (1980) recommend (2,4) or red-black trees with level linking. Huddleston and Mehlhorn (1982) show how to update these trees efficiently in an amortized sense. On pointer-based machines, Brodal (1998) shows how to implement insertion in constant worst-case time in an adaptation of these trees.

In a random binary search tree or a treap, suitably augmented, but without level linking, we note that both kinds of finger search operations take time proportional to the path distance between the nodes. The augmentation consists of maintaining with each node either the minimum and maximum keys in the subtree, or the size of the subtree. These parameters are easy to update. Furthermore, when searching for v, starting from u, one first proceeds by following parent pointers towards the root until the least common ancestor of u and v is found. At that point, one can find v by the standard search operation.

If the nodes are level-linked, then it is also possible to identify an ancestor of v that is either the least common ancestor of u and v, or a descendant of that least common ancestor, simply by checking the key values of the appropriate level neighbors of the ancestors of u when traveling towards the root. In this implementation, the complexity of the finger search operation is the path distance between u and v or less. Other possible augmentations for treaps are presented by Seidel and Aragon (1996).

We give an approach of the distributional analysis of the path distance between nodes in a random binary search tree whose keys have ranks that differ by Δ. The connection used between records and random permutations for the study of random binary search trees was developed in Devroye (1988) and, when it applies, leads to short and intuitive proofs. While the expectation of the path distance of two nodes that hold keys with ranks differing by Δ is always $O(\ln \Delta)$ as $\Delta \to \infty$, for a refined distributional analysis the location of the ranks matters, since in particular the leading constant in the expansion of the expectation of the path distance depends upon the location of the ranks. This affects the proper scaling of the quantities to obtain distributional convergence, see Theorem 1.3 below.

For simplicity we assume that the random binary search tree is build up from the keys $1, \ldots, n$ identifying the key of rank j with the key j. See, e.g., Mahmoud (1992) for the definition of random binary search trees. For $1 \leq k \leq \ell \leq n$ we denote by $A_{k\ell}$ the number of ancestors of the nodes holding the keys k and ℓ in the tree when n numbers are inserted. Note that A_{kk} is the depth of the node with rank k in the tree, $1 \leq k \leq n$. By $P_{k\ell}$ the path distance between the keys k and ℓ is denoted, that is, the number of nodes on the path (strictly) between k and ℓ, $1 \leq k < \ell \leq n$.

We denote by $\mathcal{N}(0,1)$ the standard normal distribution and by $\xrightarrow{\mathcal{L}}$ convergence in distribution. For sequences $(a_n), (b_n)$ asymptotical equivalence, $a_n/b_n \to 1$ as $n \to \infty$, is denoted by $a_n \sim b_n$. We have the following asymptotic behavior.

Theorem 1.1 *For all $1 \leq k < \ell \leq n$, where k, ℓ may depend on n, we have, as $n \to \infty$,*

$$\begin{aligned} \mathbb{E}\, A_{k\ell} &= \ln(k(\ell - k)^2(n - \ell + 1)) + O(1), \\ \frac{A_{k\ell} - \ln(k(\ell - k)^2(n - \ell + 1))}{\sqrt{\ln(k(\ell - k)^2(n - \ell + 1))}} &\xrightarrow{\mathcal{L}} \mathcal{N}(0,1). \end{aligned}$$

Theorem 1.2 *For all $1 \leq k \leq n$, where k may depend on n, we have, as $n \to \infty$,*

$$\frac{A_{kk} - \ln(k(n - k + 1))}{\sqrt{\ln(k(n - k + 1))}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

Theorem 1.3 *For all $1 \leq k < \ell \leq n$ with k, ℓ depending on n such that $\Delta := \ell - k + 1 \to \infty$ as $n \to \infty$ and $a_n := (k \wedge \Delta)\Delta^2((n - \ell + 1) \wedge \Delta)$ we have, as $n \to \infty$,*

$$\frac{P_{k\ell} - \ln a_n}{\sqrt{\ln a_n}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

Theorem 1.4 *Let P_n denote the path distance between a pair of nodes chosen uniformly at random from all possible pairs of different nodes in the tree. Then we have, as $n \to \infty$,*

$$\frac{P_n - 4 \ln n}{\sqrt{4 \ln n}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

Theorem 1.5 *There exists a constant $C > 0$ such that for all $\varepsilon > 0$ and all $1 \leq k < \ell \leq n$ with $\Delta := \ell - k + 1 \geq \Delta_0$ we have with $a_n := (k \wedge \Delta)\Delta^2((n - \ell + 1) \wedge \Delta)$:*

$$\mathbb{P}(P_{k\ell} > (1 + \varepsilon) \ln a_n) \leq C \Delta^{-\varepsilon^2/(2+3\varepsilon)}.$$

Here, for all $\delta > 0$, we can choose $\Delta_0 \geq 1$ uniformly for all $\varepsilon \in [\delta, \infty)$.

Moreover, if $\Delta \to \infty$ as $n \to \infty$, we have, for all $p \geq 1$,

$$\mathbb{E}\, P_{k\ell}^p \sim \ln^p a_n.$$

Note that exact expressions for $\mathbb{E}\, A_{kk}$ and $\mathbb{E}\, P_{k\ell}$ in terms of harmonic numbers are given in Seidel and Aragon (1996) and, for $\mathbb{E}\, A_{k\ell}$, in Prodinger (1995). The limit law in Theorem 1.4, together with additional results for the model of uniformly chosen pairs of nodes, have been derived in Mahmoud and Neininger (2003) and Panholzer and Prodinger (2003+), an exact expression for $\mathbb{E}\, P_n$ has first been given in Flajolet, Ottmann, and Wood (1985). Finally, we note that the limit law for the depth of a typical node inserted in a random binary search tree was obtained by Mahmoud and Pittel (1984), Louchard (1987), and Devroye (1988). It can be obtained from Theorem 1.2 by replacing k by a uniform$\{1, \dots, n\}$ random variable.

