THE VARIANCE OF THE HEIGHT OF DIGITAL SEARCH TREES

MICHAEL DRMOTA

Abstract. By using analytic tools it is shown that the variance $\mathbf{E}(H_n - \mathbf{E}H_n)^2$ of the height H_n of digital search trees of size n is bounded.

1. Introduction

Digital search trees are used to store data which are labeled with a (possible) infinite string of symbols from a finite alphabet (for simplicity we work only with the binary alphabet $\{0,1\}$). Suppose that we have n data of that kind. The empty string is stored in the root, while the first item occupies the right or left child of the root depending whether its first symbol is "1" or "0". The remaining items are always stored in the next available node according to the rule that we move to the right if the next symbol is "1" and we move to the left if the next symbol is "0". In the same way can also search for a specific item. (For more details see [8] and [9].)

It is usual to assume that the 0-1-strings follow the Bernoulli model with probability $p = \frac{1}{2}$. With respect to this probabilistic model many parameters of digital search trees have been very well studied, compare (again) with [8] and [9].

In this paper we want to study the height H_n^{DST} of digital search trees. Is was proved by Aldous and Shields [1] that

$$H_n^{DST} = \log_2 n + \sqrt{2\log_2 n} \cdot (1 + o(1)) \tag{1}$$

in probability as $n \to \infty$, where $\log_2$ denotes the binary logarithm. There is also a very interesting paper by Knessl and Szpankowski [6] The authors of [6] used an applied math. method called 'WKB-method' that assumes the existance of certain asymptotic expansions for $\mathbf{P}[H_n^{DST} \le h]$. They provide a perspective of the exact limiting behaviour of the distribution of H_n^{DST} and of the expected value.

The purpose of this paper is to prove the following theorem.

Theorem 1. *Let $G_k(x) = \sum_{n\ge 0} \mathbf{P}[H_n^{DST} \le k]\, x^n/n!$ be the generating functions of the distribution of the height H_n^{DST} of digital search trees of size n and let n_k be defined by*

$$G_k(n_k) = e^{n_k - 1}. \tag{2}$$

Then the expected value of H_n^{DST} is given by

$$\mathbf{E}H_n^{DST} = \max\{k : n_k \le n\} + \mathcal{O}(1) \qquad (n \to \infty). \tag{3}$$

and all centralized moments are bounded:

$$\mathbf{E}|H_n^{DST} - \mathbf{E}H_n^{DST}|^L = \mathcal{O}(1) \tag{4}$$

for all $L \ge 0$ as $n \to \infty$.

Date: October 30, 2000.

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

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

It should be noted that the sequence n_k defined implicitly in Theorem 1 is well defined, which will be shown in what follows. We want to note that the methods presented here are too weak to prove directly exact asymptotics of n_k which would lead to asymptotic expansions of $\mathbf{E}H_n^{DST}$. So in some sense it is easier to prove concentration properties like (4) than to prove asymptotic expansions for the expected value. This *phenomonon* was also observed for binary search trees, compare with [3]

However, (3) implies

$$H_n^{DST} = \max\{k : n_k \le n\} + \mathcal{O}(r(n))$$

in probabilty as $n \to \infty$, where $r(n)$ is any positive real sequence with $r(n) \to \infty$ (as $n \to \infty$). By comparing this with the above stated result (1) of Aldous and Shields [1] we get as an

Corollary .

$$\mathbf{E}H_n^{DST} = \log_2 n + \sqrt{2\log_2 n} \cdot (1 + o(1)). \tag{5}$$

The paper is organized in the following way. In Section 2 we collect some basic properties of the generating functions $G_k(x)$ which will be used to prove (3) and (4) in Section 3. (One more involved proof of an auxiliary result is postponed to Section 4.)

We also want to point out the analytic structure for generating functions is quite simliar for tries, in fact the case of tries is much simpler. Therefore we will present a short proof of corresponding (but well known) properties for tries in a final Section 5.

