Dedicated to Philippe Flajolet on the occasion of his 50^{th} birthday.

AN ANALYTIC APPROACH TO THE HEIGHT OF BINARY SEARCH TREES

MICHAEL DRMOTA

ABSTRACT. By using analytic tools it is shown that the expected value of the height H_n of binary search trees of size n is asymptotically given by $\mathbf{E}H_n = c\log n + \mathcal{O}(\log\log n)$ and its variance by $\mathbf{V}H_n = \mathcal{O}((\log\log n)^2)$, where $c = 4.31107\ldots$. The same bounds have been obtained by Devroye and Reed [3] by completely different methods.

1. INTRODUCTION

Let H_n denote the random variable which describes the height of binary search trees of size n, i.e.

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

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 $\le h$. (For a description of binary search trees see [7].)

In 1986 Devroye [1] 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 $c = 4.31107\ldots > 2$ is the solution of the equation $\left(\frac{2e}{c}\right)^c = e$. (Earlier Pittel [9] had shown that $H_n/\log n \to \gamma$ almost surely as $n \to \infty$, where $\gamma \le c$, compare also with Robson [10].) Later Devroye [2] proved quantified versions, i.e.

$$H_n - c\log n = \mathcal{O}(\sqrt{\log n\,\log\log n})$$

in probabilty. The best (general) estimates are due to Devroye and Reed [3]. They proved

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

and

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

However, it is known to Robson [11] that there is an infinite subsequence for which

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

Devroye's methods always use branching processes in one form or another. The aim of the present paper is to present a completely different approach to this topic. We use generating functions for the numbers $a_{n,h}$ and provide analytic tools for extracting estimates for $a_{n,h}$ via properties of these generating functions. We will prove the following theorems.

Theorem 1.1. *There exists a constant $L \ge 1$ such that*

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

as $n \to \infty$, where $c = 4.31107\ldots > 2$ is the solution of the equation

$$\left(\frac{2e}{c}\right)^c = e.$$

Remark . In what follows we will show that we can use $L = 6.6$. Of course this sharpens Devroye and Reed's constant $30(c-1)/\log 2 = 143.3\ldots$, see [3].

Date: May 14, 1999.

Key words and phrases. binary search tree, average case analysis.

Institut für Algebra und Diskrete Mathematik, 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.

Theorem 1.2. *The absolute centralized moments of H_n are bounded by*

$$\mathbf{E}|H_n - \mathbf{E}H_n|^k = \mathcal{O}\left(((L-1)\log\log n)^k\right) + \mathcal{O}\left(k!\left(\left(1-\frac{2}{c}\right)^{-1} + \mathcal{O}\left(L\frac{\log\log n}{\log n}\right)\right)^k\right)$$

uniformly for all $k \ge 1$ as $n \to \infty$.

Remark . Especially if one can prove that $L = 1$ then all centralized moments will be bounded.

The paper if organized in the following way. In section 2 we provide a recurrent relation for the numbers $a_{n,h}$ and identities for corresponding generating functions. In section 3 we show that proper estimates for $a_{n,h}$ (Propositions 3.1 and 3.2) imply Theorems 1.1 and 1.2. The remainder of the paper is devoted to the proofs of Propositions 3.1 and 3.2. Proposition 3.1 is the *easy* part and was already provided by Devroye [2]. We present an alternate proof via generating functions in section 4. In section 5 we establish properties of a specific retarded differential equation which are used to prove Proposition 3.2 in section 6. Finally in section 7 we collect some results which might be helpful for a proof of the conjectural relations

$$\mathbf{E}H_n = c\log n - \frac{c}{2(c-1)}\log\log n + \mathcal{O}(1) \quad \text{and} \quad \mathbf{V}H_n = \mathcal{O}(1).$$

Throughout the paper we will frequently use the notation

$$[x^n]A(x) = a_n$$

for the n-th Taylor coeffient of the power series expansion (around $x_0 = 0$) of the analytic function

$$A(x) = \sum_{k\ge 0} a_k x^k.$$

2. Combinatorial Background

We recall that $a_{n,h}$ denotes the number of permutations $\sigma \in S_n$ of n elements such that the corresponding binary search tree has height $\le h$. We obviously have $a_{n,h} \le n!$ and $a_{n,h} \le a_{n,h+1}$.

Lemma 2.1. *The numbers $a_{n,h}$ satisfy the recurrent relation*

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

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

Proof. The proof is immediate by considering a binary search tree where the root is labeled by $k+1$, $0 \le k \le n-1$. □

The following estimate will be used in the proof of Proposition 3.2.

Lemma 2.2. *We have*

$$a_{n+1,h} \le (n+1)a_{n,h}. \tag{1}$$

Proof. Obviously, (1) is satisfied for $h = 0$. Now suppose that (1) is true for some $h \ge 0$. Then, by Lemma 2.1

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

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

We now introduce the generating functions

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

which are in fact polynomial. Since $a_{n,h} \le n!$ we obviously have

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

for $0 \le x < 1$. Furthermore, since $a_{n,h} = n!$ for $n \le h$ we also obtain

$$\lim_{h\to\infty} y_h(x) = \frac{1}{1-x}. \tag{2}$$

Lemma 2.1 can be translated in the following way.

Lemma 2.3. *The functions $y_h(x)$ are recursively given by*

$$y_0(x) \equiv 1$$

and by

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

Proof. Obviously $y_0(x) \equiv 1$ and $y_h(0) = 1$. Furthermore

$$y'_{h+1}(x) = \sum_{n\ge 1} n \frac{a_{n,h+1}}{n!} x^{n-1} = \sum_{n\ge 1} \frac{a_{n,h+1}}{(n-1)!} x^{n-1}$$

and

$$y_h(x)^2 = \sum_{n\ge 0}\sum_{k=0}^{n} \frac{a_{k,h}}{k!}\frac{a_{n-k,h}}{(n-k)!} x^n = \sum_{n\ge 1}\sum_{k=0}^{n-1} \frac{a_{k,h}}{k!}\frac{a_{n-1-k,h}}{(n-1-k)!} x^{n-1}.$$

Hence, Lemma 2.1 implies $y'_{h+1}(x) = y_h(x)^2$. $\square$

Remark . Equivalently we have

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

i.e. $y_h(x)$ are approximate solutions of the differential equation $y' = y^2$ with initial condition $y(0) = 1$. The exact solution is $y(x) = 1/(1-x)$ in accordance with (2).

The distribution of the random variable H_n (describing the height of binary search trees of size n) is given by

$$\mathbf{P}[H_n = k] = \frac{a_{n,k}}{n!} - \frac{a_{n,k-1}}{n!}.$$

Lemma 2.4. *The expected value of H_n is given by*

$$\mathbf{E}H_n = \sum_{h\ge 0}\left(1 - \frac{a_{n,h}}{n!}\right)$$

and the corresponding generating function by

$$E(x) = \sum_{n\ge 0} \mathbf{E}H_n\, x^n = \sum_{h\ge 0}\left(y(x) - y_h(x)\right),$$

where $y(x) = 1/(1-x)$.

Proof. Note that $(1 - a_{n,h}/n!) = 0$ for $h \ge n$. Hence, we obtain by definition and partial summation

$$\begin{aligned}\mathbf{E}H_n &= \sum_{h\ge 0} h\mathbf{P}[H_n = h] \\ &= \sum_{h\ge 1} h\left(\frac{a_{n,h}}{n!} - \frac{a_{n,h-1}}{n!}\right) \\ &= \sum_{h\ge 0}\left(1 - \frac{a_{n,h}}{n!}\right).\end{aligned}$$

The representation of $E(x)$ is now obvious. □

3. Tail Estimates

In this section we show that proper estimates for the tails of the distribution of H_n are sufficient to provide Theorems 1.1 and 1.2. These estimates are stated in two Propositions which will be proved in section 4 (Proposition 3.1) and in section 6 (Proposition 3.2). (As already mentioned, Proposition 3.1 is due to Devroye [2], where a completely different proof is provided.)

As above, $c = 4.31107\cdots > 2$ will always denote the solution of

$$\left(\frac{2e}{c}\right)^c = e. \tag{3}$$

Furthermore, for convenience we use the abbreviation

$$d := \frac{c}{2(c-1)}. \tag{4}$$

Proposition 3.1. *There exists a constant $C_1 > 0$ such that for all $h \ge c\log n - d\log\log n$*

$$1 - \frac{a_{n,h}}{n!} \le C_1 \frac{1}{n}\frac{(2\log n)^h}{h!} \tag{5}$$

Proposition 3.2. *There exists $L \ge 1$ and a constant $C_2 > 0$ such that for all $h \le c\log n - dL\log n$*

$$\frac{a_{n,h}}{n!} \le C_2 \exp\left(-ne^{-(h+dL\,\log h)/c}\right). \tag{6}$$

We will first prove estimates for $\mathbf{E}H_n$.

Lemma 3.1. *Assume that Proposition 3.1 holds. Then we have*

$$\sum_{h\ge c\log n - d\log\log n}\left(1 - \frac{a_{n,h}}{n!}\right) = \mathcal{O}(1)$$

as $n \to \infty$.