2 Representation via Records

In a permutation $(x_1, \dots, x_n)$ of distinct numbers we define the local ranks $R_1, \dots, R_n$, where R_j denotes the rank of x_j in $\{x_1, \dots, x_j\}$. If $R_j = j$ or $R_j = 1$ we say that x_j is an up-record or down-record in $x_1, \dots, x_n$ respectively. It is well known that if the permutation is a random permutation, i.e., all $n!$ permutations are equally likely, R_j is uniformly distributed on $\{1, \dots, j\}$ for all $j = 1, \dots, n$ and that $R_1, \dots, R_n$ are independent.

We give a representation of the number $A_{k\ell}$ of ancestors of keys k and ℓ in terms of local ranks and records, so that based on the independence properties we can apply the classical central limit theorem in the version of Lindeberg-Feller.

Let us build up the random binary search tree from the numbers $1, \dots, n$ as follows: We draw independent unif$[0,1]$ random variables $T_1, \dots, T_n$, where unif$[0,1]$ denotes the uniform distribution on the interval $[0,1]$. These we use as time stamps as T_j is associated with j and denotes the time at which number j is inserted into the tree. Inserting now the numbers in order according to their time stamps, starting with the earliest, yields a random binary search tree for the keys $1, \dots, n$.

A basic property of the binary search tree is that j is an ancestor of k in the tree if and only if it is inserted before k and also before all numbers s between j and k. Now we fix $1 \leq k < \ell \leq n$ and count the ancestors $A_{k\ell}$ of the elements k and ℓ in the tree. If, for $i < k$, element i is ancestor of ℓ then it is as well ancestor of k and hence it contributes to $A_{k\ell}$ if and only if

$$T_i = \min\{T_i, T_{i+1}, \ldots, T_k\}, \quad i < k.$$

Analogously, for $i > \ell$, we get a contribution of number i to $A_{k\ell}$ if and only if $T_i = \min\{T_\ell, T_{\ell+1}, \ldots, T_i\}$, and in the case $k < i < \ell$ if $T_i = \min\{T_k, T_{k+1}, \ldots, T_i\}$ or $T_i = \min\{T_i, T_{i+1}, \ldots, T_\ell\}$. Passing to indicator functions we rewrite these events as

$$\begin{aligned}
\mathbf{1}_{\{T_i=\min\{T_i,T_{i+1},\ldots,T_k\}\}} &= \mathbf{1}_{\{T_i=\min\{T_i,\ldots,T_{k-1}\}\}} - \mathbf{1}_{\{T_k<T_i,\, T_i=\min\{T_i,\ldots,T_{k-1}\}\}} \\
&=: \mathbf{1}_{B_i} - \mathbf{1}_{C_i}, \quad i < k, \\
\mathbf{1}_{\{T_i=\min\{T_\ell,\ldots,T_i\}\}} &= \mathbf{1}_{\{T_i=\min\{T_{\ell+1},\ldots,T_i\}\}} - \mathbf{1}_{\{T_\ell<T_i,\, T_i=\min\{T_{\ell+1},\ldots,T_i\}\}} \\
&=: \mathbf{1}_{B_i} - \mathbf{1}_{C_i}, \quad i > \ell,
\end{aligned}$$

and

$$\mathbf{1}_{B_i} := \mathbf{1}_{\{T_i=\min\{T_k,T_{k+1},\ldots,T_i\}\}\cup\{T_i=\min\{T_i,T_{i+1},\ldots,T_\ell\}\}}, \quad k \le i \le \ell.$$

Note that above $\mathbf{1}_{B_i}, \mathbf{1}_{C_i}$ are differently defined for the three ranges of the index i. Altogether we obtain the representation

$$A_{k\ell} = \sum_{i=1}^{n} \mathbf{1}_{B_i} - \sum_{i=1}^{k-1} \mathbf{1}_{C_i} - \sum_{i=\ell+1}^{n} \mathbf{1}_{C_i} - 2, \tag{1}$$

where we subtract 2 referring to the convention that k and ℓ are not counted as ancestors of themselves. The main contribution comes from the sum over the $\mathbf{1}_{B_i}$, as the sums over the $\mathbf{1}_{C_i}$ will be asymptotically negligible.

To get the connection with records we introduce three auxiliary random binary search trees as follows. The binary search tree $\mathcal{T}_<$ is build up from the elements $1, \ldots, k-1$, inserted according to their time stamps $T_1, \ldots, T_{k-1}$. Analogously $\mathcal{T}_>$ is build up from the elements $\ell+1, \ldots, n$, inserted according to their time stamps $T_\ell, \ldots, T_n$ and $\mathcal{T}$ is build up from the elements $k, \ldots, \ell$, inserted according to their time stamps $T_k, \ldots, T_\ell$. Now, for $i < k$, the event B_i is equivalent for i to be an ancestor of $k-1$ in $\mathcal{T}_<$. Since $k-1$ is the largest element in $\mathcal{T}_<$, this implies that i is an up-record at the time of insertion into $\mathcal{T}_<$. Analogously, for $i > \ell$, the event B_i is equivalent for i to constitute a down-record at time of its insertion into $\mathcal{T}_>$. For $k \le i \le \ell$, event B_i is equivalent to i being up or down-record at its time of insertion into $\mathcal{T}$.

