JIRSS (2004)
Vol. 3, No. 2, pp 117-138

Profile and Height of Random Binary Search Trees

Michael Drmota

Department for Discrete Mathematics and Geometry, TU Wien, Wiedner Hauptstrasse 8-10/118, A-1040 Wien, Austria. (drmota@tuwien.ac.at)

Abstract. The purpose of this article is to survey recent results on distributional properties of random binary search trees. In particular we consider the profile and the height.

1 Introduction

Probably the most widely used sorting algorithm is the algorithm *Quicksort* which was invented by C. A. R. Hoare [14, 15]. It is the standard sorting procedure in Unix systems, and the basic idea can be described as follows:

> A list of n (different) real numbers $A = (x_1, x_2, \ldots, x_n)$ is given. Select an element (a pivot) x_j from this list. Divide the remaining numbers into sets $A_{\leq}, A_{>}$ of numbers smaller (or equal) and larger than x_j. Next apply the same procedure to each of these two sets if they contain more than one element. Finally, we end up with a sorted list of the original numbers.

Received: November 2003, Revised: November 2003
Key words and phrases: binary search tree, generating functions, height, limiting distribution, profile.

This sorting procedure can be encoded with a binary tree with n (internal) nodes.[1] The first selected element x_j is put in the root, whereas recursively $A_{\leq}$ produces a left subtree of x_j and $A_{>}$ the right subtree of x_j. (An empty string produces an empty tree which is usually encoded as an external node.)

These kinds of binary trees are also called *binary search trees* and are quite common as a data structure to store data represented by keys which can be totally ordered (compare with [18, 20]). It is then easy to search for an item by comparing it with the root and then proceeding to the left subtree if it is smaller and to the right subtree if it is larger.

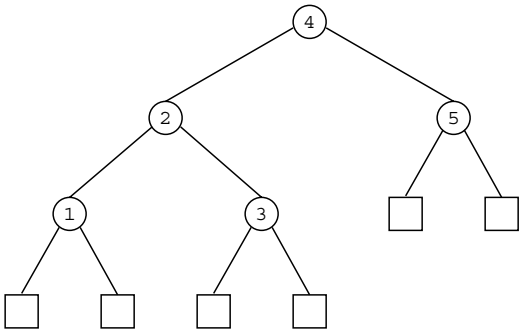

Figure 1: Binary search tree generated by the list $(4, 2, 3, 5, 1)$, where the pivot element is always the first element.

Notation

Consider a (finite) binary tree t, that is, a rooted tree, where every node has either 0 or 2 descendents. A node with 2 descendents is called *internal node* and a node with 0 descendents *external node*. For every node x the distance to the root is denoted by $h(x)$. The (internal) *path length* is given by

$$l = l_t = \sum_{x \text{ internal node of } t} h(x).$$

The number of external nodes at level k is denoted by

$$u_k = \#\{x : h(x) = k,\ x \text{ external node of } t\}$$

[1] The nodes of a (rooted) binary tree can be divided into *internal* nodes with two descendents and *external* nodes with no descendents.

and the number of internal nodes at level k by

$$v_k = \#\{x : h(x) = k,\ x \text{ internal node of } t\}.$$

Note that

$$v_k = \sum_{j>k} 2^{k-j} u_j$$

which can easily be proved by induction. The *external profile* of t is then given by the sequence $(u_k)_{k\geq 0}$ and the *internal profile* by $(v_k)_{k\geq 0}$. Observe that the (internal) path length is also given by

$$l_t = \sum_{k\geq 0} k v_k.$$

Finally, the *height* of t is defined by

$$h = h_t = \max_{x \text{ internal node of } t} h(x) = \max\{k \geq 0 : v_k > 0\}$$

and the *fill-up-level* by

$$\overline{h} = \overline{h}_t = \max\{k \geq 0 : v_k = 2^k\}.$$

Probabilistic Model

When analyzing Quicksort or binary search trees it is standard to assume that the data $x_1 = X_1, x_2 = X_2, \ldots, x_n = X_n$ are iid real random variables with a (common) continuous probability distribution. For example, this implies that their ranks form a random permutation of $\{1, 2, \ldots, n\}$. Thus, the kind of the distribution of X_j $(1 \leq j \leq n)$ has no influence on the distribution of the parameters of Quicksort or the corresponding binary search tree. It is therefore no loss of generality to assume that X_j are uniformly distributed on $[0, 1]$. Even the choice of the pivot element (in Quicksort) does not change the probabilistic structure. It is therefore common to assume that the pivot element is always the first in the corresponding list.[2]

In this context it is also natural to consider an infinte sequence $(X_n)_{n\geq 1}$ of iid random variables (uniformly distributed on $[0, 1]$) which induce a random sequence $(T_n)_{n\geq 0}$ of binary search trees. This means that T_n contains n internal nodes, where the data $X_1, X_1, \ldots, X_n$ are *stored* in a way that all (internal) nodes of the left subtree of X_j

[2]In Unix systems the pivot element is always an element in the middle position.

have smaller values than X_j and all nodes of the right subtree are larger than X_j. Furthermore, T_{n+1} is generated from T_n by inserting X_{n+1} at one of the $n+1$ external nodes of T_n in the following way. One starts at the root and goes to the left subtree if X_{n+1} is smaller than the value of the root and to the right subtree if X_{n+1} is larger. This procedure is recursively applied until one reaches an external node where X_{n+1} is inserted. By assumption any of these $n+1$ external nodes (or *free places*) is replaced by X_{n+1} with equal probability $1/(n+1)$. Thus, we also have a kind of Markov property: T_{n+1} just depends on T_n (and this in a very simple way).

In what follows we will discuss just parameters in (random) binary search trees. As explained above there is a direct correspondance to Quicksort. For example, the number of comparisions in Quicksort that are needed to sort n data is exactly the internal path length of the corresponding binary search tree.