Proof. Define h_0 and D_n by

$$h_0 := \lceil c\log n - d\log\log n\rceil = c\log n - d\log\log n + D_n.$$

Since $c > 2$ we obtain by (5) and by Stirling's formula (we note that Vinogradov's notation $a \ll b$ means that there exists an absolute constant C with $a \le Cb$)

$$
\begin{aligned}
\sum_{h \ge h_0} \left(1 - \frac{a_{n,h}}{n!}\right) &\ll \frac{1}{n} \sum_{h \ge h_0} \frac{(2\log n)^h}{h!} \\
&= \frac{1}{n} \frac{(2\log n)^{h_0}}{h_0!} \left(1 + \frac{2\log n}{h_0+1} + \frac{(2\log n)^2}{(h_0+1)(h_0+2)} + \cdots\right) \\
&\le \frac{1}{n} \frac{(2\log n)^{h_0}}{h_0!} \frac{1}{1 - \frac{2\log n}{h_0+1}} \\
&\ll \frac{1}{n} \left(\frac{2e\log n}{h_0}\right)^{h_0} \frac{1}{\sqrt{\log n}} \\
&= \frac{1}{2\sqrt{\log n}} \left(\frac{2e}{c\left(1 - \frac{d\log\log n - D_n}{c\log n}\right)}\right)^{c\log n - d\log\log n + D_n} \\
&\ll \frac{1}{2\sqrt{\log n}} \left(\frac{2e}{c}\right)^{c\log n} \left(\frac{2e}{c}\right)^{-d\log\log n} \\
&\quad \times \exp\left(\left(\frac{d\log\log n - D_n}{c\log n} + \mathcal{O}\left(\left(\frac{\log\log n}{\log n}\right)^2\right)\right)(c\log n - d\log\log n)\right) \\
&\ll \frac{1}{\sqrt{\log n}} \left(\frac{2e}{c}\right)^{-d\log\log n} e^{d\log\log n} \\
&= \frac{1}{\sqrt{\log n}} \left(\frac{c}{2}\right)^{\frac{c}{2(c-1)}\log\log n} = 1,
\end{aligned}
$$

where we also used that $\left(\frac{2e}{c}\right)^c = e$ and $\log\frac{c}{2} = \frac{c-1}{2}$. $\square$

Lemma 3.2. *Assume that Proposition 3.2 holds. Then we have*

$$\sum_{h \le c\log n - dL\log\log n} \frac{a_{n,h}}{n!} = \mathcal{O}(1)$$

as $n \to \infty$.

Proof. Similarly to the above set $h_0' := \lceil c\log n - dL\log\log n\rceil$. Obviously, there exists a constant $C_3 > 0$ such that

$$n \ge C_3 e^{(h_0' + dL\log h_0')/c}.$$

Hence, by (6) directly obtain

$$
\begin{aligned}
\sum_{1 \le h < h_0'} \frac{a_{n,h}}{n!} &\ll \sum_{1 \le h < h_0'} \exp\left(-ne^{-(h + dL\log h)/c}\right) \\
&\le \sum_{1 \le h < h_0'} \exp\left(-C_3 e^{(h_0' - h + dL(\log h_0' - \log h))/c}\right) \\
&= \sum_{0 < l < h_0'} \exp\left(-C_3 e^{(l + dL(\log h_0' - \log(h_0' - l)))/c}\right) \\
&\le \sum_{0 < l \le h_0'/2} \exp\left(-C_3 e^{(l + dL\log 2)/c}\right) + \sum_{h_0'/2 < l < h_0'} \exp\left(-C_3 e^{(l + dL\log h_0')/c}\right) \\
&= \mathcal{O}(1) + \mathcal{O}(1) = \mathcal{O}(1),
\end{aligned}
$$

where we also used the substitution $l = h_0' - h$. $\square$

Now it is easy to complete the proof of **Theorem 1.1**.

Proof. (Theorem 1.1). By applying Lemma 2.4 and 3.1 we directly obtain

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

Conversely by Lemma 3.2

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

This completes the proof of Theorem 1.1. □

The proof of **Theorem 1.2** is a little bit more technical.

Proof. (Theorem 1.2). Let $F_n(x)=\mathbf{P}[H_n\le x]=a_{n,[x]}/[x]!$ (for $x\ge 0$) be the distribution function of H_n. Then by partial integration we get

$$\begin{aligned}
\mathbf{E}|H_n-\mathbf{E}H_n|^k &= \int_0^\infty |x-\mathbf{E}H_n|^k\,dF_n(x)\\
&= \int_0^{\mathbf{E}H_n}(\mathbf{E}H_n-x)^k\,dF_n(x)+\int_{\mathbf{E}H_n}^\infty (x-\mathbf{E}H_n)^k\,dF_n(x)\\
&= k\int_0^{\mathbf{E}H_n}(\mathbf{E}H_n-x)^{k-1}\,F_n(x)+k\int_{\mathbf{E}H_n}^\infty (x-\mathbf{E}H_n)^{k-1}\,(1-F_n(x))\\
&\le k\sum_{h\le c\log n-Ld\log\log n}(\mathbf{E}H_n-h)^{k-1}\frac{a_{n,h}}{n!}\\
&\quad +k\sum_{c\log n-Ld\log\log n<h<c\log n-d\log\log n}|\mathbf{E}H_n-h|^{k-1}\\
&\quad +k\sum_{h\ge c\log n-d\log\log n}(h-\mathbf{E}H_n)^{k-1}\left(1-\frac{a_{n,h}}{n!}\right)\\
&= S_1+S_2+S_3.
\end{aligned}$$

Set $h_0=\lceil c\log n-d\log\log n\rceil$. Then the third sum can be estimated by

$$\begin{aligned}
S_3 &\ll k\sum_{h\ge h_0}(h-h_0+\mathcal{O}(1))^{k-1}\left(1-\frac{a_{n,h-1}}{n!}\right)\\
&\ll \frac{k}{n}\sum_{h\ge h_0}(h-h_0+\mathcal{O}(1))^{k-1}\frac{(2\log n)^h}{h!}\\
&\ll \frac{k}{n}\frac{(2\log n)^{h_0}}{h_0!}\sum_{l\ge 0}(l+\mathcal{O}(1))^{k-1}\left(\frac{2\log n}{h_0}\right)^l\\
&\ll k\sum_{l\ge 0}l^{k-1}\left(\frac{2\log n}{h_0}\right)^l.
\end{aligned}$$

Since

$$\sum_{l\ge 0} l^{k-1}u^l \le (k-1)!\sum_{l\ge 0}\binom{k+l-1}{k-1}u^l = \frac{(k-1)!}{(1-u)^k}$$

for $k\ge 1$ and $0\le u<1$ we finally can estimate S_3 by

$$S_3 \ll k!\left(\left(1-\frac{2}{c}\right)^{-1}+\mathcal{O}\left(L\frac{\log\log n}{\log n}\right)\right)^k.$$

The first sum S_1 is much smaller. (Set $h_0' = \lceil c\log n - d\log\log n\rceil$.)

$$\begin{aligned}
S_1 &\ll kC_4^k \sum_{h<h_0'} (h_0'-h_0)^{k-1}\frac{a_{n,h}}{n!} \\
&\ll kC_4^k \sum_{l<h_0'} l^{k-1}\exp\left(-C_3 e^{(l+dL(\log h_0' - \log(h_0'-l)))/c}\right) \\
&\ll kC_4^k \sum_{l\le h_0'/2} l^{k-1}\exp\left(-C_3 e^{(l+dL\,\log 2)/c}\right) \\
&\quad + kC_4^k \sum_{h_0'/2<l\le h_0'} l^{k-1}\exp\left(-C_3 e^{(l+dL\,\log h_0')/c}\right) \\
&\ll C_4^k k^{1+k/c}.
\end{aligned}$$

Finally the second sum S_2 is trivially estimated from above by

$$S_2 \ll k \sum_{l=1}^{(L-1)\log\log n} l^{k-1} \ll \left((L-1)\log\log n\right)^k.$$

This proves Theorem 1.2. □

4. Lower Bound for $y_h(x)$

It was already known to Robson [10] that $\mathbf{E}H_n \le c\log n$. In the first part of this section we present an extended version of an unpublished manuscript by Gourdon [5], who translated Robson's ideas to the language of generating functions. The result is a new proof of Proposition 3.1.

For convenience we will use the abbreviations

$$\begin{aligned}
y(x) &= \frac{1}{1-x}, \\
\delta_h(x) &= y(x) - y_h(x), \\
L(x) &= \log\frac{1}{1-x},
\end{aligned}$$

and

$$\alpha = e^{1/c} = 1.26\ldots.$$

Lemma 4.1. *For $0\le x<1$ we have*

$$\delta_h(x) \le 2^h \sum_{k>h}\frac{L(x)^k}{k!}. \tag{7}$$

Furthermore

$$1-\frac{a_{n,h}}{n!} \le [x^n]\, 2^h \sum_{k\ge h}\frac{L(x)^k}{k!}. \tag{8}$$

Proof. We proceed by induction. First of all we have

$$\delta_0(x) = \frac{x}{1-x} = \sum_{k>0} \frac{L(x)^k}{k!}.$$

Now observe that $\delta_h(0) = 0$ and $a_{n,h} \le n!$ implies $y_h(x) \le y(x)$. Hence

$$\begin{aligned} \delta'_{h+1}(x) &= y'(x) - y'_{h+1}(x) \\ &= y(x)^2 - y_h(x)^2 \\ &= \delta_h(x)(y(x) + y_h(x)) \\ &\le 2\delta_h(x)y(x) \end{aligned}$$

and we immediately obtain (7). Finally observe, that the only *inequality* which is used is $y_h(x) \le y(x)$ which is also true on the level of coefficients, i.e. $a_{n,h}/n! \le 1$. Thus (8) follows from (7). □

The crucial observation is that Lemma 4.1 is also true on the level of coefficients. Therefore we can apply it to obtain bounds for $a_{n,h}$. For this purpose we will use the following estimate. (A proof is given in the appendix.)

Lemma 4.2. *We have*

$$[x^n]L(x)^k = \mathcal{O}\left(\frac{(\log n)^k}{n}\right) \tag{9}$$

uniformly for all $k \ge 0$.

In fact we will provide a little bit more general estimate than stated in Proposition 3.1.