We denote by R_j the local rank of the (in time) jth element inserted into $\mathcal{T}_<$ at the time of its insertion, $1 \leq j < k$, and by R'_j, R''_j analogously the local ranks of the jth elements inserted into $\mathcal{T}$, $\mathcal{T}_>$ for $1 \leq j \leq \ell - k + 1$ and $1 \leq j \leq n - \ell$ respectively. Note that $R_1, \ldots, R_{k-1}, R'_1, \ldots, R'_{\ell-k+1}, R''_1, \ldots, R''_{n-\ell}$ are independent and R_j, R'_j, R''_j are uniform$\{1, \ldots, j\}$ distributed for $j = 1, \ldots, k-1$ and $j = 1, \ldots, \ell - k + 1$ and $j = 1, \ldots, n - \ell$ respectively. We have

$$\sum_{i=1}^{n} \mathbf{1}_{B_i} = \sum_{j=1}^{k-1} \mathbf{1}_{\{R_j = j\}} + \sum_{j=1}^{\ell-k+1} \mathbf{1}_{\{R'_j \in \{1,j\}\}} + \sum_{j=1}^{n-\ell} \mathbf{1}_{\{R''_j = 1\}}. \tag{2}$$

For the representation of $P_{k\ell}$ we denote

$$T_A := \min\{T_k, \ldots, T_\ell\}.$$

For $1 \leq i \leq n$, element i belongs to the path between k and ℓ if and only if it is ancestor of k or ℓ and $T_i \geq T_A$. Hence with $D_i := \{T_i \geq T_A\}$ we have

$$P_{k\ell} = \sum_{i=1}^{n} \mathbf{1}_{B_i \cap D_i} - \sum_{i=1}^{k-1} \mathbf{1}_{C_i \cap D_i} - \sum_{i=\ell+1}^{n} \mathbf{1}_{C_i \cap D_i} - 2. \tag{3}$$

The main contribution will come from the sum over the $\mathbf{1}_{B_i \cap D_i}$. For the corresponding representation with records we introduce

$$N_1 := |\{1 \leq j < k : T_j < T_A\}|, \quad N_2 := |\{\ell < j \leq n : T_j < T_A\}|,$$

and obtain

$$\begin{aligned} \sum_{i=1}^{n} \mathbf{1}_{B_i \cap D_i} &= \sum_{j=N_1+1}^{k-1} \mathbf{1}_{\{R_j = j\}} + \sum_{j=1}^{\ell-k+1} \mathbf{1}_{\{R'_j \in \{1,j\}\}} + \sum_{j=N_2+1}^{n-\ell} \mathbf{1}_{\{R''_j = 1\}} \\ &=: \mathcal{P}_I + \mathcal{P}_{II} + \mathcal{P}_{III}. \end{aligned} \tag{4}$$

3 Proofs

Throughout this section we denote by $H_n := \sum_{i=1}^{n} 1/i = \ln n + O(1)$ the nth harmonic number for $n \geq 1$ and $H_0 := 0$.

Proof of Theorem 1.1: We derive $\mathbb{E}\, A_{k\ell}$ using the representations (1) and (2). From the distribution of the local ranks R_j, R'_j, and R''_j we obtain

$$\mathbb{E} \sum_{i=1}^{n} \mathbf{1}_{B_i} = H_{k-1} + 2H_{\ell-k+1} - 1 + H_{n-\ell} = \ln(k(\ell-k)^2(n-\ell+1)) + O(1).$$

The remaining summands in (1) we denote by $\Upsilon := \sum_{i=1}^{k-1} \mathbf{1}_{C_i} + \sum_{i=\ell+1}^{n} \mathbf{1}_{C_i} + 2$. For $1 \leq i < k$ we have

$$\begin{aligned} \mathbb{E}\, \mathbf{1}_{C_i} &= \mathbb{P}\Big(T_k < T_i,\ T_i = \min\{T_i, \ldots, T_{k-1}\}\Big) \\ &\leq \mathbb{P}\Big(T_k, T_i \text{ are the smallest elements among } T_i, \ldots, T_k\Big) \\ &= \binom{k-i+1}{2}^{-1} \leq \frac{2}{(k-i)^2}. \end{aligned}$$

This implies $\mathbb{E} \sum_{i=1}^{k-1} \mathbf{1}_{C_i} = O(1)$. Analogously we conclude to find $\mathbb{E}\, \Upsilon = O(1)$, hence we obtain $\mathbb{E}\, A_{k\ell} = \ln(k(\ell-k)^2(n-\ell+1)) + O(1)$.

For the central limit law we write

$$\begin{aligned} \frac{A_{k\ell} - \ln(k(\ell-k)^2(n-\ell+1))}{\sqrt{\ln(k(\ell-k)^2(n-\ell+1))}} &= \frac{\sum_{i=1}^{n} \mathbf{1}_{B_i} - \ln(k(\ell-k)^2(n-\ell+1))}{\sqrt{\ln(k(\ell-k)^2(n-\ell+1))}} \\ &\quad - \frac{\Upsilon}{\sqrt{\ln(k(\ell-k)^2(n-\ell+1))}}. \end{aligned}$$

For all choices of $1 \leq k < \ell \leq n$ we have $\ln(k(\ell-k)^2(n-\ell+1)) \to \infty$ as $n \to \infty$, and, from (2) it follows that the Lindeberg-Feller condition (see Chow and Teicher (1978, p. 291)) is satisfied for $\sum_{i=1}^{n} \mathbf{1}_{B_i}$, thus the first fraction on the right hand side of the latter display tends in distribution to the standard normal distribution. Again since $\ln(k(\ell-k)^2(n-\ell+1)) \to \infty$ and $\mathbb{E}\,|\Upsilon| = O(1)$ we obtain from Markov's inequality that $\Upsilon / \ln(k(\ell-k)^2(n-\ell+1)) \to 0$ in probability as $n \to \infty$. The assertion follows. ■

Proof of Theorem 1.2: Note that for A_{kk} we have the same representation as for $A_{k\ell}$ given, for the case $k < \ell$, in (1), where we have to replace the -2 there by -1 due to the fact that we now have $\mathbf{1}_{B_k} = 1$. Hence the same arguments as in the proof of Theorem 1.1 apply. ■

Proof of Theorem 1.3: We have $P_{k\ell} = \mathcal{P}_I + \mathcal{P}_{II} + \mathcal{P}_{III} - \Upsilon'$, with $\Upsilon' := \sum_{i=1}^{k-1} \mathbf{1}_{C_i \cap D_i} + \sum_{i=\ell+1}^{n} \mathbf{1}_{C_i \cap D_i} + 2$ and $a_n := (k \wedge \Delta)\Delta^2((n-\ell+1) \wedge \Delta) \to \infty$ as $n \to \infty$. From $\mathbb{E}\,|\Upsilon'| = O(1)$ we obtain from Markov's inequality $\Upsilon'/\sqrt{\ln a_n} \to 0$ in probability. Thus it is sufficient to show

$$\frac{\mathcal{P}_I + \mathcal{P}_{II} + \mathcal{P}_{III} - \ln a_n}{\sqrt{\ln a_n}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1). \tag{5}$$

Since we want to apply the central limit theorem to the sum of indicators in (4) we will condition on the random indices N_1 and N_2. Note that we may assume $k \to \infty$

and $n-\ell+1 \to \infty$ as $n \to \infty$ since otherwise $\mathcal{P}_I$ and $\mathcal{P}_{III}$ remain bounded and do not contribute respectively.