2. Basic Properties

Let $H_n = H_n^{DST}$ denote the random variable which describes the height of digital search trees of size n and set

$$G_k(x) := \sum_{n\ge 0} \mathbf{P}[H_n \le k]\,\frac{x^n}{n!} \tag{6}$$

Then it is known (see [5, 6]) that these generating functions (which are in fact polynomials) satisfy the recurrent relation

$$G'_{k+1}(x) = G_k(x/2)^2 \tag{7}$$

with $G_0(x) \equiv 1 + x$ and $G_k(0) = 1$ (for $k \ge 1$).

First we collect some properties of $G_k(x)$.

Lemma 1. *For $x > 0$ we have*

1. $\displaystyle\sum_{n=0}^{k} \frac{x^n}{n!} \le G_k(x) \le e^x$.
2. $G_{k+1}(x) > G_k(x)$.
3. $G_k(x) > G'_k(x)$.
4. $G_k(x) - G'_k(x) \le C\dfrac{e^x}{\sqrt{x}}$ *with a uniform constant $C > 0$.*

Proof. For $n \le k$ we trivially have $\mathbf{P}[H_n \le k] = 1$ and in general $0 \le \mathbf{P}[H_n \le k] \le 1$. This proves the first property.

The second property follows from the trivial inequality $\mathbf{P}[H_n \le k+1] \ge \mathbf{P}[H_n \le k]$ which is also strict for at least one n.

Since $\mathbf{P}[H_n \le k] \ge \mathbf{P}[H_{n+1} \le k]$ for all n (with strict inequality for some n) we directly get (for $x > 0$)

$$G'_k(x) = \sum_{n\ge 0} \mathbf{P}[H_{n+1} \le k]\frac{x^n}{n!} < G_k(x),$$

i.e. the third property.

Finally, we can estimate the difference between $G_k(x)$ and $G_k'(x)$ by

$$\begin{aligned}
G_k(x) - G_k'(x) &= 1 + \sum_{n\ge 1} \mathbf{P}[H_n \le k] \left(\frac{x^n}{n!} - \frac{x^{n-1}}{(n-1)!} \right) \\
&\le 1 + \sum_{n\ge 1} \left| \frac{x^n}{n!} - \frac{x^{n-1}}{(n-1)!} \right| \\
&\le 1 + 2 \max_{n\ge 1} \frac{x^n}{n!} \\
&\ll \frac{e^x}{\sqrt{x}}.
\end{aligned}$$

□

It is now convenient to introduce the auxiliary functions

$$P_k(x) := G_k(x) e^{-x}. \tag{8}$$

These functions may be considered as the *Poisson transforms* of the sequences $\mathbf{P}[H_n \le k]$, $n \ge 0$.

Lemma 2. *The functions $P_k(x)$ satisfy*

1. $0 < P_k(x) \le 1$.
2. $P_k(0) = 1$, $\lim_{x\to\infty} P_k(x) = 0$.
3. $P_k'(x) < 0$, *i.e. $P_k(x)$ is strictly decreasing.*

Proof. The first property is obvious.

By definition we have $P_k(0) = 1$. Furthermore, since $G_k(x)$ is a polynomial we surely have $\lim_{x\to\infty} P_k(x) = 0$.

Finally, we have $P_k'(x) = e^{-x}(G_k'(x) - G(x))$. By property 3 of Lemma 1 it follows that $P_k'(x) < 0$ for $x > 0$. □

It is now clear that there uniquely exist n_k defined by

$$P_k(n_k) = \frac{1}{e}. \tag{9}$$

Of course, this definition is equivalent to (2). It turns out that the crucial part of the proof of Theorem 1 is to know something about the asymptotic behaviour of this sequence n_k. The next lemma collects some first properties.

Lemma 3. *The sequence n_k satisfies*

1. $n_k < n_{k+1} < 2n_k$.
2. $2^{k-2\sqrt{k}} \le n_k \le 2^k$.