2 Profile

We now consider the sequence $(T_n)_{n\geq 0}$ of binary search trees generated by an iid sequence $(X_n)_{n\geq 1}$ of random variables which are uniformly distributed (ud) on $[0,1]$. The internal profile will be denoted $V_{k,n}$ (that is, $V_{k,n}$ equals the number of internal nodes of T_n at level k) and the external profile by $U_{k,n}$. It turns out that $U_{k,n}$ is much easier to handle. On the other hand we have

$$V_{k,n} = \sum_{j>k} 2^{k-j} U_{j,n}. \tag{1}$$

Therefore we will work mainly with $U_{k,n}$. Corresponding results for $V_{k,n}$ are then immediate corollaries.

2.1 Expected Profile

We introduce the generating functions

$$Y_k(x,u) := \sum_{n\geq 0} \mathbf{E}\, u^{U_{k,n}} \cdot x^n. \tag{2}$$

Then we have $Y_0(x,u) = u + x/(1-x)$ and recursively

$$\frac{\partial Y_{k+1}(x,u)}{\partial x} = Y_k(x,u)^2,$$

with $Y_k(0,u) = 1$. The reason for the appearance of the square $Y_k(x,u)^2$ on the right hand side of this equation is that the two subtrees of the root of a random binary search tree (with n external nodes) can be considered as two independent binary search trees (with ℓ resp. $n-\ell$ external nodes, where every $\ell = 0, 1, \dots n$ appears with equal probabilty $1/(n+1)$). Furthermore, if one wants to count the number of nodes at level $k+1$ then one has to add the number of nodes at level k of the two subtrees of the root.

There is no known method to solve this kind or recurrence (explicitly or asymptotically). Nevertheless it can be used to derive the expected profile. By definition we have

$$Z_k(x) := \frac{\partial Y_k(x,1)}{\partial u} = \sum_{n\geq 0} \mathbf{E}\, U_{k,n} \cdot x^n.$$

Furthermore, $Z_0(x) = 1$ and by (2)

$$Z'_{k+1}(x) = 2Y_k(x,1)Z_k(x) = \frac{2}{1-x} Z_k(x),$$

with $Z_{k+1}(0) = 0$ (for $k \geq 0$). Hence,

$$Z_k(x) = \frac{2^k}{k!} \left(\frac{1}{1-x}\right)^k$$

and one obtains

$$\mathbf{E}\, U_{k,n} = \frac{2^k}{n!} s_{n,k}, \tag{3}$$

where $s_{n,k}$ are the (absolute) Stirling numbers of the first kind – in other words the number of permutations σ of n elements such that the canonical cyclic representation of σ has exactly k cycles. (It seems that this explicit formula was first observed by Lynch [19], compare also with [20]). By well known asymptotics for Stirling numbers (see [21]) we can derive the following asymptotic relations for $\mathbf{E}U_{k,n}$ and $\mathbf{E}V_{k,n}$

Theorem 2.1.1. *Set $\alpha_{n,k} = k/\log n$. Then we have*

$$\mathbf{E}\, U_{k,n} \sim \frac{n^{\alpha_{n,k}\left(1-\log(\alpha_{n,k}/2)\right)-1}}{\sqrt{2\pi k}}, \tag{4}$$

uniformly for $\alpha_{n,k} = \mathcal{O}(1)$.

Furthermore, if $1+\varepsilon \le \alpha_{n,k} = \mathcal{O}(1)$ *(where* $\varepsilon > 0$*) then*

$$\begin{aligned}\mathbf{E}\, V_{k,n} &\sim \frac{\mathbf{E}\, U_{k,n}}{\alpha_{n,k}-1} \\ &\sim \frac{n^{\alpha_{n,k}\left(1-\log(\alpha_{n,k}/2)\right)-1}}{(\alpha_{n,k}-1)\sqrt{2\pi k}},\end{aligned}$$

and for $\alpha_{n,k} \le 1-\varepsilon$,

$$\mathbf{E}\, V_{k,n} \sim 2^k.$$

Note that for k close to $2\log n$ these estimates show that

$$\mathbf{E}U_{k,n} = \frac{n}{\sqrt{4\pi\log n}} e^{-\frac{(k-2\log n)^2}{4\log n}} + \mathcal{O}\left(\frac{n}{\log n}\right)$$

and

$$\mathbf{E}V_{k,n} = \frac{n}{\sqrt{4\pi\log n}} e^{-\frac{(k-2\log n)^2}{4\log n}} + \mathcal{O}\left(\frac{n}{\log n}\right).$$

This indicates that the *mass* of a binary search tree T_n is concentrated around level $2\log n$. This fact is also reflected by the well known asymptotic formula for the expectation of the internal path length $L_n = \sum_{k\ge 0} kV_{k,n}$:

$$\begin{aligned}\mathbf{E}L_n &= 2(n+1)\sum_{h=1}^{n+1}\frac{1}{h} - 4(n+1) + 2 \\ &= 2n\log n + n(2\gamma-4) + 2\log n + 2\gamma + 1 + \mathcal{O}\left(\frac{\log n}{n}\right)\end{aligned}$$

with $\gamma = 0.57721...$ being Euler's constant. Recall that L_n is also the number of comparision in Quicksort to sort n items.

2.2 Profile Polynomials

Next we consider the so-called *random profile polynomials*

$$W_n(z) = \sum_{k\ge 0} U_{k,n} z^k.$$

By (3)

$$\mathbf{E}W_n(z) = (-1)^n \binom{-2z}{n}. \tag{5}$$

The basic property of $W_n(z)$ is that the normalized version is a martingale (see [16]).

Lemma 2.2.1. *The random (analytic) functions*

$$M_n(z) = \frac{W_n(z)}{\mathbf{E}W_n(z)}$$

constitute a martingale with respect to the natural filtration $(\mathcal{F}_n)_{n\geq 0}$ *associated to the sequence of trees* $(T_n)_{n\geq 0}$.