First we consider the case $k/\Delta > \ln k$ and $(n-\ell+1)/\Delta > \ln(n-\ell+1)$ for all sufficiently large n. We define, for $\varepsilon > 0$,

$$B_\varepsilon := \{N_1 \in [\alpha_1, \beta_1]\} \cap \{N_2 \in [\alpha_2, \beta_2]\},$$

with

$$\alpha_1 = \frac{\varepsilon}{2}\frac{k}{\Delta}, \quad \beta_1 = \frac{2}{\varepsilon}\frac{k}{\Delta}, \qquad \alpha_2 = \frac{\varepsilon}{2}\frac{n-\ell+1}{\Delta}, \quad \beta_2 = \frac{2}{\varepsilon}\frac{n-\ell+1}{\Delta}.$$

Note that the values of N_1 and N_2 depend on $T_k, \dots, T_\ell$. However, conditioned on N_1 and N_2 the permutations induced by $T_1, \dots, T_{k-1}$, by $T_k, \dots, T_\ell$, and by $T_{\ell+1}, \dots, T_n$ are independent and uniformly distributed. In particular, conditioning on N_1, N_2 preserves the independence and the distributions of $R_1, \dots, R_{k-1}, R'_1, \dots, R'_\Delta, R''_1, \dots, R_{n-\ell}$.

On B_ε we have the bounds $P^-_{k\ell} \le P_{k\ell} \le P^+_{k\ell}$ with

$$\begin{aligned}
P^-_{k\ell} &= \sum_{j=\lceil\beta_1\rceil+1}^{k-1} \mathbf{1}_{\{R_j=j\}} + \sum_{j=1}^{\Delta} \mathbf{1}_{\{R'_j\in\{1,j\}\}} + \sum_{j=\lceil\beta_2\rceil+1}^{n-\ell} \mathbf{1}_{\{R''_j=1\}},\\
P^+_{k\ell} &= \sum_{j=\lfloor\alpha_1\rfloor}^{k-1} \mathbf{1}_{\{R_j=j\}} + \sum_{j=1}^{\Delta} \mathbf{1}_{\{R'_j\in\{1,j\}\}} + \sum_{j=\lfloor\alpha_2\rfloor}^{n-\ell} \mathbf{1}_{\{R''_j=1\}}.
\end{aligned}$$

Now, we have

$$\begin{aligned}
\mathbb{E}\, P^-_{k\ell} &= \ln k - \ln\lceil\beta_1\rceil + 2\ln\Delta \\
&\quad + \ln(n-\ell+1) - \ln\lceil\beta_2\rceil + O(1) \\
&= \ln a_n + O\Big(1 + \ln\frac{1}{\varepsilon}\Big),
\end{aligned} \tag{6}$$

where, for the last equality, we distinguish the cases $k/\Delta \le 2/\varepsilon$, $k/\Delta > 2/\varepsilon$ as well as $(n-\ell+1)/\Delta \le 2/\varepsilon$ and $(n-\ell+1)/\Delta > 2/\varepsilon$. Analogously we obtain $\mathrm{Var}(P^-_{k\ell}) = \mathbb{E}\, P^-_{k\ell} + O(1+\ln(1/\varepsilon))$. Since $\varepsilon > 0$ is fixed and $a_n \to \infty$ as $n \to \infty$ we obtain from the central limit theorem in the version of Lindeberg-Feller that

$$\frac{P^-_{k\ell} - \ln a_n}{\sqrt{\ln a_n}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1), \quad n \to \infty. \tag{7}$$

Similarly we obtain $(P^+_{k\ell} - \ln a_n)/\sqrt{\ln a_n} \to \mathcal{N}(0,1)$ in distribution as $n \to \infty$. We have, for $x \in \mathbb{R}$,

$$\begin{aligned}
\mathbb{P}\Big(\frac{P_{k\ell} - \ln a_n}{\sqrt{\ln a_n}} \le x\Big) &\le \mathbb{P}(B^c_\varepsilon) + \mathbb{P}\Big(\frac{P_{k\ell} - \ln a_n}{\sqrt{\ln a_n}} \le x \Big| B_\varepsilon\Big) \\
&\le \mathbb{P}(B^c_\varepsilon) + \mathbb{P}\Big(\frac{P^-_{k\ell} - \ln a_n}{\sqrt{\ln a_n}} \le x\Big).
\end{aligned}$$

Hence denoting by Φ the distribution function of the standard normal distribution and $\psi(\varepsilon) := \limsup_{n\to\infty} \mathbb{P}(B^c_\varepsilon)$ we obtain

$$\limsup_{n\to\infty} \mathbb{P}\Big(\frac{P_{k\ell} - \ln a_n}{\sqrt{\ln a_n}} \le x\Big) \le \Phi(x) + \psi(\varepsilon),$$

and analogously

$$\begin{aligned}
\liminf_{n\to\infty} \mathbb{P}\Big(\frac{P_{k\ell} - \ln a_n}{\sqrt{\ln a_n}} \le x\Big) &\ge \liminf_{n\to\infty} \mathbb{P}(B_\varepsilon)\mathbb{P}\Big(\frac{P^+_{k\ell} - \ln a_n}{\sqrt{\ln a_n}} \le x\Big) \\
&= (1 - \psi(\varepsilon))\Phi(x).
\end{aligned}$$

Hence the central limit law is established once we have shown that $\psi(\varepsilon) \to 0$ as $\varepsilon \downarrow 0$. For this it is sufficient to show that $[\limsup_{n\to\infty} \mathbb{P}(N_i \notin [\alpha_i, \beta_i])] \to 0$ as $\varepsilon \downarrow 0$ for $i = 1, 2$. By symmetry we only need to show the case $i = 1$.

