ON ROBSON'S CONVERGENCE AND BOUNDEDNESS CONJECTURE CONCERING THE HEIGHT OF BINARY SEARCH TREES

MICHAEL DRMOTA

ABSTRACT. Let C_n denote the number of nodes in a random binary search tree (of n nodes) at the maximal level. In this paper we present a direct proof of Robson's *boundedness conjecture* saying that the expected values $\mathbf{E}\,C_n$ remain bounded as $n \to \infty$. We also prove that $\mathbf{E}\,C_n$ is asymptotically (multiplicatively) periodic which shows that Robson's *convergence conjecture* (that is, $\mathbf{E}\,C_n$ is convergent) is only true if the limiting periodic function $\tilde{C}(x)$ is constant. Interestingly it can be shown that $\tilde{C}(x)$ is almost constant in the sense that possible oscillations are very small. There are also strong indications that $\tilde{C}(x)$ is not constant which would imply a disproof of the convergence conjecture.

We present similar properties for the variance of the height $\mathbf{Var}\,H_n$, too.

Keywords: binary search tree, height distribution, average case analysis, generating functions

1. INTRODUCTION

A binary search tree T_n of n (internal) nodes is constructed from n distinct keys $x_1, \ldots, x_n$ in random order by inserting each key step by step. The first key x_1 is put into the root. Then the next key x_2 is put to left of the root if it is smaller than the first key and put to the right of the root if it is larger. In this way one proceeds further. If $x_1, \ldots, x_j$ are already "stored" then one goes to the left subtree of the root x_1 if x_{j+1} is smaller than x_1 and to the right subtree if it is larger. This procedure is recursively applied until one reaches an empty place where x_{j+1} is put there.

It is usually assumed that the keys $x_1 = X_1, \ldots, x_n = X_n$ are iid random variables with a (common) continuous distribution function. Equivalently one can assume that every permuation of (given distinct values) $x_1, \ldots, x_n$ are equally likely.

It is sometimes useful to consider the n internal nodes together with the (empty) $n+1$ external nodes. Namely, the above probabilistic model for binary search trees is also induced in the following way. One starts with T_0 consisting just of 1 external node. Now, suppose that T_n is given. Then T_{n+1} is generated from T_n by replacing randomly (with equal probability $1/(n+1)$) one of the $n+1$ external nodes by an internal node (together with two adjacent external ones).

Date: August 21, 2003.

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

The height H_n of T_n is then a random variable which has been considered by several authors. It is now known (see Reed [10]) that the expected value is given by

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

where $c = 4.31107\ldots$ is the largest solution of the equation $\left(\frac{2e}{c}\right)^c = e$. (Previous results concerning $\mathbf{E}H_n$ are due to Robson [11], Pittel [9], Devroye [2], Devroye and Reed [4], and Drmota [5].)

It has been also a long standing conjecture that the variance $\mathbf{Var}\, H_n = \mathbf{E}(H_n - \mathbf{E}H_n)^2$ remains bounded as $n \to \infty$. This conjecture has been proved independently by Reed [10] and by Drmota [6, 7].

Previoulsy Robson [13] could show that there exists an infinite subsequence n_k for which the variance $\mathbf{Var}\, H_{n_k}$ stays bounded. He also showed that boundedness of the variance is equivalent to the statement that the expected value of the number C_n of nodes in the highest level (that is, the number of nodes which constitute the height) is bounded. In a previous paper ([12]) Robson has stated two conjectures on the expeced value $\mathbf{E}\, C_n$. The *convergence conjecture* says that the sequence $\mathbf{E}\, C_n$ converges and the *boundedness conjecture* that the sequence $\mathbf{E}\, C_n$ is bounded. In view of [13] and the results of Reed [10] and Drmota [6, 7] the boundedness conjecture is true.

The purpose of the present paper is to discuss the expected values $\mathbf{E}\, C_n$ in more detail. First we present a direct proof of the boundedness conjecture. Second we prove 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 (11)) 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 that show that the sequence $\mathbf{E}\, C_n$ is increasing for $7 \le n \le 100\,000$. However, this is no contradiction to (expected) non-convergence since the oscillations are very small and the error term in (7) is surely relatively large for moderate n.

We will prove similar properties for the sequence of variances $\mathbf{Var}\, H_n$, too.

2. Results

We first introduce the polynomials

$$y_k(x) := \sum_{n\ge 0} \mathbf{P}[H_n \le k]\, x^n \qquad (k \ge 0). \tag{1}$$

These polynomials are recursively given by $y_0(x) \equiv 1$ and by

$$y_{k+1}(x) = 1 + \int_0^x y_k(t)^2\, dt \qquad (k \ge 0). \tag{2}$$

Alternatively we can characterize them by

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

and $y_k(0) = 1$.

The sequence $y_k(1)$ plays an important rôle in the analysis of the distribution behaviour of the height H_n (see [6]). It is rapidly growing and one has the limiting relation (see [7])

$$\lim_{k\to\infty} \frac{y_{k+1}(1)}{y_k(1)} = e^{1/c} = 1.2610\ldots. \tag{3}$$

More precisely, the sequence of ratios satisfies $y_{k+1}(1)/y_k(1) \ge e^{1/c}$ and decreases to its limit $e^{1/c}$.

Furthermore, let $\Psi(y)$, $y \ge 0$, denote the unique solution of the integral equation

$$y\Psi(y/e^{1/c}) = \int_0^y \Psi(z)\Psi(y-z)\,dz, \tag{4}$$

which is monotonically decreasing and satisfies $\Psi(0) = 1$, $\lim_{y\to\infty}\Psi(y) = 0$ and $\int_0^\infty \Psi(y)\,dy = 1$. (Existence and uniqueness of $\Psi(y)$ has been shown in [7]. One even knows that proper tail estimates, see Lemma 5)

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

$$C(x) := -\frac{1}{2}\sum_{k\ge 0} \frac{x}{y_k(1)}\Psi'\left(\frac{x}{y_k(1)}\right). \tag{5}$$

Due to proper tail estimates for $\Psi'(y)$ (see Lemma 5) it follows that $C(x)$ is a bounded function for $x > 0$. Furthermore, the limiting relation (3) implies that $C(x)$ is *almost periodic* in the sense that

$$C(e^{1/c}x) = C(x) + o(1) \qquad (x\to\infty). \tag{6}$$

With help of this function we can formulate our main result:

Theorem 1. *Let C_n denote the number of 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*

$$\mathbf{E}\,C_n = C(n) + o(1) \qquad (n\to\infty) \tag{7}$$

and it is asymptotically periodic in the sense that

$$\mathbf{E}\,C_{\lfloor e^{1/c}n\rfloor} = \mathbf{E}\,C_n + o(1) \qquad (n\to\infty). \tag{8}$$

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

$$\max_{n\ge n_0}\left|\mathbf{E}\,C_n - \frac{c}{2}\right| \le 10^{-4}. \tag{9}$$

and we have

$$\lim_{n\to\infty}\sum_{k=n}^{\lfloor e^{1/c}n\rfloor} \frac{\mathbf{E}\,C_k}{k} = \frac{1}{2}. \tag{10}$$

and

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

$$\tilde{C}(x) := -\frac{1}{2}\sum_{k=-\infty}^{\infty} xe^{-k/c}\Psi'\left(xe^{-k/c}\right) \tag{11}$$

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 $y_{h_0(x)}(1) \le x < y_{h_0(x)+1}(1)$ (compare with Lemma 6). 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.