Lemma 4.3. *Suppose that* $h > \log n$. *Then*

$$1 - \frac{a_{n,h}}{n!} \le C\frac{2^h(\log n)^h}{n\,h!}\,\frac{1}{1-\frac{\log n}{h+1}},$$

where $C > 0$ *is a uniform constant.*

Proof. By (8) and (9) we obtain

$$\begin{aligned} 1 - \frac{a_{n,h}}{n!} &\le 2^h \sum_{k\ge h} \frac{1}{k!}[x^n]L(x)^k \\ &\le C\frac{2^h}{n}\sum_{k\ge h}\frac{(\log n)^k}{k!} \\ &\le C\frac{2^h}{n}\frac{(\log n)^h}{h!}\frac{1}{1-\frac{\log n}{h+1}}. \end{aligned}$$

□

It is immediate that Proposition 3.1 follows from Lemma 4.3

5. A Retarded Differential Equation

For convenience we will use the abbreviation

$$\alpha = e^{1/c} = 1.26\ldots.$$

The purpose of this section is to provide asymptotic properties of the (retarded) differential equation

$$\Phi'(u) = -\frac{1}{\alpha^2}\Phi(u/\alpha)^2 \tag{10}$$

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

First of all, (10) has a unique solution $\Phi(u)$ with $\Phi(0) = 1$.

Lemma 5.1. *There exists a unique entire function $\Phi(u)$ with $\Phi(0) = 1$ which solves (10).*

Proof. It is easy to obtain a formal solution. Starting with a formal power series

$$\Phi(u) = \sum_{n\geq 1} c_n u^n$$

and inserting it into (10) we immediately obtain the recurrence

$$c_{n+1} = -\frac{\alpha^{-n-2}}{n+1}\sum_{k=0}^{n} c_k c_{n-k}.$$

By induction it follows that $|c_n| \leq 1$ for all $n \geq 0$. Hence, $\sum_{n\geq 1} c_n u^n$ represents an analytic function for $|u| < 1$ and thus provides an analytic solution of the functional equation (10). Now (10) ensures that $\Phi'(u)$ (and consequently $\Phi(u)$) has an analytic continuation to the disc $|u| < \alpha$ which satisfies (10). Repeating this argument it follows that this solution $\Phi(u)$ can be analytically continued to an entire function. Obviouly there is at most one entire solution. □

It is clear that $\Phi(u)$ is strictly decreasing for $u \geq 0$. However, it seems to be a non-trivial problem to decide whether $\Phi(u)$ remains positive or not (for $u \geq 0$). The key is to introduce the auxiliary function

$$f(x) := e^{(c-1)x}\left(1 - e^x\Phi(e^x)\right), \tag{11}$$

where $\Phi(u)$ is the unique analytic solution of (10) with initial condition $\Phi(0) = 1$. First of all the differential equation (10) translates to a differential equation for $f(x)$ which is more easy to handle.

Lemma 5.2. *Let $f(x)$ be defined by (11). Then $f(x)$ satisfies the retarded differential equation*

$$f'(x) = c\big(f(x) - f(x - c^{-1})\big) + e^{-(c-1)x} f(x - c^{-1})^2. \tag{12}$$

Proof. First, set $\Psi(u) := u\Phi(u)$. Then we have

$$\begin{aligned} u\Psi'(u) &= u\Phi(u) + u^2\Phi'(u) \\ &= u\Phi(u) - \left(\frac{u}{\alpha}\Phi\left(\frac{u}{\alpha}\right)\right)^2 \\ &= \Psi(u) - \Psi(u/\alpha)^2. \end{aligned}$$

Hence, $F(u) := 1 - \Psi(u)$ satisfies

$$\begin{aligned} uF'(u) &= -(1 - F(u)) + (1 - F(u/\alpha))^2 \\ &= F(u) - 2F(u/\alpha) + F(u/\alpha)^2. \end{aligned}$$

Next set $g(x) := F(e^x)$. Fere we have $g'(x) = e^x F'(e^x)$ and consequently

$$g'(x) = g(x) - 2g(x - c^{-1}) + g(x - c^{-1})^2. \tag{13}$$

Finally, $f(x)$ (defined by (11)) is given by $f(x) = e^{(c-1)x} g(x)$. Inserting the representation $g(x) = e^{-(c-1)x} f(x)$ in (13) directly provides (12). □

The crucial point is to know the asymptotic behaviour of $\Phi(u)$ for (real) $u \to \infty$. The proof of the following Proposition is very involved and therefore is divided into several steps.

Proposition 5.1. *Let $\Phi(u)$ be the unique analytic solution of (10) with initial condition $\Phi(0) = 1$. Then for (real) $u \geq 0$ the function $u\Phi(u)$ is monotonically increasing and*

$$1 - u\Phi(u) \sim C_5 \frac{\log u}{u^{c-1}} \tag{14}$$

as $u \to \infty$ for some real constant $C_5 > 0$.

Proof. **Step 1.** *$f'(x) > 0$ for all real x.*
Since
$$\Phi(u) = 1 - \alpha^{-2}u + \mathcal{O}(u^2) \quad \text{and} \quad \Phi'(u) = -\alpha^{-2} + \mathcal{O}(u)$$
as $u \to 0$ it follows that
$$\lim_{x\to-\infty} e^x\Phi(e^x) = \lim_{x\to-\infty} e^x\Phi'(e^x) = 0.$$
Hence, there exists a real number x_0 such that
$$f'(x) = e^{(c-1)x}\left((c-1)(1 - e^x\Phi(e^x)) - e^x\Phi'(e^x)\right) > 0$$
for all $x \le x_0$. Assume for a moment that
$$x_1 := \inf\{x \in \mathbf{R} : f'(x) \le 0\} < \infty.$$
Then we surely have $f'(x_1) = 0$. Furthermore, since $f'(x) > 0$ for all $x < x_1$ it follows that $f(x_1) > f(x_1 - c^{-1})$. Thus
$$f'(x_1) = c(f(x_1) - f(x_1 - c^{-1})) + e^{(c-1)x_1}f(x_1 - c^{-1})^2 > 0$$
leads to a contradiction.

Step 2. *$f(x) > 0$ for all real x.*
Since $\lim_{x\to-\infty} e^x\Phi(e^x) = 0$ there surely exists a real number x_2 such that $f(x) > 0$ for all $x \le x_2$. However, by Step 1 we have $f'(x) > 0$ for all real x, which implies $f(x) > 0$ for $x > x_2$, too.

Step 3. *$u\Phi(u) < 1$ for all real $u \ge 0$.*
Since $f(x) = e^{(c-1)x}(1 - e^x\Phi(e^x)) > 0$ for all $x \in \mathbf{R}$ it directly follows that $e^x\Phi(e^x) < 1$ for all $x \in \mathbf{R}$. Consequently $u\Phi(u) < 1$ for all real $u > 0$. For $u = 0$ the inequality $u\Phi(u) < 1$ is trivial.

Step 4. *$u\Phi(u)$ is monotonically increasing for all real $u \ge 0$.*
As in the proof of Lemma 5.2 set $\Psi(u) := u\Phi(u)$. Then $\Psi(u)$ satisfies the differential equation
$$u\Psi'(u) = \Psi(u) - \Psi(u/\alpha)^2.$$
Furthermore, $\Psi'(u) = \Phi(u) + u\Phi'(u)$ implies
$$\lim_{u\to0} \Psi'(u) = 1.$$
Therefore $\Psi(u) = u\Phi(u)$ is monotonically increasing at least for sufficiently small (real) $u \ge 0$. Now assume for a moment that
$$u_1 = \inf\{u > 0 : \Psi'(u) \le 0\} < \infty.$$
Then $u_1 > 0$, $\Psi'(u_1) = 0$, and $\Psi'(u) > 0$ for $u < u_1$. Thus, $\Psi(u_1/\alpha) < \Psi(u_1)$. Furthermore, by Step 3 $\Psi(u_1) < 1$. Hence,
$$\Psi(u_1/\alpha)^2 < \Psi(u_1/\alpha) < \Psi(u_1)$$
and consequently
$$u_1\Psi'(u_1) = \Psi(u_1) - \Psi(u_1/\alpha) > 0$$
which leads to a contradiction.

Step 5. *$\Phi(u) > 0$ for all $u \ge 0$.*
Since $\Phi(0) = 1$ it follows that $u\Phi(u) > 0$ for sufficiently small real $u > 0$. Hence, by Step 4 $u\Phi(u) > 0$ for all real $u > 0$ which implies $\Phi(u) > 0$.

Step 6. *$\Phi''(u) > 0$ for all real $u \ge 0$.*
By (10) we have
$$\begin{aligned}\Phi''(u) &= -\frac{2}{\alpha^3}\Phi(u/\alpha)\Phi'(u/\alpha)\\ &= \frac{2}{\alpha^5}\Phi(u/\alpha)\Phi(u/\alpha^2)^2.\end{aligned}$$
Hence, Step 5 implies $\Phi''(u) > 0$ for all $u \ge 0$.

Step 7. $(u\Phi(u))'' < 0$ *for all real* $u \geq 0$.
As above set $\Psi(u) = u\Phi(u)$. From $u\Psi(u)' = \Psi(u) - \Psi(u/\alpha)^2$ it follows that

$$u\Psi''(u) = -\frac{2}{\alpha}\Psi(u/\alpha)\Psi'(u/\alpha) < 0.$$

This proves Step 7.

Step 8. $\Phi(u) \geq 1/(1+u)$ *for all real* $u \geq 0$.
By Step 4 the function $u\Phi(u)$ in monotonically increasing. Hence, $u\Phi(u) \geq (u/\alpha)\Phi(u/\alpha)$ and consequently by (10)

$$-\Phi'(u) \leq \Phi(u)^2 \quad \text{resp.} \quad -\frac{\Phi'(u)}{\Phi(u)} \leq 1.$$

for all $u \geq 0$. Hence,

$$\frac{1}{\Phi(u)} - \frac{1}{\Phi(0)} \leq u$$

or $\Phi(u) \geq 1/(1+u)$.