Proof. By property 2 of Lemma 1 we have

$$P_{k+1}(n_k) > P_k(n_k) = \frac{1}{e}.$$

Since $P_{k+1}(x)$ is strictly decreasing (property 3 of Lemma 2), it directly follows that $n_{k+1} > n_k$.

Now we use (7) and property 4 of Lemma 1 to get the estimate

$$\begin{aligned} G_{k+1}(2n_k) &= G'_{k+1}(2n_k) + \mathcal{O}\left(\frac{e^{2n_k}}{\sqrt{n_k}}\right) \\ &= G_k(n_k)^2 + \mathcal{O}\left(\frac{e^{2n_k}}{\sqrt{n_k}}\right) \\ &= e^{2n_k}\left(\frac{1}{e^2} + \mathcal{O}\left(\frac{1}{\sqrt{n_k}}\right)\right) \\ &< e^{2n_k}\frac{1}{e} \end{aligned}$$

for sufficiently large k. Thus

$$P_{k+1}(2n_k) < \frac{1}{e},$$

which implies that $n_{k+1} < 2n_k$. (The first few k can be checked by hand.)

The proof of the second property is a little bit more involved. Starting from the asymptotic relation

$$\begin{aligned} G_l(x) &= \sum_{n=0}^{l} \frac{x^n}{n!} + \sum_{n>l} \mathbf{P}[H_n \le k]\,\frac{x^n}{n!} \\ &= \sum_{n=0}^{l} \frac{x^n}{n!} + \mathcal{O}\left(\frac{x^{l+1}}{(l+1)!}\right), \end{aligned}$$

which is at least true for $|x| \le l/2$, we first observe that

$$P_l(x) = e^{-x}G_l(x) = 1 - \mathcal{O}\left(\frac{x^{l+1}}{(l+1)!}\right).$$

By (7) and the relation $G_k(x) \ge G'_k(x)$ it follows that $P_{k+1}(x) \ge P_k(x/2)^2$. Hence, we get for $k \ge l$ and $|x| \le l2^{k-l-1}$

$$P_k(x) \ge P_l(2^{l-k}x)^{2^{k-l}} = \left(1 - \mathcal{O}\left(\frac{x^{l+1}}{(l+1)!2^{(l+1)(k-l)}}\right)\right)^{2^{k-l}}.$$

Hence, by choosing $l = [\sqrt{k}] - 1$ and $x = 2^{k-2[\sqrt{k}]}$ one gets

$$\begin{aligned} P_k(x) &\ge 1 - \mathcal{O}\left(\frac{1}{[\sqrt{k}]!2^{[\sqrt{k}]^2}}\right)^{k-[\sqrt{k}]} \\ &\ge 1 - \mathcal{O}\left(\frac{2^{[\sqrt{k}]}}{[\sqrt{k}]!}\right) \\ &\ge \frac{1}{e} \end{aligned}$$

for sufficiently large k and consequently

$$n_k \ge 2^{k-\sqrt{k}}.$$

□

These properties suggest that $n_{k+1} \sim 2n_k$ which is actually true (compare with (3) and (5)). However, for our purpose it is sufficient to use the following growth condition.

Proposition 1. *The sequence n_k satisfies*

$$\liminf_{k\to\infty} \frac{n_{k+1}}{n_k} \ge \sqrt{2}. \tag{10}$$

The proof if Proposition 1 is the most difficult part of the proof of Theorem 1 and will postponed to section 4.

3. Centralized Moments

In order to prove (3) and (4) we make use of *tail estimates*. We start with the following easy observation.