As announced there is a similar theorem for the variance. Set

$$V(x) := \sum_{k\ge 0}(2k+1)\left(1 - \Psi\left(\frac{x}{y_k(1)}\right)\right) - \left(\sum_{k\ge 0}\left(1 - \Psi\left(\frac{x}{y_k(1)}\right)\right)\right)^2 \tag{12}$$

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:

$$V(e^{1/c}x) = V(x) + o(1) \qquad (x \to \infty). \tag{13}$$

Theorem 2. *The variance* $\mathbf{Var}\, H_n$ *remains bounded for* $n \to \infty$. *It is asymptotically given by*

$$\mathbf{Var}\, H_n = V(n) + o(1) \qquad (n \to \infty) \tag{14}$$

and it is asymptotically periodic in the sense that

$$\mathbf{Var}\, H_{\lfloor e^{1/c} n\rfloor} = \mathbf{Var}\, H_n + o(1) \qquad (n \to \infty). \tag{15}$$

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

$$\max_{n\ge n_1} |\mathbf{Var}\, H_n - v_0| \le 10^{-3}, \tag{16}$$

and we have

$$\lim_{n\to\infty} \sum_{k=n}^{\lfloor e^{1/c} n\rfloor} \frac{\mathbf{Var}\, H_k}{k} = \frac{v_0}{c}, \tag{17}$$

in which

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

and

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

3. The Boundedness Property

In this section we present a short proof of the property that $\mathbf{E}\, C_n$ remains bounded as $n \to \infty$.

Lemma 1. *We have*

$$\mathbf{E}\, C_n = \frac{n+1}{2}\left(\mathbf{E}H_{n+1} - \mathbf{E}H_n\right) \tag{18}$$

and

$$\sum_{n\ge 0} \mathbf{E}\, C_n\, x^n = \frac{1}{2(1-x)} + \frac{1}{2}\sum_{k\ge 0} y_k(x)\left(1 + (x-1)y_k(x)\right). \tag{19}$$

Remark . Note that (7) and (18) reprove that

$$\mathbf{E}\, H_n \sim c \log n.$$

Proof. Let D_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 (18).

Next we use the representation

$$\begin{aligned}\sum_{n\ge 0} \mathbf{E}H_n\, x^n &= \sum_{n\ge 0}\sum_{k\ge 0} (1 - \mathbf{P}[H_n \le k])\; x^n\\ &= \sum_{k\ge 0}\left(\frac{1}{1-x} - y_k(x)\right)\end{aligned}$$

and (18) to obtain

$$\begin{aligned}\sum_{n\ge 0} \mathbf{E}D_n\, x^n &= \sum_{n\ge 0}(n+1)\left(\mathbf{E}H_{n+1} - \mathbf{E}H_n\right)\, x^n\\ &= \left((1-x)\sum_{n\ge 0}\mathbf{E}H_n\, x^n\right)'\\ &= \sum_{k\ge 0}\left(1-(1-x)y_k(x)\right)'\\ &= 1 + \sum_{k\ge 1}\left(y_k(x) + (x-1)y_{k-1}(x)^2\right)\\ &= 1 + \sum_{k\ge 1}\left(y_k(x) - y_{k-1}(x)\right) + \sum_{k\ge 1} y_{k-1}(x)\left(1+(x-1)y_{k-1}(x)\right)\\ &= \frac{1}{1-x} + \sum_{k\ge 0} y_k(x)\left(1+(x-1)y_k(x)\right),\end{aligned}$$

which proves (19). ∎

Lemma 2. *Set $a_{n,k} := \mathbf{P}[H_n \le k]$. Then for $n \ge 1$ we have*

$$\mathbf{E}\, C_n = \frac{n+1}{2}\sum_{k\ge 0}(a_{n,k} - a_{n+1,k}) \tag{20}$$

and

$$\mathbf{E}\, C_n = \frac{1}{2} + \frac{1}{2}\sum_{k\ge 0}\sum_{m=0}^{n-1} a_{m,k}(a_{n-m-1,k} - a_{n-m,k}). \tag{21}$$

Proof. (20) follows from

$$\mathbf{E} H_n = \sum_{k\ge 0}(1-a_{n,k})$$

and from (18), and (21) is just a translation of (19). ■

In order to estimate the expected value $\mathbf{E}\, C_n$ we make use of the following tail estimates which have been (implicitly) established in [6].

Lemma 3. *Set* $h_0(n) := \max\{k \ge 0 : y_k(1) \le n\}$. *Then there exists a constant* $C > 0$ *such that*

$$\mathbf{P}[H_n \le k] \le Ce^{-(h_0(n)-k)/c} \quad \text{for } k \le h_0(n) \tag{22}$$

and

$$\mathbf{P}[H_n > k] \le Ce^{-(k-h_0(n))/c} \quad \text{for } k \ge h_0(n). \tag{23}$$

Proof. In [6] it was shown that

$$\mathbf{P}[H_n \le k] \le C\frac{y_k(1)}{n} \quad \text{for } n \ge y_k(1)$$

and

$$\mathbf{P}[H_n \le k] \le C\frac{n}{y_k(1)} \quad \text{for } n \le y_k(1).$$

Since $y_{k+1}(1) \ge e^{1/c}y_k(1)$ these inequalities immediately translate to (22) and (23). ■

We want to note that these tail estimates can easily be used to show that

$$\mathbf{E} H_n = \sum_{k\ge 0}(1-a_{n,k}) = h_0(n) + \mathcal{O}(1). \tag{24}$$

Thus, (22) and (23) directly yield exponential tails of the form

$$\mathbf{P}[|H_n - \mathbf{E}H_n| > \eta] \le C'e^{-\eta/c} \tag{25}$$

for some constanst $C' > 0$. Obviously, (25) implies boundedness of all centralized moments (such as the variance).

It is now quite easy to show that $\mathbf{E}\, C_n$ remains bounded.