Step 9. $1 - u\Phi(u) = \mathcal{O}(u^{-1.7})$ *as* $u \to +\infty$.
As in the proof of Lemma 5.2 set $F(u) := 1 - u\Phi(u) = 1 - \Psi(u)$. By Step 3 we know that $F(u) > 0$ and by Step 8 that

$$\lim_{u\to\infty} F(u) = 0.$$

Hence there exist $u_0 > 0$ such that $0 < F(u) < \varepsilon := 10^{-3}$ for $u \geq u_0$, i.e. for $u \geq \alpha u_0$ we have

$$\begin{aligned} uF'(u) &= F(u) - 2F(u/\alpha) + F(u/\alpha)^2 \\ &< F(u) - 2F(u/\alpha)(1-\varepsilon/2). \end{aligned}$$

By Step 7 $F''(u) = -(u\Phi(u))'' > 0$. Consequently

$$F(u/\alpha) \geq F(u) - u(1-1/\alpha)F'(u)$$

which gives

$$uF'(u) < F(u) - 2(F - u(1-1/\alpha)F'(u))(1-\varepsilon/2)$$

or

$$\frac{F'(u)}{F(u)} < -\frac{1-\varepsilon}{1-2(1-1/\alpha)(1-\varepsilon/2)}\,\frac{1}{u} < -1.7\,\frac{1}{u}$$

Now integration yields

$$\log F(u) - \log F(\alpha u_0) < -1.7\,(\log u - \log(\alpha u_0))$$

resp. $F(u) = \mathcal{O}(u^{-1.7})$ as $u \to \infty$.

Step 10. *We have* $\displaystyle\int_{-\infty}^{x} e^{-(c-1)t} f(t-c^{-1})^2\,dt = D + \mathcal{O}(e^{-\delta t})$ *as* $x \to \infty$ *for some positive real constants* δ, D.
By definition $f(x) \leq e^{(c-1)x}$. Therefore the integral

$$\int_{-\infty}^{0} e^{-(c-1)t} f(t-c^{-1})^2\,dt$$

is surely convergent. Furthermore, by Step 9

$$e^{-(c-1)t} f(t-c^{-1})^2 = e^{2-2/c} e^{(c-1)t} F(e^{t-c^{-1}})^2 = \mathcal{O}(e^{3.32t-3.4t}) = \mathcal{O}(e^{-0.08t})$$

as $t \to \infty$. Now Step 10 follows immediately with

$$D = \int_{-\infty}^{\infty} e^{-(c-1)t} f(t-c^{-1})^2\,dt > 0$$

and $\delta = 0.08 > 0$.

Step 11. *We have* $f(x) = c\int_{x-c^{-1}}^{x} f(t)\,dt + D + \mathcal{O}(e^{-\delta x})$ *as* $x\to\infty$.

Since $\lim_{x\to-\infty} f(x) = 0$ and the integrals

$$\int_{-\infty}^{x} f(t)\,dt$$

converge for all $x\in\mathbf{R}$ Step 11 follows directly from the differential equation for $f(x)$ (12) and from Step 10.

Step 12. *For every* $m\ge 0$ *there exists a real constant* C_m' *such that*

$$f(x) \le (2+2^{-m})cDx + C_m' \quad \text{for } x\ge 0.$$

As in the proof of Step 11 it follows that

$$f(x) \le c\int_{x-c^{-1}}^{x} f(t)\,dt + D \tag{15}$$

for all $x\in\mathbf{R}$. Starting with $f(x)\le e^{(c-1)x}$ it follows inductively that for every integer $k\ge 0$

$$f(x) \le \left(\frac{c}{c-1}\left(1-e^{-(c-1)/c}\right)\right)^k e^{(c-1)x} + kD \le e^{3.4x-0.35k} + kD.$$

One just has to observe that

$$\int_{x-c^{-1}}^{x} e^{(c-1)t}\,dt = \frac{c}{c-1}\left(1-e^{-(c-1)/c}\right)e^{(c-1)x}.$$

Choosing $k=\lfloor 3cx\rfloor$ (for $x\ge 0$) we immediately get

$$f(x) \le 3cDx + \mathcal{O}(1) \le (2+2^{-0})cDx + C_0'.$$

Now suppose that $f(x)\le (2+2^{-m})cDx + C_m'$ (for $x\ge 0$) for some $m\ge 0$. Then iterated use of (15) gives

$$f(x) \le (2+2^{-m})cDx + \left(D - \frac{(2+2^{-m})cD}{2c}\right)k + C_m' = (2+2^{-m})cDx - 2^{-m-1}Dk + C_m'$$

for $x\ge k/c$. Choosing $k=\lfloor cx\rfloor$ we obtain

$$\begin{aligned} f(x) &\le (2+2^{-m})cDx - 2^{-m-1}cDx + \mathcal{O}(2^{-m}) + C_m' \\ &\le (2+2^{-m-1})cDx + C_{m+1}'. \end{aligned}$$

This proves Step 12.

Step 13. *For every* $m\ge 0$ *there exists a real constant* C_m'' *such that*

$$f(x) \ge (2-2^{-m})cDx + C_m'' \quad \text{for } x\ge 0.$$

By Step 11 and the monotonicity of $f(x)$ (see Step 1) it follows that

$$f(x) \ge f(x-c^{-1}) + D + \mathcal{O}(e^{-\delta x})$$

for $x\ge 0$. Setting $a_k := f(kc)$ then we have $a_{k+1}\ge a_k + D + \mathcal{O}(e^{-(c\delta)k})$ and consequently $a_k\ge kD+\mathcal{O}(1)$. Obviously this implies

$$f(x) \ge cDx + \mathcal{O}(1) \ge (2-2^{-0}) + C_0''.$$

for $x\ge 0$. Now suppose that $f(x)\ge (2-2^{-m})cDx + C_m''$ (for $x\ge 0$) for some $m\ge 0$. Then iterated use of Step 11 gives

$$f(x) \ge (2-2^{-m})cDx + 2^{-m-1}Dk + C_m'' + \mathcal{O}\left((1+\gamma+\gamma^2+\cdots+\gamma^k)e^{-\delta x}\right)$$

for $x\ge k/c$, where

$$\gamma = \frac{c}{\delta}\left(e^{\delta/c}-1\right) \le e^{0.0093}.$$

Choosing $k = \lfloor cx \rfloor$ we obtain

$$\begin{aligned} f(x) &\geq (2-2^{-m})cDx + 2^{-m-1}cDx + \mathcal{O}(2^{-m}) + C''_m + \mathcal{O}\left(e^{-\delta x + 0.0093cx}\right) \\ &\leq (2-2^{-m-1})cDx + C''_{m+1}. \end{aligned}$$

Note that $\delta = 0.08$ and $-\delta + 0.0093c \leq -0.0399 < 0$. This completes the proof of Step 13.

Step 14. *$f(x) \sim 2cDx$ as $x \to \infty$.*

Combine Step 12 and 13.

Step 15. *Completion of the proof.*

Step 4 says that $u\Phi(u)$ is monotonically increasing. Furthermore, Step 14 directly translates to (14). ☐

Remark . Note that one crucial step was an estimate for $F'(u)/F(u)$ (Step 9) leading to polynomial growth. This part of the analysis looks similar to the analysis of specific differential equations by Jacquet and Szpankowski [6] which are related to digital search trees.

It is clear that much more precise information can be derived for the properties of $\Phi(u)$, e.g. an estimate of the kind

$$\Phi(u) = \frac{1}{u} - C_5 \frac{\log u}{u^c} + \mathcal{O}\left(\frac{1}{u^c}\right) \quad (u \to \infty). \tag{16}$$

In fact it is very likely that $u\Phi(u)$ has a representation of the form $u\Phi(u) = f(u^{-(c-1)}, \log u \cdot u^{-(c-1)})$, where $f(v,w)$ has a power series expansion around $(0,0)$. The next lemma supports this conjecture.

Lemma 5.3. *Suppose that $1 < \alpha < 2$. Then there exist infinitely many solutions $\Phi(u)$, $u > u_0$ of (10) satisfying $\Phi(u) \sim 1/u$ as $u \to \infty$. More precisely:*

- *If $1 < \alpha < e^{1/c}$ and if $\beta < \gamma$ denote the two real solutions of $1+\beta = 2\alpha^\beta$, $1+\gamma = 2\alpha^\gamma$, then there exist infinitely many functions $f(v,w)$ with $f(0,0) = 1$ which are analytic around $(0,0)$ such that $\Phi(u) = f(u^{-\beta}, u^{-\gamma})/u$ satisfies (10)*
- *If $e^{1/c} < \alpha < 2$ and if $\beta, \gamma = \overline{\beta}$ denote the two complex solutions of $1+\beta = 2\alpha^\beta$, $1+\gamma = 2\alpha^\gamma$, then there exist infinitely many functions $f(v,w)$ with $f(0,0) = 1$ which are analytic around $(0,0)$ such that $\Phi(u) = f(u^{-\beta}, u^{-\gamma})/u$ satisfies (10).*
- *If $\alpha = e^{1/c}$ and if $\beta = c-1$ denotes the unique real solution of $1+\beta = 2\alpha^\beta$ then there exist infinitely many functions $f(v,w)$ with $f(0,0) = 1$ which are analytic around $(0,0)$ such that $\Phi(u) = f(u^{-\beta}, \log u \cdot u^{-\beta})/u$ satisfies (10)*

Proof. Suppose that $f(v,w)$ is a differentiable function satisfying

$$\beta v f_v(v,w) + \gamma w f_w(v,w) + f(v,w) = f(\alpha^\beta v, \alpha^\gamma w)^2 \tag{17}$$

then it is clear that $\Phi(u) = f(u^{-\beta}, u^{-\gamma})/u$ satisfies (10). Now suppose that $f(v,w)$ has a power series expansion

$$f(v,w) = \sum_{j,k \geq 0} d_{jk} v^j w^k.$$

If $f(v,w)$ satisfies (17) then $d_{00} = f(0,0) = 1$. Furthermore, by differentiating with respect to u it follows that

$$\beta f_v(0,0) + f_v(0,0) = 2f(0,0) f_v(0,0) \alpha^\beta.$$

Hence, if $d_{10} = f_v(0,0) \neq 0$ then $1+\beta = 2\alpha^\beta$. Similarly if $d_{01} = f_w(0,0) \neq 0$ then $1+\gamma = 2\alpha^\gamma$.

