AN ANALYTIC APPROACH TO THE HEIGHT OF BINARY SEARCH TREES II

MICHAEL DRMOTA

Abstract. It is shown that all centralized absolute moments $\mathbf{E}|H_n - \mathbf{E}H_n|^\alpha$ ($\alpha \ge 0$) of the height H_n of binary search trees of size n and of the saturation level H'_n are bounded. The methods used rely on the analysis of a *retarded* differential equation of the form $\Phi'(u) = -\alpha^{-2}\Phi(u/\alpha)^2$ with $\alpha > 1$. The method can also be extended to prove the same result for the height of m-ary search trees. Finally the limiting behaviour of the distribution of the height of binary search trees is precisely determined.

1. Introduction

A binary search tree is a binary tree, in which each internal node[1] is associated with a key, where the keys are drawn from some totally ordered set, say $\{1, 2, \ldots, n\}$. The first key is associated with the root. Now, the next key is placed in the left child of the root if it is smaller than the key of the root and it is sent to the right child of the root if it is larger than the key of the root. In this way we proceed further by inserting key by key. More precisely, the algorithms repeats itself recursively in the subtrees (and thus all subtrees are in fact binary search trees). So starting from a permutation of $\{1, 2, \ldots, n\}$ we get a binary tree with n internal nodes such that the keys of the left subtree of any given node x are smaller than the key of x and the keys of the right subtree are larger than the key of x.

Binary search trees are widely used to store (totally ordered) data, and many parameters have been discussed in the literature. (The monograph of Mahmoud [10] gives a very good overview of the state of the art.) Usually it is assumed that every permutation of $\{1, 2, \ldots, n\}$ is equally likely and hence any parameter of binary search trees may be considered as a random variable.

An alternative way of looking at this probabilistic model is a "Markov chain" of binary trees $(T_n)_{n\ge0}$ describing the evolution of a binary search tree by inserting one key after another. T_0 has no internal nodes. It consists of exactly one external node which is the root. T_1 has one internal node which is the root and two external nodes. Now T_2 is generated from T_1 by replacing one of the two external nodes by an additional internal one (with two external nodes as left and right children) with equal probability 1/2. In that way we proceed further. T_{n+1} is generated from T_n by replacing one of the $n+1$ external nodes by an additional internal one (and two external nodes as left and right children) with equal probability $1/(n+1)$. It is an easy exercise to show that for any fixed n the probability distribution of T_n of this Markov chain $(T_n)_{n\ge0}$ is exactly the same as the probability distribution induced by equally likely permutations of $\{1, 2, \ldots, n\}$ as above.

Date: April 29, 2002.

Key words and phrases. binary search tree, average case analysis, height, saturation level, generating functions.

Institut für Geometrie, TU Wien, Wiedner Hauptstrasse 8-10/118, A-1040 Wien, Austria, email: drmota@tuwien.ac.at.

This research was supported by the Austrian Science Foundation FWF, grant P10187-MAT.

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

In this paper we first consider the height and the saturation level. The height H_n is the largest distance of an internal node from the root (or the highest level containing internal nodes), and the saturation level H_n' is the maximal level without external nodes, i.e., the binary trees is complete[2] until level H_n'. (We will also consider the height of m-ary search trees and describe the distribution of the height of binary search trees.)

In 1986 Devroye [2] 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 $\left(\frac{2e}{c}\right)^c = e$. (Earlier Pittel [13] had shown that $H_n/\log n \to \gamma$ almost surely as $n \to \infty$, where $\gamma \le c$, compare also with Robson [16]. Later Devroye [3] 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 $\mathbf{V}H_n$ is bounded. In fact, he could prove (see [17]) 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 of the number of nodes at level $h = H_n$ is bounded (see [18]). The best bounds (before 1999) were given by two completely different methods by Devroye and Reed [5] and later by Drmota [7]. They (both) proved

$$\mathbf{E}H_n = c\log n + \mathcal{O}(\log\log n) \tag{1}$$

and

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

Eventually, Reed [14, 15] settled Robson's conjecture

$$\mathbf{V}H_n = \mathcal{O}(1).$$

His approach is related to that of [5], moreover he could also show that

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

A second proof of Robson's conjecture was given by the author [8]. (In the present paper we present a much more detailed proof leading to more precise estimates for higher moments and we will provide several extensions.)

The saturation level H_n' of a binary search tree is defined to be the maximal level h' such that for all levels h up to h' there are no external nodes, i.e. the binary search tree has 2^h (internal) nodes for all levels $h \le h'$ but less than $2^{h'+1}$ (internal) nodes at level $h'+1$. For example, it follows from Biggins [1] that

$$\frac{H_n'}{c'\log n} \to 1 \quad a.s.,$$

where $c' = 0.373365\ldots$ is the other real solution of the equation $\left(\frac{2e}{c'}\right)^{c'} = e$. (Compare also with Mahmoud [10].)

This result indicates some *duality* between the height H_n and the saturation level H_n' of binary search trees. The analysis given below supports this observation (in the combinatorial and in the analytic part).

The concept of m-ary search trees is a direct generalization of that of binary search trees ($m = 2$), see e.g. [11]. Here every internal node stores at least one and at most $m-1$ keys. If an internal node contains exactly $m-1$ keys $x_1 < x_2 < \cdots < x_{m-1}$ then it may have (at most) m subtrees, also containing internal nodes, such that all keys in the j-th subtrees are greater than x_{j-1} and less than x_j. Of course, all these subtrees are again m-ary search trees.

[2] A binary tree is complete if every level $h \ge 0$ contains exactly 2^h nodes.

Devroye [4] has proved the following asymptotic property concerning the height $H_{m,n}$ of m-ary search trees. Let $c = c_m > 1/(h_m - 1)$ ($h_m = \sum_{i=1}^m 1/i$) be the minimal solution of the equations

$$\beta + c\log(m!) = c\sum_{i=1}^{m-1} \log(\beta + i),$$

where β is linked to c by

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

Then we have

$$\frac{H_{m,n}}{c_m \log n} \to 1 \quad a.s.$$

2. Results

The purpose of this paper is to present a second proof of Robson's conjecture. i.e., $\mathbf{V}\, H_n = \mathcal{O}(1)$, which is independent from the proof of Reed [14, 15]. A variation of this method provides a similar result for the saturation level. In both cases a *retarded* differential equation of the form $\Phi'(u) = -\alpha^{-2}\Phi(u/\alpha)^2$ (with some $\alpha > 1$) plays a central rôle. A similar differential equation (see (57)) is related to the height of m-ary search trees. However, it turns out that this equation is much more difficult to handle for $m > 2$. For $m = 2$ and $\alpha = e^{1/c} = 1.26107\ldots$ we could use a relatively simple trick to get a proper solution, compare with [7] and [8]. For $\alpha = 16$ there exists a non-trivial solution $\Phi(u) = (u^{-1} + u^{-\frac{3}{4}})e^{-u^{-3/4}}$ (see Lemma 11) which is suitable to prove the case of the saturation level in binary search trees. For general m-ary search trees we will use a related integral equation (58) for the inverse Laplace transform $\Psi(y)$ of $\Phi(u)$ which can be solved by a contraction argument after a proper scaling. It is very likely that the solution $\Psi(y)$ of the integral equation is closely related to the distribution of the height of m-ary search trees. At least for the binary case this can be worked out, see Theorem 4.

Theorems 1 and 2 concern the height and the saturation level of binary search trees.

Theorem 1. *Let $y_h(x)$ ($h \ge 0$) be (polynomials) recursively defined by $y_0(x) \equiv 1$ and by*

$$y_{h+1}(x) := 1 + \int_0^x y_h(t)^2\, dt \qquad (h \ge 0). \tag{3}$$

Then the expected value of the height H_n of binary search trees of size n is given by

$$\mathbf{E}H_n = \max\{h : y_h(1) \le n\} + \mathcal{O}(1) \qquad (n \to \infty), \tag{4}$$

and there exists a constant $C > 0$ such that for all integers $k > 0$ and $n \ge 0$

$$\mathbf{E}|H_n - \mathbf{E}H_n|^k \le C\,(k+1)!\left(1 - \frac{2}{c}\right)^{-k}. \tag{5}$$

Theorem 2. *Let $z_h(x)$ ($h \ge 0$) be recursively defined by $z_0(x) \equiv 1$ and by*

$$z_{h+1}(x) := 1 + \int_0^x z_h(t)\left(\frac{2}{1-t} - z_h(t)\right) dt \qquad (h \ge 0).$$

Then the expected value $\mathbf{E}H_n'$ of the saturation level of binary search trees is given by

$$\mathbf{E}H_n' = \max\{h : z_h(1 - n^{-1}) \le (1 - 2/e)n\} + \mathcal{O}(1) \qquad (n \to \infty), \tag{6}$$

and all centralized moments of H_n' are bounded:

$$\mathbf{E}|H_n' - \mathbf{E}H_n'|^k = \mathcal{O}(1) \qquad (n \to \infty). \tag{7}$$

The constant $1 - 2/e = 0.264\ldots$ appearing in (6) is not essential. It can be replaced by any constant between 0 and 1 and the theorem remains true.

The following result for m-ary search trees is, of course, a generalization of Theorem 1. However, the estimates for the centralized moments are more explicit in Theorem 1. Therefore, we decided to state Theorem 3 separately.

Theorem 3. *Let $m \ge 2$ be a fixed integer and let $y_h(x)$ $(h \ge 0)$ be (polynomials) recursively defined by $y_0(x) \equiv 1$ and by*

$$y_{h+1}^{(m-1)}(x) = (m-1)!y_h(x)^m \qquad (h \ge 0),$$

where

$$y_{h+1}(0) = y_{h+1}'(0) = \cdots = y_{h+1}^{(m-2)}(0) = 1.$$

Then the expected value of the height $H_{m,n}$ of m-ary search trees of size n is given by

$$\mathbf{E}H_{m,n} = \max\{h : y_h(1) \le n\} + \mathcal{O}(1) \qquad (n \to \infty), \tag{8}$$

and all centralized moments of H_n are bounded:

$$\mathbf{E}|H_{m,n} - \mathbf{E}H_{m,n}|^k = \mathcal{O}(1) \qquad (n \to \infty). \tag{9}$$

Remark 1. Note that we also prove a relation for the expected value $\mathbf{E}H_n$ and $\mathbf{E}H_n'$ which are, however, implicit and *do not* reprove the limiting relations $\mathbf{E}H_n \sim c \log n$ and $\mathbf{E}H_n' \sim c' \log n$. In view of the following proof it seems to be *easier* just to prove the boundedness of $\mathbf{V}H_n$ and $\mathbf{V}H_n'$ (and of all centralized moments.)[3] However, we can combine this theorem with Reed's result (2) to obtain tight asymptotics for

$$y_h(1) = e^{h/c + \frac{3}{2(c-1)}\log h + \mathcal{O}(1)}. \tag{10}$$

Interestingly the asymptotic behaviour of $y_h(1)$ was posed as an unsolved problem by C. Ponder [12] without stating any connection to binary search trees.

We finally present a theorem on the limiting behaviour of the distribution of H_n. As for the moments the result is in some sense implicit since it is only precise in terms of $y_h(1)$.

Theorem 4. *Let $y_h(x)$ be defined by (3). Then there exists a monotonically decreasing function $\Psi(y)$, $y \ge 0$, with $\Psi(0) = 1$ and $\lim_{y\to\infty} \Psi(y) = 0$ satisfying the integral equation*

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

such that

$$\mathbf{P}[H_n \le h] = \Psi(n/y_h(1)) + o(1) \quad (n \to \infty),$$

where the $o(1)$-error term is uniform for all $h \ge 0$.

The paper is organized in the following way. Section 3 is devoted to the proof of Theorem 1, section 4 to the proof of Theorem 2 and section 5 to the proof of Theorem 3. All three proofs have a similar structure. However, they get more and more involved. In a final section 6 we prove Theorem 4.

[3] It should be mentioned that in [7] there is an outline of a possible proof of Robson's conjecture which relies more or less on the conjecture that $\mathbf{E}H_n = c\log n - \frac{c}{2(c-1)}\log\log n + \mathcal{O}(1)$ which turned out to be wrong due to Reed's result [14, 15]. In fact, the purpose of [7] was to find tight asymptotics for $\mathbf{E}H_n$. However, it seems that it is more or less unmanageable to get more than (1) by the methods of [7]. After all one gets the impression that it is *an intrinsic property* of the height of binary search trees that *any expansion* of the form $\mathbf{E}H_n = h(n) + \mathcal{O}(1)$ with some *manageable* function $h(n)$ leads to a proof of Robson's conjecture. This had been made explicit by Reed [14, 15] (compare with (2)) and some months later by the author with the implicit expansion (4) presented here.

3. Height

3.1. **Combinatorial Background.** Let H_n be the height of binary search trees of size n. So, its distribution function is

$$\mathbf{P}[H_n \le h] = \frac{a_{n,h}}{n!},$$

where $a_{n,h}$ denotes the number of permutations $\sigma \in S_n$ of n elements such that the corresponding binary search tree has height $\le h$.

The following two lemmata collect some properties of the numbers $a_{n,h}$ and their (exponential) generating functions

$$y_h(x) := \sum_{n \ge 0} \frac{a_{n,h}}{n!} x^n. \tag{11}$$

Since $a_{n,h} = 0$ for $n \ge 2^h$ these generating functions are in fact polynomials.

The corresponding proofs of Lemmata 1 and 2 can be found in [7].

Lemma 1. *The numbers $a_{n,h}$ satisfy the recurrence equation*

$$a_{n,h+1} = \sum_{k=0}^{n-1} \binom{n-1}{k} a_{k,h} a_{n-1-k,h} \tag{12}$$

with initial conditions $a_{0,0} = 1$ and $a_{n,0} = 0$ for $n > 0$. Furthermore we have

$$\frac{a_{n+1,h}}{(n+1)!} \le \frac{a_{n,h}}{n!} \tag{13}$$

for all $h \ge 0$ and $n \ge 0$.

Lemma 2. *The functions $y_h(x)$ are recursively given by*

$$y_0(x) \equiv 1$$

and by

$$y'_{h+1}(x) = y_h(x)^2, \quad y_h(0) = 1. \tag{14}$$

Remark 2. Note that the relation (14) is just a reformulation of (12). Thus, Lemma 2 follows from Lemma 1. Furthermore (13) may be interpreted as $\mathbf{P}[H_{n+1} \le h] \le \mathbf{P}[H_n \le h]$ which is quite obvious if one thinks of the Markov chain approach mentioned in the Introduction.

Remark 3. It is clear that (14) can be reformulated to

$$y_{h+1}(x) = 1 + \int_0^x y_h(t)^2 dt,$$

i.e. these functions are exactly the same as those which are introduced in Theorem 1.