Lemma 4. *We have, as* $n \to \infty$,

$$\mathbf{E}\, C_n = \mathcal{O}(1). \tag{26}$$

Proof. As above, set $a_{n,k} := \mathbf{P}[H_n \le k]$. Since $a_{0,k} = 1$ and $a_{n+1,k} \le a_{n,k}$ we have for every $L \le n$

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

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

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

First, by using the tail estimates (22) and (23) from Lemma 3 we have

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

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

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

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

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

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

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

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

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

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

Consequently

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

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

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

and

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

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

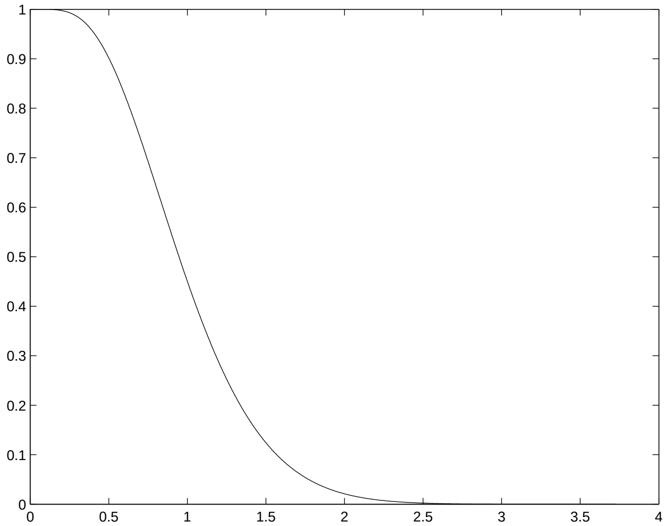

FIGURE 1. Picture of $\Psi(y)$

4. Asymptotics for the Solution of a Fixed Point Equation

In section 2 we already mentioned the fixed point equation (4) which has been discussed in [7]. In this section we also show that the derivative $\Psi'(y)$ exists and has proper tail estimates which will be used for the proof of Theorem 1.

Lemma 5. *There uniquely exists a function $\Psi(y)$, $y \ge 0$, with the following properties:*

1. $y\Psi(y/e^{1/c}) = \int_0^y \Psi(z)\Psi(y-z)\,dz$
2. $\int_0^\infty \Psi(y)\,dy = 1.$
3. $\Psi(y) - 1 \sim c_1 y^{c-1}\log y$ *as* $y \to 0+$ *for some constant* c_1.
4. *For every* $\gamma < (c\log 2)/(c\log 2 - 1)$ *there exists* $C > 0$ *and* y_0 *such that* $\Psi(y) \le e^{-Cy^\gamma}$ *for* $y \ge y_0$.
5. $\Psi(y)$, $0 \le y < \infty$, *is decreasing.*
6. $\Psi(y) = \int_0^\infty e^{-zy^{c-1}}\,dG(z)$ *for a proper distribution function* $G(z)$, $z \ge 0$.
7. $\Psi(y)$ *is continuously differentiable for* $y > 0$ *and the derivative* $\Psi'(y)$ *is bounded by*
$$0 \le -\Psi'(y) \le C_1 y^\beta e^{-Cy^\gamma}, \tag{27}$$
for some $\beta > 0$.

Proof. Existence and Uniqueness of $\Psi(y)$ (together with the stated properties 2.–5.) have been established in [7, Lemma23]. The representation 6. is proved in [1].

By property 6. it follows that the derivative $\Psi'(y)$ exists. By differentiation property 1. we obtain that $\Psi'(y)$ satisfies the functional equation

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

However, it seems to be a non-trivial problem to establish the bounds (27) directly. Therefore, we proceed indirectly.

First, we will solve the functional equation

$$\Psi(y/e^{1/c}) - \Psi(y) + y/e^{1/c} R(y/e^{1/c}) = \int_0^y \Psi(z) R(y-z)\, dz \tag{28}$$

and derive certain properties of a one-dimensional variety of continuous solutions. In a second step we will show that one of these solutions has the property that

$$1 + \int_0^y R(z)\, dz = \Psi(y).$$

Thus, $\Psi'(y)$ is continuous and equals $R(y)$.

Let $\mathcal{R}$ denote the set of continuous functions $R(y)$, $y > 0$, such that $R(y) = \mathcal{O}(y^{c-3})$ for $y > 0$ and consider the mapping

$$\mathcal{F} : \mathcal{R} \to \mathcal{R}$$

defined by

$$\mathcal{F}(R)(y) := \frac{1}{y} \int_0^{ye^{1/c}} \Psi(z) R(ye^{1/c} - z)\, dz + \frac{1}{y} \left(\Psi(ye^{1/c}) - \Psi(y) \right).$$

It is easy to establish that $\mathcal{F}$ is indeed a mapping from $\mathcal{R}$ to $\mathcal{R}$, and obviously, a fixed point of $\mathcal{F}$ is a solution of (28). Furthermore, observe that $\mathcal{R}$ adjusted with the metric

$$d(R_1, R_2) := \sup_{y>0} \left| \frac{R_1(y) - R_2(y)}{y^{c-3}} \right|$$

constitutes a complete metric space.

We will now show that $\mathcal{F}$ is a contraction on $\mathcal{R}$. Let $R_1, R_2 \in \mathcal{R}$ with $d(R_1, R_2) = \delta > 0$. Then

$$\begin{aligned}
|\mathcal{F}(R_1)(y) - \mathcal{F}(R_2)(y)| &\le \frac{1}{y} \int_0^{ye^{1/c}} \Psi(z) \left| R_1(ye^{1/c} - z) - R_2(ye^{1/c} - z) \right| dz \\
&\le \frac{1}{y} \int_0^{ye^{1/c}} |R_1(z) - R_2(z)|\, dz \\
&\le \frac{\delta}{y} \int_0^{ye^{1/c}} z^{c-1} \\
&= \delta \frac{e^{(c-2)/c}}{c-2} y^{c-3}.
\end{aligned}$$

and consequently $d(\mathcal{F}(R_1), \mathcal{F}(R_2)) \le L \cdot d(R_1, R_2)$ with $L = (e^{(c-2)/c})/(c-2) < 1$. Thus, there is a unique solution $R_0 \in \mathcal{R}$ of (28). However, at the moment it is not clear whether $R_0 = \Psi'$ or not. The reason is that R_0 is not the unique solution of (28).

We next show that the equation

$$y/e^{1/c} R(y/e^{1/c}) = \int_0^y \Psi(z) R(y-z)\, dz \tag{29}$$

has infinitely many solutions. Let $\mathcal{T}$ denote the set of non-negative continuous functions

$$R(y) = y^\beta + \mathcal{O}\left(y^2\right) \qquad (y > 0),$$

where $0 < \beta < 1$ is the solution of the equation

$$e^{(\beta+1)/c} = \beta + 1$$

and c is a fixed real number. If we adjust $\mathcal{T}$ with the metric

$$\overline{d}(R_1, R_2) := \sup_{y>0} \left| \frac{R_1(y) - R_2(y)}{y^2} \right|$$

then $\mathcal{T}$ is again a complete metric space. As above, it now follows that the mapping $\mathcal{G} : \mathcal{T} \to \mathcal{T}$, defined by

$$\mathcal{G}(R)(y) := \frac{1}{y} \int_0^{ye^{1/c}} \Psi(z) R(ye^{1/c} - z)\, dz,$$

is a contraction with Lipschitz constant $\overline{L} = e^{3/c}/3 < 1$. Thus, there is a unique fixed point $R_1 \in \mathcal{T}$ of $\mathcal{G}$. Consequently, all functions of kind

$$R(y) = R_0(y) + \lambda R_1(y) \qquad (\lambda \in \mathbb{R})$$

are solutions of (28).