Conversely if we suppose that $1+\beta = 2\alpha^\beta$ and $1+\gamma = 2\alpha^\gamma$ then we can arbitrarily choose d_{01} and d_{10} and obtain a formal solution of (17) by comparing the power series expansions on both sides of (17) and extracting a recursion for the coefficients d_{jk}. (Of course, $d_{00} = 1$.) As above, it is also possible to show that this formal solution is analytic around $(0,0)$.

By using this procedure, we directly obtain infinitely many different solutions of (10) if $\alpha \neq e^{1/c}$.

If $\alpha = e^{1/c}$ we suppose that $f(v,w)$ is a solution of

$$\beta v f_v(v,w) - v f_w(v,w) + \beta w f_w(v,w) + f(v,w) = f(\alpha^\beta v, \alpha^\beta w - \log\alpha \cdot \alpha^\beta v)^2. \tag{18}$$

Here, $\Phi(u) = f(u^{-\beta}, \log u \cdot u^{-\beta})/u$ satisfies (10). And if we suppose that $\beta = c-1$ is the (only) real solution of $1+\beta = \alpha^\beta$ then we again find infinitely many solutions of (18) by choosing d_{01} and d_{10} independently and arbitarily. □

Especially it might be conjectured that on of these solutions of (10) for $\alpha = e^{1/c}$ is exactly that solution which is analytic around $u=0$ with initial condition $\Phi(0)=1$.

Another conjecture which is related to the above one is that there exists a function $\Psi(y)$, $0 \le y < \infty$, satisfying

1. $\Psi(y) \sim 1 - y^{c-1} + c' \log y \cdot y^{c-1}$, $(y \to 0+)$.
2. $\lim_{y\to\infty} e^{(1+\delta)y}\Psi(y) = 0$ for some $\delta > 0$.
3. $\Psi(y)$ is strictly decreasing.
4. $0 < \int_0^\infty \Psi(y)\,dy < \infty$.
5. $\int_0^y \Psi(z)\Psi(y-z)\,dy = y\Psi(y/\alpha)$, $(0 \le y < \infty)$.

Here we can use

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

It is easy to show that this function $\Phi(u)$ satisfies the following properties:

1. $\Phi(u)$ is entire.
2. $0 < \Phi(u) < \infty$.
3. $u\Phi(u)$ is strictly increasing for $u \ge 0$.
4. There exists a constant $C_5' > 0$ such that
$$\Phi(u) = \frac{1}{u} - C_5' \frac{\log u}{u^c} + \mathcal{O}\left(\frac{1}{u^c}\right) \qquad (u \to \infty)$$
5. $\Phi'(u) = -\alpha^{-2}\Phi(u/\alpha)^2$.

One advantage of knowing the existence of $\Psi(y)$ is that the Taylor coefficients of

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

are approximately given by

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

is h is close to $c\log n$. (The importance of these functions $\tilde{y}_h(x)$ will be clear in the next section.) Maybe this gives some hint for a possible limiting distribution of $H_n - \mathbf{E}H_n$.

6. Upper Bound for $y_h(x)$

The most important tool for the proof of Proposition 3.2 is the set of functions $\tilde{y}_h(x)$ $(h \ge 0)$ defined by

$$\tilde{y}_h(x) := \alpha^h \Phi(\alpha^h(1-x)), \tag{19}$$

where $\alpha = e^{1/c} = 1.26\ldots$ and $\Phi(u)$ is the analytic solution of (10) with initial condition $\Phi(0)=1$. The reason why we discuss these functions is that they simulate the recurrence of $y_h(x)$.

Lemma 6.1. *Suppose that $\tilde{y}_h(x)$ is definded by (19), where $\Phi(u)$ is a solution of (10). Then*

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

Proof. By using (10) we directly obtain

$$\begin{aligned}\tilde{y}'_{h+1}(x) &= \alpha^{h+1}\Phi'(\alpha^{h+1}(1-x))(-1)\alpha^{h+1}\\ &= -\alpha^{2h+2}\Phi'(\alpha^{h+1}(1-x))\\ &= \alpha^{2h}\Phi(\alpha^h(1-x))^2\\ &= \tilde{y}_h(x)^2.\end{aligned}$$

□

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 $\tilde{y}_h(0) = \alpha^h\Phi(\alpha^h) < 1$. This means that $\tilde{y}_h(x) < y_h(x)$. Nevertheless it will be possible to provide an upper bound for $y_h(x)$ by using a *shifted* function $\tilde{y}_{h+D}(x)$.

The following lemma lists all properties of $\tilde{y}_h(x)$ which will be used in the sequel.

Lemma 6.2. *The functions $\tilde{y}_h(x)$, $h \ge 0$, $x \ge 0$, defined by (19) satisfy*

1. $\tilde{y}_h(0) < 1$.
2. $1 - \tilde{y}_h(0) \sim \dfrac{C_5}{c} h \left(\dfrac{2}{c}\right)^h \quad (h \to \infty)$.
3. $\tilde{y}_h(1) = \alpha^h$.
4. $\tilde{y}_{h+r}(x) \ge \tilde{y}_h(x)$ *and* $\tilde{y}'_{h+r}(x) \ge \tilde{y}'_h(x)$ *for* $r \ge 0$.
5. $\tilde{y}'_{h+1}(x) = \tilde{y}_h(x)^2$.

Proof. Since $\tilde{y}_h(0) = \alpha^h\Phi(\alpha^h)$ 1. and 2. follows from Proposition 5.1.
3. is trivial since $\tilde{y}_h(1) = \alpha^h\Phi(0) = \alpha^h$.
The first part of 4. is an immediate consequence Step 4 of the proof of Proposition 5.1. Since $y'_h(x) = y_{h-1}(x)^2$ (see 5.) the second part of 4. follows, too.
Finally, 5. is Lemma 6.1. □

We already mentioned that $\tilde{y}_h(x) \le y_h(x)$ since $\tilde{y}_h(0) \le y_h(0) = 1$. The essential point is that these functions $\tilde{y}_h(x)$ can also be used to find upper bounds for $y_h(x)$. We present an upper bound of the following form.

Lemma 6.3. *There exists $L > 1$ and constants $C_6 > 0$, $C_7 > 0$, $C_8 > 0$ such that*

$$y_h(x) \le \tilde{y}_{h+dL\log h + C_7}(x) \quad \text{for } x \ge 1 - \alpha^{-h-C_6\sqrt{h}} \tag{20}$$

and sufficiently large $h \ge C_8$.

With help of Lemma 6.3 **Proposition 3.2** follows immediately:

Proof. (Proposition 3.2). Set $x = 1+\eta$ (with $\eta \ge 0$). By Lemma 2.2 we have

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

Hence, by Lemma 6.3

$$\frac{a_{n,h}}{n!} \le \frac{\eta}{(1+\eta)^n - 1}\alpha^{h+dL\log h + C_7}\Phi\left(-\eta\alpha^{h+dL\log h+C_7}\right).$$

Finally setting $\eta = \alpha^{-h-dL\log h}$ we directly get

$$\frac{a_{n,h}}{n!} \le \frac{1}{(1+\alpha^{-h-dL\log h})^n - 1}\alpha^{C_7}\Phi\left(-\alpha^{C_7}\right) \ll \exp\left(-n\alpha^{-h-dL\log h}\right)$$

if $h \ge c\log n - dL\log\log n$. $\square$

It remains to prove Lemma 6.3. Of course we proceed by induction. The next lemma shows that it is sufficient to prove (20) just for $x = 1-\alpha^{-h-C_6\sqrt{h}}$.

Lemma 6.4. *Suppose that*

$$y_h(x) \le \tilde{y}_{h+dL\log h+C_7}(x) \quad \text{for } x \ge 1-\alpha^{-h+C_6\sqrt{h}}$$

(for some constants $L>1$, $C_6>0$ $C_7>0$) and that

$$y_{h+1}(1-\alpha^{-h-1+C_6\sqrt{h+1}}) \le \tilde{y}_{h+1+dL\log(h+1)+C_7}(1-\alpha^{-h-1+C_6\sqrt{h+1}})$$

then

$$y_{h+1}(x) \le \tilde{y}_{h+1+dL\log(h+1)+C_7}(x) \quad \text{for } x \ge 1-\alpha^{-h-1+C_6\sqrt{h+1}}$$

Proof. Set $x_h = 1-\alpha^{-h+C_6\sqrt{h}}$. Then $x_{h+1} \ge x_h$ and we obtain by applying $y_{h+1}'(x) = y_h(x)^2$ and 4.–5. of Lemma 6.2 (for $x \ge x_{h+1}$)

$$\begin{aligned}
y_{h+1}(x) &= y_{h+1}(x_{h+1}) + \int_{x_{h+1}}^x y_h(t)^2 dt \\
&\le \tilde{y}_{h+1+dL\log(h+1)+C_7}(x_{h+1}) + \int_{x_{h+1}}^x \tilde{y}_{h+dL\log h+C_7}(t)^2\,dt \\
&\le \tilde{y}_{h+1+dL\log(h+1)+C_7}(x_{h+1}) + \int_{x_{h+1}}^x \tilde{y}_{h+dL\log(h+1)+C_7}(t)^2\,dt \\
&= \tilde{y}_{h+1+e\log(h+1)+C_7}(x_{h+1}) + \int_{x_{h+1}}^x \tilde{y}_{h+1+dL\log(h+1)+C_7}'(t)\,dt \\
&= \tilde{y}_{h+1+dL\log(h+1)+C_7}(x_{h+1}) \\
&\quad + \tilde{y}_{h+1+dL\log(h+1)+C_7}(x) - \tilde{y}_{h+1+dL\log(h+1)+C_7}(x_{h+1}) \\
&= \tilde{y}_{h+1+dL\log(h+1)+C_7}(x).
\end{aligned}$$

This completes the proof of Lemma 6.4 $\square$

Finally we show (inductively) that

$$y_h(x) \le \tilde{y}_{h+dL\log h+C_7}(x) \quad \text{for } x \ge 1-\alpha^{-h+C_6\sqrt{h}}$$

for $h < H$ implies

$$y_H(1-\alpha^{-H+C_6\sqrt{H}}) \le \tilde{y}_{H+dL\log H+C_7}(1-\alpha^{-H+C_6\sqrt{H}}), \tag{21}$$

compare with Lemma 6.8. For this purpose we need some preliminaries.