Lemma 4. *We have*

$$\mathbf{P}[H_n \le k] \le 2P_k(n), \tag{11}$$
$$\mathbf{P}[H_n > k] \le 2(1 - P_k(n)). \tag{12}$$

Proof. By applying the monoticity property $\mathbf{P}[H_n \le k] \ge \mathbf{P}[H_{n+1} \le k]$ we obtain the inequality

$$\begin{aligned} G_k(n) &\ge \sum_{l=0}^{n} \mathbf{P}[H_l \le k]\, \frac{n^l}{l!} \\ &\ge \mathbf{P}[H_n \le k] \sum_{l=0}^{n} \frac{n^l}{l!} \\ &\ge \mathbf{P}[H_n \le k]\, \frac{e^n}{2}, \end{aligned}$$

which proves (11).

In as similar way we can prove (12):

$$\begin{aligned} e^n - G_k(n) &\ge \sum_{l \ge n} \mathbf{P}[H_l > k]\, \frac{n^l}{l!} \\ &\ge \mathbf{P}[H_n > k] \sum_{l \ge n} \frac{n^l}{l!} \\ &\ge \mathbf{P}[H_n > k]\, \frac{e^n}{2}. \end{aligned}$$

□

Lemma4 suggest that $\mathbf{P}[H_n \le k] = 2P_k(n) + o(1)$ as $n, k \to \infty$. It also seems that this is a *general* property of height distributions and might be proved by *analytic depoissonization* (see [7]).

The next step is to provide proper bounds for $P_k(x)$.

Lemma 5. *There exists a sequence c_k of positive numbers with $\lim_{k\to\infty} c_k = 1$ such that*

$$P_k(x) \ge \left(1 - \frac{c_k}{n_k}\right) e^{-c_k x/n_k} \qquad (\textit{for } x \le n_k),$$
$$P_k(x) \le \left(1 - \frac{c_k}{n_k}\right) e^{-c_k x/n_k} \qquad (\textit{for } x \ge n_k).$$

Proof. Let c_k be defined by

$$\left(1 - \frac{c_k}{n_k}\right) e^{-c_k} = \frac{1}{e}.$$

Obviously, we have $\lim_{k\to\infty} c_k = 1$. Next set

$$\tilde{G}(x) := \left(1 - \frac{c_k}{n_k}\right) e^{(1-c_k/n_k)x}.$$

By definition we have

$$\tilde{G}(n_k) = G_k(n_k) = e^{n_k - 1}$$

and

$$\tilde{G}'(x) = \tilde{G}(x/2)^2.$$

We will now show by induction that the difference

$$D_l(x) := G_l(x) - \tilde{G}(x)$$

has exactly one zero $x_l > 0$ (for $l \ge 0$).

For convenience we set $G_{-1}(x) \equiv 1$ and prove this property even for $l = -1$. (Note that the recurrence $G'_{k+1}(x) = G_k(x/2)^2$ is also true for $k = -1$.)

For

$$D_{-1}(x) = 1 - \left(1 - \frac{c_k}{n_k}\right) e^{(1-c_k/n_k)x}$$

we obviously have $D_{-1}(0) > 0$, $D'_{-1}(x) < 0$, and $\lim_{x\to\infty} D_{-1}(x) = -\infty$. Thus, there exists a unique zero $x_{-1} > 0$.

Next observe that (for $l \ge -1$)

$$D'_{l+1}(x) = \left(G_l(x/2) + \tilde{G}(x/2)\right) D_l(x/2). \tag{13}$$

Furthermore, $D_{l+1}(0) > 0$ and $\lim_{x\to\infty} D_{l+1}(x) = -\infty$. Now, by (13) $D_{l+1}(x)$ is increasing for $0 \le x \le 2x_l$ and decreasing for $x \ge 2x_l$. Consequently there exists a unique zero $x_{l+1} > 2x_l > 0$ of $D_{l+1}(x)$.