We denote by $B_{n,u}$ a binomial $B(n,u)$ distributed random variable, $n \ge 0$, $u \in [0,1]$. Since N_1 has the mixed $B(k-1, T_A)$ distribution with $T_A = \min\{T_k, \dots, T_\ell\}$ we obtain with Chebyshev's inequality, for $k \ge 4$ and Δ sufficiently large such that $\varepsilon/\Delta \le 1$,

$$\begin{aligned}
\mathbb{P}\Big(N_1 < \frac{\varepsilon k}{2\Delta}\Big) &\le \mathbb{P}\Big(T_A < \frac{\varepsilon}{\Delta}\Big) + \mathbb{P}\Big(B_{k-1,\varepsilon/\Delta} \le \frac{\varepsilon k}{2\Delta}\Big) \\
&\le 2\varepsilon + \mathbb{P}\Big(\Big|B_{k-1,\varepsilon/\Delta} - \frac{\varepsilon(k-1)}{\Delta}\Big| \ge \frac{\varepsilon k}{4\Delta}\Big) \\
&\le 2\varepsilon + \frac{16}{\varepsilon(k/\Delta)} \\
&\le 2\varepsilon + \frac{16}{\varepsilon \ln k} \\
&\to 2\varepsilon,
\end{aligned}$$

as $n \to \infty$. Similarly we obtain for sufficiently large Δ,

$$\begin{aligned}
\mathbb{P}\Big(N_1 > \frac{2k}{\varepsilon\Delta}\Big) &\le \mathbb{P}\Big(T_A > \frac{1}{\varepsilon\Delta}\Big) + \mathbb{P}\Big(B_{k-1,1/(\varepsilon\Delta)} \ge \frac{2k}{\varepsilon\Delta}\Big) \\
&\le 2e^{-1/\varepsilon} + \frac{\varepsilon}{\ln k} \\
&\to 2e^{-1/\varepsilon},
\end{aligned}$$

as $n \to \infty$. Hence we obtain $[\limsup_{n\to\infty} \mathbb{P}(N_1 \notin [\alpha_1, \beta_1])] \le 2(\varepsilon + e^{-1/\varepsilon}) \to 0$ as $\varepsilon \downarrow 0$.

In the second case we assume that $k/\Delta \le \ln k$ and $(n-\ell+1)/\Delta > \ln(n-\ell+1)$ for all n sufficiently large. Now we replace α_i, β_i by

$$\alpha_1' = 0, \quad \beta_1' = \ln^2 k, \qquad \alpha_2' = \alpha_2, \quad \beta_2' = \beta_2,$$

and define $B_\varepsilon, P_{k\ell}^-, P_{k\ell}^+$ as in the first case but with the α_i, β_i replaces by α_i', β_i', $i = 1, 2$. The argument is now applied as in the first case. The only difference to be shown is that we have $\limsup_{n\to\infty} \mathbb{P}(N_1 \notin [\alpha_1', \beta_1']) = 0$: We have

$$\mathbb{P}(N_1 \notin [\alpha_1', \beta_1']) = \mathbb{P}(N_1 > \ln^2 k) \le \frac{\mathbb{E}\, N_1}{\ln^2 k} = \frac{(k-1)/\Delta}{\ln^2 k} \le \frac{1}{\ln k} \to 0,$$

as $n \to \infty$.

The case $k/\Delta > \ln k$ and $(n-\ell+1)/\Delta \le \ln(n-\ell+1)$ is covered by the previous case by symmetry. In the remaining case $k/\Delta \le \ln k$ and $(n-\ell+1)/\Delta \le \ln(n-\ell+1)$ we replace α_i, β_i by

$$\alpha_1'' = \alpha_1', \quad \beta_1'' = \beta_1', \qquad \alpha_2'' = 0, \quad \beta_2'' = \ln(n-\ell+1)^2,$$

and define $B_\varepsilon, P_{k\ell}^-, P_{k\ell}^+$ as in the first case but with the α_i, β_i replaced by α_i'', β_i'', $i = 1, 2$. The argument is again applied as in the first case and $\limsup_{n\to\infty} \mathbb{P}(N_i \notin [\alpha_i'', \beta_i'']) = 0$ follows for $i = 1, 2$ as in the second case.

This finishes the proof of the limit law since for a given sequence $(k, \ell) = (k(n), \ell(n))$ with $\ell(n) - k(n) \to \infty$ we decompose into four subsequences according to whether $k/\Delta \le \ln k$ or $k/\Delta > \ln k$ and $(n-\ell+1)/\Delta \le \ln(n-\ell+1)$ or $(n-\ell+1)/\Delta > \ln(n-\ell+1)$. Each of the subsequences satisfies, by the previous arguments, the limit law (5), hence the whole sequence satisfies the limit law. ■

Proof of Theorem 1.4: We denote by (K, L) the ranks of the pair of nodes chosen uniformly at random from all possible pairs of distinct nodes in the tree, where we may assume that $K < L$. We define the set

$$B := \left\{K < \frac{n}{\ln n}\right\} \cup \left\{n - L < \frac{n}{\ln n}\right\} \cup \left\{L - K < \frac{n}{\ln n}\right\}$$

and note that $\mathbb{P}(B) \to 0$ as $n \to \infty$. On B^c we will condition on $(K, L) = (k, \ell)$. For these (k, ℓ) we have $\ln(k(\ell - k + 1)^2(n - \ell + 1)) = 4 \ln n + O(\ln \ln n)$. Hence application of Theorem 1.3 yields $(P_{k\ell} - 4 \ln n)/\sqrt{4 \ln n} \to \mathcal{N}(0, 1)$ in distribution.