Lemma 6.5. *Suppose that $y_{h_0}(x) \le \tilde{y}_{h_0+D}(x)$ for $x \ge x_0$ and for some constant $D>0$. Then we have for $x \ge x_0$ and $h \ge h_0$*

$$y_h(x) - \tilde{y}_{h+D}(x) \le \sum_{k=0}^{h-h_0} \max\left(y_{h-k}(x_0) - \tilde{y}_{h+D-k}(x_0), 0\right)\frac{2^k(L(x)-L(x_0))^k}{k!}. \tag{22}$$

Proof. Set $\tilde{\delta}_{h,D} := y_h(x) - \tilde{y}_{h+D}(x)$. Then

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

or

$$\tilde{\delta}_{h+1,D}(x) = \tilde{\delta}_{h+1,D}(x_0) + \int_{x_0}^{x} \tilde{\delta}_{h,D}(t)(y_h(t) + \tilde{y}_{h+D}(t))\,dt.$$

Since $y_h(x) \le 1/(1-x)$ and $\tilde{y}_{h+D}(x) \le 1/(1-x)$ we obtain by induction

$$\begin{aligned}
\tilde{\delta}_{h+1,D}(x) &\le \max\left(\tilde{\delta}_{h+1,D}(x_0), 0\right) \\
&\quad + \int_{x_0}^{x} \sum_{k=0}^{h-h_0} \max\left(\tilde{\delta}_{h-k,D}(x_0), 0\right) \frac{2^k (L(t) - L(x_0))^k}{k!} (y_h(t) + \tilde{y}_{h+D}(t))\,dt \\
&\le \max\left(\tilde{\delta}_{h+1,D}(x_0), 0\right) \\
&\quad + \sum_{k=0}^{h-h_0} \max\left(\tilde{\delta}_{h-k,D}(x_0), 0\right) \int_{x_0}^{x} \frac{2^k (L(t) - L(x_0))^k}{k!} \frac{2}{1-t}\,dt \\
&= \sum_{k=0}^{h+1-h_0} \max\left(\tilde{\delta}_{h+1-k,D}(x_0), 0\right) \frac{2^k (L(x) - L(x_0))^k}{k!}.
\end{aligned}$$

□

Lemma 6.6. *If $r = \mathcal{O}(\sqrt{h})$ then*

$$\sum_{k=0}^{h} \frac{(h-r)^k}{k!} = e^{h-r}\,\Psi\left(\frac{r}{\sqrt{h-r}}\right)\left(1 + \mathcal{O}\left(h^{-1/2}\right)\right) \tag{23}$$

as $h \to \infty$, where

$$\Psi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\,dt.$$

Conversely, if $\sqrt{h} \le r \le r^{2/3}$ then we have

$$\begin{aligned}
\sum_{k=0}^{h} \frac{(h-r)^k}{k!} &= e^{h-r} - \frac{(h-r)^{h+1}}{(h+1)!}\,\frac{h}{r}\left(1 - \frac{h}{r^2} + \mathcal{O}\left(\frac{h^2}{r^4}\right)\right) \qquad (24)\\
&= e^{h-r}\left(1 - e^{-r^2/(2(h-r))}\frac{\sqrt{h-r}}{\sqrt{2\pi}\,r}\left(1 - \frac{h}{r^2} + \mathcal{O}\left(\frac{h^2}{r^4}\right) + \mathcal{O}(h^{-\frac{1}{2}})\right)\right)
\end{aligned}$$

as $h \to \infty$.

Proof. (23) follows from the Gaussian approximation of the Poisson distribution (with probabilities $p_k = e^{-(h-r)}(h-r)^k/k!$.) The second assertion (24) follows from

$$\sum_{k=0}^{h} \frac{(h-r)^k}{k!} = e^{h-r} - \sum_{k=h+1}^{\infty} \frac{(h-r)^k}{k!}$$

and from a quantified comparions of the series

$$\sum_{k=h+1}^{\infty} \frac{(h-r)^k}{k!} = \frac{(h-r)^{h+1}}{(h+1)!}\left(1 + \frac{h-r}{h+2} + \frac{(h-r)^2}{(h+2)(h+3)} + \cdots\right)$$

with the geometric series

$$\frac{(h-r)^{h+1}}{(h+1)!}\left(1 + \frac{h-r}{h} + \frac{(h-r)^2}{h^2} + \cdots\right) = \frac{(h-r)^{h+1}}{(h+1)!}\,\frac{h}{r}.$$

Finally the asymptotic representation of (24) comes from the Gaussian approximation of the Poisson distribution (and is easily checked by Stirling's approximation formula.) □

Lemma 6.7. *If $r = \mathcal{O}(h^{2/3})$ then*

$$\sum_{k=0}^{h}(h-r-k)\frac{(h-r)^k}{k!} = \frac{(h-r)^{h+1}}{h!} \tag{25}$$

$$= e^{h-r}\, e^{-r^2/(2(h-r))}\sqrt{\frac{h-r}{2\pi}}\left(1+\mathcal{O}\left(\frac{r}{h}\right)+\mathcal{O}\left(\frac{r^3}{h^2}\right)\right)$$

as $h \to \infty$.

Proof. Obviously we have

$$\sum_{k=0}^{h}(h-r-k)\frac{(h-r)^k}{k!} = \sum_{k=0}^{h}\frac{(h-r)^{k+1}}{k!} - \sum_{k=1}^{h}\frac{(h-r)^k}{(k-1)!}$$

$$= \frac{(h-r)^{h+1}}{h!}.$$

Again the asymptotic representation comes from the Gaussian approximation of the Poisson distribution. □

Now we complete the final step.

Lemma 6.8. *Suppose that $L > 1$ such that there exist $C_6 > 1$, $\gamma > 1$ with*

$$C_6\sqrt{\gamma}\left(1-\Psi\left(\frac{C_6(\sqrt{\gamma}-1)}{\sqrt{\gamma}-1}\right)\right) - \frac{\sqrt{\gamma}-1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{C_6(\sqrt{\gamma}-1)}{\sqrt{\gamma}-1}\right)^2\right) \ge \frac{C_6}{\gamma^{(L-1)/2}} \tag{26}$$

then there exists a constant $C_7 > 0$ such that the following holds. If

$$y_h(x) \le \tilde{y}_{h+dL\log h+C_7}(x) \quad \textit{for } x \ge 1-\alpha^{-h+C_6\sqrt{h}} \tag{27}$$

for all $h < H$ then

$$y_H(1-\alpha^{-H+C_6\sqrt{H}}) \le \tilde{y}_{H+dL\log H+C_7}(1-\alpha^{-H+C_6\sqrt{H}}). \tag{28}$$

Remark . We can use $C_6 = \gamma = 2$ and $L = 6.6$, for example.

Proof. Let us suppose that H ($\ge H_0$) is sufficiently large (H_0 will be specified in the sequel) and that (27) is satisfied. Furthermore, let h_0 be given by $h_0 = [H/\gamma]$. Then by Lemma 6.5

$$y_H(x) - \tilde{y}_{H+D}(x) \le \sum_{k=0}^{H-h_0}\max\left(y_{H-k}(x_0)-\tilde{y}_{H+D-k}(x_0),0\right)\frac{2^k(L(x)-L(x_0))^k}{k!},$$

where $D = dL\log(h_0+r') + C_7$, $r' = C_6\sqrt{h_0}$, and $x_0 = 1-\alpha^{h_0}$. By (27) we have

$$y_{H-k}(x_0) \le \tilde{y}_{H+D-k}(x_0)$$

if $k \ge H-h_0-r'$ and by 2. of Lemma 6.2

$$\max\left(y_{H-k}(x_0)-\tilde{y}_{H+D-k}(x_0),0\right) \le \alpha^{h_0}\frac{C_5}{c}(H-h_0+D-k)\left(\frac{2}{c}\right)^{H-h_0+D-k}(1+o(1))$$

for $k < H-h_0-r'$. Furthermore, if we set $x' = 1-\alpha^{-H+r}$, where $r = C_6\sqrt{H}$, we have

$$2^k(L(x')-L(x_0))^k = \left(\frac{2}{c}\right)^k(H-h_0-r)^k$$

and consequently

$$y_H(x') - \tilde{y}_{H+D}(x') \le \alpha^{h_0}\left(\frac{2}{c}\right)^{H-h_0+D}\sum_{k=0}^{H-h_0-r'}\frac{C_5}{c}(H-h_0+D-k)\frac{(H-h_0-r)^k}{k!}(1+o(1)).$$

Now

$$\sum_{k=0}^{H-h_0-r'} (H-h_0+D-k)\frac{(H-h_0-r)^k}{k!} = \sum_{k=0}^{H-h_0-r'} (H-h_0-r-k)\frac{(H-h_0-r)^k}{k!} + (r+D)\sum_{k=0}^{H-h_0-r'} \frac{(H-h_0-r)^k}{k!}$$

and by applying Lemma 6.6 and 6.7 (with $H-h_0-r'$ instead of h and $r-r'$ instead of r) we obtain

$$\begin{aligned}\sum_{k=0}^{H-h_0-r'} &(H-h_0+D-k)\frac{(H-h_0-r)^k}{k!}(1+o(1)) \\ &= e^{H-h_0-r}(r+D)\Psi\left(\frac{r-r'}{\sqrt{H-h_0-r}}\right)(1+o(1)) \\ &\quad + e^{H-h_0-r}\exp\left(-\frac{(r-r')^2}{2(H-h_0-r)}\right)\sqrt{\frac{H-h_0-r}{2\pi}}(1+o(1))\end{aligned}$$

or (by using $\alpha = 2e/c$)

$$\begin{aligned}y_H(x') - \tilde{y}_{H+D}(x') &\le \alpha^{H-r}\frac{C_5}{c}\left(\frac{2}{c}\right)^{r+D} \\ &\times \left((r+D)\Psi\left(\frac{r-r'}{\sqrt{H-h_0-r}}\right) + \exp\left(-\frac{(r-r')^2}{2(H-h_0-r)}\right)\sqrt{\frac{H-h_0-r}{2\pi}}\right)(1+o(1))\end{aligned}$$