3.2. **Analytic Tools.** The most important tool for the proof of Theorem 1 is the set of auxiliary functions $\tilde{y}_h(x)$ $(h \ge 0)$ defined by

$$\tilde{y}_h(x) := \alpha^h \Phi(\alpha^h(1-x)), \tag{15}$$

where $\Phi(u)$ is the solution of the (retarded) differential equation

$$\Phi'(u) = -\frac{1}{\alpha^2}\Phi(u/\alpha)^2 \tag{16}$$

with the initial condition $\Phi(0) = 1$, and

$$\alpha = e^{1/c} = 1.26107\ldots.$$

These functions emulate the recurrence equation (14) of $y_h(x)$. In fact, by using (16) one easily gets $\tilde{y}'_{h+1}(x) = \tilde{y}_h(x)^2$.

This approach was already used in [7]. In Lemma 3 we collect some properties of the solution of (16). It is then an easy exercise to translate these properties to corresponding properties of $\tilde{y}_h(x)$, see Lemma 4. (The proofs can be found in [7]).

Lemma 3. *There exists a unique entire function $\Phi(u)$ with $\Phi(0) = 1$ which solves (16). This solution $\Phi(u)$ is strictly decreasing for (real) $u \ge 0$ and the function $u\Phi(u)$ is strictly increasing for (real) $u \ge 0$. Especially we have $0 < \Phi(u) < 1/u$ for (real) $u \ge 0$ and there exists a real constant $C_1 > 0$ such that*

$$1 - u\Phi(u) \sim C_1 \frac{\log u}{u^{c-1}} \tag{17}$$

as $u \to \infty$.

Remark 4. It will be clear from the following analysis that any (proper) solution $\Phi(u)$ of (16) with α satisfying $1 < \alpha \le e^{1/c}$ is sufficient to prove boundedness of all centralized moments of H_n. (However, for $\alpha = e^{1/c}$ we get sharper bounds than for $\alpha < e^{1/c}$, compare also with the proof of Theorem 2.) In [7] a more or less direct proof of Lemma 3 has been given (which can be modified to prove corresponding properties for $\alpha < e^{1/c}$). In section 5 we provide a different approach. We consider an integral equation (82) for a function $\Psi(y)$ so that the Laplace transform $\Phi(u)$ of $\Psi(y)$ satisfies (16) and has all necessary properties to prove Theorem 1 (and similarly Theorem 3).

Lemma 4. *The functions $\tilde{y}_h(x)$, $h \ge 0$, $x \ge 0$, defined by (15) satisfy*

1. $0 < \tilde{y}_h(0) < 1$.
2. $1 - \tilde{y}_h(0) \sim \frac{C_1}{c} h \left(\frac{2}{c}\right)^h$ *$(h \to \infty)$.*
3. $\tilde{y}_h(1) = \alpha^h$.
4. $\tilde{y}_{h+r}(x) \ge \tilde{y}_h(x)$ *for all $x \ge 0$ and $r \ge 0$.*
5. $\tilde{y}'_{h+1}(x) = \tilde{y}_h(x)^2$.

The only (but important) difference between $y_h(x)$ and $\tilde{y}_h(x)$ is that they do not have the same initial condition. We have $\tilde{y}_h(0) = \alpha^h \Phi(\alpha^h) < 1 = y_h(0)$. Furthermore $\tilde{y}_0(x) < y_0(x)$ for $0 \le x < 1$. Thus it follows by induction that $\tilde{y}_h(x) < y_h(x)$ for $0 \le x \le 1$. Next observe that the power series expansion of $\tilde{y}_h(x)$ is given by

$$\tilde{y}_h(x) = \sum_{n \ge 0} \frac{\alpha^{(n+1)h}}{n!} (-1)^n \Phi^{(n)}(\alpha^h)\, x^n.$$

Set $\tilde{\Phi}(u) = \Phi(-u)$. Then $\tilde{\Phi}(u)$ satisfies the differential equation $\tilde{\Phi}'(u) = \alpha^{-2}\tilde{\Phi}(u/\alpha)^2$ Since $\tilde{\Phi}(u) = \Phi(-u) > 0$ for all real u it follows by induction that $\tilde{\Phi}^{(n)}(u) > 0$ for all integers $n \ge 0$ and for all real u. Thus,

$$(-1)^n \Phi^{(n)}(\alpha^h) = \tilde{\Phi}^{(n)}(\alpha^h) > 0$$

for all n, which implies that all Taylor coefficients of $\tilde{y}_h(x)$ are positive. Hence, $\tilde{y}_h(x)$ grows faster than any polynomial and consequently there always exists $x_h > 1$ such that $\tilde{y}_h(x) \ge y_h(x)$ for $x \ge x_h$. In fact, we can be much more precise.

Lemma 5. *For every non-negative integer h and for every (real) $D \ge 0$ there exists $x_{h,D} > 0$ such that*

$$\tilde{y}_{h+D}(x) < y_h(x) \qquad (0 \le x < x_{h,D}), \tag{18}$$

and

$$y_h(x) < \tilde{y}_{h+D}(x) \qquad (x > x_{h,D}). \tag{19}$$

Furthermore we have

$$x_{h+1,D} > x_{h,D}. \tag{20}$$

Proof. We proceed by induction. Since $\tilde{y}_D(x)$ is strictly increasing and satisfies $0 < \tilde{y}_D(0) < 1$ and $\lim_{x\to\infty} \tilde{y}_D(x) = \infty$ the assertion is surely true for $h = 0$. Now suppose that (18) and (19) are satisfied for some $h \ge 0$, i.e. the difference

$$\delta_{h,D}(x) := y_h(x) - \tilde{y}_{h+D}(x)$$

has a unique zero $x_{h,D} > 0$ such that $\delta_{h,D}(x) > 0$ for $0 \le x < x_{h,D}$ and $\delta_{h,D}(x) < 0$ for $x > x_{h,D}$. Now we have

$$\begin{aligned} \delta'_{h+1,D}(x) &= y'_{h+1}(x) - \tilde{y}'_{h+1+D}(x) \\ &= y_h(x)^2 - \tilde{y}_{h+D}(x)^2 \\ &= \delta_{h,D}(x)(y_h(x) + \tilde{y}_{h+D}(x)). \end{aligned}$$

Hence, $\delta_{h+1,D}(x)$ is increasing for $0 \le x < x_{h,D}$ and decreasing for $x > x_{h,D}$. Since $\delta_{h+1,D}(0) > 0$ and $\lim_{x\to\infty} \delta_{h+1,D}(x) = -\infty$ there exists a unique zero $x_{h+1,D} > x_{h,D}$ of $\delta_{h+1,D}(x)$ such that $\delta_{h+1,D}(x) > 0$ for $0 \le x < x_{h+1,D}$ and $\delta_{h+1,D}(x) < 0$ for $x > x_{h+1,D}$. □

By using Lemma 5 we can prove the following property which is indeed the key to the proof of Theorem 1.

Lemma 6. *Let e_h be defined by $e_h := c \cdot \log y_h(1)$. Then we have $e_{h+1} \ge e_h + 1$. Moreover,*

$$\tilde{y}_{e_h}(x) \le y_h(x) \qquad (0 \le x \le 1), \tag{21}$$

and

$$y_h(x) \le \tilde{y}_{e_h}(x) \qquad (x \ge 1). \tag{22}$$

Proof. Since

$$\tilde{y}_{e_h}(1) = \alpha^{e_h} = y_h(1)$$

we have $x_{h,e_h-h} = 1$. Thus Lemma 5 directly implies (21) and (22).

Finally, by (20) we have

$$x_{h+1,e_h+1-(h+1)} = x_{h+1,e_h-h} > x_{h,e_h-h}$$

which implies $y_{h+1}(1) > \tilde{y}_{e_h+1}(1)$. Hence, 4. of Lemma 4 gives $e_{h+1} \ge e_h + 1$. □

Now we are able to prove tight upper and lower bounds for $a_{n,h}$.

Theorem 5. *There exists an absolute constant $C_2 > 0$ such that*

$$\frac{a_{n,h}}{n!} \le C_2 \alpha^{-4(c\log n - e_h)}, \tag{23}$$

for all non-negative integers h with $e_h \le c \log n$, and an absolute constant $C_3 > 0$ such that

$$1 - \frac{a_{n,h}}{n!} \le C_3(e_h - c\log n + 1)\left(\frac{2}{c}\right)^{e_h - c\log n}, \tag{24}$$

for all non-negative integers h with $e_h \ge c \log n$.

Proof. Suppose that $x > 1$. Then by Lemma 6 and (13) we have

$$\begin{aligned} \tilde{y}_{e_h}(x) &\ge y_h(x) \\ &\ge \sum_{k=0}^{n} \frac{a_{k,h}}{k!} x^k \\ &\ge \frac{a_{n,h}}{n!} \sum_{k=0}^{n} x^k \\ &\ge \frac{a_{n,h}}{n!} \frac{x^{n+1} - 1}{x - 1}, \end{aligned}$$

and consequently

$$\frac{a_{n,h}}{n!} \le \frac{x-1}{x^{n+1}-1}\tilde{y}_{e_h}(x) = \frac{x-1}{x^{n+1}-1}\alpha^{e_h}\Phi\left(-(x-1)\alpha^{e_h}\right).$$

Finally setting $x = 1+\alpha^{-e_h}$ and using the inequality

$$(1+y)^{n+1}-1 \ge \binom{n+1}{4}y^4 \gg (ny)^4,$$

we directly get

$$\frac{a_{n,h}}{n!} \le \frac{1}{(1+\alpha^{-e_h})^{n+1}-1}\Phi(-1) \ll \frac{1}{(n\alpha^{-e_h})^4}$$

if $e_h \le c\log n$. This proves (23). (We note that Vinogradov's notation $a \ll b$ means that there exists an absolute constant C with $a \le Cb$.)

The proof of (24) runs along the same lines. Here we use $x = 1-\frac{1}{n} < 1$. Again by Lemma 6 and (13) we have

$$\begin{aligned}\frac{1}{1-x} - \tilde{y}_{e_h}(x) &\ge \frac{1}{1-x} - y_h(x)\\ &= \sum_{k\ge 0}\left(1-\frac{a_{k,h}}{k!}\right)x^k\\ &\ge \sum_{k=n}^{\infty}\left(1-\frac{a_{k,h}}{k!}\right)x^k\\ &\ge \left(1-\frac{a_{n,h}}{n!}\right)\sum_{k=n}^{\infty}x^k\\ &= \left(1-\frac{a_{n,h}}{n!}\right)\frac{x^n}{1-x}.\end{aligned}$$

Thus

$$1-\frac{a_{n,h}}{n!} \le x^{-n}\left(1-(1-x)\tilde{y}_{e_h}(x)\right).$$

Finally, by using 2. of Lemma 4 we directly get

$$\begin{aligned}1-\frac{a_{n,h}}{n!} &\ll 1-\alpha^{e_h-c\log n}\Phi(\alpha^{e_h-c\log n})\\ &\ll (e_h-c\log n+1)\left(\frac{2}{c}\right)^{e_h-c\log n}\end{aligned}$$

if $e_h \ge c\log n$. □

3.3. **Proof of Theorem 1.** In this section we show that the estimates for the tails of the distribution of H_n (provided in Theorem 5) are sufficient to prove Theorem 1.

First of all we note the following.

Lemma 7. *We have*

$$\sum_{h:e_h\le c\log n}\frac{a_{n,h}}{n!} = \mathcal{O}(1) \qquad (n\to\infty) \tag{25}$$

and

$$\sum_{h:e_h\ge c\log n}\left(1-\frac{a_{n,h}}{n!}\right) = \mathcal{O}(1) \qquad (n\to\infty). \tag{26}$$

Proof. By Theorem 5 we just have to show that

$$\sum_{h:e_h\le c\log n}\alpha^{-4(c\log n-e_h)} = \mathcal{O}(1) \qquad (n\to\infty) \tag{27}$$

and

$$\sum_{h:e_h\ge c\log n} (e_h - c\log n + 1)\left(\frac{2}{c}\right)^{e_h - c\log n} = \mathcal{O}(1) \qquad (n\to\infty). \tag{28}$$

However, this is almost trivial. By Lemma 6 we have $e_{h+1}-e_h \ge 1$. Hence, the sum in (27) is bounded above by

$$\sum_{j\ge 0}\alpha^{-4j} = \mathcal{O}(1)$$

and the sum in (28) by

$$\sum_{j\ge 0}(j+1)\left(\frac{2}{c}\right)^j = \mathcal{O}(1).$$

□

Now it is easy to prove the estimate (4) for the expected value $\mathbf{E}H_n$: We directly obtain

$$\begin{aligned}\mathbf{E}H_n &= \sum_{h\ge 0}\left(1-\frac{a_{n,h}}{n!}\right)\\ &= \sum_{h:e_h<c\log n}\left(1-\frac{a_{n,h}}{n!}\right) + \sum_{h:e_h\ge c\log n}\left(1-\frac{a_{n,h}}{n!}\right)\\ &= \max\{h : e_h \le c\log n\} + \mathcal{O}(1).\end{aligned}$$

The proof of the bounds for the centralized moments (5) is a little bit more technical. Let $F_n(x) = \mathbf{P}[H_n \le x] = a_{n,\lfloor x\rfloor}/\lfloor x\rfloor!$ (for $x\ge 0$) be the distribution function of H_n. Then by integration by parts we get

$$\begin{aligned}\mathbf{E}|H_n-\mathbf{E}H_n|^k &= \int_0^\infty |x-\mathbf{E}H_n|^k\,dF_n(x)\\ &= \int_0^{\mathbf{E}H_n}(\mathbf{E}H_n-x)^k\,dF_n(x) + \int_{\mathbf{E}H_n}^\infty (x-\mathbf{E}H_n)^k\,dF_n(x)\\ &= k\int_0^{\mathbf{E}H_n}(\mathbf{E}H_n-x)^{k-1}\,F_n(x)\,dx + k\int_{\mathbf{E}H_n}^\infty (x-\mathbf{E}H_n)^{k-1}\,(1-F_n(x))\,dx\\ &= S_1+S_2.\end{aligned}$$

Set $h_0 = \max\{h : e_h \le c\log n\}$. Then $|\mathbf{E}H_n - h_0| \le C_4$ for some constant $C_4>0$. Hence, the first integral S_1 can be estimated by

$$\begin{aligned}S_1 &\ll k\sum_{h\le h_0}(h_0-h+C_4+1)^{k-1}\frac{a_{n,h}}{n!} + k\int_0^{C_4}(x+C_4+1)^{k-1}\,dx\\ &\ll k\sum_{h\le h_0}(h_0-h+C_4+1)^{k-1}\alpha^{-4(c\log n-e_h)} + (2C_4+1)^k\\ &\ll k\sum_{j\ge 0} j^{k-1}\alpha^{-4j} + (2C_4+1)^k.\end{aligned}$$

Since

$$\begin{aligned}\sum_{j\ge 0} j^{k-1}u^j &\le (k-1)!\sum_{j\ge 0}\binom{k+j-1}{k-1}u^j\\ &= \frac{(k-1)!}{(1-u)^k}\end{aligned}$$

for $k\ge 1$ and $0\le u<1$, we can finally estimate S_1 by

$$S_1 \ll k!\left(1-\frac{1}{\alpha^4}\right)^{-k}.$$

Similarly we can treat the second integral S_2:

$$\begin{aligned}
S_2 &\le k \sum_{h\ge h_0} (h-h_0+C_4+1)^{k-1}\left(1-\frac{a_{n,h}}{n!}\right) + (2C_4+1)^k \\
&\ll k \sum_{h\ge h_0} (h-h_0+C_4+1)^{k-1}(e_h - c\log n + 1)\left(\frac{2}{c}\right)^{e_h - c\log n} + (2C_4+1)^k \\
&\ll k \sum_{j\ge 0} j^k \left(\frac{2}{c}\right)^j + (2C_4+1)^k \\
&\ll (k+1)!\left(1-\frac{2}{c}\right)^{-k}.
\end{aligned}$$

Since $\alpha^4 > c/2$ this completes the proof of Theorem 1.