Our next aim is to show that the Laplace transforms of R_0 and R_1 exist and constitute entire functions. For this purpose it suffices to show that $R_0(y)$ and $R_1(y)$ decrease to 0 (as $y \to \infty$) faster than exponentially. We fix some α with $1 < \alpha < e^{1/c}$. Then $\gamma := (\log 2)/(\log 2 - \log \alpha)$ satisfies $1 < \gamma < (c \log 2)/(c \log 2 - 1)$. Thus, we know that for some constant $C > 0$ and $y \geq y_0$

$$\Psi(y) \leq e^{-Cy^\gamma}.$$

We first show that there is another constant $C_1 > 0$ such that for all $y \geq 0$

$$0 \leq -R_0(y) \leq C_1 y e^{-Cy^\gamma}. \tag{30}$$

We set

$$R^{(0)}(y) := \begin{cases} -y^{c-1} & \text{for } 0 \leq y \leq 1 \\ -e^{C - Cy^\gamma} & \text{for } y > 1, \end{cases}$$

and inductively $R^{(i+1)} = \mathcal{F}(R^{(i)})$. Since $\mathcal{F}$ is a contraction it follows that $\lim_{i\to\infty} R^{(i)} = R_0$ and that there is uniform constant C_2 such that $0 \leq -R^{(i)}(y) \leq C_2 y^{c-3}$ for all $i \geq 0$ and $y > 0$. Thus there exists $y_1 \geq y_0/(e^{1/c} - 1)$ and a constant $C_1 > 0$ such that the function $y \mapsto ye^{-Cy^\gamma}$ is decreasing for $y \geq y_1$, that

$$0 \leq -R^{(i)}(y) \leq C_1 y e^{-Cy^\gamma}$$

for $0 \leq y \leq y_1$, that

$$0 \leq -R^{(0)}(y) \leq C_1 y e^{-Cy^\gamma}$$

(even) for $y \geq y_1$, and that

$$\frac{1}{y}\left(\Psi(y) - \Psi(e^{1/c} y)\right) \leq \left(1 - \frac{e^{2/c}}{2}\right) C_1 y e^{-Cy^\gamma}.$$

Now we proceed by induction and suppose that we already know

$$0 \leq -R^{(i)}(y) \leq C_1 y e^{-Cy^\gamma}$$

for all $y \geq 0$. It is sufficient to consider the case $y \geq y_1$. If $0 \leq z \leq y_0$ we have (since $y_1 \geq y_0/(e^{1/c}-1)$)

$$\begin{aligned} -\Psi(z)R^{(i)}(e^{1/c}y - z) &\leq C_1(e^{1/c}y - z)e^{-C(e^{1/c}y-z)^\gamma} \\ &\leq C_1(e^{1/c}y - z)e^{-Cy^\gamma} \end{aligned}$$

If $y_0 \leq z \leq e^{1/c}y$ we also get

$$\begin{aligned} -\Psi(z)R^{(i)}(e^{1/c}y - z) &\leq C_1(e^{1/c}y - z)e^{-Cz^\gamma - C(e^{1/c}y-z)^\gamma} \\ &\leq C_1(e^{1/c}y - z)e^{-2C(e^{1/c}y/2)^\gamma} \\ &= C_1(e^{1/c}y - z)e^{-C(e^{1/c}/\alpha)y^\gamma} \\ &\leq C_1(e^{1/c}y - z)e^{-Cy^\gamma}. \end{aligned}$$

Thus, in all cases we obtain for $y \geq y_1$

$$\begin{aligned} -\frac{1}{y}\int_0^{e^{1/c}y} \Psi(z)R^{(i)}(e^{1/c}y - z)\,dz &\leq C_1 e^{-Cy^\gamma}\frac{1}{y}\int_0^{e^{1/c}y}(e^{1/c}y - z)\,dz \\ &= \frac{e^{2/c}}{2}C_1 y e^{-Cy^\gamma} \end{aligned}$$

and consequently (for $y \geq y_1$)

$$\begin{aligned} -R^{(i+1)}(y) &= -\frac{1}{y}\int_0^{e^{1/c}y} \Psi(z)R^{(i)}(e^{1/c}y - z)\,dz \\ &\quad + \frac{1}{y}\Big(\Psi(y) - \Psi(e^{1/c}y)\Big) \\ &\leq C_1 y e^{-Cy^\gamma}. \end{aligned}$$

Of course, this also proves (30).

For R_1 we use a similar approach. We define

$$\overline{R}^{(0)}(y) := \begin{cases} y^\beta & \text{for } 0 \leq y \leq 1 \\ -e^{C-Cy^\gamma} & \text{for } y > 1, \end{cases}$$

and inductively $\overline{R}^{(i+1)} = \mathcal{G}(\overline{R}^{(i)})$. Again the goal is to prove an inequality of the kind

$$0 \leq \overline{R}^{(i)}(y) \leq C_3 y^\beta e^{-Cy^\gamma}$$

for all $i \geq 0$ and $y \geq 0$. We do not work out all the details. We just mention the *crucial* relation

$$\frac{1}{y}\int_0^{e^{1/c}y}(e^{1/c}y - z)^\beta\,dz = \frac{e^{(\beta+1)/c}}{\beta+1}y^\beta = y^\beta.$$

Thus, we also have

$$0 \leq R_1(y) \leq C_3 y^\beta e^{-Cy^\gamma}. \tag{31}$$

Now, let

$$S_0(u) := \int_0^\infty R_0(y)e^{-yu}\,dy$$

and

$$S_1(u) := \int_0^\infty R_1(y)e^{-yu}\,dy$$

denote the Laplace transforms of R_0 and R_1. Since $S_1(0) > 0$ there exists λ_0 such that

$$S_0(0) + \lambda_0 S_1(0) = -1.$$

The major step of the proof of Lemma 5 is now to show that $R(y) := R_0(y) + \lambda_0 R_1(y)$ is exactly the derivative of $\Psi(y)$.

Let

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

the Laplace transform of Ψ which satisfies the differential equation

$$-e^{2/c}\Phi'(e^{1/c}u) = \Phi(u)^2 \tag{32}$$

with initial condition $\Phi(0) = 1$. It is easy to show that (32) has a unique entire solution. (Note that $\Phi(u)$ is surely an entire function because of the tail estimates of $\Psi(y)$.) One just has to observe that the coefficients of the Taylor series $\Phi(u) = \sum_{k\ge 0} c_k u^k$ satisfy the recurrence

$$c_{k+1} = -e^{-(k+2)/c}\sum_{\ell=0}^{k} c_\ell c_{k-\ell}.$$

Thus, they are uniquely determined by $c_0 = \Phi(0) = 1$.

Similarly if we assume that $R(y)$ is a solution of (28) for which the Laplace transform $S(u)$ is analytic. It then follows from

$$e^{1/c}\Phi(e^{1/c}u) - \Phi(u) - e^{2/c}S'(e^{1/c}u) = \Phi(u)S(u) \tag{33}$$

that the Taylor coefficients of $S(u) = \sum_{k\ge 0} d_k u^k$ satisfy the recurrence

$$d_{k+1} = c_k - e^{-(k+1)/c}c_k - e^{-(k+1)/c}\sum_{\ell=0}^{k} d_\ell c_{k-\ell}.$$

Consequently, they are (again) uniquely determined by $d_0 = S(0)$.