Especially, we have $D_k(n_k) = 0$, i.e. $x_k = n_k$. Thus, $D_k(x) > 0$ for $0 \le x < n_k$ and $D_k(x) < 0$ for $x > n_k$. However, this is just a reformulation of Lemma 5. □

Corollary 1. *We have*

$$P_k(n) = \mathcal{O}\left(e^{-n/(2n_k)}\right) \qquad (\textit{for } n \ge n_k),$$
$$1 - P_k(n) = \mathcal{O}\left(\frac{n}{n_k}\right) \qquad (\textit{for } n \le n_k).$$

Corollary 2. *Set* $k_0(n) := \max\{k : n_k \le n\}$ *Then for every* $L \ge 0$ *and* $\delta > 0$ *we have uniformly, as* $n \to \infty$,

$$\sum_{k:n\ge n_k} P_k(n)(k_0(n) - k + \delta)^L = \mathcal{O}(1),$$
$$\sum_{k:n\le n_k} (1 - P_k(n))(k - k_0(n) + \delta)^L = \mathcal{O}(1).$$

Proof. By definition we have $n_{k_0(n)+1} > n \ge n_{k_0(n)}$. Furthermore, Proposition 1 implies that there exist $\eta > 0$ such that $n_{k+1} \ge (1+\eta)n_k$. Hence, for k with $n \ge n_k$ we have

$$\frac{n}{n_k} \ge (1+\eta)^{k_0(n)-k-1}$$

and consequently

$$\sum_{k:n\ge n_k} P_k(n)(k_0(n) - k + \delta)^L \ll \sum_{k\le k_0(n)} e^{-\frac{1}{2}(1+\eta)^{k_0(n)-k-1}} (k_0(n) - k + \delta)^L$$
$$\le \sum_{l\ge 0} e^{-\frac{1}{2}(1+\eta)^{l-1}} (l+\delta)^L = \mathcal{O}(1).$$

Similarly, we have for k with $n \le n_k$

$$\frac{n}{n_k} \le (1+\eta)^{-(k_0(n)-k)}$$

which leads to

$$\sum_{k:n\le n_k} (1 - P_k(n))(k - k_0(n) + \delta)^L \ll \sum_{k\ge k_0(n)} (1+\eta)^{-(k_0(n)-k)}(k_0(n) - k + \delta)^L$$
$$\le \sum_{l\ge 0} (1+\eta)^{-l}(l+\delta)^L = \mathcal{O}(1).$$

This completes the proof of Corollary 2. □

These estimates are sufficient to prove the proposed expansion (3) for the expected value $\mathbf{E}H_n$. By applying Lemma 4 and Corollary 2 (with $L=0$) we derive

$$\begin{aligned}
\mathbf{E}H_n &= \sum_{k\ge 0}(1-\mathbf{P}[H_n\le k]) \\
&= \sum_{k:n\ge n_k}(1-\mathbf{P}[H_n\le k]) + \sum_{k:n<n_k}\mathbf{P}[H_n>k] \\
&= \max\{k : n_k\le n\} + \mathcal{O}(1) - \sum_{h:n\ge n_k}\mathbf{P}[H_n\le k] + \sum_{h:n<n_k}\mathbf{P}[H_n>k] \\
&= \max\{k : n_k\le n\} + \mathcal{O}(1) = k_0(n) + \mathcal{O}(1).
\end{aligned}$$

Similarly, we can treat the case of centralized moments. Without loss of generality we can assume that $L\ge 1$ is an integer. Here we get (with $\delta = |\mathbf{E}H_n - k_0(n)| + 1$)

$$\begin{aligned}
\mathbf{E}|H_n-\mathbf{E}H_n|^L &\ll \sum_{k\le \mathbf{E}H_n}\mathbf{P}[H_n\le k]\,(\mathbf{E}H_n - k + 1)^{L-1} \\
&\quad + \sum_{k\ge \mathbf{E}H_n}\mathbf{P}[H_n\ge k]\,(k-\mathbf{E}H_n+1)^{L-1} \\
&\ll \sum_{k\le k_0(n)} P_k(n)\,(k_0(n)-k+\delta)^{L-1} \\
&\quad + \sum_{k\ge k_0(n)}(1-P_k(n))\,(k-k_0(n)+\delta)^{L-1} \\
&\quad + \mathcal{O}(1) = \mathcal{O}(1).
\end{aligned}$$

This completes the proof of (4).