4. Saturation Level

4.1. **Combinatorial Background.** Let $b_{n,h}$ denote the number of permutations $\sigma \in S_n$ of n elements such that the corresponding binary search tree has saturation level $> h$, i.e.

$$\mathbf{P}[H'_n > h] = \frac{b_{n,h}}{n!}.$$

Then these numbers $b_{n,h}$ satisfy the same recurrence as the numbers $a_{n,k}$.

Lemma 8. *The numbers $b_{n,h}$ satisfy the recurrence equation*

$$b_{n,h+1} = \sum_{k=0}^{n-1} \binom{n-1}{k} b_{k,h} b_{n-1-k,h} \tag{29}$$

with initial conditions $b_{0,0} = 0$ and $b_{n,0} = 1$ for $n > 0$. Furthermore we have

$$\frac{b_{n+1,h}}{(n+1)!} \ge \frac{b_{n,h}}{n!} \tag{30}$$

for all $h \ge 0$ and $n \ge 0$.

Proof. The proof of the recurrence relation (29) is immediate by considering a binary search tree where the root is labeled by $k+1$, $0 \le k \le n-1$.

Next observe that (30) is the same as $\mathbf{P}[H'_{n+1} > h] \ge \mathbf{P}[H'_n > h]$. However, this inequality is immediately clear by applying the "Markov chain approach" mentioned in the Introduction. □

Remark 5. It should be remarked that (30) can be proved by induction, too. Obviously, (30) is satisfied for $h = 0$. Now suppose that (30) is true for some $h \ge 0$. Then, by (29)

$$\begin{aligned}
b_{n+1,h+1} &= \sum_{k=0}^{n} \binom{n}{k} b_{k,h} b_{n-k,h} \\
&= \sum_{k=0}^{n-1} \binom{n}{k} b_{k,h} b_{n-k,h} + b_{n,h} \\
&\ge \sum_{k=0}^{n-1} \frac{n}{n-k}\binom{n-1}{k} b_{k,h}(n-k) b_{n-k-1,h} + b_{n,h+1} \\
&= n b_{n,h+1} + b_{n,h+1} = (n+1) b_{n,h+1}.
\end{aligned}$$

Hence, (30) is also satisfied for $h+1$.

We introduce the generating functions

$$\overline{y}_h(x) = \sum_{n\ge 0} \frac{b_{n,h}}{n!} x^n = \sum_{n\ge 0} \mathbf{P}[H'_n > h]\, x^n. \tag{31}$$

The recurrence equation (29) can be rewritten to (compare with Lemma 1 and Lemma 2)

$$\overline{y}'_{h+1}(x) = \overline{y}_h(x)^2$$

but with initial condition $\overline{y}_0(x) = x/(1-x)$ and initial values $\overline{y}_h(0) = 0$ for $h > 0$. These functions are no longer polynomials. Even, the *converse* functions

$$z_h(x) = \frac{1}{1-x} - \overline{y}_h(x) = \sum_{n\ge 0} \mathbf{P}[H'_n \le n]\, x^n$$

are no polynomials, either. They satisfy the following recurrence equation.

Lemma 9. *The functions $z_h(x)$ are recursively given by*

$$z_0(x) \equiv 1$$

and by

$$z'_{h+1}(x) = z_h(x) \left(\frac{2}{1-x} - z_h(x) \right) \qquad (z_h(0) = 1) \tag{32}$$

for $h > 0$.

Remark 6. Equivalently we have

$$z_{h+1}(x) = 1 + \int_0^x z_h(t) \left(\frac{2}{1-x} - z_h(x) \right) dt,$$

i.e. these functions are exactly the same as those which are introduced in Theorem 2.

For notational convenience, let $[x^n]A(x)$ denote the n-th coefficient of the power series expansion of $A(x)$ at $x_0 = 0$. We will also use the notation

$$A(x) \le_c B(x)$$

if $[x^n]A(x) \le [x^n]B(x)$ for all $n \ge 0$. Note that if all coefficients are non-negative then $A(x) \le_c B(x)$ implies $A(x) \le B(x)$ for all $x \ge 0$. Furthermore there are quite simple rules for "$\le_c$", e.g. if $0 \le_c A(x) \le_c B(x)$ and $0 \le_c C(x) \le_c D(x)$ then $0 \le_c A(x)C(x) \le_c B(x)D(x)$.

In what follows we will also make use of the following upper bound for $z_h(x)$ which is stated in terms of "$\le_c$".

Lemma 10. *For every $h \ge 0$ we have*

$$z_h(x) \le_c 2^h \sum_{k=0}^{h} \frac{\left(\log \frac{1}{1-x}\right)^k}{k!}. \tag{33}$$

Proof. Obviously, we have equality for $h = 0$. Now we use the inequality

$$z'_{h+1}(x) = z_h(x) \left(\frac{2}{1-x} - z_h(x) \right) \le_c \frac{2}{1-x} z_h(x) \tag{34}$$

and proceed by induction. Note that the inequality "$\le_c$" is preserved by integration since $z_h(0) = 1$ for all $h \ge 0$. □

Remark 7. Lemma 10 can also be used to obtain a lower bound for the expected value $\mathbf{E}H'_n$ of the form

$$\mathbf{E}H'_n \ge c' \log n + \frac{c'}{2(1-c')} \log\log n + \mathcal{O}(1).$$

We only have to adapt the corresponding proof of [7]. Here it is crucial that inequality (33) holds in the sense of "$\le_c$".

4.2. **Analytic Background.** We again use the (retarded) differential equation

$$\overline{\Phi}'(u) = -\frac{1}{\overline{\alpha}^2}\overline{\Phi}(u/\overline{\alpha})^2 \tag{35}$$

but now with

$$\overline{\alpha} \ge e^{1/c'} = 14.5625\ldots$$

Interestingly there is an explicit solution $\overline{\Phi}(u)$ of (35) for $u > 0$ with $\overline{\alpha} = 16$.

Lemma 11. *The function*

$$\overline{\Phi}(u) = \frac{1+u^{1/4}}{u}e^{-u^{1/4}} \tag{36}$$

is solution of the differential equation (35) with $\overline{\alpha} = 16$ for $u > 0$ and satisfies $\lim_{u\to 0+} u\overline{\Phi}(u) = 1$.

As above, we define a set of auxiliary functions $\tilde{z}_h(x)$ $(h \ge 0)$ by

$$\tilde{z}_h(x) := \frac{1}{1-x} - 16^h\overline{\Phi}(16^h(1-x)), \tag{37}$$

where $\overline{\Phi}(u)$ is given by (36). They emulate the recurrence equation (32) of $z_h(x)$.

Lemma 12. *The functions $\tilde{z}_h(x)$, $h \ge 0$, $x \ge 0$, defined by (37) satisfy*

1. $0 < \tilde{z}_h(0) < 1$.
2. $1 - \tilde{y}_h(0) = \mathcal{O}\left(2^h e^{-2^h}\right)$ *$(h \to \infty)$.*
3. $\tilde{z}_h(1-16^{-h}) = \left(1-\frac{2}{e}\right)\cdot 16^h$.
4. $\tilde{z}_{h+r}(x) \ge \tilde{z}_h(x)$ *for all* $x \ge 0$ *and* $r \ge 0$.
5. $\tilde{z}'_{h+1}(x) = \tilde{z}_h(x)\left(\frac{2}{1-x} - \tilde{z}_h(x)\right)$.

Again, the difference between $z_h(x)$ and $\tilde{z}_h(x)$ is that they do not have the same initial condition. We have $\tilde{z}_h(0) < z_h(0) = 1$. This implies that $\tilde{z}_h(x) < z_h(x)$ for $0 \le x \le \overline{x}_h$ (for some $\overline{x}_h > 0$). By Lemma 10 $z_h(x)$ grows as a power of $\log\frac{1}{1-x}$ as $x \to 1-$. Since $\tilde{z}_h(x)$ grows as $(1-x)^{-1/2}$ it follows that $\lim_{x\to 1-}(z_h(x) - \tilde{z}_h(x)) = -\infty$. Actually, we can say a little bit more.

Lemma 13. *For every non-negative integer h and for every (real) $D \ge 0$ there exists $0 < \overline{x}_{h,D} < 1$ such that*

$$\tilde{z}_{h+D}(x) < z_h(x) \qquad (0 \le x < \overline{x}_{h,D}), \tag{38}$$

and

$$z_h(x) < \tilde{z}_{h+D}(x) \qquad (\overline{x}_{h,D} < x < 1). \tag{39}$$

Furthermore

$$\overline{x}_{h+1,D} > \overline{x}_{h,D}. \tag{40}$$

Proof. We proceed by induction. Since $\tilde{z}_D(x)$ is strictly increasing and satisfies $0 < \tilde{z}_D(0) < 1$ and $\lim_{x\to 1-}\tilde{z}_D(x) = \infty$ the assertion is surely true for $h = 0$. Now suppose that (38) and (39) are satisfied for some $h \ge 0$, i.e. the difference

$$\delta_{h,D}(x) := z_h(x) - \tilde{z}_{h+D}(x)$$

has a unique zero $\overline{x}_{h,D} > 0$ such that $\delta_{h,D}(x) > 0$ for $0 \le x < x_{h,D}$ and $\delta_{h,D}(x) < 0$ for $x > x_{h,D}$. Now we have

$$\begin{aligned}
\delta'_{h+1,D}(x) &= z'_{h+1}(x) - \tilde{z}'_{h+1+D}(x) \\
&= z_h(x)\left(\frac{2}{1-x} - z_h(x)\right) - \tilde{z}_{h+D}(x)\left(\frac{2}{1-x} - \tilde{z}_{h+D}(x)\right) \\
&= \delta_{h,D}(x)\left(\frac{2}{1-x} - z_h(x) - \tilde{z}_{h+D}(x)\right).
\end{aligned}$$

Recall that by definition $z_h(x) = \sum_{n\ge 0} \mathbf{Pr}[H'_n \le k]\, x^n \le 1/(1-x)$ and $\tilde{z}_{h+D}(x) \le 1/(1-x)$, compare with (37). Hence, $\delta_{h+1,D}(x)$ is increasing for $0 \le x < \overline{x}_{h,D}$ and decreasing for $x > \overline{x}_{h,D}$. Since $\delta_{h+1,D}(0) > 0$ and $\lim_{x\to\infty} \delta_{h+1,D}(x) = -\infty$ there exists a unique zero $\overline{x}_{h+1,D} > \overline{x}_{h,D}$ of $\delta_{h+1,D}(x)$ such that $\delta_{h+1,D}(x) > 0$ for $0 \le x < \overline{x}_{h+1,D}$ and $\delta_{h+1,D}(x) < 0$ for $x > \overline{x}_{h+1,D}$. □

By using Lemma 13 we can prove the following property which is the key to the proof of Theorem 2.

Lemma 14. *For every $h \ge 0$ there uniquely exists f_h such that*

$$\tilde{z}_{f_h}\left(1 - 16^{-f_h}\right) = z_h\left(1 - 16^{-f_h}\right). \tag{41}$$

They satisfy

$$f_{h+1} \ge f_h + \frac{1}{10}. \tag{42}$$

Moreover,

$$\tilde{z}_{f_h}(x) \le z_h(x) \qquad (0 \le x \le 1 - 16^{-f_h}), \tag{43}$$

and

$$z_h(x) \le \tilde{z}_{f_h}(x) \qquad (1 - 16^{-f_h} \le x < 1). \tag{44}$$

Proof. By Lemma 10

$$z_h\left(1 - 16^{-v}\right) \le 2^h \sum_{k=0}^{h} \frac{(\log 16)^k\, v^k}{k!}.$$

If $v \log 16 \ge h$ then the summands increase with k and we have

$$z_h\left(1 - 16^{-v}\right) \le (h+1) 2^h \frac{(\log 16)^h\, v^h}{h!}.$$

This means that $z_h\left(1 - 16^{-v}\right)$ is of polynomial growth with respect to v (if $v \log 16 \ge h$ which is no restriction). Consequently there exists $v_0 > 0$ such that

$$z_h\left(1 - 16^{-v_0}\right) \le \left(1 - \frac{2}{e}\right) 16^{v_0}.$$

Hence, by 3. of Lemma 12 we have $z_h\left(1 - 16^{-v_0}\right) \le \tilde{z}_{v_0}\left(1 - 16^{-v_0}\right)$. Since $z_h(0) = 1$ we have the converse inequality for $v = 0$. Thus, by continuity there exists f_h with

$$z_h\left(1 - 16^{-f_h}\right) = \left(1 - \frac{2}{e}\right) 16^{f_h} = \tilde{z}_{f_h}\left(1 - 16^{-f_h}\right).$$

Now we can apply Lemma 13 which implies that f_h (satisfying (41)) is uniquely determined and that (43) and (44) hold.

For the proof of (42) we need several steps. Firstly we prove that $f_{h+1} > f_h$. From

$$\overline{x}_{h+1, f_h + 1 - (h+1)} = \overline{x}_{h+1, f_h - h} > \overline{x}_{h, f_h - h}$$

it follows that

$$z_{h+1}\left(1 - 16^{-f_h}\right) > \tilde{z}_{f_h+1}\left(1 - 16^{-f_h}\right).$$

If we assume that $f_{h+1} \le f_h$ this also implies

$$z_{h+1}\left(1 - 16^{-f_{h+1}}\right) > \tilde{z}_{f_h+1}\left(1 - 16^{-f_{h+1}}\right).$$

However, since

$$z_{h+1}\left(1 - 16^{-f_{h+1}}\right) = \tilde{z}_{f_{h+1}}\left(1 - 16^{-f_{h+1}}\right)$$

this would imply $f_{h+1} \ge f_h + 1$. This is of course a contradiction to the assumption $f_{h+1} \le f_h$. Thus, we get $f_{h+1} > f_h$ as proposed.