Furthermore, the entire function

$$S(u) = u\Phi(u) - 1$$

satisfies (33) and has (initial value) $S(0) = -1$. Thus,

$$u\Phi(u) - 1 = S_0(u) + \lambda_0 S_1(u)$$

and consequently $R(y) = R_0(y) + \lambda_0 R_1(y)$ satisfies

$$1 + \int_0^y R(z)\,dz = \Phi(y).$$

This shows that $\Phi(y)$ is continuously differentiable and $\Phi'(y) = R(y)$ has the proposed properties. ∎

5. Almost Constancy Phenomena

We will now have a more precise look at the functions $C(x)$ and $V(x)$. As already indicated they can be approximated with help of the following functions:

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

(which was alredy defined in (11)) and

$$\tilde{V}(x) := \sum_{k=-\infty}^{\infty} \left(E(xe^{-k/c}) + E(xe^{-(k+1)/c})\right)\Psi(xe^{-k/c}),$$

where

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

Lemma 6. *The functions $\tilde{C}(x)$ and $\tilde{V}(x)$ are bounded for $x>0$ and multiplicatively periodic:*

$$\tilde{C}(e^{1/c}x) = \tilde{C}(x), \quad \tilde{V}(e^{1/c}x) = \tilde{V}(x).$$

Furthermore, let $h_0(x)$, $x>0$, be uniquely defined by

$$y_{h_0(x)}(1) \leq x < y_{h_0(x)+1}(1).$$

Then we have, as $x\to\infty$

$$C(x) = \tilde{C}\left(\frac{x}{y_{h_0(x)}}\right) + o(1) \tag{34}$$

and

$$V(x) = \tilde{V}\left(\frac{x}{y_{h_0(x)}}\right) + o(1). \tag{35}$$

Proof. First of all, the tail estimates for $\Psi'(y)$ of Lemma 5 show that $\tilde{C}(x)$ is a bounded function, and by definition we have $\tilde{C}(e^{1/c}x) = \tilde{C}(x)$.

Next we show that

$$C(x) = -\frac{1}{2}\sum_{\ell\geq -h_0(n)} \frac{x}{y_{h_0(n)+\ell}(1)}\Psi'\left(\frac{x}{y_{h_0(n)+\ell}(1)}\right)$$

is close to

$$\tilde{C}\left(\frac{x}{y_{h_0(x)}}\right) = -\frac{1}{2}\sum_{\ell=-\infty}^{\infty} \frac{x}{y_{h_0(n)}e^{\ell/c}}\Psi'\left(\frac{x}{y_{h_0(n)}e^{\ell/c}}\right).$$

Note that $y_{h_0(n)+\ell}(1) \geq y_{h_0(n)}(1)e^{\ell/c}$ for $\ell\geq 0$ (and $y_{h_0(n)+\ell}(1) \leq y_{h_0(n)}(1)e^{\ell/c}$ for $\ell \leq 0$). Thus

$$\begin{aligned}\frac{x}{y_{h_0(n)+\ell}(1)} &\leq \frac{x}{y_{h_0(n)}(1)e^{\ell/c}}\\ &\leq \frac{y_{h_0(n)+1}}{y_{h_0(n)}}e^{-\ell/c}\\ &\leq Ce^{-\ell/c}\end{aligned}$$

for $\ell \ge 0$ and an absolute constant $C > 0$. Similarly we have

$$\frac{x}{y_{h_0(n)+\ell}(1)} \ge C' e^{-\ell/c}$$

for $\ell \le 0$.

Now, fix some $\varepsilon > 0$. Due to the tail estimates of $\Psi'(y)$ from Lemma 5 (and the above considerations) there exist $L = L(\varepsilon) > 0$ such that for all $x \ge 1$

$$\left| \sum_{|\ell|>L} \frac{x}{y_{h_0(n)+\ell}(1)} \Psi'\left(\frac{x}{y_{h_0(n)+\ell}(1)}\right) \right| < \varepsilon$$

and

$$\left| \sum_{|\ell|>L} \frac{x}{y_{h_0(n)} e^{\ell/c}} \Psi'\left(\frac{x}{y_{h_0(n)} e^{\ell/c}}\right) \right| < \varepsilon$$

Furthermore, for $|\ell| \le L$ we have

$$c_1 \le \frac{x}{y_{h_0(n)+\ell}(1)} \le c_2$$

and

$$c_1 \le \frac{x}{y_{h_0(n)} e^{\ell/c}} \le c_2$$

for certain constants $c_1, c_2 > 0$ (depending on ε).

Hence, by applying the limiting relation (3) we obtain for every ℓ with $|\ell| \le L$

$$\lim_{x\to\infty} \left(\frac{x}{y_{h_0(n)+\ell}(1)} \Psi'\left(\frac{x}{y_{h_0(n)+\ell}(1)}\right) - \frac{x}{y_{h_0(n)} e^{\ell/c}} \Psi'\left(\frac{x}{y_{h_0(n)} e^{\ell/c}}\right) \right) = 0.$$

Consequently,

$$\limsup_{x\to\infty} \left| C(x) - \tilde{C}\left(\frac{x}{y_{h_0(x)}}\right) \right| \le \varepsilon.$$

Since $\varepsilon > 0$ was arbitrary, we have thus proved that, as $x \to \infty$,

$$C(x) = \tilde{C}\left(\frac{x}{y_{h_0(x)}}\right) + o(1).$$

This completes the proof of the properties of $\tilde{C}(x)$ in Lemma 6.

The proof of the corresponding properties of $\tilde{V}(x)$ is similar, however, it is convenient to introduce another auxiliary function:

$$\overline{V}(x) := \sum_{k\ge 0} (2k+1)\left(1 - \Psi(xe^{-k/c})\right) - \left(\sum_{k\ge 0}\left(1 - \Psi(xe^{-k/c})\right)\right)^2.$$

As above it follows that (compare also with [7])

$$V(x) = \overline{V}\left(\frac{n}{y_{h_0(x)}} e^{h_0(x)/c}\right) + o(1).$$

Next, it is another easy exercise to derive an alternate representation for $\overline{V}(x)$:

$$\overline{V}(x) = \sum_{k\ge 0} \left(E(xe^{-k/c}) + E(xe^{-(k+1)/c}\right) \Psi(xe^{-k/c}).$$

Since $E(x) = \mathcal{O}\left(x^{c-1}\right)$ as $x \to 0+$ it finally follows that

$$\tilde{V}(x) = \overline{V}(x) + o(1)$$

as $x \to \infty$. Thus

$$\begin{aligned} V(x) &= \overline{V}\left(\frac{n}{y_{h_0(x)}} e^{h_0(x)/c}\right) + o(1) \\ &= \tilde{V}\left(\frac{n}{y_{h_0(x)}}\right) + o(1) \end{aligned}$$

and the proof of Lemma 6 is completed. ■

Since $\tilde{C}(x)$ and $\tilde{V}(x)$ are periodic in the sense that $\tilde{C}(e^{1/c}x) = \tilde{C}(x)$ and $\tilde{V}(e^{1/c}x) = \tilde{V}(x)$, we get another verification of the oscillation properties of $\mathbf{E}\, C_n$ and $\mathbf{Var}\, H_n$. Furthermore, Lemma 6 also shows that $\lim_{x\to\infty} C(x)$ exists if and only if $\tilde{C}(x)$ is constant (and similary for $V(x)$).