Proof. With the help of the above description of how T_{n+1} evolves from T_n one has

$$\begin{aligned}\mathbf{E}(U_{k,n+1}|\mathcal{F}_n) &= (U_{k,n}+2)\frac{U_{k-1,n}}{n+1} + (U_{k,n}-1)\frac{U_{k,n}}{n+1} \\ &+ U_{k,n}\left(1 - \frac{U_{k-1,n}+U_{k,n}}{n+1}\right) \\ &= \frac{2U_{k-1,n}}{n+1} + \frac{nU_{k,n}}{n+1}.\end{aligned}$$

Hence

$$\mathbf{E}(W_{n+1}(z)|\mathcal{F}_n) = \frac{2z+n}{n+1}W_n(z).$$

Consequently

$$\mathbf{E}(M_{n+1}(z)|\mathcal{F}_n) = M_n(z),$$

which completes the proof of the martingale property. □

Hence, for positive values of z, the martingale converges to an almost sure limit $M(z)$ that has quite interesting properties.

Lemma 2.2.2. **([16])** *Let* $c' = 0.37\ldots$ *and* $c = 4.31\ldots$ *be the two solutions of the equation* $c\log(\frac{2e}{c}) = 1$ *and set* $z_c^- = c'/2 = 0.186\ldots$ *and* $z_c^+ = c/2 = 2.155\ldots$. *Then the limiting martingle* $M(z)$ *is positive a.s. for positive real* $z \in (z_c^-, z_c^+)$ *and* $M(z) = 0$ *if* $z \notin [z_c^-, z_c^+]$.

Interestingly $M_n(z)$ converges, too, for certain complex values of z.

Lemma 2.2.3. **([1])** *For any compact set* $C \subseteq \{z \in \mathbf{C} : |z-1| < 1/\sqrt{2}\}$ *the martingale* $M_n(z)$ *converges a.s. uniformly to its limit* $M(z)$ *(which is again an analytic function).*

Furthermore, the distribution of $M(z)$ can be characterized by a fixed point equation.

Lemma 2.2.4. ([3]) *Suppose that z is positive and real. Then the distribution of $M(z)$ equals the distribution of*

$$zU^{2z-1}M(z) + z(1-U)^{2z-1}\overline{M}(z),$$

where $\overline{M}(z)$ is a copy of $M(z)$ and U is ud on $[0,1]$ such that $M(z), \overline{M}(z)$, and U are independent.

2.3 Asymptotic Properties of the Profile

The idea behind the following result for the profile $U_{k,n}$ is that the ratio $U_{k,n}/\mathbf{E}\,U_{k,n}$ can be estimated by the ratio $M_n(z) = W_n(z)/\mathbf{E}\,W_n(z)$, where $z = k/(2\log n)$. Note that $z = k/(2\log n)$ is exactly the saddle point of the function $\mathbf{E}\,W_n(z)z^{-k}$. In fact, with the help of the martingale properties of $M_n(z)$ the following results can be derived.

Theorem 2.3.1. ([1, 3]) *Suppose that $k = k(n) = 2z\log n + o(\sqrt{n})$ and $z \in (z_c^-, z_c^+)$. Then*

$$\lim_{n\to\infty} \frac{U_{k,n}}{\mathbf{E}U_{k,n}} = M(z)$$

in probability.

Furthermore, we have a.s. that

$$\frac{U_{k,n}}{\mathbf{E}U_{k,n}} = M\left(\frac{k}{2\log n}\right) + o(1)$$

uniformly for $1.2 \le k/\log n \le 2.8$.

The idea behind the proof is the following one. We know that

$$W_n(z) \sim M(z)\cdot \mathbf{E}W_n(z) \qquad a.s. \tag{6}$$

if z is complex and $|z-1| < 1/\sqrt{2}$. Consequently we can use Cauchy's formula to evaluate $U_{k,n}$:

$$\begin{aligned} U_{k,n} &= \frac{1}{2\pi i}\int_{|z|=k/(2\log n)} \frac{W_n(z)}{z^{k+1}}\,dz \\ &\approx M\left(\frac{k}{2\log n}\right)\cdot\frac{1}{2\pi i}\int_{|z|=k/(2\log n)} \frac{\mathbf{E}\,W_n(z)}{z^{k+1}}\,dz \\ &= M\left(\frac{k}{2\log n}\right)\cdot \mathbf{E}\,U_{k,n}. \end{aligned}$$

Of course, there are some *missing details*, see [1, 3].

In particular it follows that $U_{k,n}/\mathbf{E}\,U_{k,n}$ has a limiting distribution $M(z)$ if $k = k(n) = 2z\log n + o(\sqrt{\log n})$ and $n \to \infty$.

Furthermore, by using the fact that $W_n'(1) = \sum_{k\geq 0} kU_{k,n}$ and that $L_n = \sum_{k\geq 0} kV_{k,n} = \sum_{k\geq 0} kU_{k,n} - 2n$ it also follows that

$$M_n'(1) = \frac{L_n - \mathbf{E}\,L_n}{n+1}$$

Thus,

$$\frac{L_n - \mathbf{E}\,L_n}{n+1} \to M'(1) \quad a.s.$$

Note also that $M'(1)$ has the same distribution as

$$U\,M'(1) + (1-U)\overline{M}'(1) + 2U\log U + 2(1-U)\log(1-U) + 1,$$

where $\overline{M}'(1)$ is a copy of $M'(1)$ and U is ud on $[0,1]$ such that $M'(1), \overline{M}'(1), U$ are independent, compare with [27]. The distribution of $M'(1)$ is also called *Quicksort distribution*. It is known that there exists a density ([28]), which is a bounded C^∞ function, tail estimates are available, and orders of convergence are estimated (compare with [11, 12, 13, 17]). However, no explicit representations for the limiting distribution are known.

3 Height

The distribution of the height H_n of binary search trees has turned out to be an interesting (and difficult) problem. We start with some history.