Denoting by Φ the distribution function of $\mathcal{N}(0,1)$ and by σ the distribution of (K,L) we obtain, for all $x \in \mathbb{R}$,

$$\begin{aligned}
&\left|\mathbb{P}\left(\frac{P_n - 4\ln n}{\sqrt{4\ln n}} \le x\right) - \Phi(x)\right| \\
&\le \quad \mathbb{P}(B) + \int \left|\mathbb{P}\left(\frac{P_{k\ell} - 4\ln n}{\sqrt{4\ln n}} \le x\right) - \Phi(x)\right| d\sigma(k,\ell) \\
&\to \quad 0,
\end{aligned}$$

by dominated convergence. The assertion follows. ■

To prepare for the proof of Theorem 1.5 we provide the following tail estimate:

Lemma 3.1 *Let Y_j, $1 \le j \le n$ be independent and Y_j be Bernoulli $B(p_j)$ distributed for $0 \le p_j \le 1$, and $\mu = \sum_{j=1}^n p_j$. Then we have, for all $\varepsilon > 0$,*

$$\mathbb{P}\left(\sum_{j=1}^n Y_j \ge \mu + \varepsilon\right) \le \exp\left(-\frac{\varepsilon^2}{2\mu+\varepsilon}\right).$$

Proof: The proof relies on Chernoff's bounding technique. The details follow the proof of Theorem L1 in Devroye (1988). ■

Corollary 3.2 *Let X_j, X_j' be Bernoulli $B(1/j)$ distributed, $j \ge 1$, $Z_1 = 1$ and Z_j be $B(2/j)$ distributed, $j \ge 2$, such that all random variables are independent. Then for all $1 \le q \le s, \Delta \ge 1, 1 \le r \le t$ we have with $\alpha := s\Delta^2 t/(qr)$,*

$$\mathbb{P}\left(\sum_{j=q}^s X_j + \sum_{j=1}^{\Delta} Z_j + \sum_{j=r}^t X_j' - \ln\alpha \ge \varepsilon\right) \le \exp\left(-\frac{(\varepsilon-7)^2}{\varepsilon+6+2\ln\alpha}\right).$$

Proof: We apply Lemma 3.1 and note that from $\ln(n+1) \le H_n \le 1 + \ln n$ for $n \ge 1$, we obtain

$$\ln(\alpha) - 7 \le H_s - H_{q-1} + 2H_\Delta - 1 + H_t - H_{r-1} \le \ln(\alpha) + 3.$$

The assertion follows. ■

Proof of Theorem 1.5: First we prove the tail bound, where we distinguish several cases for the ranges of k and $n-\ell+1$. We abbreviate a_n as in Theorem 1.5. Let ε be arbitrarily given.

For $k \geq \Delta^{1+\varepsilon}$, $n-\ell+1 \geq \Delta^{1+\varepsilon}$ we have with the representations (3) and (4), and X_j, X'_j, and Z_j as in Corollary 3.2

$$\begin{aligned}
&\mathbb{P}(P_{k\ell} > (1+\varepsilon)\ln a_n) \\
&\quad\leq\ \mathbb{P}(\mathcal{P}_I + \mathcal{P}_{II} + \mathcal{P}_{III} > (1+\varepsilon)\ln a_n) \\
&\quad\leq\ \mathbb{P}\left(\left\{N_1 < \frac{k-1}{\Delta^{1+\varepsilon}}\right\} \cup \left\{N_2 < \frac{n-\ell}{\Delta^{1+\varepsilon}}\right\}\right) \qquad (8)\\
&\qquad + \mathbb{P}\left(\sum_{j=\lfloor (k-1)/\Delta^{1+\varepsilon}\rfloor+1}^{k-1} X_j + \sum_{j=1}^{\Delta} Z_j + \sum_{j=\lfloor (n-\ell)/\Delta^{1+\varepsilon}\rfloor+1}^{n-\ell} X'_j > (1+\varepsilon)\ln a_n\right).
\end{aligned}$$

Using that N_1 is $B(k-1, T_A)$ distributed and $T_A = \min\{T_k, \ldots, T_\ell\}$ we obtain

$$\mathbb{P}\left(N_1 < \frac{k-1}{\Delta^{1+\varepsilon}}\right) \leq \mathbb{P}\left(T_A < \Delta^{-(1+\varepsilon/2)}\right) + \mathbb{P}\left(B_{k-1,1/\Delta^{1+\varepsilon/2}} < \frac{k-1}{\Delta^{1+\varepsilon}}\right). \qquad (9)$$

The first summand in (9) is bounded by

$$\mathbb{P}\left(T_A < \Delta^{-(1+\varepsilon/2)}\right) = 1 - (1-\Delta^{-(1+\varepsilon/2)})^{\Delta} \leq \Delta^{-\varepsilon/2}.$$

For the second summand in (9) we use Okamoto's inequality (Okamoto, 1958), which states that $\mathbb{P}(B_{n,u} \leq ny) \leq \exp(-n(u-y)^2/(2u(1-u)))$ for all $y \leq u \leq 1/2$. For $y := \Delta^{-(1+\varepsilon)}$ and $u := \Delta^{-(1+\varepsilon/2)}$ we obtain, for Δ sufficiently large,

$$\begin{aligned}
\mathbb{P}\left(B_{k-1,1/\Delta^{1+\varepsilon/2}} < \frac{k-1}{\Delta^{1+\varepsilon}}\right) &\leq \exp\left(-(k-1)\frac{\left(\Delta^{-(1+\varepsilon/2)} - \Delta^{-(1+\varepsilon)}\right)^2}{2\Delta^{-(1+\varepsilon/2)}}\right) \\
&\leq \exp\left(-\frac{k-1}{8\Delta^{1+\varepsilon/2}}\right) \\
&\leq \exp\left(-\frac{k+1}{\Delta^{1+\varepsilon}}\frac{\Delta^{\varepsilon/2}}{24}\right) \\
&\leq \exp\left(-\frac{\Delta^{\varepsilon/2}}{24}\right) \\
&\leq 24\Delta^{-\varepsilon/2},
\end{aligned}$$