We will next show that the functions $\tilde{C}(x)$ and $\tilde{V}(x)$ are (at least) almost constant.

Lemma 7. *We have*

$$\max_x \left| \tilde{C}(x) - \frac{c}{2} \right| \le 10^{-4}$$

and

$$\max_x \left| \tilde{V}(x) - v_0 \right| \le 10^{-3},$$

where

$$v_0 = c \int_0^\infty (E(u) + E(ue^{-1/c}))\Psi(u) \frac{du}{u} = 2.085\ldots$$

Proof. By definition the function $C_1(x) = \tilde{C}(x)$ is periodic with period $1/c$ and the (complex) Fourier coefficients are given by

$$\begin{aligned} c_h &= c \int_0^{1/c} C_1(x) e^{-2\pi i c h x}\, dx \\ &= -\frac{c}{2} \int_{-\infty}^{\infty} e^x \Psi'(e^x) e^{-2\pi i c h x}\, dx \\ &= -\frac{c}{2} \int_0^{\infty} \Psi'(y) e^{-2\pi i c h \log y}\, dy \\ &= -\frac{c}{2} F_1(-2\pi i c h), \end{aligned}$$

where

$$F_1(t) = \int_{-\infty}^{\infty} e^x \Psi'(e^x) e^{-itx}\, dx$$

denotes the Fourier transform of $e^x\Psi'(e^x)$. Figure 2 shows $|F_1(t)|$ for $0 \le t \le 30$. This picture indicates that the Fouriercoefficients c_h are very small. In fact, by numerical calculations we (surely) have $|c_1| \le 10^{-5}$ and $|c_2| \le 3 \cdot 10^{-5}$. We now

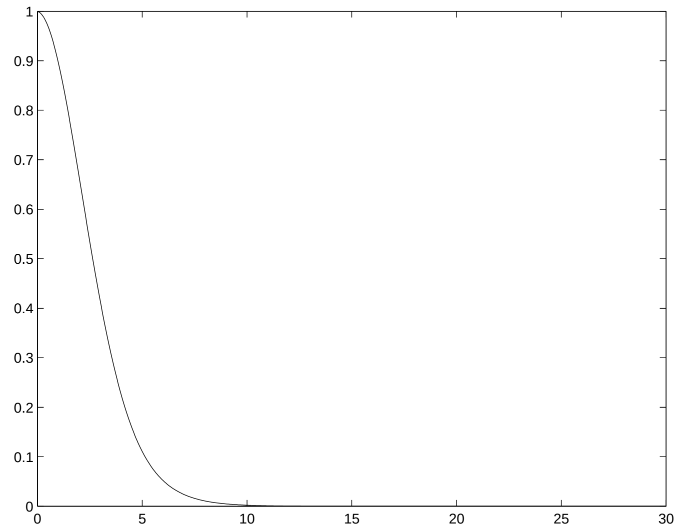

FIGURE 2. Picture of $|F_1(t)|$

give a theoretical justification for the property that $F_1(t)$ is small. By using property 6. of Lemma 5 we directly get

$$\begin{aligned} F_1(t) &= -\int_0^\infty \int_0^\infty (c-1)ze^{(c-1)x}e^{-ze^{(c-1)x}}e^{-itx}\,dx\,dG(z) \\ &= -\int_0^\infty z^{it/(c-1)}\int_0^\infty e^{-v}v^{-it/(c-1)}\,dv\,dG(z) \\ &= -\Gamma\left(1-\frac{it}{c-1}\right)\int_0^\infty z^{it/(c-1)}\,dG(z). \end{aligned}$$

By Stirling's formula we have (for real s)

$$\Gamma(1+is) \sim \sqrt{2\pi s}\,e^{-\frac{\pi}{4}s}.$$

Consequently it follows that

$$2\sum_{h=3}^{\infty}|c_h| \le c\sum_{h=2}^{\infty}|F_1(-2\pi ich)| \le 10^{-6}.$$

Hence, the maximal deviation of $\tilde{C}(x)$ from $c_0 = c/2$ is bounded by 10^{-4}.

A similar procedure works for $\tilde{V}(x)$. Here we have to consider the Fouriertransform

$$F_2(t) = \int_{-\infty}^{\infty}(E(e^x)+E(e^{x-1/c}))\Psi(e^x)e^{-itx}\,dx.$$

The Fouriercoefficients of $\tilde{V}(e^x)$ are then given by $d_h = cF_2(2\pi ch)$. With help of numerical calculations it follows that $|d_1| \le 3\cdot 10^{-5}$ and $|d_2| \le 5\cdot 10^{-5}$

As above it follows that

$$F_2(t) = \frac{\Gamma\left(-\frac{it}{c-1}\right)}{c-1}\int_0^\infty\int_0^\infty \sum_{k\ge 0}(2-\delta_{k,0})\left(w^{\frac{it}{c-1}}-(w+z(2/c)^k)^{\frac{it}{c-1}}\right)dG(z)\,dG(w).$$

Thus, we can again estimate the deviation of $\tilde{V}(x)$ from $v_0 = d_0 = cF_2(0) = 2.085\ldots$ and obtain (after proper numerical calculations) a (crude) upper bound 10^{-3}. ■

Note that Lemma 7 provides just upper bounds for the deviation from the mean. By calculating $\tilde{C}(x)$ and $\tilde{V}(x)$ directly one observes that these bounds are surely far away from beeing optimal. And these calculations cannot decide, either, whether $\tilde{C}(x)$ or $\tilde{V}(x)$ are constant or not. The accuracy of the numerical calculations (the author uses) is not sufficient to answer this question. The problem is that the calculations for the tail of $\Psi(y)$ are very sensitive. And the tail is, of course, important for the order of magnitude of $F_1(t)$ and $F_2(t)$. Nevertheless, one gets the impression that $F_{1,2}(t)$ are non-zero for all t which would implies that $c_h \neq 0$ and $d_h \neq 0$ for all integers h and consequently $\tilde{C}(x)$ and $\tilde{V}(x)$ are not constant.

6. Proof of Theorem 1

The unique solution $\Psi(y)$ of the fixed point equation (4) (compare with Lemma 5) is also very important for the distribution of the height H_n. The following theorem is one of the main results of [7].

Theorem 3. *Let $\Psi(y)$ be the unique solution of the fixed point equation (4) (with side conditions $\Psi(0) = 1$, $\lim_{y\to\infty} \Psi(y) = 0$, and $\int_0^\infty \Psi(y)\,dy = 1$). Then, as $n \to \infty$,*

$$\mathbf{P}[H_n \le k] = \Psi\left(\frac{n}{y_k(1)}\right) + o(1), \tag{36}$$

where the error term is uniform for all $k \ge 0$.