In 1986 Devroye [4] proved that the expected value $\mathbf{E}H_n$ satisfies the asymptotic relation $\mathbf{E}H_n \sim c\log n$ (as $n \to \infty$), where $c = 4.31107\ldots$ is the (largest real) solution of the equation $c\log\left(\frac{2e}{c}\right) = 1$. (Earlier Pittel [22] had shown that $H_n/\log n \to \gamma$ almost surely as $n \to \infty$, where $\gamma \leq c$, compare also with Robson [24]. Later Devroye [5] provided a first bound for the error term; he proved $H_n - c\log n = \mathcal{O}(\sqrt{\log n \log\log n})$ in probability.) Based on numerical data Robson conjectured that the variance $\mathbf{Var}H_n$ is bounded. In fact, he could prove (see [25]) that there is an infinite subsequence for which

$$\mathbf{E}|H_n - \mathbf{E}H_n| = \mathcal{O}(1),$$

and that his conjecture is equivalent to the assertion that the expected value $\mathbf{E}\,V_{H_n,n}$ of the number of nodes at level $k = H_n$ is bounded (see [26] or section 3.2). The best bounds (before 1999) were given using two completely different methods by Devroye and Reed [6] and later by Drmota [7]. They (both) proved

$$\mathbf{E}H_n = c\log n + \mathcal{O}(\log\log n) \tag{7}$$

and

$$\mathbf{Var}H_n = \mathbf{E}(H_n - \mathbf{E}H_n)^2 = \mathcal{O}((\log\log n)^2).$$

Eventually, Reed [23] settled Robson's conjecture:

$$\mathbf{Var}H_n = \mathcal{O}(1) \qquad (n \to \infty).$$

Reed's approach is related to that of [6], moreover he could also show that

$$\mathbf{E}H_n = c\log n - \frac{3c}{2(c-1)}\log\log n + \mathcal{O}(1). \tag{8}$$

A second proof of Robson's conjecture was given (independently) by the author [8] (just a few months later than Reed).

Before stating results on the distribution of the height of binary search trees we want to present a first flavour of this problem. It is clear that $V_{k,n} > 0$ is equivalent to $H_n > k$. By (1) and (4) it follows that $\mathbf{E}\,V_{n,k} < 1$ if $k > c\log n - \frac{c}{2(c-1)}\log\log n + \mathcal{O}\,(1)$, where $c = 4.31107\ldots$ is the (largest real) solution of the equation $c\log\left(\frac{2e}{c}\right) = 1$. Hence, one might expect that H_n is *concentrated* around $c_n := c\log n - \frac{c}{2(c-1)}\log\log n$. We can be even more precise. Since

$$\mathbf{Pr}[H_n > k] \le \mathbf{Pr}[V_{k,n} > 0] \le \mathbf{E}U_{k,n}$$

we get (with the help of (4))

$$\begin{aligned}\mathbf{E}H_n &= \sum_{k\ge 0}\mathbf{Pr}[H_n > k]\\ &\le c_n + \sum_{k\ge c_n}\mathbf{E}V_{k,n}\\ &\le c_n + \mathcal{O}\,(1)\,.\end{aligned}$$

This estimate would be optimal if $\mathbf{E}\,V_{k,n}^2 = \mathcal{O}\,(1)$ for $k > c_n$. However, this is not true. And this is really the crux of the matter. As mentioned above, see (8), the expected height is definitely smaller.

3.1 Distribution of the Height

Let

$$y_k(x) = \sum_{n\geq 0} \mathbf{Pr}[H_h \leq k] \cdot x^n.$$

Then $y_0(x) \equiv 1$ and

$$y'_{k+1}(x) = y_k(x)^2$$

with initial condition $y_{k+1}(0) = 1$. Obviously, $y_k(x)$ are polynomials of degree $2^k - 1$ and have a limit $y(x) = 1/(1-x)$ (for $0 \leq x < 1$). Thus, there is a singularity for $x = 1$, and in fact the main characteristics can be formulated in terms of the (singular) sequence $y_k(1)$.

Theorem 3.1.1. ([9]) *There exists a monotonically decreasing function $\Psi(y)$, $y \geq 0$, with $\Psi(0) = 1$ and $\lim_{y\to\infty} \Psi(y) = 0$ satisfying the integral equation*

$$y\Psi(y/e^{1/c}) = \int_0^y \Psi(z)\Psi(y-z)\,dz, \tag{9}$$

such that

$$\mathbf{Pr}[H_n \leq k] = \Psi(n/y_k(1)) + o(1) \quad (n \to \infty), \tag{10}$$

with the $o(1)$-error term being uniform for all $k \geq 0$. Furthermore, there exist constants $C, \eta > 0$ such that

$$\mathbf{Pr}[|H_n - h_n| \geq y] \leq Ce^{-\eta y}, \qquad (y > 0), \tag{11}$$

where $h_n = \max\{k : y_k(1) \leq n\}$.

Especially, it follows from Theorem 3.1.1 that the expected value of the height H_n of binary search trees of size n is given by

$$\mathbf{E}H_n = \max\{k : y_k(1) \leq n\} + \mathcal{O}(1) \qquad (n \to \infty). \tag{12}$$

and that all centralized moments are bounded

$$\mathbf{E}|H_n - \mathbf{E}H_n|^r = \mathcal{O}(1) \qquad (n \to \infty). \tag{13}$$

If one combines (8) and (12) one gets

$$y_k(1) = e^{k/c + \frac{3}{2(c-1)}\log k + \mathcal{O}(1)}. \tag{14}$$

It would be (very) interesting to find a direct proof of (14) or even tighter estimates.

Note that (13) solves Robson's conjecture, however, in a quite implicit way. Interestingly, Theorem 3.1.1 does not provide any precise information on the magnitude of $y_k(1)$ and thus (via (12)) no quantitative bound for the expected height $\mathbf{E}H_n$. The exact order of $\mathbf{E}H_n$ was given by Reed [23] by improving the previous bound (7) by Devroye and Reed [6] and by the author [7].

We also want to mention that there are similar results for the height of m-ary search trees and also for the fill-up-level, see [2].