4. Proof of Proposition 1

In the proof of Lemma 5 we have used the fact that a function $\tilde{G}(x)$ of the kind

$$\tilde{G}(x) = (1-\gamma)e^{(1-\gamma)x}$$

satisfies

$$\tilde{G}'(x) = \tilde{G}(x/2)^2.$$

This fixed point property was used to prove Lemma 5 by induction. Unfortunately, the estimates of Lemma 5 are not sufficient to prove Proposition 1 or at least an estimate of the kind $n_{k+1}\ge (1+\eta)n_k$ (for some $\eta>0$). The reason is the following one.

By definition we have $P_{k+1}(n_{k+1}) = e^{-1}$ and consequently

$$P_k(n_{k+1}/2) = e^{-1/2} + \mathcal{O}\left(n_k^{-1/2}\right) \sim e^{-1/2}.$$

Hence, if we know for some x_0 that $P_k(x_0)\ge e^{-1/2}$ it follows that $x_0\le n_{k+1}/2$.

The problem with the estimate of Lemma 5 is that we only get

$$P_k(n_k/2) \ge \left(1-\frac{c_k}{n_k}\right)e^{-c_k/2} \sim e^{-1/2}$$

and observe the trivial inequality $n_k/2\le n_{k+1}/2$. So we need tighter bounds for $P_k(x)$ resp. $G_k(x)$.

For this purpose we use a system of auxiliary functions. First of all let m_k be defined by the relation

$$\left(1+\frac{n_k}{m_k}\right)^{m_k-1} = e^{n_k-1}. \tag{14}$$

It is easy to see that m_k is well defined and of size

$$m_k = \frac{n_k^2}{2} + \mathcal{O}(n_k).$$

Now set (for $0 \le l \le k$)

$$\tilde{G}_{k,l}(x) := c_{k,l}\left(1 + \frac{x}{2^{l-k}m_k}\right)^{2^{l-k}m_k - 1}, \tag{15}$$

where $c_{k,k} := 1$ and

$$c_{k,l} := \left(1 - \frac{1}{m_k}\right)^{\frac{1}{2^{k-l}}}\left(1 - \frac{2}{m_k}\right)^{\frac{1}{2^{k-l-1}}} \cdots \left(1 - \frac{2^{k-l-1}}{m_k}\right)^{\frac{1}{2}}. \tag{16}$$

for $0 \le l < k$.

Lemma 6. *The functions $\tilde{G}_{k,l}(x)$ $(0 \le l \le k)$ satisfy*

1. $0 \le \tilde{G}_{k,l}(x) \le e^x$.
2. $0 < \tilde{G}_{k,l}(0) \le 1$.
3. $\tilde{G}'_{k,l+1}(x) = \tilde{G}_{k,l}(x/2)^2$.
4. $\tilde{G}_{k,k}(n_k) = e^{n_k - 1} = G_k(n_k)$.

Proof. By definition we have $0 \le c_{k,l} \le 1$. Thus, it immediately follows that $0 \le \tilde{G}_{k,l}(x) \le e^x$ and that $0 < \tilde{G}_{k,l}(0) \le 1$.