where we used that $(k+1)/\Delta^{1+\varepsilon} \geq 1$. Note that for this estimate Δ can be chosen uniformly large for all $\varepsilon \in [\delta, \infty)$, $\delta > 0$. By symmetry we obtain the same bound for $\mathbb{P}(N_2 < (n-\ell)/\Delta^{1+\varepsilon})$. The second summand in (8) we estimate with Corollary 3.2 for Δ sufficiently large:

$$\begin{aligned}
&\mathbb{P}\Bigg(\sum_{j=\lfloor (k-1)/\Delta^{1+\varepsilon}\rfloor+1}^{k-1} X_j + \sum_{j=1}^{\Delta} Z_j + \sum_{j=\lfloor (n-\ell)/\Delta^{1+\varepsilon}\rfloor+1}^{n-\ell} X'_j > (1+\varepsilon)\ln a_n\Bigg)\\
&\quad\le\ \mathbb{P}\Bigg(\sum_{j=\lfloor (k-1)/\Delta^{1+\varepsilon}\rfloor+1}^{k-1} X_j + \sum_{j=1}^{\Delta} Z_j + \sum_{j=\lfloor (n-\ell)/\Delta^{1+\varepsilon}\rfloor+1}^{n-\ell} X'_j - \ln\Delta^{4+2\varepsilon} > 2\varepsilon\ln\Delta\Bigg)\\
&\quad\le\ \exp\left(-\frac{(2\varepsilon\ln\Delta-7)^2}{2\ln\Delta^{4+2\varepsilon}+6+2\varepsilon\ln\Delta}\right)\\
&\quad\le\ \exp\left(\frac{28\varepsilon}{8+6\varepsilon}-\frac{4\varepsilon^2}{9+6\varepsilon}\ln\Delta\right)\\
&\quad\le\ e^5\Delta^{-\varepsilon^2/(3+2\varepsilon)}. \qquad (10)
\end{aligned}$$

Collecting the estimates we obtain $\mathbb{P}(P_{k\ell}\ge(1+\varepsilon)\ln a_n)\le 200\Delta^{-\varepsilon^2/(3+2\varepsilon)}$.

For the case $\Delta\le k\le\Delta^{1+\varepsilon}$ and $n-\ell+1\ge\Delta^{1+\varepsilon}$ we estimate

$$\begin{aligned}
&\mathbb{P}(P_{k\ell} > (1+\varepsilon)\ln a_n)\\
&\quad\le\ \mathbb{P}\left(N_2 < \frac{n-\ell}{\Delta^{1+\varepsilon}}\right)\\
&\qquad\quad + \mathbb{P}\Bigg(\sum_{j=1}^{\lfloor\Delta^{1+\varepsilon}\rfloor} X_j + \sum_{j=1}^{\Delta} Z_j + \sum_{j=\lfloor (n-\ell)/\Delta^{1+\varepsilon}\rfloor+1}^{n-\ell} X'_j > (1+\varepsilon)\ln a_n\Bigg),
\end{aligned}$$

and both summands can be estimated as in the previous case.

The same estimates apply to the cases $k\ge\Delta^{1+\varepsilon}$ and $\Delta\le n-\ell+1\le\Delta^{1+\varepsilon}$ as well as $\Delta\le k\le\Delta^{1+\varepsilon}$ and $\Delta\le n-\ell+1\le\Delta^{1+\varepsilon}$. The remaining cases are where either $k<\Delta$ or $n-\ell+1<\Delta$. If $k<\Delta$ and $n-\ell+1\ge\Delta^{1+\varepsilon}$ then

$$\begin{aligned}
&\mathbb{P}(P_{k\ell} > (1+\varepsilon)\ln a_n)\\
&\quad\le\ \mathbb{P}\left(N_2 < \frac{n-\ell}{\Delta^{1+\varepsilon}}\right)\\
&\qquad\quad + \mathbb{P}\Bigg(\sum_{j=1}^{k-1} X_j + \sum_{j=1}^{\Delta} Z_j + \sum_{j=\lfloor (n-\ell)/\Delta^{1+\varepsilon}\rfloor+1}^{n-\ell} X'_j > (1+\varepsilon)\ln a_n\Bigg),
\end{aligned}$$

where the first summand is bounded as before and the second one has the upper bound

$$\begin{aligned}
&\mathbb{P}\Bigg(\sum_{j=1}^{k-1} X_j + \sum_{j=1}^{\Delta} Z_j + \sum_{j=\lfloor (n-\ell)/\Delta^{1+\varepsilon}\rfloor+1}^{n-\ell} X_j' - \ln(k\Delta^{3+\varepsilon}) > 2\varepsilon\ln\Delta\Bigg) \\
&\quad\leq\ \exp\left(-\frac{(2\varepsilon\ln\Delta-7)^2}{2\ln(k\Delta^{3+\varepsilon})+6+2\varepsilon\ln\Delta}\right),
\end{aligned}$$

which leads to the bound given in (10) since $k\Delta^{3+\varepsilon} \leq \Delta^{4+2\varepsilon}$. For the case $k \leq \Delta$ and $\Delta \leq n-\ell+1 \leq \Delta^{1+\varepsilon}$ we estimate

$$\mathbb{P}(P_{k\ell} > (1+\varepsilon)\ln a_n) \leq \mathbb{P}\Bigg(\sum_{j=1}^{k-1} X_j + \sum_{j=1}^{\Delta} Z_j + \sum_{j=1}^{\lfloor \Delta^{1+\varepsilon}\rfloor} X_j' > (1+\varepsilon)\ln a_n\Bigg),$$

and, for the case $k \leq \Delta$ and $n-\ell+1 \leq \Delta$,

$$\mathbb{P}(P_{k\ell} > (1+\varepsilon)\ln a_n) \leq \mathbb{P}\Bigg(\sum_{j=1}^{k-1} X_j + \sum_{j=1}^{\Delta} Z_j + \sum_{j=1}^{n-\ell} X_j' > (1+\varepsilon)\ln a_n\Bigg),$$

and estimate as before. The remaining cases with $n-\ell+1 \leq \Delta$ are covered by symmetry.