In what follows we present a sketch of the proof of Theorem 3.1.1. The first step of the proof is to solve the fixed point equation (9). In particular it can be shown (see [9]) that there exists a unique solution $\Psi(y)$ provided that $\Psi(y)$ is decreasing and $\int_0^\infty \Psi(y)\,dy = 1$. Furthermore, one has $1 - \Psi(y) \sim c_1 y^{c-1} \log y$ as $y \to 0+$ for some constant c_1 and $\Psi(y) = \mathcal{O}\left(e^{-Cy^\gamma}\right)$ as $y \to \infty$ for some $C > 0$ and some $\gamma > 1$. With the help of the function $\Psi(y)$ we define auxiliary functions

$$\tilde{y}_k(x) := \int_0^\infty \Psi(ye^{-k/c})e^{-y(1-x)}\,dy, \tag{15}$$

where k is an arbitrary real (not necessarily an integral) number. In some sense these functions *simulate* the above polynomials $y_k(x)$. They saytisfy

$$\tilde{y}'_{k+1}(x) = \tilde{y}_k(x)^2$$

with initial condition

$$1 - \tilde{y}_k(0) \sim \frac{C_1}{c} k \left(\frac{2}{c}\right)^k \quad (k \to \infty).$$

Furthermore, by definition $\tilde{y}_k(1) = e^{k/c}$, and $\tilde{y}_k(x)$ has a power series expansion $\tilde{y}_k(x) = \sum_{n\ge 0} a_{nk} x^n$ with positive coefficients $a_{nk} > 0$ that are asymptotically given by

$$a_{nk} = \Psi(ne^{-k/c}) + o(1),$$

where the $o(1)$ error term is uniform for all integers $n \ge 0$ and all real numbers $k \ge 0$. Finally, there is a crucial *intersection property*. For every integer $k \ge 0$ and for every real number l the difference $y_k(x) - \tilde{y}_l(x)$ has exactly one zero $x_{k,l}$ on the positive real line. Furthermore, these zeros satisfy $x_{k+1,l+1} > x_{k,l}$. (Interestingly the proof is immediate by induction on k.)

With the help of these auxiliary functions $\tilde{y}_k(x)$ we obtain proper tail estimates of the distribution of H_n.

Lemma 3.1.1. *Set $e_k = c \log y_k(1)$. Then $e_{k+1} \geq e_k + 1$ and there exists a constant $C < 0$ such that*

$$\mathbf{Pr}[H_n \leq k] \leq C e^{-(c \log n - e_k)/c}$$

and

$$\mathbf{Pr}[H_n > k] \leq C e^{-(e_k - c \log n)/c}.$$

Proof. Set $\alpha = e^{1/c}$. By definition $\tilde{y}_{e_k}(1) = y_k(1)$. Thus, $\tilde{y}_{e_k}(x) \leq y_k(x)$ for $0 \leq x \leq 1$ and $\tilde{y}_{e_k}(x) \geq y_k(x)$ for $x \geq 1$. In particular it follows that $\tilde{y}_{e_k+1}(x) \leq y_{k+1}(x)$ for $0 \leq x \leq 1$ and consequently $\alpha^{e_k+1} \leq \alpha^{e_{k+1}}$ which gives $e_{k+1} \geq e_k + 1$.

Suppose that $x \geq 1$. Then we get (by using the trivial inequality $\mathbf{Pr}[H_n \leq k] \geq \mathbf{Pr}[H_{n+1} \leq k]$)

$$\tilde{y}_{e_k}(x) \geq y_k(x) \geq \sum_{l=0}^{n} \mathbf{Pr}[H_l \leq k]\, x^l \geq \mathbf{Pr}[H_n \leq k]\, \frac{x^{n+1} - 1}{x - 1}.$$

Choosing $x = 1 + \alpha^{-e_k}$ and using the definition of $\tilde{y}_{e_k}(x)$ we obtain the upper bound

$$\begin{aligned} \mathbf{Pr}[H_n \leq k] &\leq \frac{1}{(1 + \alpha^{-e_k})^{n+1} - 1} \int_0^\infty \Psi(z) e^z \, dz \qquad (16) \\ &\ll \frac{1}{n \alpha^{-e_k}} = \alpha^{-(c \log n - e_k)}. \end{aligned}$$

In the same fashion we have for $0 < x < 1$

$$\begin{aligned} \frac{1}{1-x} - \tilde{y}_{e_k}(x) &\geq \frac{1}{1-x} - y_k(x) \\ &\geq \sum_{l=n}^{\infty} (1 - \mathbf{Pr}[H_l \leq k])\, x^l \\ &\geq (1 - \mathbf{Pr}[H_n \leq k])\, \frac{x^n}{1-x}. \end{aligned}$$

Finally, setting $x = 1 - 1/n$ we directly get

$$1 - \mathbf{Pr}[H_n \leq k] \ll 1 - \alpha^{e_k - c \log n} \Phi(\alpha^{e_k - c \log n}), \qquad (17)$$

where $\Phi(u) = \int_0^\infty \Psi(z)e^{-zu}\,du$ denotes the Laplace transform of $\Psi(z)$. Since $1-\Psi(y) \sim c_1 y^{c-1}\log y$ we have

$$1 - u\Phi(u) \sim c_2(\log u)/u^{c-1} \ll 1/u.$$

Hence

$$1 - \mathbf{Pr}[H_n \le k] \ll \alpha^{-(e_k - c\log n)}. \quad \square$$

Obviously, the tail estimate (11) follows from Lemma 3.1.1. In order to complete the proof of Theorem 3.1.1 we have to refine the methods a little bit. The idea is to approximate $y_k(x)$ by $\tilde{y}_{e_k}(x)$. (Recall that e_k was defined such that $y_k(1) = \tilde{y}_{e_k}(1)$.) In order to do this we use the fact (see [9]) that the sequence $y_{k+1}(1)/y_k(1)$ converges and its limit is given by

$$\lim_{k\to\infty} \frac{y_{k+1}(1)}{y_k(1)} = \alpha = e^{1/c}. \tag{18}$$

It then follows that

$$y_k'(1) = y_{k-1}(1)^2 \sim y_k(1)^2\alpha^{-2} = \tilde{y}_{e_k-1}(1)^2 = \tilde{y}_{e_k}'(1),$$