Furthermore by 2. of Lemma 6.2

$$\tilde{y}_{H+D}(x') = \alpha^{H-r}\left(1 - \frac{C_5}{c}(r+D)\left(\frac{2}{c}\right)^{r+D}(1+o(1))\right).$$

Hence, for sufficiently large $H \ge H_0$

$$\begin{aligned}y_H(x') &= (y_H(x') - \tilde{y}_{H+D}(x')) + \tilde{y}_{H+D}(x') \\ &\le \alpha^{H-r} - \alpha^{H-r}\left(\frac{2}{c}\right)^{r+D}\frac{C_5}{c}\sqrt{h_0} \\ &\quad \times \left(C_6\sqrt{\gamma}\left(1 - \Psi\left(\frac{C_6(\sqrt{\gamma}-1)}{\sqrt{\gamma-1}}\right)\right) - \frac{\sqrt{\gamma-1}}{\sqrt{2\pi}}\exp\left(-\frac{C_6^2(\sqrt{\gamma}-1)^2}{2(\gamma-1)}\right)\right)(1+o(1)).\end{aligned}$$

By 2. of Lemma 6.2 and by

$$(dL\log H + C_7) - D = dL\log\gamma\cdot(1+o(1)) \quad \text{and} \quad \left(\frac{2}{c}\right)^{d\log x} = \frac{1}{\sqrt{x}}$$

we obtain

$$\begin{aligned}\tilde{y}_{H+dL\log H+C_7}(x') &= \alpha^{H-r}\left(1 - \frac{C_5}{c}(r+dL\log H+C_7)\left(\frac{2}{c}\right)^{r+dL\log H+C_7}(1+o(1))\right) \\ &= \alpha^{H-r} - \alpha^{H-r}\left(\frac{2}{c}\right)^{r+D}\frac{C_5}{c}\sqrt{\gamma h_0}\frac{C_6}{\gamma^{L/2}}(1+o(1)).\end{aligned}$$

Since we have assumed that (26) holds it follows that $y_H(x') \le \tilde{y}_{H+dL\log H+C_7}(x')$ if $H \ge H_0$ is sufficiently large. Finally we can choose C_7 sufficiently large such that the assertion of Lemma 6.8 holds for $H < H_0$, too. □

7. What remains?

In Theorem 1.1 we proved very precise bounds for the expected value $\mathbf{E}H_n$. However there remains a gap. It is not known which constant $L \ge 1$ is optimal. In this section we collect some clues and facts which indicate that it should be possible to choose $L = 1$. Of course then we would have

$$\mathbf{E}H_n = c\log n - \frac{c}{2(c-1)}\log\log n + \mathcal{O}(1). \tag{29}$$

and all centralized moments would be bounded.

We will first state a conjecture which implies (29) and present a *wrong proof* of the conjecture. However it seems that this *proof* is not far from reality. Finally we show that modified functions $\overline{y}_h(x)$ actually satisfy the estimate of the conjecture.

Conjecture. *There exists a constant $C > 0$ such that*

$$y_h(x) \le \tilde{y}_{h+d\log h+C}(x) \tag{30}$$

for $x \ge 1-\alpha^{-h-d\log h}$.

It is clear that this conjecture implies (29) and consequently that all centralized moments of H_n are bounded. (We can choose $L = 1$ in the proof of Proposition 3.2. So Theorems 1.1 and 1.2 hold for $L = 1$.) It would be also sufficient to find a constant $C > 0$ and a sequence $0 < x_h \le 1$ with $x_h \le x_{h+1}$ such that

$$y_h(x_h) \le \tilde{y}_{h+d\log h+C}(x_h).$$

By an analogue of Lemma 6.4 it then follows that $y_h(x) \le \tilde{y}_{h+d\log h+C'}(x)$ for $x \ge x_h$ (maybe with a larger constant C' in order to ensure $y_{h_0}(x) \le \tilde{y}_{h_0+d\log h_0+C'}(x)$ for $x \ge x_{h_0}$ as an initial condition).

Note that the above conjecture is equivalent to the statement that the lower bound for $y_h(x)$ (presented in Lemma 4.1)

$$y_h(x) \ge \frac{1}{1-x} - 2^h \sum_{k>h} \frac{L(x)^k}{k!}. \tag{31}$$

is tight for $x \le 1-\alpha^{-n-d\log n}$. Especially for $x = 1-\alpha^{-n-d\log n}$ we have

$$y_h(x) \ge \alpha^{h+d\log h} - \alpha^{h+d\log h}\frac{1}{(c-1)\sqrt{2\pi}}(1+o(1)).$$

This implies an lower bound for

$$y_h(1) \ge y_h(1-\alpha^{-h-d\log h}) \gg \alpha^{h+d\log h}$$

which meets the conjectural upper bound

$$y_h(1) \le \tilde{y}_{h+d\log h+C}(1) = \alpha^{h+d\log h+C} \ll \alpha^{h+d\log h}.$$

The lower bound (31) also shows that (30) cannot hold for $x \ge 1-\alpha^{-h+C\sqrt{h}}$. For $1-\alpha^{-h+C\sqrt{h}} \le x \le 1-\alpha^{-h-d\log h}$ the function $y_{h+d\log h+C}(x)$ is surely smaller than $y_h(x)$.

Next we try to *prove* the Conjecture by induction. Unfortunately it is impossible to verify the initial condition.

A Wrong Proof of the Conjecture. First of all it follows by induction that

$$y_{h+1}(x) - y_h(x) \le \frac{1}{2}\frac{(2L(x))^{h+1}}{(h+1)!} \tag{32}$$

since $y_1(x) - y_0(x) = x \le L(x)$ and $(y_{h+1}(x) - y_h(x))' \le 2(y_h(x) - y_{h-1}(x))/(1-x)$.

Now set $x_h = 1-\alpha^{-\overline{h}-d\log h+r}$, where r will be chosen in the sequel, and suppose that

$$y_h(x) \le \tilde{y}_{h+d\log h+C}(x) \quad \text{for } x \ge x_h$$

for some h. Hence

$$\begin{aligned}
y_{h+1}(x_{h+1}) &= (y_{h+1}(x_{h+1}) - y_h(x_{h+1})) + y_h(x_{h+1}) \\
&\le \frac{1}{2}\frac{(2L(x_{h+1}))^{h+1}}{(h+1)!} + \tilde{y}_{h+d\log h+C}(x_{h+1}) \\
&= \frac{1}{2\sqrt{2\pi}}\,\alpha^{h+1+d\log(h+1)-r}\left(\frac{2}{c}\right)^r(1+o(1)) \\
&\quad + \alpha^{h+1+d\log(h+1)-r}\left(1-\frac{C_5}{c}(r+C-1)\left(\frac{2}{c}\right)^{r+C-1}(1+o(1))\right) \\
&= \alpha^{h+1+d\log(h+1)-r} \\
&\quad\times\left(1-\left(\frac{C_5}{2}(r+C-1)-\frac{1}{2\sqrt{2\pi}}\left(\frac{c}{2}\right)^C\right)\left(\frac{2}{c}\right)^{r+C}(1+o(1))\right).
\end{aligned}$$

Since

$$\tilde{y}_{h+1+d\log(h+1)+C}(x_{h+1}) = \alpha^{h+1+d\log(h+1)-r}\left(1-\frac{C_5}{c}(r+C)\left(\frac{2}{c}\right)^{r+C}(1+o(1))\right)$$

it is possible *to choose the constant* r sufficiently large such that

$$y_{h+1}(x_{h+1}) \le \tilde{y}_{h+1+d\log(h+1)+C}(x_{h+1}).$$

Finally an analogon to Lemma 6.4 proves $y_{h+1}(x) \le \tilde{y}_{h+1+d\log(h+1)+C}(x)$ for $x \ge x_{h+1}$. □

As already mentioned the *flaw* in this proof is that if we choose r sufficiently large (depending on C) we would get an upper bound for $y_h(x)$ which contradics the lower bound (31).

Nevertheless there is some hope that a proof of this kind will succeed. The point is to find $h_0 > 0$, $C > 0$, $0 < C' < \frac{1}{2}$, $r > 0$ (empirically) such that

$$\begin{aligned}
y_{h_0}(x) &\le \tilde{y}_{h_0+d\log h_0+C}(x) \quad \text{for } x \ge 1-\alpha^{-h_0-d\log h_0+r}, \\
y_{h_0+1}(x) - y_{h_0}(x) &\le C'\frac{(2L(x))^{h_0+1}}{(h_0+1)!} \qquad \text{for } 0 \le x \le 1, \text{ and} \\
\frac{C_5}{2}(r+C-1) &\ge \frac{C'}{\sqrt{2\pi}}\left(\frac{c}{2}\right)^C + \frac{C_5}{c}(r+C)+\varepsilon,
\end{aligned}$$

where $\varepsilon > 0$ has to chosen in a way that all error terms of the form $(1+o(1))$ are covered by it for $h \ge h_0$. (It was beyond the author's ability in numerics to settle this question.)

Finally we want to present another sequence of functions $\overline{y}_h(x)$ with $\tilde{y}_h(x) \le \overline{y}_h(x) \le y_h(x)$ satisfying an analogue of the above conjecture.

Set $\overline{y}_0(x) \equiv 0$ and define $\overline{y}_h(x)$ for $h > 0$ inductively by

$$\overline{y}'_{h+1}(x) = \overline{y}_h(x)^2 \quad \text{with} \quad \overline{y}_h(0) = 1-\left(\frac{2}{c}\right)^h.$$

Here it is possible to prove the following upper bound.