Now we assume that $f_{h+1} < f_h + \frac{1}{10}$. This would imply that

$$\begin{aligned}\tilde z_{f_{h+1}}\left(1-16^{-f_{h+1}}\right) &= \left(1-\frac{2}{e}\right)16^{f_{h+1}} \\ &\le \left(1-\overline{\Phi}(1)\right)16^{f_h+\frac{1}{10}}.\end{aligned}$$

Next observe that

$$\left(1-\overline{\Phi}(1)\right)16^{\frac{1}{10}} < \left(1-16\overline{\Phi}(16)\right).$$

Thus

$$\begin{aligned}\tilde z_{f_{h+1}}\left(1-16^{-f_{h+1}}\right) &< \left(1-16\overline{\Phi}(16)\right)16^{f_h} \\ &= \tilde z_{f_h+1}\left(1-16^{-f_h}\right) \\ &< z_{h+1}\left(1-16^{-f_h}\right),\end{aligned}$$

which implies that

$$\begin{aligned}z_{h+1}\left(1-16^{-f_{h+1}}\right) &> z_{h+1}\left(1-16^{-f_h}\right) \\ &\ge \tilde z_{f_{h+1}}\left(1-16^{-f_{h+1}}\right).\end{aligned}$$

This is of course a contradiction and so we have finally proved (42). □

We are now able to prove tight upper and lower bounds for $b_{n,h}$.

Theorem 6. *There exists an absolute constant $\overline{C}_2 > 0$ such that*

$$1-\frac{b_{n,h}}{n!} \le \overline{C}_2 4^{-\left(\frac{\log n}{\log 16}-f_h\right)}, \tag{45}$$

for all non-negative integers h with $f_h \le (\log n)/(\log 16)$, and an absolute constant $\overline{C}_3 > 0$ such that

$$\frac{b_{n,h}}{n!} \le \overline{C}_3 2^{f_h-\frac{\log n}{\log 16}} \exp\left(-16^{f_h-\frac{\log n}{\log 16}}\right), \tag{46}$$

for all non-negative integers h with $f_h \ge (\log n)/(\log 16)$.

Proof. Suppose that $n \ge 16^{f_h}$ and consider

$$x = 1-\frac{1}{n} \ge 1-16^{-f_h}.$$

By Lemma 14 we have $z_h(x) \le \tilde z_{f_h}(x)$ and consequently

$$\begin{aligned}\tilde z_{f_h}(x) &\ge z_h(x) \\ &\ge \sum_{k=0}^{n}\left(1-\frac{b_{k,h}}{k!}\right)x^k \\ &\ge \left(1-\frac{b_{n,h}}{n!}\right)\sum_{k=0}^{n} x^k \\ &= \left(1-\frac{b_{n,h}}{n!}\right)\frac{1-x^{n+1}}{1-x}.\end{aligned}$$

Hence

$$\begin{aligned}1-\frac{b_{n,h}}{n!} &\le \tilde z_{f_h}\left(1-\frac{1}{n}\right)\frac{\frac{1}{n}}{1-\left(1-\frac{1}{n}\right)^{n+1}} \\ &\ll \frac{16^{f_h}}{n}\left(1-\overline{\Phi}\left(\frac{16^{f_h}}{n}\right)\right) \\ &\ll \left(\frac{16^{f_h}}{n}\right)^{1/2} \\ &= 4^{-\left(\frac{\log n}{\log 16}-f_h\right)}.\end{aligned}$$

The proof of (46) runs along the same lines. Here we assume that $n \le 16^{f_h}$ and use

$$x = 1 - \frac{1}{n} \le 1 - 16^{-f_h}.$$

By Lemma 14 and (30) we have

$$\begin{aligned}\frac{1}{1-x} - \tilde{z}_{f_h}(x) &\ge \frac{1}{1-x} - z_h(x)\\ &= \sum_{k\ge 0} \frac{b_{k,h}}{k!} x^k\\ &\ge \sum_{k=n}^{\infty} \frac{b_{k,h}}{k!} x^k\\ &\ge \frac{b_{n,h}}{n!} \sum_{k=n}^{\infty} x^k\\ &= \frac{b_{n,h}}{n!} \frac{x^n}{1-x}.\end{aligned}$$

Thus

$$\frac{b_{n,h}}{n!} \le x^{-n}\left(1-(1-x)\tilde{z}_{f_h}(x)\right) = x^{-n}(1-x)16^{f_h}\overline{\Phi}(16^{f_h}(1-x))$$

Finally, by using 2. of Lemma 12 we directly get

$$\begin{aligned}\frac{b_{n,h}}{n!} &\ll \frac{16^{f_h}}{n}\overline{\Phi}\left(\frac{16^{f_h}}{n}\right)\\ &\ll \left(\frac{16^{f_h}}{n}\right)^{1/4} \exp\left(-\frac{16^{f_h}}{n}\right)\\ &= 2^{f_h - \frac{\log n}{\log 16}} \exp\left(-16^{f_h - \frac{\log n}{\log 16}}\right)\end{aligned}$$

if $f_h \ge (\log n)/(\log 16)$. □

4.3. **Proof of Theorem 2.** It is now an easy task to prove Theorem 2 along similar lines as in the preceding proof of Theorem 1. We just use the estimates of Theorem 6 instead of Theorem 5.

As above we first have to make the following observation.

Lemma 15. *We have*

$$\sum_{h:n\ge 16^{f_h}} \left(1 - \frac{b_{n,h}}{n!}\right) = \mathcal{O}(1) \qquad (n\to\infty), \tag{47}$$

and

$$\sum_{h:n\le 16^{f_h}} \frac{b_{n,h}}{n!} = \mathcal{O}(1) \qquad (n\to\infty). \tag{48}$$

Proof. By Theorem 6 we just have to show that

$$\sum_{h:n\ge 16^{f_h}} 4^{-\left(\frac{\log n}{\log 16} - f_h\right)} = \mathcal{O}(1) \qquad (n\to\infty), \tag{49}$$

and

$$\sum_{h:n\le 16^{f_h}} 2^{f_h - \frac{\log n}{\log 16}} \exp\left(-16^{f_h - \frac{\log n}{\log 16}}\right) = \mathcal{O}(1) \qquad (n\to\infty). \tag{50}$$

By Lemma 14 we have $f_{h+1} - f_h \ge \frac{1}{10}$. Hence, the sum in (49) is bounded above by

$$\sum_{j\ge 0} 4^{-j/10} = \mathcal{O}(1)$$

and the sum in (50) by

$$\sum_{j\ge 0} 2^{j/10}\exp\left(-16^{j/10}\right) = \mathcal{O}(1).$$

□

So we can prove the estimate (6) for the expected value $\mathbf{E}H_n'$: We directly obtain

$$\begin{aligned}\mathbf{E}H_n' &= \sum_{h\ge 0}\frac{b_{n,h}}{n!}\\ &= \sum_{h:n>16^{f_h}}\frac{b_{n,h}}{n!} + \sum_{h:n\le 16^{f_h}}\frac{b_{n,h}}{n!}\\ &= \max\{h : f_h \le (\log n)/(\log 16)\} + \mathcal{O}(1)\\ &- \sum_{h:n>16^{f_h}}\left(1-\frac{b_{n,h}}{n!}\right) + \sum_{h:n\le 16^{f_h}}\frac{b_{n,h}}{n!}\\ &= \max\{h : f_h \le (\log n)/(\log 16)\} + \mathcal{O}(1).\end{aligned}$$

We finally have to check that

$$\max\{h : f_h \le (\log n)/(\log 16)\} = \max\{h : z_h(1-n^{-1}) \le (1-2/e)n\}. \tag{51}$$

For this purpose note that $1-16^{-f_h}$ is a zero of the function

$$z_h(x) - (1-2/e)/(1-x).$$

It is an easy exercise to show that for every η with $0<\eta<1$ there exists a unique zero of $z_h(x)-\eta/(1-x)$. (We just have to adapt the proof of Lemma 13. Note that one proves this assertion for all η at once.) Since $z_h(0)-\eta>0$ it follows that

$$z_h(x) < (1-2/e)\frac{1}{1-x} \quad \text{for } x > 1-16^{-f_h}$$

and

$$z_h(x) > (1-2/e)\frac{1}{1-x} \quad \text{for } x < 1-16^{-f_h}.$$

Now set $h_0 := \max\{h : f_h \le (\log n)/(\log 16)\}$, i.e., $16^{f_{h_0}} \le n$ and $16^{f_{h_0+1}} > n$. Hence

$$z_{h_0}\left(1-\frac{1}{n}\right) \le \left(1-\frac{2}{e}\right)n$$

and

$$z_{h_0+1}\left(1-\frac{1}{n}\right) > \left(1-\frac{2}{e}\right)n$$

and consequently

$$h_0 = \max\{h : z_h(1-n^{-1}) \le (1-2/e)n\}$$

as proposed.

The proof of the bounds for the centralized moments (7) is now an easy exercise.

5. m-Ary Search Trees

5.1. **Combinatorial Background.** Let $a_{n,h}$ denote the number of permutations $\sigma\in S_n$ of n elements such that the corresponding m-ary search tree has height $\le h$, i.e., $\mathbf{P}[H_{m,n}\le h] = a_{n,h}/n!$. Their (exponential) generating functions are given by

$$y_h(x) := \sum_{n\ge 0}\frac{a_{n,h}}{n!}x^n \tag{52}$$

and are again polynomials.

As in the case of binary search trees we can prove the following two properties.

Lemma 16. *The numbers $a_{n,h}$ satisfy the recurrence equation*

$$a_{n,h+1} = (m-1)! \sum_{n_1+n_2+\cdots+n_m=n-m+1} \frac{(n-m+1)!}{n_1!n_2!\cdots n_m!} a_{n_1,h}a_{n_2,h}\cdots a_{n_m,h} \tag{53}$$

with initial conditions $a_{0,0} = 1$ and $a_{n,0} = 0$ for $n > 0$. Furthermore we have

$$\frac{a_{n+1,h}}{(n+1)!} \le \frac{a_{n,h}}{n!} \qquad (h \ge 0). \tag{54}$$

Proof. Firstly, the recurrence (53) easily follows by considering an m-ary search trees of height $\le h+1$ as a root with keys $(1 \le) j_1 < j_2 < \cdots < j_{m-1} (\le n)$ and m subtrees of height $\le h$ with $n_1 = j_1 - 1, n_2 = j_2 - j_1 - 2, \dots, n_{m-1} = j_{m-1} - j_{m-2} - 1, n_m = n - j_{m-1}$ keys.

The inequality (54) can be restated as

$$\mathbf{P}[H_{m,n+1} \le h] \le \mathbf{P}[H_{m,n} \le h] \tag{55}$$

There is a "Markov chain model" for m-ary search (which is similar to that mentioned in the Introduction for binary search trees, see [4]). There $T_{m,n+1}$ is obtained from $T_{m,n}$ by inserting an additional key. Thus, one always has $H_{m,n+1} \ge H_{m,n}$ and consequently (55) and (54) follow immediately.[4] □

Lemma 17. *The functions $y_h(x)$ are recursively given by*

$$y_0(x) \equiv 1$$

and by

$$y_{h+1}^{(m-1)}(x) = (m-1)!\, y_h(x)^m \tag{56}$$

with $y_0(x) \equiv 1$ and initial values $y_h(0) = y_h'(0) = \cdots = y_h^{(m-2)}(0) = 1$ for $h > 0$.

Proof. The recurrence equation (56) is just a reformulation of (53). □

5.2. **Analytic Background.** Our aim is to find proper solutions $\Phi(u)$ of the (retarded) differential equation

$$\Phi^{(m-1)}(u) = (-1)^m \frac{(m-1)!}{\alpha^m} \Phi\left(\frac{u}{\alpha}\right)^m \tag{57}$$

for some $\alpha > 1$ and $\Phi(0) = 1$. Then the functions

$$\tilde{y}_h(x) = \alpha^h \Phi(\alpha^h(1-x))$$

emulate the recurrence equation for $y_h(x)$, i.e., they satisfy

$$\tilde{y}_{h+1}^{(m-1)}(x) = (m-1)!\, \tilde{y}_h(x)^m.$$

We can easily prove that (57) has entire solutions $\Phi(u)$ for any $\alpha > 1$ by showing that there exists a formal power series solution which converges in the whole complex plane. However, the essential difference between the case $m > 2$ and the case $m = 2$ is that we have $m - 2 > 0$ degrees of freedom if $m > 2$. For any choice of complex numbers $a_1, a_2, \dots, a_{m-2}$ there exists a unique solution $\Phi(u)$ of (57) with $\Phi(0) = 1, \Phi'(0) = a_1, \dots, \Phi^{(m-2)}(0) = a_{m-2}$. It seems that there is only one $(m-2)$-tuple $(a_1, a_2, \dots, a_{m-2})$ such that the corresponding solution of (57) satisfies $\Phi(u) \sim 1/u$ as $u \to \infty$. However, if we proceed as in [7] we get no hint how to choose $a_1, a_2, \dots, a_{m-2}$.

[4] In the Appendix the reader will find a combinatorial proof of (54) which is quite involved. This is strange since it seems that the simplicity of the *probabilistic argument* has no counterpart there.

Therefore we use a different approach to this problem, we consider a related integral equation

$$\int\limits_{z_1+\cdots+z_m=y, z_i>0} \Psi(z_1)\cdots\Psi(z_m)\, d\mathbf{z} = \frac{y^{m-1}}{(m-1)!}\Psi(y/\alpha). \tag{58}$$

We will solve this integral equation via a contraction argument (compare with Lemma 18).

The crucial observation is that the Laplace transform

$$\Phi(u) = \int_0^\infty \Psi(y)e^{-uy}\, dy$$

of $\Psi(y)$ is a solution of (57). Moreover asymptotic properties of $\Psi(y)$ easily translate to asymptotic properties for $\Phi(u)$, e.g. $\lim_{y\to 0}\Psi(y) = 1$ translates to $\Phi(u) \sim 1/u$ as $u \to \infty$.