and inductively, one gets for every fixed $l \ge 0$ that $y_k^{(l)}(1) \sim \tilde{y}_{e_k}^{(l)}(1)$ as $k \to \infty$. Thus, $y_k(x)$ can be properly approximated by $\tilde{y}_{e_k}(x)$ in a neighbourhood of $x = 1$ in the complex plane. Together with some further (technical but easy) estimates (compare with [9]) it follows via Cauchy's formula that

$$\begin{aligned}
\mathbf{Pr}[H_n \le k] &= \frac{1}{2\pi i}\int_{|x|=1} \frac{y_k(x)}{x^{n+1}}\,dx \\
&= \frac{1}{2\pi i}\int_{|x|=1} \frac{\tilde{y}_{e_k}(x)}{x^{n+1}}\,dx + o(1) \\
&= \Psi(n/\alpha^{e_k}) + o(1) \\
&= \Psi(n/y_k(1)) + o(1),
\end{aligned}$$

which completes the proof of Theorem 3.1.1.

3.2 Nodes of Largest Distance

Let $C_n = V_{H_n,n}$ denote the number of internal nodes in a random binary search tree (of n nodes) at the maximal level H_n. It turns

out that the expected values $\mathbf{E}\, C_n$ is of particular interest. First of all, it has been shown by Robson [26] that $\mathbf{E}\, C_n$ remains bounded (as $n \to \infty$) if and only if $\mathbf{Var}\, H_n$ remains bounded. Since the second assertion has been verified (see (13)) it thus follows that $\mathbf{E}\, C_n = \mathcal{O}\,(1)$ (as $n \to \infty$). Robson has also conjectured that the sequence $\mathbf{E}\, C_n$ has a limit. At least, numerically it looks convergent. It is increasing for $7 \le n \le 100\,000$.

In what follows we indicate a direct proof of the property $\mathbf{E}\, C_n = \mathcal{O}\,(1)$. Furthermore we observe that $\mathbf{E}\, C_n$ is asymptotically (multiplicatively) periodic which shows that Robson's convergence conjecture is only true if a corresponding limiting periodic function $\tilde{C}(x)$ (see (25)) is constant. Interestingly $\tilde{C}(x)$ looks constant (numerically) and it can be shown that the possible oscillation are very small. However, there are strong indications that $\tilde{C}(x)$ is not constant. Thus, we are confronted here with a new almost constancy phenomenon. Interestingly this observation seems to be in contrast to Robson's numerical experiments.

With the help of the sequence $y_k(1)$ and the derivative of the function $\Psi(y)$ one can introduce the function

$$C(x) := -\frac{1}{2} \sum_{k \ge 0} \frac{x}{y_k(1)} \Psi' \left(\frac{x}{y_k(1)} \right). \tag{19}$$

Due to proper tail estimates for $\Psi'(y)$ (that are similar to those of $\Psi(x)$) it follows that $C(x)$ is a bounded function for $x > 0$. Furthermore, the limiting relation (18) implies that $C(x)$ is *almost periodic* in the sense that

$$C(e^{1/c} x) = C(x) + o(1) \qquad (x \to \infty). \tag{20}$$

With the help of this function we can formulate the following result:

Theorem 3.2.1. ([10]) *Let C_n denote the number of internal nodes in T_n at level H_n. Then the sequence $\mathbf{E}\, C_n$ remains bounded for $n \to \infty$. It is asymptotically given by*

$$\mathbf{E}\, C_n = C(n) + o(1) \qquad (n \to \infty) \tag{21}$$

and it is asymptotically periodic in the sense that

$$\mathbf{E}\, C_{\lfloor e^{1/c} n \rfloor} = \mathbf{E}\, C_n + o(1) \qquad (n \to \infty). \tag{22}$$

Furthermore, the sequence $\mathbf{E}\, C_n$ *is almost constant. There exists* n_0 *such that*

$$\max_{n\ge n_0}\left|\mathbf{E}\, C_n - \frac{c}{2}\right| \le 10^{-4}. \tag{23}$$

and we have

$$\lim_{n\to\infty} \sum_{k=n}^{\lfloor e^{1/c}n\rfloor} \frac{\mathbf{E}\, C_k}{k} = \frac{1}{2}. \tag{24}$$

The periodicity behaviour of $\mathbf{E}\ C_n$ can be stated in a little bit more precise form. Set

$$\tilde{C}(x) := -\frac{1}{2}\sum_{k=-\infty}^{\infty} xe^{-k/c}\Psi'\left(xe^{-k/c}\right). \tag{25}$$

Then $\tilde{C}(x)$ is in fact (multiplicatively) periodic, that is, $\tilde{C}(e^{1/c}x) = \tilde{C}(x)$ and we have, as $x\to\infty$,

$$C(x) = \tilde{C}\left(\frac{x}{y_{h_0(x)}}\right) + o(1) \qquad (x\to\infty),$$

where $h_0(x)$ is uniquely defined by $y_{h_0(x)}(1) \le x < y_{h_0(x)+1}(1)$. Consequently

$$\mathbf{E}\, C_n = \tilde{C}\left(\frac{n}{y_{h_0(n)}}\right) + o(1) \qquad (n\to\infty).$$

Thus, it follows that the limits $\lim_{x\to\infty} C(x)$ and limit $\lim_{n\to\infty}\mathbf{E}\, C_n$ exist if and only if $\tilde{C}(x)$ is constant. In fact, $\tilde{C}(x)$ equals $\frac{c}{2}$ up to at least 4 decimals and there are strong indications that $\tilde{C}(x)$ is not constant.