In view of (18) and (20) this suggests that

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

Whereas the second approximation step is easy to verify, the first one cannot be directly checked. Therefore we will use (21) instead in order to prove the above approximation $\mathbf{E}\, C_n = C(n) + o(1)$ rigorously.

Proof. (Theorem 1) For convenience, set

$$A_{n,k} := \sum_{m=0}^{n-1} a_{m,k}(a_{n-m-1,k} - a_{n-m,k}). \tag{37}$$

In the proof of Lemma 4 we have (implicitly) proved that

$$A_{n,k} = \mathcal{O}\left(e^{-(h_0(n)-k)/c}\right) \quad \text{for } k \le h_0(n)$$

and

$$A_{n,k} = \mathcal{O}\left(e^{-(k-h_0(n))/c}\right) \quad \text{for } k \ge h_0(n).$$

Thus, for any given $\varepsilon > 0$ there exists $L = L(\varepsilon)$ such that for all $n \ge 1$

$$\sum_{|k-h_0(n)|\ge L} A_{n,k} \le \varepsilon.$$

Note that

$$\text{Ł} = \mathcal{O}\left(\log \frac{1}{\varepsilon}\right).$$

Furthermore, there exist constants $c_1 = c_1(\varepsilon), c_2 = c_2(\varepsilon) > 0$ such that

$$c_1 \le \frac{n}{y_k(1)} \le c_2$$

for all n, k with $|k - h_0(n)| \le L$.

The next step is to show that for k with $|k - h_0(n)| \le L$ we have uniformly, as $n \to \infty$,

$$A_{n,k} = -\int_0^{n/y_k(1)} \Psi(z)\Psi'\left(\frac{n}{y_k(1)} - z\right) dz + o(1). \tag{38}$$

First of all, we have for all $\ell \ge 1$

$$A_{n,k} \le \sum_{j=0}^{\lfloor n/\ell \rfloor - 1} a_{j\ell,k}\left(a_{n-(j+1)\ell,k} - a_{n-j\ell,k}\right) + a_{\lfloor n/\ell \rfloor \ell,k}\left(1 - a_{n-\lfloor n/\ell \rfloor \ell,k}\right) \tag{39}$$

and similarly

$$A_{n,k} \ge \sum_{j=0}^{\lfloor n/\ell \rfloor - 1} a_{(j+1)\ell,k}\left(a_{n-(j+1)\ell,k} - a_{n-j\ell,k}\right) + a_{n,k}\left(1 - a_{n-\lfloor n/\ell \rfloor \ell,k}\right) \tag{40}$$

Since both bounds are of almost the same shape we just consider the first one. First of all we replace $a_{n,k}$ by $\Psi(n/y_k(1)) + o(1)$: and suppose that $l = \lfloor n\varepsilon \rfloor$:

$$
\begin{aligned}
&\sum_{j=0}^{\lfloor n/\ell\rfloor-1} a_{j\ell,k}\left(a_{n-(j+1)\ell,k} - a_{n-j\ell,k}\right) + a_{\lfloor n/\ell\rfloor\ell,k}\left(1 - a_{n-\lfloor n/\ell\rfloor\ell,k}\right)\\
&= \sum_{j=0}^{\lfloor n/\ell\rfloor-1}\left(\Psi\left(\frac{j\ell}{y_k(1)}\right) + o(1)\right)\left(a_{n-(j+1)\ell,k} - a_{n-j\ell,k}\right)\\
&+ \left(\Psi\left(\frac{\lfloor n/\ell\rfloor\ell}{y_k(1)}\right) + o(1)\right)\left(1 - a_{n-\lfloor n/\ell\rfloor\ell,k}\right)\\
&= \sum_{j=0}^{\lfloor n/\ell\rfloor-1}\Psi\left(\frac{j\ell}{y_k(1)}\right)\left(a_{n-(j+1)\ell,k} - a_{n-j\ell,k}\right) + \Psi\left(\frac{j\ell}{y_k(1)}\right)\left(1 - a_{n-\lfloor n/\ell\rfloor\ell,k}\right) + o\left(1\right)\\
&= \sum_{j=0}^{\lfloor n/\ell\rfloor-1}\Psi\left(\frac{j\ell}{y_k(1)}\right)\left(\Psi\left(\frac{n-(j+1)\ell}{y_k(1)}\right) - \Psi\left(\frac{n-j\ell}{y_k(1)}\right) + o(1)\right)\\
&+ \Psi\left(\frac{j\ell}{y_k(1)}\right)\left(1 - \Psi\left(\frac{n-\lfloor n/\ell\rfloor\ell}{y_k(1)}\right) + o(1)\right) + o\left(1\right)\\
&= \sum_{j=0}^{\lfloor n/\ell\rfloor-1}\Psi\left(\frac{j\ell}{y_k(1)}\right)\left(\Psi\left(\frac{n-(j+1)\ell}{y_k(1)}\right) - \Psi\left(\frac{n-j\ell}{y_k(1)}\right)\right)\\
&+ \Psi\left(\frac{j\ell}{y_k(1)}\right)\left(1 - \Psi\left(\frac{n-\lfloor n/\ell\rfloor\ell}{y_k(1)}\right)\right) + o\left(1\right) + o\left(\frac{n}{\ell}\right)
\end{aligned}
$$

By the mean value theorem we know that

$$
\begin{aligned}
&-\int_{j\ell}^{(j+1)\ell}\Psi\left(\frac{z}{y_k(1)}\right)\Psi'\left(\frac{n-z}{y_k(1)}\right)dz\\
&\qquad = y_k(1)\Psi\left(\frac{\zeta}{y_k(1)}\right)\left(\Psi\left(\frac{n-(j+1)\ell}{y_k(1)}\right) - \Psi\left(\frac{n-j\ell}{y_k(1)}\right)\right)
\end{aligned}
$$

for some $\zeta \in [j\ell, (j+1)\ell]$. Consequently

$$
\begin{aligned}
&\Psi\left(\frac{j\ell}{y_k(1)}\right)\left(\Psi\left(\frac{n-(j+1)\ell}{y_k(1)}\right) - \Psi\left(\frac{n-j\ell}{y_k(1)}\right)\right)\\
&\qquad = -\frac{1}{y_k(1)}\int_{j\ell}^{(j+1)\ell}\Psi\left(\frac{z}{y_k(1)}\right)\Psi'\left(\frac{n-z}{y_k(1)}\right)dz\\
&\qquad + \mathcal{O}\left(\frac{\ell}{y_k(1)}\left(\Psi\left(\frac{n-(j+1)\ell}{y_k(1)}\right) - \Psi\left(\frac{n-j\ell}{y_k(1)}\right)\right)\right).
\end{aligned}
$$