Lemma 18. *Let $m \ge 2$ be a fixed integer and $1 < \alpha < e^{1/c_m}$. Then there exists a function $\Psi(y)$, $y \ge 0$, with the following properties:*

1. *$\Psi(y) = 1 + \mathcal{O}(y^\beta)$ as $y \to 0+$ for some $\beta > 1$.*
2. *$\Psi(y) = \mathcal{O}(e^{-Cy^\gamma})$ as $y \to \infty$ for some $C > 0$ and $\gamma > 1$.*
3. *$\Psi(y)$, $0 \le y < \infty$, is decreasing.*
4. $\displaystyle\int_0^\infty \Psi(y)\, dy = 1.$
5. $\displaystyle\int\limits_{z_1+\cdots+z_m=y, z_i>0} \Psi(z_1)\cdots\Psi(z_m)\, d\mathbf{z} = \frac{y^{m-1}}{(m-1)!}\Psi(y/\alpha),\ (0 \le y < \infty).$

Remark 8. Note that the integral

$$\int\limits_{z_1+\cdots+z_m=y, z_i>0} \Psi(z_1)\cdots\Psi(z_m)\, d\mathbf{z}$$

is exactly the m-fold convolution $(\Psi * \Psi * \cdots * \Psi)(y)$, e.g. for $m = 3$ we have

$$\int\limits_{z_1+z_2+z_3=y, z_i>0} \Psi(z_1)\Psi(z_2)\Psi(z_3)\, d\mathbf{z} = \int_0^y \Psi(z_1)\int_0^{y-z_1}\Psi(z_2)\Psi(y-z_1-z_2)\, dz_2\, dz_1.$$

Proof. We first show that if $1 < \alpha < e^{1/c_m}$ then there exists $\tilde\beta > 1$ with

$$m!\alpha^{\tilde\beta} < \prod_{i=0}^{m-1}(\tilde\beta + i). \tag{59}$$

By considering local expansions for α^β and $\prod_{i=0}^{m-1}(\beta+i)$ for α close to 1 it follows that there exists $\beta = \beta(\alpha) > 1$ such that

$$m!\alpha^{\beta} = \prod_{i=0}^{m-1}(\beta + i) \tag{60}$$

and

$$m!\log\alpha\,\alpha^{\beta} < \prod_{i=0}^{m-1}(\beta + i)\cdot\sum_{i=0}^{m-1}\frac{1}{\beta+i}. \tag{61}$$

Hence, by (61) and by the implicit function theorem there exists a unique local expansion of the solution $\beta = \beta(\alpha)$ of (60), and it also follows that there exists $\tilde\beta > \beta$ satisfying (59).

Now, if we use the substitution $c = (\log \alpha)^{-1}$ then (60) and (61) are equivalent to

$$\beta + c \log(m!) = c \sum_{i=1}^{m-1} \log(\beta + i), \tag{62}$$

and

$$\frac{1}{c} < \sum_{i=1}^{m-1} \frac{1}{\beta + i}. \tag{63}$$

However, $c = c_m$ is the smallest solution of (62) and

$$\frac{1}{c} = \sum_{i=1}^{m-1} \frac{1}{\beta + i}. \tag{64}$$

Thus, for every $\alpha < e^{1/c_m}$ we surely have (61) and consequently there exists $\beta > 1$ and $\tilde{\beta} > \beta$ satisfying (60) and (59).

Now let $\mathcal{F}$ denote the set of functions $\Psi(y)$, $y > 0$, with the following properties:

1. $\Psi(y) = 1 - y^\beta + \mathcal{O}(y^{\tilde{\beta}})$ as $y \to 0+$.
2. $\Psi(y) \ge 0$, $0 \le y < \infty$.
3. $\Psi(y)$, $0 \le y < \infty$, is decreasing.

It is clear that $\mathcal{F}$ with the distance

$$d(\Psi_1, \Psi_2) := \sup_{y>0} |(\Psi_1(y) - \Psi_2(y)) y^{-\tilde{\beta}}|$$

is a complete metric space. Now we show that the operator I, defined by

$$(I\Psi)(y) := \frac{(m-1)!}{\alpha^{m-1} y^{m-1}} \int\limits_{z_1 + \cdots + z_m = \alpha y, z_i > 0} \Psi(z_1) \cdots \Psi(z_m)\, d\mathbf{z},$$

is a contraction on $\mathcal{F}$.

Firstly, we prove that $I\Psi \in \mathcal{F}$ for all $\Psi \in \mathcal{F}$. Suppose that $\Psi \in \mathcal{F}$. Then

$$\Psi(z_1) \cdots \Psi(z_m) = 1 - \sum_{j=1}^m z_j^\beta + \sum_{j=1}^m \mathcal{O}(z_j^{\tilde{\beta}}).$$

Since

$$\int\limits_{z_1 + \cdots + z_m = \alpha y, z_i > 0} d\mathbf{z} = \frac{(\alpha y)^{m-1}}{(m-1)!}$$

and

$$\begin{aligned} \int\limits_{z_1 + \cdots + z_m = \alpha y, z_i > 0} z_j^\beta\, d\mathbf{z} &= \int_0^{\alpha y} \frac{(\alpha y - z_j)^{m-2}}{(m-1)!} z_j^\beta\, dz_j \\ &= \frac{(\alpha y)^{m-1+\beta}}{(m-1)!} \int_0^1 (1-v)^{m-2} v^\beta\, dv \\ &= \frac{(\alpha y)^{m-1+\beta}}{m-1} \frac{1}{(\beta+1)(\beta+2)\cdots(\beta+m-1)} \end{aligned}$$

it immediately follows (by using (60)) that

$$\begin{aligned} (I\Psi)(y) &= \frac{(m-1)!}{\alpha^{m-1} y^{m-1}} \int\limits_{z_1 + \cdots + z_m = \alpha y, z_i > 0} \Psi(z_1) \cdots \Psi(z_m)\, d\mathbf{z} \\ &= 1 - m \frac{(m-1)!}{\alpha^{m-1} y^{m-1}} \frac{(\alpha y)^{m-1+\beta}}{m-1} \frac{1}{(\beta+1)(\beta+2)\cdots(\beta+m-1)} + \mathcal{O}(y^{\tilde{\beta}}) \\ &= 1 - y^\beta + \mathcal{O}(y^{\tilde{\beta}}). \end{aligned}$$

Furthermore, it is clear that $(I\Psi)(y) \ge 0$ and by using the representation

$$(I\Psi)(y) = (m-1)! \int\limits_{x_1+\cdots+x_m=1, x_i>0} \Psi(\alpha y x_1)\cdots \Psi(\alpha y x_m)\, d\mathbf{x}$$

it is also clear that $(I\Psi)(y) \ge 0$ and that $(I\Psi)(y)$ is decreasing.

Now suppose that $\Psi_1, \Psi_2 \in \mathcal{F}$ with $d(\Psi_1, \Psi_2) = \delta$. Then it follows from $0 \le \Psi_j(y) \le 1$ that

$$\begin{aligned}|\Psi_1(z_1)\cdots\Psi_1(z_m) - \Psi_2(z_1)\cdots\Psi_2(z_m)| &\le \sum_{j=1}^m |\Psi_1(z_j) - \Psi_2(z_j)| \\ &\le \delta \sum_{j=1}^m z_j^{\tilde\beta}\end{aligned}$$

and consequently

$$\begin{aligned}|(I\Psi_1)(y) - (I\Psi_2)(y)| &\le \delta \frac{(m-1)!}{\alpha^{m-1}y^{m-1}} \sum_{j=1}^m \int\limits_{z_1+\cdots+z_m=\alpha y, z_i>0} z_j^{\tilde\beta}\, d\mathbf{z} \\ &= \delta \frac{m!\alpha^{\tilde\beta}}{(\tilde\beta+1)(\tilde\beta+2)\cdots(\tilde\beta+m-1)} y^{\tilde\beta}\end{aligned}$$

which implies

$$d(I\Psi_1, I\Psi_2) \le L \cdot d(\Psi_1, \Psi_2)$$

with

$$L = \frac{m!\alpha^{\tilde\beta}}{(\tilde\beta+1)(\tilde\beta+2)\cdots(\tilde\beta+m-1)}.$$

By (59) we have $L < 1$ and thus, $I : \mathcal{F} \to \mathcal{F}$ is a contraction.

By Banach's fixed point theorem there exists a unique fixed point $\Psi \in \mathcal{F}$. By definition, this fixed point satisfies properties 1., 3., and 5. of Lemma 18.

Next we prove 2. This fixed point Ψ may be obtained by starting with the function

$$\Psi_0(y) = \max\{1 - y^\beta, 0\}$$

and setting $\Psi_{k+1} := I\Psi_k$. Then $\Psi = \lim_{k\to\infty} \Psi_k$. By keeping track of the contraction I it follows that there exists $C_0 > 0$ and $y_0 > 0$ such that

$$\sup_{k\ge 0} \Psi_k(y) < 1 - y^\beta + C_0 y^{\tilde\beta} < 1 \quad \text{for } 0 < y \le y_0. \tag{65}$$

Set

$$\gamma := \frac{\log m}{\log m - \log \alpha}.$$

and

$$\eta := \min_{1\le l\le m-1} \frac{\alpha - \alpha^{\frac{\log l}{\log m}}}{m-l}.$$

Observe that $0 < \eta \le \frac{\alpha-1}{m-1}$. Now choose $C > 0$ (sufficiently small) such that

$$e^{-Cy^\gamma} > 1 - y^\beta + C_0 y^{\tilde\beta} \quad \text{for } \eta y_0 \le y \le y_0. \tag{66}$$

We prove inductively that

$$\Psi_k(y) \le e^{-Cy^\gamma} \quad \text{for all } y \ge y_0. \tag{67}$$

Obviously, (67) is satisfied for $k = 0$.

Now suppose that (67) holds for some $k \ge 0$. By (65) and (66) we also have $\Psi_k(y) \le e^{-Cy^\gamma}$ for $\eta y_0 \le y \le y_0$. Our aim is to show that

$$\Psi_k(z_1)\Psi_k(z_2)\cdots\Psi_k(z_m) \le e^{-Cy^\gamma} \tag{68}$$

for $z_1, \ldots, z_m > 0$ with $z_1 + \cdots + z_m = \alpha y$ and $y \ge y_0$.

It is then clear that (68) implies

$$\Psi_{k+1}(y) = \frac{(m-1)!}{(\alpha y)^{m-1}} \int\limits_{z_1+\cdots+z_m=\alpha y, z_i>0} \Psi_k(z_1 \cdots \Psi_k(z_m)\, d\mathbf{z} \le e^{-Cy^\gamma}$$

for $y \ge y_0$, as proposed.

If $z_i \ge \eta y_0$ for all i then we directly get

$$\begin{aligned}\Psi_k(z_1)\Psi_k(z_2)\cdots\Psi_k(z_m) &\le e^{-C(z_1^\gamma+z_2^\gamma+\cdots+z_m^\gamma)} \\ &\le e^{-Cm((z_1+z_2+\cdots+z_m)/m)^\gamma} \\ &= e^{-C((z_1+z_2+\cdots+z_m)/\alpha)^\gamma} \\ &= e^{-Cy^\gamma}.\end{aligned}$$

Next suppose that $z_i \ge \eta y_0$ for $1 \le i \le l$ and $z_i < \eta y_0$ for $l+1 \le i \le m$ with some $1 \le l \le m-1$. (By symmetry this is no loss of generality. Furthermore, since $m\eta y_0 \le \alpha y$ (for $y \ge y_0$) it is impossible that $z_i < \eta y_0$ for all i.) Here we have

$$\begin{aligned}\Psi_k(z_1)\Psi_k(z_2)\cdots\Psi_k(z_m) &\le \Psi_k(z_1)\Psi_k(z_2)\cdots\Psi_k(z_l) \\ &\le e^{-C(z_1^\gamma+z_2^\gamma+\cdots+z_l^\gamma)} \\ &\le e^{-Cl((z_1+z_2+\cdots+z_l)/l)^\gamma} \\ &= e^{-C(m/l)^{\gamma-1}(y-(z_{l+1}+\cdots+z_m)/\alpha)^\gamma} \\ &\le e^{-C(m/l)^{\gamma-1}(y-(m-l)\eta y_0/\alpha)^\gamma}.\end{aligned}$$

Since

$$\eta\frac{m-l}{\alpha} \le 1 - \alpha^{\frac{\log l}{\log m}-1} = 1 - \left(\frac{l}{m}\right)^{1-\gamma^{-1}}$$

and $y_0 \le y$ it follows that

$$\eta y_0 \frac{m-l}{\alpha} \le y - \left(\frac{l}{m}\right)^{1-\gamma^{-1}} y$$

or

$$\left(\frac{m}{l}\right)^{\gamma-1}\left(y-(m-l)\frac{\eta y_0}{\alpha}\right)^\gamma \ge y^\gamma$$

which implies that

$$\Psi_k(z_1)\Psi_k(z_2)\cdots\Psi_k(z_m) \le e^{-Cy^\gamma}$$

even in this remaining case. This completes the proof of (68) and consequently the inductive proof of (67).

We finally mention that a linear substitution $y \to \kappa y$ (with $\kappa > 0$) again leads to a solution of the integral equation $I\Psi = \Psi$, and so we can adjust $\kappa > 0$ such that $\int_0^\infty \Psi(\kappa y)\, dy = 1$. □

As already mentioned the Laplace transform $\Phi(u)$ of $\Psi(y)$ is now a proper solution of (57).

Lemma 19. *Let $\Psi(y)$ be as in Lemma 18. Then the Laplace transform*

$$\Phi(u) = \int_0^\infty \Psi(y)e^{-uy}\, dy$$

is an entire function and satisfies the following properties.

1. $\Phi(0) = 1$.
2. $\Phi(u)$ *is decreasing and* $u\Phi(u)$ *is increasing for real* $u > 0$.
3. $0 < (-1)^j\Phi^{(j)}(u) < j!u^{-j-1}$ *for real* $u > 0$.
4. $1 - u\Phi(u) = \mathcal{O}(u^{-\beta})$ *as* $u \to \infty$ *for some* $\beta > 1$.
5. $\Phi^{(m-1)}(u) = (-1)^m\frac{(m-1)!}{\alpha^m}\Phi\left(\frac{u}{\alpha}\right)^m$.

Proof. Since $\Psi(y) = \mathcal{O}(e^{-Cy^\gamma})$ as $y \to \infty$ for some $\gamma > 1$ it is clear that the Laplace transform $\Phi(u)$ is an entire function. Furthermore, by 4. of Lemma 18, we have $\Phi(0) = 1$. Moreover, since $\Psi(y)$ is non-negative, it follows by definition that $\Phi(u)$ is decreasing.

By integration by parts we get for any $u > 0$

$$u\Phi(u) = 1 - \int_0^\infty \Psi'(y)e^{-uy}\,dy.$$

Since $\Psi'(y) \le 0$ for $y > 0$ it follows that $u\Phi(u)$ is increasing for $u > 0$.

Furthermore, by differentiation we obtain

$$\Phi^{(j)}(u) = (-1)^j \int_0^\infty \Psi(y)y^j e^{-uy}\,dy,$$

and consequently

$$\begin{aligned} 0 < (-1)^j\Phi^{(j)}(u)u^{j+1} &= \int_0^\infty \Psi(z/u)z^j e^{-z}\,dz \\ &< \int_0^\infty z^j e^{-z}\,dz = j!. \end{aligned}$$

Next, the expansion $\Psi(y) = 1 - \mathcal{O}(y^\beta)$ as $y \to 0+$ directly translates to

$$\Phi(u) = \frac{1}{u} - \mathcal{O}\left(\frac{1}{u^{\beta+1}}\right)$$

as $u \to \infty$.