Interestingly there is a similar theorem for the variance of the height H_n. Set

$$\begin{aligned} V(x) \;&:=\; \sum_{k\ge 0}(2k+1)\left(1-\Psi\left(\frac{x}{y_k(1)}\right)\right) \\ &-\; \left(\sum_{k\ge 0}\left(1-\Psi\left(\frac{x}{y_k(1)}\right)\right)\right)^2 \end{aligned} \tag{26}$$

This function has similar properties as $C(x)$. $V(x)$ is a bounded function for $x>0$ and it is almost periodic in the above sense:

$$V(e^{1/c}x) = V(x) + o(1) \qquad (x\to\infty). \tag{27}$$

Theorem 3.2.2. **([10])** *The variance* $\mathbf{Var}\, H_n$ *remains bounded for* $n \to \infty$*. It is asymptotically given by*

$$\mathbf{Var}\, H_n = V(n) + o(1) \qquad (n \to \infty) \tag{28}$$

and it is asymptotically periodic in the sense that

$$\mathbf{Var}\, H_{\lfloor e^{1/c} n \rfloor} = \mathbf{Var}\, H_n + o(1) \qquad (n \to \infty). \tag{29}$$

Furthermore, the sequence $\mathbf{Var}\, H_n$ *is almost constant. There exists* n_1 *such that*

$$\max_{n \geq n_1} |\mathbf{Var}\, H_n - v_0| \leq 10^{-3}, \tag{30}$$

and we have

$$\lim_{n \to \infty} \sum_{k=n}^{\lfloor e^{1/c} n \rfloor} \frac{\mathbf{Var}\, H_k}{k} = \frac{v_0}{c}, \tag{31}$$

in which

$$v_0 = c \int_0^\infty (E(u) + E(ue^{-1/c}))\Psi(u)\, \frac{du}{u} = 2.085687\ldots,$$

and

$$E(u) = \sum_{k \geq 0} \left(1 - \Psi(ue^{-k/c})\right).$$

First, we note that there is an intimate relation between the sequence $\mathbf{E}\, C_n$ and the sequence $\mathbf{E}\, H_n$.

Lemma 3.2.1. *We have*

$$\mathbf{E}\, C_n = \frac{n+1}{2} \left(\mathbf{E} H_{n+1} - \mathbf{E} H_n\right). \tag{32}$$

Proof. Let $D_n = U_{H_n+1,n}$ denote that number of external nodes at level $H_n + 1$, i.e. there are no further (external or internal) nodes at higher level. Then $D_n = 2C_n$.

We now use the property that a random binary search trees T_{n+1} with $n+1$ internal nodes is obtained from T_n by replacing (with equal probability $1/(n+1)$ one of the $n+1$ external noded of T_n by an internal one (with two adjacent external ones). Thus

$$\begin{aligned} \mathbf{E}(H_{n+1}|T_n) &= (H_n + 1)\frac{D_n}{n+1} + H_n \left(1 - \frac{D_n}{n+1}\right) \\ &= \frac{D_n}{n+1} + H_n \end{aligned}$$

and consequently

$$\mathbf{E}H_{n+1} = \frac{\mathbf{E}D_n}{n+1} + \mathbf{E}H_n.$$

This proves (32). □

Alternatively we have

$$\mathbf{E}\,C_n = \frac{n+1}{2}\sum_{k\geq 0}\left(a_{n,k} - a_{n+1,k}\right), \tag{33}$$

where $a_{n,k} := \mathbf{P}[H_n \leq k]$, and (after applying the recurrence for $a_{n,k}$ hidden in the relation $y'_{k+1}(x) = y_k(x)^2$)

$$\mathbf{E}\,C_n = \frac{1}{2} + \frac{1}{2}\sum_{k\geq 0}\sum_{m=0}^{n-1} a_{m,k}(a_{n-m-1,k} - a_{n-m,k}). \tag{34}$$

Now observe that due to $a_{0,k} = 1$ and $a_{n+1,k} \leq a_{n,k}$ we have for every $L \leq n$

$$\begin{aligned}\sum_{m=0}^{n-1} a_{m,k}(a_{n-m-1,k} - a_{n-m,k}) &\leq \sum_{m=0}^{L-1}(a_{n-m-1,k} - a_{n-m,k}) \\ &+ a_{L,k}\sum_{m=L}^{n-1}(a_{n-m-1,k} - a_{n-m,k}) \\ &= (a_{n-L} - a_{n,k}) + a_{L,k}(1 - a_{n-L,k}).\end{aligned}$$

In particular we will work with $L = \lfloor \frac{n}{2} \rfloor$ and obtain the upper bound

$$\begin{aligned}\mathbf{E}\,C_n &\leq \frac{1}{2} + \frac{1}{2}\sum_{k\geq 0}(a_{\lceil n/2\rceil,k} - a_{n,k}) + \frac{1}{2}\sum_{k\geq 0} a_{\lfloor n/2\rfloor,k}(1 - a_{\lceil n/2\rceil,k}) \\ &= 1 + S_1 + S_2.\end{aligned}$$

Set $h_0(n) := \max\{k \geq 0 : y_k(1) \leq n\}$. First, by using the tail estimates from Lemma 3.1.1 we have

$$\begin{aligned}a_{\lceil n/2\rceil,k} - a_{n,k} &\leq a_{\lceil n/2\rceil,k} \\ &\leq Ce^{-(h_0(\lceil n/2\rceil)-k)/c}\end{aligned}$$

for $k \leq h_0(\lceil n/2\rceil)$ and

$$\begin{aligned}a_{\lceil n/2\rceil,k} - a_{n,k} &\leq 1 - a_{n,k} \\ &\leq Ce^{-(k-h_0(n))/c}\end{aligned}$$

for $k \geq h_0(n)$. Thus,

$$\left(\sum_{k \leq \lceil n/2 \rceil} + \sum_{k \geq h_0(n)} \right) (a_{\lceil n/2 \rceil, k} - a_{n,k}) = \mathcal{O}(1).$$

Since $y_{k+1}(1)/y_k(1) \geq e^{1/c}$ and $e^{3/c} > 2$ it directly follows that

$$\max\{k : y_k(1) \leq n\} - \max\{k : y_k(1) \leq \lceil n/2 \rceil\} \leq 3.$$

Hence, there are at most 2 terms (of magnitude ≤ 1) missing and consequently $S_1 = \mathcal{O}(1)$.