To show the second claim of the Theorem, $\mathbb{E}\, P_{k\ell}^p \sim \ln^p a_n$, we fix $p \geq 1$ and $\delta \in (0,1)$. Then, by the first part, there is a $C > 0$ with $\mathbb{P}(P_{k\ell} \geq (1+\varepsilon)\ln a_n) \leq C\Delta^{-\varepsilon^2/(3+2\varepsilon)}$ for all Δ sufficiently large and all $\varepsilon \geq \delta$. We obtain

$$\begin{aligned}
\mathbb{E}\, P_{k\ell}^p &= \mathbb{E}\left[P_{k\ell}^p\big(\mathbf{1}_{\{P_{k\ell}\leq(1+\delta)\ln a_n\}} + \mathbf{1}_{\{P_{k\ell}>(1+\delta)\ln a_n\}}\big)\right] \\
&\leq (1+\delta)^p \ln^p a_n + \int_{(1+\delta)^p\ln^p a_n}^{\infty} \mathbb{P}(P_{k\ell}^p \geq t)\, dt \\
&\leq (1+\delta)^p \ln^p a_n + C\int_{(1+\delta)^p\ln^p a_n}^{\infty} \exp\left(-\frac{\varepsilon^2}{3+2\varepsilon}\ln\Delta\right) dt,
\end{aligned}$$

with $\varepsilon = \varepsilon(t) = (t^{1/p}/\ln a_n) - 1$.

Note that for any convex function $f : [t_0,\infty) \to \mathbb{R}$, $t_0 \in \mathbb{R}$, differentiable in t_0 with $f'(t_0) > 0$, we have

$$\int_{t_0}^{\infty} \exp(-f(t))\, dt \leq \frac{\exp(-f(t_0))}{f'(t_0)}.$$

This follows estimating $f(t) \geq f(t_0) + f'(t_0)(t-t_0)$ for all $t \geq t_0$ and evaluating the resulting integral.

Now, the function $f(t) = (\varepsilon^2/(3+2\varepsilon))\ln\Delta$ with $\varepsilon = \varepsilon(t)$ given above and $t_0 = (1+\delta)^p \ln^p a_n$ has the latter form. Hence an explicite calculation yields

$$\begin{aligned}
\int_{(1+\delta)^p \ln^p a_n}^{\infty} \exp\left(-\frac{\varepsilon^2}{3+2\varepsilon}\ln\Delta\right) dt &\leq \frac{\exp(-f(t_0))}{f'(t_0)} \\
&= \frac{p(1+\delta)^{p-1}\ln^p a_n}{(6\delta+2\delta^2)\ln\Delta}\Delta^{-\delta^2/(3+2\delta)} \\
&= O\left(\frac{\ln^{p-1}\Delta}{\Delta^{\delta^2/(3+2\delta)}}\right),
\end{aligned}$$

which gives a vanishing contribution as $\Delta \to \infty$.

Hence we obtain

$$\limsup_{n\to\infty} \frac{\mathbb{E}\, P_{k\ell}^p}{\ln^p a_n} \leq (1+\delta)^p$$

for all $\delta > 0$, hence $\limsup_{n\to\infty} \mathbb{E}\, P_{k\ell}^p / \ln^p a_n \leq 1$.

For the lower bound we choose $c \in \mathbb{R}$. Then for all n sufficiently large such that $a_n > \exp(c^2)$ we have

$$\begin{aligned}
\frac{\mathbb{E}\, P_{k\ell}^p}{\ln^p a_n} &\geq \frac{1}{\ln^p a_n}\mathbb{E}\left[\mathbf{1}_{\{(P_{k\ell}-\ln a_n)/\sqrt{\ln a_n}\geq c\}} P_{k\ell}^p\right] \\
&\geq \left(1+\frac{c}{\sqrt{\ln a_n}}\right)^p \mathbb{P}\left(\frac{P_{k\ell}-\ln a_n}{\sqrt{\ln a_n}} \geq c\right) \\
&\to 1-\Phi(c),
\end{aligned}$$

as $n \to \infty$, by Theorem 1.3, where Φ denotes the distribution function of the standard normal distribution. With $c \to -\infty$ we obtain $\liminf_{n\to\infty} \mathbb{E}\, P_{k\ell}^p / \ln^p a_n \geq 1$. ■

Acknowledgment

We thank the referees for their careful reading.

References

[1] Brodal, G. S. (1998) Finger search trees with constant insertion time. *Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms (San Francisco, CA, 1998)*, 540–549, ACM, New York.

[2] Brown, M. R. and Tarjan, R. E. (1980) Design and analysis of a data structure for representing sorted lists. *SIAM J. Comput.* **9**, 594–614.

[3] Chow, Y. S. and Teicher, H. (1978) Probability Theory. Springer-Verlag, New York-Heidelberg.

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

[5] Flajolet, P., Ottmann, T. and Wood, D. (1985) Search trees and bubble memories. *RAIRO Inform. Théor.* **19**, 137–164.

[6] Huddleston, S. and Mehlhorn, K. (1982) A new data structure for representing sorted lists. *Acta Inform.* **17**, 157–184.

[7] Louchard, G. (1987) Exact and asymptotic distributions in digital and binary search trees. *RAIRO Inform. Théor. Appl.* **21**, 479–495.

[8] Mahmoud, H. M. (1992) *Evolution of random search trees.* Wiley-Interscience Series in Discrete Mathematics and Optimization. John Wiley & Sons, Inc., New York.

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

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

[11] Okamoto, M. (1958) Some inequalities relating to the partial sum of binomial probabilities. *Ann. Inst. Statist. Math.* **10**, 29–35.

[12] Panholzer, A. and Prodinger, H. (2003+) Spanning tree size in random binary search trees. *Ann. Appl. Probab.*, accepted for publication.

[13] Prodinger, H. (1995) Multiple Quickselect—Hoare's Find algorithm for several elements. *Inform. Process. Lett.* **56**, 123–129.

[14] Seidel, R. and Aragon, C. R. (1996) Randomized search trees. *Algorithmica* **16**, 464–497.