Finally the integral equation for $\Psi(y)$ induces the proposed differential equation for $\Phi(u)$. □

As already mentioned we will work with the auxiliary functions

$$\tilde{y}_h(x) := \alpha^h\Phi(\alpha^h(1-x)). \tag{69}$$

The properties of $\Phi(u)$ can be translated to corresponding properties of $\tilde{y}_h(x)$. The proof is immediate.

Lemma 20. *The functions $\tilde{y}_h(x)$, $h \ge 0$, $x \ge 0$, defined by (69) satisfy*

1. $\tilde{y}_h(0) < 1 = 0!, \tilde{y}_h'(0) < 1!, \ldots, \tilde{y}_h^{(m-2)}(0) < (m-2)!$.
2. $1 - \tilde{y}_h(0) = \mathcal{O}(\alpha^{-\beta h})$ *$(h \to \infty)$.*
3. $\tilde{y}_h(1) = \alpha^h$.
4. $\tilde{y}_{h+r}(x) \ge \tilde{y}_h(x)$ *for all* $x \ge 0$ *and* $r \ge 0$.
5. $\tilde{y}_{h+1}^{(m-1)}(x) = (m-1)!\tilde{y}_h(x)^m$.

We again note that although $\tilde{y}_{h+D}(0) < y_h(0)$ there surely exists $x_0 > 0$ with $\tilde{y}_{h+D}(x_0) = y_h(x_0)$, moreover we have $\tilde{y}_{h+D}(x) - y_h(x) \to \infty$ as $x \to \infty$. This is due to the fact that $y_h(x)$ is a polynomial and $\tilde{y}_{h+D}(x)$ is a power series with positive coefficients. (Observe that $[x^n]\tilde{y}_h(x) = \frac{1}{n!}\int_0^\infty y^n e^{-y}\Psi(y\alpha^{-h})\,dy > 0$.)

We can prove an analogue to Lemma 5.

Lemma 21. *For every non-negative integer h and for every (real) $D \ge 0$ there exists $x_{h,D} > 0$ such that*

$$\tilde{y}_{h+D}(x) < y_h(x) \qquad (0 \le x < x_{h,D}) \tag{70}$$

and

$$y_h(x) < \tilde{y}_{h+D}(x) \qquad (x > x_{h,D}). \tag{71}$$

Furthermore, we have

$$x_{h+1,D} > x_{h,D}. \tag{72}$$

Proof. We proceed by induction. Since $\tilde{y}_D(x)$ is strictly increasing and satisfies $0 < \tilde{y}_D(0) < 1$ and $\lim_{x\to\infty} \tilde{y}_D(x) = \infty$ the assertion is surely true for $h = 0$. Now suppose that (70) and (71) are satisfied for some $h \ge 0$, i.e. the difference

$$\delta_{h,D}(x) := y_h(x) - \tilde{y}_{h+D}(x)$$

has a unique zero $x_{h,D} > 0$ such that $\delta_{h,D}(x) > 0$ for $0 \le x < x_{h,D}$ and $\delta_{h,D}(x) < 0$ for $x > x_{h,D}$. Now we have

$$\begin{aligned}\delta_{h+1,D}^{(m-1)}(x) &= y_{h+1}^{(m-1)}(x) - \tilde{y}_{h+1+D}^{(m-1)}(x)\\ &= (m-1)!(y_h(x)^m - \tilde{y}_{h+D}(x)^m)\\ &= (m-1)!\delta_{h,D}(x)\sum_{k=0}^{m-1} y_h(x)^k \tilde{y}_{h+D}(x)^{m-1-k}.\end{aligned}$$

Thus, $\delta_{h+1,D}^{(m-2)}(x)$ is increasing for $0 \le x < x_{h,D}$ and decreasing for $x > x_{h,D}$. Since $\delta_{h+1,D}^{(m-2)}(0) > 0$ and (by the same reasoning as above) $\lim_{x\to\infty} \delta_{h+1,D}^{(m-2)}(x) = -\infty$ there exists a unique zero $x_1 > x_{h,D}$ of $\delta_{h+1,D}^{(m-2)}(x)$ such that $\delta_{h+1,D}^{(m-2)}(x) > 0$ for $0 \le x < x_1$ and $\delta_{h+1,D}^{(m-2)}(x) < 0$ for $x > x_1$. In the same way it follows that $\delta_{h+1,D}^{(m-3)}(x)$ is increasing for $0 \le x < x_1$ and decreasing for $x > x_1$. Again we have $\delta_{h+1,D}^{(m-3)}(0) > 0$ and $\lim_{x\to\infty} \delta_{h+1,D}^{(m-3)}(x) = -\infty$. Thus there exists a unique zero $x_2 > x_1$ of $\delta_{h+1,D}^{(m-3)}(x)$ such that $\delta_{h+1,D}^{(m-3)}(x) > 0$ for $0 \le x < x_2$ and $\delta_{h+1,D}^{(m-3)}(x) < 0$ for $x > x_1$. Repeating this procedure we finally get that there is a unique zero $x_{h+1,D} > x_{h,D}$ of $\delta_{h+1,D}(x)$ such that $\delta_{h+1,D}(x) > 0$ for $0 \le x < x_{h+1,D}$ and $\delta_{h+1,D}(x) < 0$ for $x > x_{h+1,D}$. □

We now proceed as in the binary case and obtain the following property.

Lemma 22. *Let e_h be defined by $e_h := \log y_h(1)/\log\alpha$. Then we have $e_{h+1} \ge e_h + 1$. Moreover,*

$$\tilde{y}_{e_h}(x) \le y_h(x) \qquad (0 \le x \le 1), \tag{73}$$

and

$$y_h(x) \le \tilde{y}_{e_h}(x) \qquad (x \ge 1). \tag{74}$$

We also get tight upper and lower bounds for $a_{n,h}$.

Theorem 7. *There exists an absolute constant $C_2 > 0$ such that*

$$\frac{a_{n,h}}{n!} \le C_2 \left(\alpha^{-\beta}\right)^{e_h - c\log n} \tag{75}$$

for all non-negative integers h with $e_h \le c\log n$, and an absolute constant $C_3 > 0$ such that

$$1 - \frac{a_{n,h}}{n!} \le C_3 \left(\alpha^{-\beta}\right)^{e_h - c\log n} \tag{76}$$

for all non-negative integers h with $e_h \ge c\log n$.

Proof. As in the binary case we get

$$\frac{a_{n,h}}{n!} \le \frac{1}{(1+\alpha^{-e_h})^n - 1}\Phi(-1).$$

Since

$$(1+y)^n - 1 \gg (ny)^{\alpha^\beta}$$

we directly get the proposed bound (75).

The proof of (76) runs along the same lines. As in the binary case we get

$$\begin{aligned}1 - \frac{a_{n,h}}{n!} &\ll 1 - \alpha^{e_h - c\log n}\Phi(\alpha^{e_h - c\log n})\\ &\ll \left(\alpha^{-\beta}\right)^{e_h - c\log n}\end{aligned}$$

if $e_h \ge c\log n$. □

5.3. **Proof of Theorem 3.** It is now clear that the estimates for the tails of the distribution of $H_{m,n}$ (provided in Theorem 7) are sufficient to prove Theorem 3.

Firstly we get

$$\sum_{h:e_h \le c\log n} \frac{a_{n,h}}{n!} = \mathcal{O}(1) \qquad (n\to\infty) \tag{77}$$

and

$$\sum_{h:e_h \ge c\log n} \left(1 - \frac{a_{n,h}}{n!}\right) = \mathcal{O}(1) \qquad (n\to\infty). \tag{78}$$

which implies that the expected value of the height $H_{m,n}$ of m-ary search trees is given by

$$\mathbf{E}H_{m,n} = \max\{h : y_h(1) \le n\} + \mathcal{O}(1) \qquad (n\to\infty), \tag{79}$$

and that all centralized moments of $H_{m,n}$ are bounded.

6. The Limiting Distribution of H_n

In this section we prove Theorem 4 on the limiting distribution of the height H_n of binary search trees. (The situation looks similar for $m > 2$. However, it seems that there are further technical difficulties. Therefore we just consider the binary case.) It turns out that it is sufficient to prove the following property.

Proposition 1. *Let $y_h(x)$ be defined by (3) and $\Psi(y)$, $y \ge 0$, the function given by Lemma 23. Then*

$$\mathbf{P}[H_n \le h] \sim \Psi(n/y_h(1)) \tag{80}$$

uniformly for h such that $\log y_h(1) - \log n$ belongs to a bounded set, as $n\to\infty$.

If we combine (80) with the tail estimates of Theorem 5 we obtain

$$\sup_{h\ge 0} |\mathbf{P}[H_n \le h] - \Psi(n/y_h(1))| = o(1) \quad (n\to\infty),$$

and thus Theorem 4 follows.

It is clear that a very precise form of the distribution of H_n (together with the tail estimates of Theorem 6) leads to precise asymptotics for the moments of H_n. However, to get this, we need a more precise asymptotic expression for $y_h(1)$ than that stated in (10).

Hypothesis 1. There exists a constant D such that as $h \longrightarrow \infty$

$$y_h(1) = e^{h/c + \frac{3}{2(c-1)}\log h + D + o(1)}. \tag{81}$$

Theorem 8. *If Hypothesis 1 is true then there exist continuous periodic functions Δ_1 and Δ_2 with period 1 such that*

$$\mathbf{E}H_n = c\log n - \frac{3c}{2(c-1)}\log\log n + \Delta_1\left(c\log n - \frac{3c}{2(c-1)}\log\log n\right) + o(1)$$

and

$$\mathbf{V}H_n = \Delta_2\left(c\log n - \frac{3c}{2(c-1)}\log\log n\right) + o(1)$$

as $n\to\infty$.

While Hypothesis 1 is plausible given (10), it seems that proving it will require a significant new idea.

The organization of this section is as follows. In section 6.1, we present some analytic preliminaries which are used in section 6.2 to prove Proposition 1 and thus Theorem 4. In section 6.3 we prove Theorem 8.

6.1. Analytic Background.

In section 5.2 we proved that for every α with $1 < \alpha < e^{1/c}$ there exists a solution $\Psi(y) = \Psi(y,\alpha)$ of the integral equation

$$y\Psi(y/\alpha) = \int_0^y \Psi(z)\Psi(y-z)\,dz. \tag{82}$$

We now show that this property is also satisfied for $\alpha = e^{1/c}$.

Lemma 23. *Let $\alpha = e^{1/c}$ and $\beta = c-1$ Then there uniquely exists a function $\Psi(y)$, $y \ge 0$, with the following properties:*

1. *$\Psi(y) - 1 \sim c_1 y^\beta \log y$ as $y \to 0+$ for some constant c_1.*
2. *$\Psi(y) = \mathcal{O}(e^{-Cy^\gamma})$ as $y \to \infty$ for some $C > 0$ and some $\gamma > 1$.*
3. *$\Psi(y)$, $0 \le y < \infty$, is decreasing.*
4. *$\displaystyle\int_0^\infty \Psi(y)\,dy = 1$.*
5. *$\displaystyle y\Psi(y/\alpha) = \int_0^y \Psi(z)\Psi(y-z)\,dz$*

Furthermore, the Laplace transform

$$\Phi(u) = \int_0^\infty e^{-uy}\Psi(y)\,dy$$

of $\Psi(y)$ is exactly the solution of (16) described in Lemma 3.

Proof. In [7] it was shown that for every $\alpha > 1$ the retarded differential equation

$$\Phi'(u) = -\frac{1}{\alpha^2}\Phi\left(\frac{u}{\alpha}\right)^2$$

has a unique entire solution $\Phi(u) = \Phi(u,\alpha)$ with initial condition $\Phi(0,\alpha) = 1$. The idea of the proof is to consider a Taylor series expansion

$$\Phi(u,\alpha) = \sum_{k\ge 0}(-1)^k c_k(\alpha) u^k.$$

Starting with $c_0(\alpha) = 1$ we get the recurrence

$$c_{k+1}(\alpha) = \frac{\alpha^{-k}}{k+1}\sum_{l=0}^{k} c_l(\alpha)c_{k-l}(\alpha).$$

It is an easy exercise to show that $\sum_{k\ge 0}(-1)^k c_k(\alpha)u^k$ constitutes an entire function (compare with [7]). We also get that $c_k(\alpha)$ is a polynomial in $1/\alpha$ with non-negative coefficients. Thus, $c_k(\alpha)$ is decreasing as a function in $\alpha > 1$. Furthermore, for every $\alpha_0 > 1$ and every compact set $K \subset \mathbb{C}$ the limit relation

$$\lim_{\alpha\to\alpha_0} \Phi(u,\alpha) = \Phi(u,\alpha_0) \tag{83}$$

is uniform for $u \in K$.

We also know from Lemmata 18 and 19 that for $\alpha < e^{1/c}$

$$\Phi(u,\alpha) = \int_0^\infty \Psi(y,\alpha)e^{-uy}\,dy,$$

where $\Psi(y) = \Psi(y,\alpha)$ is constructed in Lemma 18. We now show that the limit

$$\Psi_0(y) := \lim_{\alpha\to e^{1/c}-} \Psi(y,\alpha)$$

exists and that $\Phi(u, e^{1/c})$ is the Laplace transform of $\Psi_0(y)$.

For this purpose we first show that

$$|\Phi(it,\alpha)| \le \frac{2}{|t|} \tag{84}$$

for all $\alpha \le e^{1/c}$. By integration by parts we have

$$\begin{aligned}\Phi(it,\alpha) &= \int_0^\infty \Psi(y,\alpha)e^{-ity}\,dy\\ &= -\frac{1}{it} - \frac{1}{it}\int_0^\infty \Psi'(y,\alpha)e^{-ity}\,dy.\end{aligned}$$

Furthermore, for all $\alpha < e^{1/c}$

$$\left|\int_0^\infty \Psi'(y,\alpha)e^{-ity}\,dy\right| \le \int_0^\infty (-\Psi'(y,\alpha))\,dy = 1.$$

This proves (84) for $\alpha < e^{1/c}$. However, by (83), we get the same bound for $\alpha = e^{1/c}$. By using the retarded differential equation for $\Phi(u)$ this also proves that

$$|\Phi'(it,\alpha)| \le \frac{4}{t^2} \tag{85}$$

for all $\alpha \le e^{1/c}$.