In order to estimates the second sum S_2 we proceed in a similar way. For $k \leq h_0(\lfloor n/2 \rfloor)$ we have

$$\begin{aligned} a_{\lfloor n/2 \rfloor, k}(1 - a_{\lceil n/2 \rceil, k}) &\leq a_{\lfloor n/2 \rfloor, k} \\ &\leq C e^{-(h_0(\lfloor n/2 \rfloor) - k)/c}. \end{aligned}$$

Consequently

$$\sum_{k \leq h_0(\lfloor n/2 \rfloor)} a_{\lfloor n/2 \rfloor, k}(1 - a_{\lceil n/2 \rceil, k}) = \mathcal{O}(1).$$

Similarly for $k \geq h_0(\lceil n/2 \rceil)$ we get

$$\begin{aligned} a_{\lfloor n/2 \rfloor, k}(1 - a_{\lceil n/2 \rceil, k}) &\leq 1 - a_{\lceil n/2 \rceil, k} \\ &\leq C e^{-(k - h_0(\lceil n/2 \rceil))/c} \end{aligned}$$

and

$$\sum_{k \geq h_0(\lceil n/2 \rceil)} a_{\lfloor n/2 \rfloor, k}(1 - a_{\lceil n/2 \rceil, k}) = \mathcal{O}(1).$$

Since $h_0(\lceil n/2 \rceil) - h_0(\lfloor n/2 \rfloor) \leq 1$ there is at most one term (of magnitude ≤ 1) missing and we finally have proved that $S_2 = \mathcal{O}(1)$, too.

In order to obtain the more precise relation (21) we have to use Theorem 2.1.1 and (34) (and quite involed calculations, see [10]). We just want to note that in view of (32) and (33) the representaion (21) is not unexpected:

$$\begin{aligned} \mathbf{E}\, C_n &\approx \frac{n+1}{2} \sum_{k \geq 0} \left(\Psi\left(\frac{n}{y_k(1)} \right) - \Psi\left(\frac{n+1}{y_k(1)} \right) \right) \\ &\approx -\frac{1}{2} \sum_{k \geq 0} \frac{n}{y_k(1)} \Psi'\left(\frac{n}{y_k(1)} \right) \\ &= C(n). \end{aligned}$$

Whereas the second approximation step is easy to verify, the first one cannot be directly checked. Therefore one has to use (34) in order to prove the above approximation $\mathbf{E}\, C_n = C(n) + o(1)$ rigorously.

References

[1] Chauvin, B., Drmota, J., and Jabbour-Hattab, J. (2001), The profile of binary search trees. Ann. Applied Probab., **11**, 1042–1062.

[2] Chauvin, B. and Drmota, M. (2004), The random bisection problem, travelling waves, and the distribution of the height of binary search trees (manuscript).

[3] Chauvin, B., Klein, T., Marckert, J.-F., and Rouault, A. (2004), Martingales and profile of binary search trees. Electr. J. Prob., submitted.

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

[5] Devroye, L. (1987), Branching processes in the analysis of the height of trees. Acta Inform., **24**, 277–298.

[6] Devroye, L. and Reed, B. (1995), On the variance of the height of random binary search trees. SIAM J. Comput., **24**, 1157–1162.

[7] Drmota, M. (2001), An analytic approach to the height of binary search trees. Algorithmica, **29**, 89-119.

[8] Drmota, M. (2002), The variance of the height of binary search trees. Theoret. Comput. Sci., **270**, 913–919.

[9] Drmota, M. (2003), An analytic approach to the height of binary search trees II. J. Assoc. Comput. Mach., **50**, 333–374.

[10] Drmota, M. (2004), On Robson's convergence and boundedness conjecture concerning the height of binary search trees. Theoret. Comput. Sci., to appear.

[11] Fill, J. A. and Janson, S. (2000), Smoothness and decay properties of the limiting Quicksort density function. in Mathematics and Computer Science: Algorithms, Trees, Combinatorics and Probabilities, (eds.: D. Gardy and A. Mokkadem), Trends in Mathematics, Berlin: Birkhäuser, 53–64.

[12] Fill, J. A. and Janson, S. (2001), Approximating the limiting Quicksort distribution. Random Structures Algorithms, **19**, 376–406.

[13] Fill, J. A. and Janson, S. (2002), Quicksort asymptotics. J. Algorithms, **44**, 4–28.

[14] Hoare, C. A. R. (1961), Algorithm 64: Quicksort. Communic. Assoc. Comput. Masch., **4**, 321.

[15] Hoare, C. A. R. (1962), Quicksort. Computer Journal, **5**, 10–15.

[16] Jabbour-Hattab, J. (2001), Martingales and large deviations for binary search trees. Random Struct. Algorithms, **19**, 112–127.

[17] Knessl, C. and Szpankowski, W. (1999), Quicksort algorithms again revisited. Discete Math. Theor. Comput. Sci., **3**, 43–64.

[18] Knuth, D. E. (1998), The Art of Computer Programming. Vol. 3: Sorting and Searching, 2nd ed. Reading, Massachussetts: Addison-Wesley.

[19] Lynch, W. (1965), More combinatorial problems on certain trees. The Computer J. **7**, 299–302.

[20] Mahmoud, H. M. (1992), Evolution of Random Search Trees. New York: John Wiley & Sons.

[21] Moser, L. and Wyman, M. (1958), Asymptotic development of the Stirling numbers of the first kind. J. London Math. Soc., **33**, 133–146.

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

[23] Reed, B. (2003), The height of a random binary search tree. J. Assoc. Comput. Mach., **50**, 306–332.

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

[25] Robson, J. M. (1997), On the concentration of the height of binary search trees. ICALP 97 Proceedings, LNCS, **1256**, 441–448.

[26] Robson, J. M. (2002), Constant bounds on the moments of the height of binary search trees. Theoret. Comput. Sci., **276**, 435–444.

[27] Rösler, U. (1991), A limit theorem for "Quicksort". Informatique théorique et Applications/Theoretical Informatics and Applications, **25**, 85–100.

[28] Tan, K. H. and Hadjicostas, P. (1995), Some properties of a limiting distribution of Quicksort. Statistics Probab. Letters, **25**, 87–94.