Now we suppose that $l = \lfloor y_k(1)\varepsilon \rfloor$ which gives

$$\begin{aligned} A_{n,k} &\le -\frac{1}{y_k(1)}\int_0^n \Psi\left(\frac{z}{y_k(1)}\right)\Psi'\left(\frac{n-z}{y_k(1)}\right)\,dz \\ &\quad + o(1) + o\left(\frac{n}{\ell}\right) + \mathcal{O}\left(\frac{\ell}{y_k(1)}\right) \\ &= -\int_0^{n/y_k(1)} \Psi(z)\,\Psi'\left(\frac{n}{y_k(1)} - z\right)\,dz \\ &\quad + o(1) + o\left(\frac{n}{y_k(1)}\varepsilon\right) + \mathcal{O}(\varepsilon)\,. \end{aligned}$$

As already mentioned, we obtain a lower bound of the same kind by starting with (40) instead of (39). Thus,

$$\begin{aligned} A_{n,k} = &-\int_0^{n/y_k(1)} \Psi(z)\,\Psi'\left(\frac{n}{y_k(1)} - z\right)\,dz \\ &+ o(1) + o\left(\frac{n}{y_k(1)}\varepsilon\right) + \mathcal{O}(\varepsilon)\,. \end{aligned}$$

Summing over all k and using the fact that

$$-\int_0^{n/y_k(1)} \Psi(z)\,\Psi'\left(\frac{n}{y_k(1)} - z\right)\,dz = \mathcal{O}\left(e^{-(h_0(n)-k)/c}\right) \quad \text{for } k \le h_0(n)$$

and

$$-\int_0^{n/y_k(1)} \Psi(z)\,\Psi'\left(\frac{n}{y_k(1)} - z\right)\,dz = \mathcal{O}\left(e^{-(k-h_0(n))/c}\right) \quad \text{for } k \ge h_0(n).$$

(which follows as above by using the tail estimates for $\Psi(y)$ and $\Psi'(y)$ from Lemma 5 instead of (22) and (23) from Lemma 3) we end up with

$$\begin{aligned} \sum_{k\ge 0} A_{n,k} = &-\sum_{k\ge 0}\int_0^{n/y_k(1)} \Psi(z)\,\Psi'\left(\frac{n}{y_k(1)} - z\right)\,dz \\ &+ o(L) + o(Lc_2\varepsilon) + \mathcal{O}\left(\varepsilon\log\frac{1}{\varepsilon}\right), \end{aligned}$$

where the o-terms tend to 0 as $n \to \infty$. Since $\varepsilon > 0$ was arbitrary this also shows that

$$\sum_{k\ge 0} A_{n,k} = -\sum_{k\ge 0}\int_0^{n/y_k(1)} \Psi(z)\,\Psi'\left(\frac{n}{y_k(1)} - z\right)\,dz + o(1)$$

as $n \to \infty$.

Now, since

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

we thus get

$$\sum_{k\ge 0} A_{n,k} = \sum_{k\ge 0}\left(\Psi\left(\frac{n}{y_k(1)}\right)\right) - \Psi\left(\frac{n}{y_k(1)e^{1/c}}\right) - \sum_{k\ge 0}\frac{n}{y_k(1)e^{1/c}}\Psi'\left(\frac{n}{y_k(1)e^{1/c}}\right).$$

Hence, in order to complete the proof of Theorem 1 (and in view of (21) and (37)) we just have to show that

$$\sum_{k\ge 0}\left(\Psi\left(\frac{n}{y_k(1)}\right)-\Psi\left(\frac{n}{y_k(1)e^{1/c}}\right)\right)=-1+o(1) \tag{41}$$

and

$$\sum_{k\ge 0}\frac{n}{y_k(1)e^{1/c}}\Psi'\left(\frac{n}{y_k(1)e^{1/c}}\right)=\sum_{k\ge 0}\frac{n}{y_k(1)}\Psi'\left(\frac{n}{y_k(1)}\right)+o(1) \tag{42}$$

as $n\to\infty$.

The idea of the proof of (41) is to approximate $y_k(1)e^{1/c}$ by $y_{k+1}(1)$. Let $\varepsilon>0$ be given. By another use of the tail estimates for $\Psi(y)$ of Lemma 5 it follows that there exists L such that

$$\sum_{|k-h_0(n)|\ge L}\left(\Psi\left(\frac{n}{y_k(1)}\right)\right)-\Psi\left(\frac{n}{y_k(1)e^{1/c}}\right)\le\varepsilon.$$

Furthermore, we have for k with $|k-h_0(n)|\le L$ (with some constants $c_1,c_2>0$)

$$c_1\le\frac{n}{y_k(1)}\le c_2.$$

with some constants $c_1,c_2>0$ (depending on ε). Hence, by using the limiting relation (3) it follows that

$$\begin{aligned}\sum_{|k-h_0(n)|\le L}&\left|\Psi\left(\frac{n}{y_k(1)e^{1/c}}\right)\right)-\Psi\left(\frac{n}{y_{k+1}(1)}\right|\\ &=\mathcal{O}\left(\sum_{|k-h_0(n)|\le L}\frac{n}{y_k(1)e^{1/c}}\left|1-\frac{y_k(1)e^{1/c}}{y_{k+1}(1)}\right|\right)\\ &=o(1)\end{aligned}$$

as $n\to\infty$. Consequently (by another use of the tail estimates of $\Psi(y)$)

$$\begin{aligned}\sum_{k\ge 0}\left(\Psi\left(\frac{n}{y_k(1)}\right)-\Psi\left(\frac{n}{y_k(1)e^{1/c}}\right)\right)&=\sum_{k\ge 0}\left(\Psi\left(\frac{n}{y_k(1)}\right)-\Psi\left(\frac{n}{y_{k+1}(1)}\right)\right)\\ &\quad+\sum_{k\ge 0}\left(\Psi\left(\frac{n}{y_{k+1}(1)}\right)-\Psi\left(\frac{n}{y_k(1)e^{1/c}}\right)\right)\\ &=-1+o(1)+\mathcal{O}(\varepsilon).\end{aligned}$$

Since $\varepsilon>0$ can be chosen arbitrarily small, (41) follows.

The proof of (42) is quite similar. We only have to use corresponding tail estimates for $\Psi'(y)$ and the property that $\Psi''(y)$ is bounded. This completes the proof of Theorem 1. ■

7. Proof of Theorem 2

By definition we have

$$\mathbf{Var}\,H_n=\sum_{k\ge 0}(2k+1)(1-a_{n,k})-\left(\sum_{k\ge 0}(1-a_{n,k})\right)^2,$$

where (as above) $a_{n,k} = \mathbf{P}[H_n \le k]$. Thus, with $V(x)$ from (5) we get

$$\mathbf{Var}\, H_n = V(n) + o(1).$$

We just have to proceed as in the proof of Theorem 1. (We apply the approximation of Theorem 3 for those k which are close to $h_0(n)$ and estimate the remaining ones with help of the tail estimates of Lemma 3 and Lemma 5.)

References

[1] B. Chauvin and M. Drmota, *The random bisection problem, travellling waves, and the distribution of the height of binary search trees*, manuscript.

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

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

[6] M. Drmota, *The Variance of the Height of Binary Search Trees*, Theoret. Comput. Sci. **270** (2002), 913–919.

[7] M. Drmota, *An Analytic Approach to the Height of Binary Search Trees. II*, J. Assoc. Comput. Mach. **50** (2003), 333–374.

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

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

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

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

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

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