Next observe that the representation

$$\begin{aligned}-\alpha^2\Phi'(\alpha u,\alpha) &= \Phi(u,\alpha)^2\\ &= \int_0^\infty (\Psi * \Psi)(y,\alpha)e^{-uy}\,dy\\ &= \int_0^\infty y\Psi(y/\alpha,\alpha)e^{-uy}\,dy,\end{aligned}$$

shows that

$$y\Psi(y/\alpha,\alpha) = \frac{1}{2\pi}\int_{-\infty}^\infty (-\alpha^2\Phi'(it\alpha,\alpha))e^{ity}dt.$$

Thus, using (85) we can define a function $\Psi_0(y)$ for $y > 0$ by

$$\Psi_0(y/\alpha) = \frac{1}{2\pi y}\int_{-\infty}^\infty (-e^{2/c}\Phi'(ite^{1/c},e^{1/c}))e^{ity}dt. \tag{86}$$

By (83) and (85) this function is the limit $\lim_{\alpha\to e^{1/c}-}\Psi(y,\alpha)$. This limit is also uniform for $y \in [a,b]$ with $a > 0$ and $b < \infty$. Hence, we also get

$$y\Psi_0(y) = \int_0^y \Psi_0(z)\Psi_0(y-z)\,dz.$$

Moreover, (86) implies that $\Phi(u,e^{1/c})$ is the Laplace transform of $\Psi_0(y)$, too. Since $\Phi(u,e^{1/c})$ is uniquely given, this implies that $\Psi_0(y) = \Psi(y,e^{1/c})$ is unique, too.

In order to complete the proof of Lemma 23 we only have to check the proposed properties 1. and 2. By Lemma 3 we know that, as $u \to \infty$,

$$1 - u\Phi(u) \sim C_1\frac{\log u}{u^{c-1}},$$

or

$$\Phi(u) = \frac{1}{u} - C_1\frac{\log u}{u^c}(1+o(1)).$$

By using well known properties of the Laplace transform (see [6, pp. 208] with a slight generalization taking into account the logarithmic factor or the dual version of [9, p. 446, Theorem 4]) it follows that this kind of asymptotic relation of $\Phi(u)$ for $u \to \infty$ translates to a corresponding asymptotic property of $\Psi(y)$ for $y \to 0$. (Here we also use the fact that $\Psi(y)$ is decreasing.) This proves 1.

Finally, since $c_k(\alpha)$ is decreasing, it follows that

$$\Phi(u,e^{1/c}) \le \Phi(u,\alpha)$$

for $u < 0$ and any $\alpha < e^{1/c}$. Since we know that $\Psi(y,\alpha) = \mathcal{O}(e^{-Cy^\gamma})$ for some $C > 0$ and some $\gamma > 1$ it follows that

$$\Phi(u,\alpha) = \mathcal{O}\left(e^{C_2(-u)^{\gamma/(\gamma-1)}}\right)$$

as $u \to -\infty$ for some constant $C_2 > 0$. Thus, we get the same upper bound for $\Phi(u, e^{1/c})$ and consequently an upper bound for the inverse Laplace transform

$$\Psi_0(y) = \mathcal{O}(e^{-C'y^\gamma})$$

as $y \to \infty$ (for some constant $C' > 0$). □

Now we can prove asymptotic expansions for the coefficients of the auxiliary functions $\tilde{y}_h(x) = \alpha^h \Phi(\alpha^h(1-x))$.

Lemma 24. *We have*

$$[x^n]\,\tilde{y}_h(x) \sim \Psi(n\alpha^{-h})$$

uniformly for h such that $h - c\log n$ belongs to a bounded set, as $n \to \infty$.

Proof. By definition we get

$$\begin{aligned}\tilde{y}_h(x) &= \alpha^h \Phi(\alpha^h(1-x)) \\ &= \int_0^\infty e^{xy-y}\Psi(y\alpha^{-h})\,dy \\ &= \sum_{n\ge 0}\left(\frac{1}{n!}\int_0^\infty y^n e^{-y}\Psi(y\alpha^{-h})\,dy\right) x^n.\end{aligned}$$

If $h - c\log n = \mathcal{O}(1)$ then $n\alpha^{-h}$ is bounded, too, i.e., $C_1 \le n\alpha^{-h} \le C_2$ for some constants $C_1, C_2 > 0$. Hence, by the Laplace method and by using the relation

$$\frac{1}{n!}\int_0^\infty y^n e^{-y}\,dy = 1,$$

we obtain (by expanding locally around $y = n$)

$$[x^n]\,\tilde{y}_h(x) = \frac{1}{n!}\int_0^\infty y^n e^{-y}\Psi(y\alpha^{-h})\,dy \sim \Psi(n\alpha^{-h}).$$

□

6.2. **Proof of Proposition 1.** The idea of the proof of Proposition 1 (and consequently of Theorem 4) is that $y_h(x)$ (defined by (11)) can be properly approximated by $\tilde{y}_{e_h}(x) = y_h(1)\Phi(y_h(1)(1-x))$ (where $e_h = c\cdot\log y_h(1)$), especially the coefficients $[x^n]y_h(x)$ are asymptotically given by $[x^n]\tilde{y}_{e_h}(x)$.

In a first step we estimate $y_h(x)$ and $\tilde{y}_h(x)$ for complex values x of modulus 1.

Lemma 25. *For complex values $x \ne 1$ with $|x| = 1$ we have*

$$|y_h(x)| \le \frac{2}{|1-x|}$$

and

$$|\tilde{y}_h(x)| \le \frac{2}{|1-x|}.$$

Proof. We recall that the coefficients of $y_h(x)$ are decreasing, i.e., $a_{n+1,h}/(n+1)! \le a_{n,h}/n!$. Thus, we obtain

$$\begin{aligned}|(1-x)y_h(x)| &= \left|1 - \sum_{n=1}^\infty \left(\frac{a_{n-1,h}}{(n-1)!} - \frac{a_{n,h}}{n!}\right)x^n\right| \\ &\le 1 + \sum_{n=1}^\infty \left(\frac{a_{n-1,h}}{(n-1)!} - \frac{a_{n,h}}{n!}\right) \\ &= 2.\end{aligned}$$

It is clear that we get the same bounds for $\tilde{y}_h(x)$ if the coefficients of $\tilde{y}_h(x)$ are decreasing, too, i.e., $[x^{n+1}]\tilde{y}_h(x) \le [x^n]\tilde{y}_h(x)$. In order to prove this property, we first observe that the m-th derivative of $\tilde{y}_h(x)$ can be represented by

$$\tilde{y}_h^{(n)}(x) = \sum_{l=1}^{n!} \prod_{j=1}^{n+1} \tilde{y}_{h-d_{lj}}(x), \tag{87}$$

where $d_{lj} \ge 1$ are (specific) integers. This follows by induction. Firstly, by 5. of Lemma 4 we have $\tilde{y}_h'(x) = \tilde{y}_{h-1}(x)\tilde{y}_{h-1}(x)$. Furthermore,

$$\frac{d}{dx}\left(\prod_{j=1}^{n+1} \tilde{y}_{h-d_{lj}}(x)\right) = \sum_{k=1}^{n+1} \tilde{y}_{h-d_{lk}-1}(x)^2 \prod_{1\le j\le n+1, j\ne k} \tilde{y}_{h-d_{lj}}(x).$$

Thus, we obtain the proposed representation (87). Moreover, since

$$\tilde{y}_{h-d_{lk}-1}(0)^2 < \tilde{y}_{h-d_{lk}}(0),$$

we also obtain

$$\frac{d}{dx}\left(\prod_{j=1}^{n+1} \tilde{y}_{h-d_{lj}}(x)\right)_{x=0} < (n+1)\prod_{j=1}^{n+1} \tilde{y}_{h-d_{lj}}(0),$$

and consequently

$$\tilde{y}_h^{(n+1)}(0) < (n+1)\tilde{y}_h^{(n)}(0),$$

which is equivalent to $[x^{n+1}]\tilde{y}_h(x) < [x^n]\tilde{y}_h(x)$. □

In a second step we show that $y_h(x) \sim \tilde{y}_{e_h}(x)$ for x close to 1. We start with the following easy but important property.

Lemma 26. *Let $y_h(x)$ be defined by (11). Then we have*

$$\frac{y_{h+2}(1)}{y_{h+1}(1)} \le \frac{y_{h+1}(1)}{y_h(1)}. \tag{88}$$

Consequently the sequence $y_{h+1}(1)/y_h(1)$ converges and its limit is given by

$$\lim_{h\to\infty} \frac{y_{h+1}(1)}{y_h(1)} = e^{1/c}. \tag{89}$$

Proof. Let $\gamma \in (0,1)$ be a fixed constant and let $z_h(x) = z_h(x,\gamma)$ be defined by $z_0(x) = 1/(1-x)$ for $x \le 1-\gamma$, by $z_0(x) = 1/\gamma$ for $x > 1-\gamma$, and recursively by

$$z_{h+1}(x) = 1 + \int_0^x z_h(t)^2\, dt \quad \text{for } h \ge 0.$$

Of course, by induction it follows that $z_h(x) = 1/(1-x)$ for $x \le 1-\gamma$ and that

$$z_h(x) = \frac{1}{\gamma} y_h\left((x-1+\gamma)/\gamma\right) \quad \text{for } x > 1-\gamma.$$

Now we proceed as in the proof of Lemma 5 and obtain that the difference $z_h(x) - y_{h+1}(x)$ has exactly one zero $x_h(\gamma)$ in the range $x > 0$, i.e. $z_h(x) > y_{h+1}(x)$ for $0 < x < x_h(\gamma)$ and $z_h(x) < y_{h+1}(x)$ for $x > x_h(\gamma)$. Furthermore $x_{h+1}(\gamma) > x_h(\gamma)$. We now apply this property for $\gamma = y_h(1)/y_{h+1}(1)$. Since $z_h(1,\gamma) = y_h(1)/\gamma$ it follows that

$$z_h\left(1, y_h(1)/y_{h+1}(1)\right) = y_{h+1}(1)$$

or

$$x_h(y_h(1)/y_{h+1}(1)) = 1.$$

Consequently

$$x_{h+1}(y_h(1)/y_{h+1}(1)) > 1$$

and thus

$$z_{h+1}\left(1, y_h(1)/y_{h+1}(1)\right) = \frac{y_{h+1}(1)^2}{y_h(1)} \ge y_{h+2}(1)$$

as proposed.

Since the sequence $y_{h+1}(1)/y_h(1)$ is decreasing and non-negative is it thus convergent. We already know that $\log y_h(1) \sim h/c$ (compare with (10)). Hence the limit is given by (89). □

Lemma 27. *For every constant $K > 0$ we have*

$$y_h(x) \sim \tilde{y}_{e_h}(x)$$

uniformly for all complex values x with $|x-1| \le K/y_h(1)$.

Proof. The idea is to consider the Taylor series expansions of $y_h(x)$ and $\tilde{y}_{e_h}(x)$ locally around $x = 1$:

$$y_h(x) = \sum_{n\ge 0} \frac{y_h^{(n)}(1)}{n!}(x-1)^n \quad \text{and} \quad \tilde{y}_{e_h}(x) = \sum_{n\ge 0} \frac{\tilde{y}_{e_h}^{(n)}(1)}{n!}(x-1)^n. \tag{90}$$

Firstly, we show that

$$y_h^{(n)}(1) \le \tilde{y}_{e_h}^{(n)}(1) \tag{91}$$

for all $n \ge 0$. By (87) and a corresponding relation for the derivatives of $\tilde{y}_{e_h}(x)$ it suffices to show that

$$y_{h-d}(1) \le \tilde{y}_{e_h-d}(1) \tag{92}$$

for all $d \ge 0$. (92) is equivalent to

$$e_{h-d} \le e_h - d,$$

which is satisfied since we know that $e_{h+1} \ge e_h + 1$ for all $h \ge 0$. This proves (91).

Next we provide upper bounds for

$$\tilde{y}_{e_h}^{(n)}(1) = (-1)^n \alpha^{e_h(n+1)} \Phi^{(n)}(0).$$

Since

$$(-1)^n \Phi^{(n)}(0) = \int_0^\infty \Psi(y) y^n \, dy$$

and $\Psi(y) = \mathcal{O}(e^{-Cy^\gamma})$ for some $\gamma > 1$ it follows that

$$|\Phi^{(n)}(0)| \ll e^{c_1 n \log n}$$

for some constant $c_1 < 1$. Thus

$$\left| \frac{\tilde{y}_{e_h}^{(n)}(1)}{n!} \right| \ll \alpha^{e_h(n+1)} e^{-c_2 n \log n}$$

for some constant $c_2 > 0$. Consequently, for every $K > 0$ and for every $\varepsilon > 0$ there exists K_1 such that

$$\sum_{n > K_1} \left| \frac{\tilde{y}_{e_h}^{(n)}(1)}{n!} \right| \left(\frac{K}{\alpha^{e_h}} \right)^n < \varepsilon.$$

By (91) we get the same estimate if we replace $\tilde{y}_{e_h}^{(n)}(1)$ by $y_h^{(n)}(1)$.

Now, by using (89) we have uniformly for all $0 \le k \le K_1$

$$y_{h-k}(1) = \alpha^{e_{h-k}} \sim \alpha^{e_h - k} = \tilde{y}_{e_h-k}(1),$$

and hence uniformly for all $n \le K_1$

$$y_h^{(n)}(1) \sim \tilde{y}_{e_h}^{(n)}(1)$$

as $h \to \infty$.

By combining these properties with the Taylor series expansions (90) it immediately follows that $y_h(x) \sim \tilde{y}_{e_h}(x)$ (as $h \to \infty$) uniformly for complex values of x with $|x-1| \le K/\alpha^{e_h}$. □

In a second step we compare the coefficients of $y_h(x)$ and $\tilde{y}_{e_h}(x)$. A direct combination of Lemma 24 and Lemma 28 proves Proposition 1.

Lemma 28. *We have*

$$[x^n]\,\tilde{y}_{e_h}(x) \sim [x^n]\,y_h(x)$$

uniformly for $n, h \to \infty$ such that $e_h - c\log n$ belongs to a bounded set.

Proof. First of all we note that

$$[x^n]\,y_h(x) = \frac{1}{n}[x^{n-1}]\,y_{h-1}(x)^2,$$

and of course

$$[x^n]\,\tilde{y}_{e_h}(x) = \frac{1}{n}[x^{n-1}]\,\tilde{y}_{e_h-1}(x)^2.$$

Hence, by Cauchy's formula

$$\begin{aligned}
[x^n]\,y_h(x) &= \frac{1}{n}\,\frac{1}{2\pi i}\int_{|x|=1} y_{h-1}(x)^2\,dx \\
&= \frac{1}{n}\,\frac{1}{2\pi i}\left(\int\limits_{|x|=1,|x-1|\le K/n} + \int\limits_{|x|=1,|x-1|>K/n}\right) y_{h-1}(x)^2\,dx \\
&= I_1 + I_2.
\end{aligned}$$

By Lemma 25 we have

$$\begin{aligned}
|I_2| &\le \frac{1}{2\pi n}\int\limits_{|x|=1,|x-1|>K/n} \frac{4}{|1-x|^2}\,|dx| \\
&\ll \frac{1}{K}.
\end{aligned}$$

Similarly we have

$$\begin{aligned}
[x^n]\,\tilde{y}_{e_h}(x) &= \frac{1}{n}\,\frac{1}{2\pi i}\int_{|x|=1} \tilde{y}_{e_h-1}(x)^2\,dx \\
&= \frac{1}{n}\,\frac{1}{2\pi i}\left(\int\limits_{|x|=1,|x-1|\le K/n} + \int\limits_{|x|=1,|x-1|>K/n}\right) \tilde{y}_{e_h-1}(x)^2\,dx \\
&= I_1' + I_2'
\end{aligned}$$

with

$$I_2' \ll \frac{1}{K}.$$

Now Lemma 27 (resp. an analogue for $y_{h-1}(x)$ and $\tilde{y}_{e_h-1}(x) \sim \tilde{y}_{e_{h-1}}(x)$) implies

$$I_1 = I_1' + o(\alpha^{2e_h}/n^2) = I_1' + o(1)$$

uniformly for n, h in question. This completes the proof of the lemma. □

6.3. **Proof of Theorem 8.** We start with the following observation.