Lemma 7.1. *There exist constants $C_9, C_{10} > 0$ such that*

$$\overline{y}_h(x) \le \tilde{y}_{h+d\log h+C_9}(x) \tag{33}$$

for $x \ge 1-\alpha^{-h+C_{10}\sqrt{h}}$.

Proof. Similarly to Lemma 6.5 we have

$$\overline{y}_h(x) - \tilde{y}_{h+D}(x) \le \sum_{k=0}^{h}\max\left(\overline{y}_{h-k}(0)-\tilde{y}_{h+D-k}(0),0\right)\frac{(2L(x))^k}{k!}. \tag{34}$$

Set $D = d\log h$ and $x = x_h = 1-\alpha^{-h+C_{10}\sqrt{h}}$. ($C_{10} > 0$ will be chosen in a moment.) Now there exists a constant $C_{11} > 0$ such that for sufficiently large h

$$\overline{y}_{h-k}(0) = 1 - \left(\frac{2}{c}\right)^{h-k} < \tilde{y}_{h-k}(0) = 1 - \frac{C_5}{c}\,\frac{h-k+D}{\sqrt{h}}\left(\frac{2}{c}\right)^{h-k}(1+o(1))$$

if $k \ge h - C_{11}\sqrt{h}$ and

$$\max\left(\overline{y}_{h-k}(0) - \tilde{y}_{h-k}(0), 0\right) \le \frac{C_5}{c}\,\frac{h-k+D}{\sqrt{h}}\left(\frac{2}{c}\right)^{h-k}$$

if $k < h - C_{11}\sqrt{h}$. Hence, we have

$$\begin{aligned}\overline{y}_h(x_h) &= (\overline{y}_h(x_h) - \tilde{y}_{h+D}(x_h)) + \tilde{y}_{h+D}(x_h)\\ &\le \sum_{k=0}^{h-C_{11}\sqrt{h}} \frac{C_5}{c}\,\frac{h-k+D}{\sqrt{h}}\left(\frac{2}{c}\right)^{h-k}\frac{(2(h-C_{10}\sqrt{h})^k}{c^k\,k!}\\ &\quad + \alpha^{h-C_{10}\sqrt{h}}\left(1 - \frac{C_5}{c}\,\frac{C_{10}\sqrt{h}+D}{\sqrt{h}}\left(\frac{2}{c}\right)^{C_{10}\sqrt{h}}(1+o(1))\right).\end{aligned}$$

Now similar considerations as in the proof of Lemma 6.8 (and especially by an application of the second part of (24)) we get

$$\begin{aligned}\overline{y}_h(x_h) &\le \alpha^{h-C_{10}\sqrt{h}} - \alpha^{h-C_{10}\sqrt{h}}\\ &\quad\times \frac{C_5}{c\sqrt{2\pi}}e^{\frac{1}{2}(C_{10}-C_{11})^2}\left(\frac{C_{10}}{C_{10}-C_{11}}\left(1+\mathcal{O}\left(\frac{1}{C_{10}-C_{11}}\right)\right) - 1 + o(1)\right)\left(\frac{2}{c}\right)^r.\end{aligned}$$

Now choose $C_{10} > 0$ sufficiently large such that there exists $C_{12} > 0$ with

$$\frac{C_{10}}{C_{10}-C_{11}}\left(1+\mathcal{O}\left(\frac{1}{C_{10}-C_{11}}\right)\right) - 1 + o(1) \ge C_{12} > 0.$$

Finally choose $C_9 > 0$ such that

$$C_{10}\left(\frac{2}{c}\right)^{C_9} < \frac{1}{\sqrt{2\pi}}e^{\frac{1}{2}(C_{10}-C_{11})^2}\,C_{12}.$$

Thus, for sufficiently large h we have

$$\begin{aligned}\overline{y}_h(x_h) &\le \tilde{y}_{h+d\log h+C_9}(x_h)\\ &= \alpha^{h-C_{10}\sqrt{h}}\left(1 - \frac{C_5}{c}\,\frac{C_{10}\sqrt{h}+d\log h+C_9}{\sqrt{h}}\left(\frac{2}{c}\right)^{r+C_9}(1+o(1))\right).\end{aligned}$$

Finally, an analogon to Lemma 6.4 completes the proof of Lemma 7.1. □

Now the question is to estimate $y_h(x) - \overline{y}_h(x)$. However, ot seems that the estimates presented in this paper are not sufficient to prove the Conjecture via Lemma 7.1 but maybe an additional idea will help.

Nevertheless, at least from the author's point of view, there is some chance to settle the above problem (whether L equals 1) via methods which are similar to the presented ones.

Acknowledgement. I heartly want to thank Philippe Flajolet for his selfless and inspiring work. Without him the present and several other papers of mine would not exist.

References

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

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

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

[4] M. Drmota and M. Soria, *Marking in combinatorial constructions: Generating functions and limiting distributions*, Theor. Comput. Sc. **144** (1995), 67–99.

[5] X. Gourdon, unpublished manuscript, 1995.

[6] P. Jacquet and W. Szpankowski, *Aysmptotic behaviour of the Lempel-Ziv parsing scheme and digital search trees*, Theor. Computer Science **144** (1995), 161–197.

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

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

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

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

[11] J. M. Robson, personal communication.

8. Appendix

In what follows we present a proof of **Lemma 4.2**.

Proof. If $k = \mathcal{O}(\log n)$ then we can use [4, Theorem 6] saying that

$$[x^n]L(x)^k = \frac{\log^k n}{n\Gamma(\frac{k}{\log n})} \exp\left(\frac{k}{\log n}\right)\left(1 + \mathcal{O}\left(k\left(\frac{\log\log n}{\log n}\right)^2\right)\right). \tag{35}$$

For the reader's convenience we repeat the proof. We will use Cauchy's formula

$$[x^n]L(x)^k = \frac{1}{2\pi i}\int_\gamma \frac{L(x)^k}{z^{n+1}}\,dz$$

for the following path of integration $\gamma = \gamma_1 \cup \gamma_2$:

$$\gamma_1 = \left\{x = 1 + \frac{t}{n} : |t| = 1, \Re t \le 0, \text{ or } 0 < \Re t \le \log^2 n, \Im t = \pm 1\right\}$$

$$\gamma_2 = \left\{x : |x| = \left|1 + \frac{\log^2 n + i}{n}\right|, \arg\left(1 + \frac{\log^2 n + i}{n}\right) \le |\arg(x)| \le \pi\right\}.$$

Then we have for $x \in \gamma_1$

$$\begin{aligned} L(x)^k x^{-(n+1)} &= (\log n)^k \left(1 - \frac{\log(-t)}{\log n} + \frac{1}{\log n} + \mathcal{O}(\log^{-2} n)\right)^k \left(1 + \frac{t}{n}\right)^{-n-1} \\ &= (\log n)^k e^{-t}(-t)^{-\frac{k}{\log n}} \exp\left(\frac{k}{\log n}\right) \cdot \\ &\quad \cdot \left(1 + \mathcal{O}\left(\frac{|t|^2}{n}\right) + \mathcal{O}\left(k\left(\frac{\log(|t|+1)}{\log n}\right)^2\right)\right) \end{aligned}$$

Since

$$\frac{1}{2\pi i}\int_{\gamma'} e^{-t}(-t)^{-\frac{k}{\log n}}\,dt = \frac{1}{\Gamma(\frac{k}{\log n})} + \mathcal{O}(e^{-\log^2 n}),$$

where $\gamma' = \{t \mid |t| = 1, \Re t \le 0, \text{ or } 0 < \Re t \le \log^2 n, \Im t = \pm 1\}$, and

$$\int_{\gamma_2} \frac{L(x)^k}{x^{n+1}}\,dx = \mathcal{O}(e^{-\log^2 n})$$

we directly get (35).

In the case $k \ge C\log n$ (where $C > 0$ is a fixed constant) we use again Cauchy's formula but the following path of integration $\gamma = \gamma_1 \cup \gamma_2$:

$$\gamma_1 = \left\{ x = 1 - \frac{1}{n} + i\frac{t}{n} : -\sqrt{2n-1} \le t \le \sqrt{2n-1} \right\}$$
$$\gamma_2 = \left\{ x : |x| = 1, \arg\left(1 - \frac{1}{n} + i\frac{\sqrt{2n-1}}{n}\right) \le |\arg(x)| \le \pi \right\}.$$

For $x \in \gamma_1$ we have (for some constant $c' > 0$)

$$\begin{aligned} |L(x)^k| &\le (\log n)^k \left(1 - c\frac{|t|}{\log n}\right)^k \\ &\le (\log n)^k e^{-c'|t|k/\log n} \\ &\le (\log n)^k e^{-c'C|t|}. \end{aligned}$$

and

$$x^{-n} = \mathcal{O}(1).$$

Hence

$$\int_{\gamma_1} L(x)^k x^{-n-1}\,dx = \mathcal{O}\left(\frac{(\log n)^k}{n}\right).$$

Finally, for $x \in \gamma_2$ we trivially have

$$L(x)^k = \mathcal{O}\left(\frac{(\log n)^k}{n}\right) \quad \text{and} \quad |x| = 1 + \mathcal{O}\left(\frac{(\log n)^2}{n}\right),$$

which completes the proof of Lemma 4.2 □

Remark . Since

$$[x^n]L(x)^k = \frac{n!}{k!}|s_{n,k}|,$$

where $s_{n,k}$ denote the Stirling numbers of the first kind, Lemma 4.2 provides a uniform upper bound for theses Stirling numbers. This estimate is only best possible if k is of order $\log n$. Note that there exists an implicit but full asymptotic expansion of $s_{n,k}$ provided by Moser and Wyman [8]. Therefore it should be possible to extract the upper bound (9) by solving a specific implicit function asymptotically. However, we feel that it is appropriate to provide a direct proof of (9).