Lemma 29. *Let* $\Psi(y)$ *be as in Lemma 23 and set*

$$E_1(x) := \sum_{h\ge 0} \left(1-\Psi(x/\alpha^h)\right),$$

$$E_2(x) := \sum_{h\ge 0} (2h+1)\left(1-\Psi(x/\alpha^h)\right),$$

and

$$V(x) := E_2(x) - E_1(x)^2.$$

Then there exist continuous periodic functions δ_1, δ_2 *with period* 1 *such that, as* $x\to\infty$,

$$E_1(x) = c\,\log x + \delta_1(c\,\log x) + o(1)$$

and

$$V(x) = \delta_2(c\,\log x) + o(1).$$

Proof. Let $\Psi_0(y)=1$ for $0\le y<1$ and $\Psi_0(y)=0$ for $y\ge 1$. Then

$$E_1(x) = \sum_{h\ge 0}(1-\Psi_0(x/\alpha^h)) + \sum_{h\ge 0}\left(\Psi_0(x/\alpha^h)-\Psi(x/\alpha^h)\right).$$

Now observe that the function

$$G(x) := \sum_{h\in\mathbb{Z}}\left(\Psi_0(x/\alpha^h)-\Psi(x/\alpha^h)\right)$$

exists for all $x>0$ and that $G(x/\alpha)=G(x)$. Thus,

$$G(x) = \delta_0(c\log x),$$

where δ_0 is a periodic function with period 1. Furthermore, for $x>1$ we have

$$\begin{aligned}\sum_{h\ge 0}\left(\Psi_0(x/\alpha^h)-\Psi(x/\alpha^h)\right) &= G(x) - \sum_{h>0}\Psi(x\alpha^h)\\ &= G(x)+o(1).\end{aligned}$$

Finally,

$$\begin{aligned}\sum_{h\ge 0}(1-\Psi_0(x/\alpha^h)) &= \lfloor c\log x\rfloor + 1\\ &= c\log x + (1-\{c\log x\}),\end{aligned}$$

where $\{z\}=z-\lfloor z\rfloor$ denotes the fractional part of z. Thus, with

$$\delta_1(z) := \delta_0(z)+1-\{z\}$$

we obtain the proposed representation for $E_1(x)$. By definition it is also clear that $\delta_1(z)$ is continuous for non-integral z. It is also easy to check that $\lim_{z\to 0}\delta_1(z)=\delta_1(0)$, and thus, δ_1 is continuous.

In a similar way we can treat $V(x)$. Firstly, it follows by simple manipulations that

$$V(x) = \sum_{h\ge 0}\left(E_1(x/\alpha^h)+E_1(x/\alpha^{h+1})\right)\Psi(x/\alpha^h).$$

Now observe that, as $x\to 0+$, we have $E_1(x)=\mathcal{O}(x^\beta)$ and $E_2(x)=\mathcal{O}(x^\beta)$. Thus, if we define

$$H(x) := \sum_{h\in\mathbb{Z}}\left(E_1(x/\alpha^h)+E_1(x/\alpha^{h+1})\right)\Psi(x/\alpha^h),$$

then $H(x)$ converges for every $x>0$ and we have $H(x/\alpha)=H(x)$. Consequently,

$$H(x) = \delta_2(c\log x),$$

for a continuous periodic function δ_2 with period 1. Furthermore, since $E_1(x) = \mathcal{O}(\log x)$ we have, as $x \to \infty$,

$$V(x) = H(x) + o(1) = \delta_2(c \log x) + o(1)$$

as proposed. □

The next lemma is a little bit stronger (but not as beautiful) as Theorem 8.

Lemma 30. *For every $n \ge 1$ let $h = h_0(n)$ be an integer satisfying $n/2 \le \alpha^{e_h} \le n$. Then we have, as $n \to \infty$,*

$$\mathbf{E}H_n = c \log n - (e_{h_0(n)} - h_0(n)) + \delta_1\big(c \log n - (e_{h_0(n)} - h_0(n))\big) + o(1)$$

and

$$\mathbf{V}H_n = \delta_2\big(c \log n - (e_{h_0(n)} - h_0(n))\big) + o(1).$$

Proof. First of all note that $E_1(x)$ equals the expected value of a discrete random variable X with $\mathbf{P}[X \le h] = \Psi(x/\alpha^h)$ and similarly $V(x)$ equals the variance of X. Since $\Psi(y) = 1 + \mathcal{O}(y^\beta \log y)$ as $y \to 0$ and $\Psi(y) = \mathcal{O}(e^{-Cy^\gamma})$ as $y \to \infty$, there are only finitely many terms in the defining sum of $E_1(x)$ and $V(x)$ which determine the asymptotic behaviour for $x \to \infty$. More precisely, for any $\varepsilon > 0$ there exists $K > 0$ and x_0 such that

$$\left| E_1(x) - \sum_{|h - c \log x| \le K} \big(1 - \Psi(x/\alpha^h)\big) \right| < \varepsilon$$

for $x \ge x_0$ (and similarly for $E_2(x)$ and $V(x)$).

By Theorem 4 and Lemma 26 we have

$$\mathbf{P}[H_n \le h] = \frac{a_{n,h}}{n!} \sim \Psi(n/\alpha^{e_h}) \sim \Psi((n/\alpha^{e_{h_0(n)} - h_0(n)})/\alpha^h)$$

uniformly for h with $e_h - c \log h = \mathcal{O}(1)$. Furthermore, by Theorem 5 we also have (for properly adjusted K)

$$\left| \mathbf{E}H_n - \sum_{h: |e_h - c \log n| \le K} (1 - \mathbf{P}[H_n \le h]) \right| < \varepsilon.$$

Hence,

$$\mathbf{E}H_n = E_1(n/\alpha^{e_{h_0(n)} - h_0(n)}) + o(1)$$

as $n \to \infty$ and consequently

$$\mathbf{E}H_n = c \log n - (e_{h_0(n)} - h_0(n)) + \delta_1\big(c \log n - (e_{h_0(n)} - h_0(n))\big) + o(1).$$

Similarly, we get the proposed relation for $\mathbf{V}H_n$. □

It is now an easy exercise to derive Theorem 8 from Lemma 30. We only have to observe that Hypothesis 1 implies

$$e_{h_0(n)} - h_0(n) = \frac{3c}{2(c-1)} \log \log n + D_2 + o(1)$$

for some constant D_2.

Acknowledgement. The author wants to thank three anonymous referees for their careful reading and for many valuable suggestions to improve the presentation of the paper. Furthermore, he is grateful to Michael Saks for his patience and advice during the refereeing process.

References

[1] J. D. Biggins, *How fast does a general branching random walk spread ?*, in: IMA Vol. Math. Appl. **84** (1997), 19–39.

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

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

[4] L. Devroye, *On the height of random m-ary search trees*, Random Struc. Algorithms **1** (1990), 191–203.

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

[6] G. Doetsch, *Theorie und Anwendung der Laplace-Transformation*, Springer, Berlin, 1937.

[7] M. Drmota, *An Analytic Approach to the Height of Binary Search Trees*, Algorithmica **29** (2001), 89-119.

[8] M. Drmota, *The Variance of the Height of Binary Search Trees*, Theoret. Comput. Sci., to appear.

[9] W. Feller, *An Introduction to Probability Theory and Its Applications, Vol. II*, 2^{nd} ed., J. Wiley, New York, 1971.

[10] H. M. Mahmoud, *Evolution of Random Search Trees*, John Wiley & Sons, New York, 1992.

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

[12] C. Ponder, *Unsolved Problems* (R. Guy ed.), Amer. Math. Monthly **93** (1986), 280–281.

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

[14] B. Reed, *How tall is a tree*, Proceedings of STOC 2000, 479–483.

[15] B. Reed, *The height of a random binary search tree*, J. Assoc. Comput. Mach., THIS VOLUME.

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

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

[18] J. M. Robson, *Constant bounds on the moments of the height of binary search trees*, Theoret. Comput. Sci, to appear.

Appendix

Let $a_{n,h}$ denote the number of permutations $\sigma \in S_n$ of n elements such that the corresponding m-ary search tree has height $\le h$. In what follows we want to present a *purely combinatorial* proof of (54):

$$\frac{a_{n+1,h}}{(n+1)!} \le \frac{a_{n,h}}{n!} \qquad (h \ge 0, n \ge 0).$$

This proof is quite involved and – interestingly – the simplicity of the *probabilistic argument* in Lemma 16 is completely hidden here.

Proof. Before proving (54) we need an auxiliary result, namely that

$$a_{n,h+1} \ge (m-2)! \sum_{n_1+n_2+\cdots+n_{m-1}=n-m+2} \frac{(n-m+2)!}{n_1!n_2!\cdots n_{m-1}!} a_{n_1,h}a_{n_2,h}\cdots a_{n_{m-1},h} \tag{93}$$

for all $n \ge m-2$ and all $h \ge 0$. This inequality is equivalent to the inequality[5]

$$y_{h+1}^{(m-2)}(x) \ge_c (m-2)!\, y_h(x)^{m-1}. \tag{94}$$

We now prove (94) by induction on h, more precisely we show inductively that

$$y_{h+1}^{(m-2)}(x) \ge_c (m-2)!\, y_h(x)^{m-1} \quad \text{and} \quad y_h(x)^2 \ge y_h'(x). \tag{95}$$

Evidently, (95) is true for $h=0$ as can be seen from $y_1(x)^2 = 1 \ge_c 0 = y_1'(x)$ and from

$$y_1^{(m-1)}(x) = (m-2)! + (m-1)!x \ge_c (m-2)! = (m-2)!y_0(x)^{m-1}.$$

[5] Similarly to the above $A(x) \ge_c B(x)$ denotes that $[x^n]A(x) \ge [x^n]B(x)$ for all $n \ge 0$.

Now suppose that (95) is true for some $h \ge 0$. Then we get

$$y_{h+1}^{(m-2)}(x) \ge_c (m-2)!\, y_h(x)^{m-1} \ge_c (m-2)!\, y_h(x)^{m-3} y_h'(x)$$

and consequently (by integration and checking the zeroth coefficient)

$$y_{h+1}^{(m-3)}(x) \ge_c (m-3)!\, y_h(x)^{m-2}.$$

In the same way we can proceed and we finally obtain (by induction)

$$y_{h+1}^{(k)}(x) \ge_c k!\, y_h(x)^{k+1} \quad \text{for } 1 \le k \le m-2.$$

Thus, it follows that

$$\begin{aligned}\left(y_{h+1}(x)^2\right)^{(m-2)} &= \sum_{k=0}^{m-2} \binom{m-2}{k} y_{h+1}^{(k)}(x) y_{h+1}^{(m-2-k)}(x)\\ &\ge_c \sum_{k=0}^{m-2} \binom{m-2}{k} k!\, y_h(x)^{k+1} (m-2-k)!\, y_h(x)^{m-1-k}\\ &= (m-1)! y_h(x)^m = y_{h+1}^{(m-1)}(x)\end{aligned}$$

which implies (by integration, $m-2$ times)

$$y_{h+1}(x)^2 \ge_c y_{h+1}'(x). \tag{96}$$

By multiplying (96) with $(m-1)! y_{h+1}(x)^{m-2}$ we get

$$(m-1)! y_{h+1}(x)^m = y_{h+2}^{(m-1)}(x) \ge_c (m-1)! y_{h+1}(x)^{m-2} y_{h+1}'(x)$$

and (again by integration)

$$y_{h+2}^{(m-2)}(x) \ge_c (m-2)!\, y_{h+1}(x)^{m-1},$$

which implies (95) for $h+1$.

Now we are able to complete the proof of (54) by induction on n. It is clear that (54) is true for $n \le m-1$ (for all $h \ge 0$) and for $h = 0$ (for all $n \ge 0$) since we know that $a_{n,0} = \delta_{n,0}$ and

$$a_{n,h} = n! \quad \text{for } n \le m-1 \text{ and } h \ge 0.$$

So let us assume that (54) holds for some $n \ge m-1$ and all $h \ge 0$. Then we get (also by using (93))

$$\begin{aligned}\frac{a_{n+1,h+1}}{(n+1)!} &= \frac{(m-1)!}{(n+1)n\cdots(n-m+3)} \sum_{n_1+n_2+\cdots+n_m=n-m+2} \frac{a_{n_1,h}}{n_1!}\frac{a_{n_2,h}}{n_2!}\cdots\frac{a_{n_m,h}}{n_m!}\\ &= \frac{(m-1)!}{(n+1)n\cdots(n-m+3)} \sum_{n_1+n_2+\cdots+n_{m-1}=n-m+2} \frac{a_{n_1,h}}{n_1!}\frac{a_{n_2,h}}{n_2!}\cdots\frac{a_{n_{m-1},h}}{n_{m-1}!}\\ &+ \frac{(m-1)!}{(n+1)n\cdots(n-m+3)} \sum_{n_1+\cdots+n_m=n-m+2, n_m>0} \frac{a_{n_1,h}}{n_1!}\frac{a_{n_2,h}}{n_2!}\cdots\frac{a_{n_m,h}}{n_m!}\\ &\le \frac{(m-1)!}{(n+1)n\cdots(n-m+3)} \frac{n(n-1)\cdots(n-m+3)}{(m-2)!}\frac{a_{n,h+1}}{n!}\\ &+ \frac{(m-1)!}{(n+1)n\cdots(n-m+3)} \sum_{n_1+\cdots+(n_m-1)=n-m+1} \frac{a_{n_1,h}}{n_1!}\frac{a_{n_2,h}}{n_2!}\cdots\frac{a_{n_m-1,h}}{(n_m-1)!}\\ &= \frac{a_{n,h+1}}{n!}\frac{m-1}{n+1} + \frac{(m-1)!}{(n+1)n\cdots(n-m+3)} \frac{n(n-1)\cdots(n-m+2)}{(m-1)!}\frac{a_{n,h+1}}{n!}\\ &= \frac{a_{n,h+1}}{n!}\left(\frac{m-1}{n+1} + \frac{n-m+2}{n+1}\right) = \frac{a_{n,h+1}}{n!},\end{aligned}$$

which completes the proof of (54). □