Next observe that by definition

$$c_{k,l}^2 = c_{k,l+1}\left(1 - \frac{2^{k-l-1}}{m_k}\right).$$

Hence, we get

$$\begin{aligned}
\tilde{G}'_{k,l+1}(x) &= c_{k,l+1}\frac{2^{l+1-k}m_k - 1}{2^{l+1-k}m_k}\left(1 + \frac{x}{2^{l+1-k}}\right)^{2^{l+1-k}m_k - 2} \\
&= c_{k,l+1}\left(1 - \frac{2^{k-l-1}}{m_k}\right)\left(1 + \frac{x/2}{2^{l-k}}\right)^{2(2^{l-k}m_k - 1)} \\
&= \tilde{G}_{k,l}(x/2)^2.
\end{aligned}$$

Finally, by definition ((14) and (15)) we trivially have $\tilde{G}_{k,k}(n_k) = e^{n_k - 1}$. $\square$

This properties show that these functions $\tilde{G}_l(x)$ have almost the same properties as the original functions $G_l(x)$. The next lemma shows that these functions can also be used to get bounds for $G_k(x)$ which are tight enough to prove Proposition 1.

Lemma 7. *We have*

$$\begin{aligned}
G_k(x) &\ge \tilde{G}_{k,k}(x) \qquad (\textit{for } 0 \le x \le n_k), \\
G_k(x) &\le \tilde{G}_{k,k}(x) \qquad (\textit{for } x \ge n_k).
\end{aligned}$$

Proof. The proof is almost the same as that of Lemma 5. We will now show by induction that the differences

$$\Delta_l(x) := G_l(x) - \tilde{G}_{k,l}(x)$$

have exactly one zero $x_l > 0$ (for $0 \le l \le k$).

For convenience we again extend the range to $l = -1$, set $G_{-1}(x) \equiv 1$ and start by proving this property for $l = -1$. (We only have to define $\tilde{G}_{k,-1}$ by $\tilde{G}_{k,-1}(x/2)^2 = \tilde{G}_{k,0}(x)$. Note that $0 < \tilde{G}_{k,-1}(0) < 1$ and that $\tilde{G}_{k,-1}(x)$ is strictly increasing to infinity.)

For

$$\Delta_{-1}(x) = G_{-1}(x) - \tilde{G}_{k,-1}(x)$$

we hence have $\Delta_{-1}(0) > 0$, $\Delta'_{-1}(x) < 0$, and $\lim\limits_{x\to\infty} \Delta_{-1}(x) = -\infty$. Thus, there exists a unique zero $x_{-1} > 0$.

Next observe that (for $l \ge -1$)

$$\Delta'_{l+1}(x) = \left(G_l(x/2) + \tilde{G}_{k,l}(x/2)\right) \Delta_l(x/2). \tag{17}$$

Furthermore, $\Delta_{l+1}(0) > 0$ and $\lim\limits_{x\to\infty} \Delta_{l+1}(x) = -\infty$. Now, by (13) $\Delta_{l+1}(x)$ is increasing for $0 \le x \le 2x_l$ and decreasing for $x \ge 2x_l$. Consequently there exists a unique zero $x_{l+1} > 2x_l > 0$ of $\Delta_{l+1}(x)$.

Especially, we have $\Delta_k(n_k) = 0$, i.e. $x_k = n_k$. Thus, $\Delta_k(x) > 0$ for $0 \le x < n_k$ and $\Delta_k(x) < 0$ for $x > n_k$, which completes the proof of the Lemma. □

So it is easy to complete the proof of Proposition 1.

Proof. (Proposition 1) By definition and the relation $m_k = n_k^2/2 + \mathcal{O}(n_k)$ we have

$$\tilde{G}_{k,k}(x) = e^{x-(x/n_k)^2+\mathcal{O}(x^2/n_k^3)+\mathcal{O}(x^3/n_k^4)}.$$

Hence, for $x = n_k/\sqrt{2+\varepsilon}$ we get

$$\tilde{G}_{k,k}(n_k/\sqrt{2+\varepsilon}) = e^{n_k/\sqrt{2+\varepsilon}-1/(2+\varepsilon)+\mathcal{O}(1/n_k)}.$$

By combining this estimate with the first part of Lemma 7 we obtain

$$P_k(n_k/\sqrt{2+\varepsilon}) \ge e^{-1/(2+\varepsilon)+\mathcal{O}(1/n_k)}.$$

We recall that

$$P_k(n_{k+1}/2) = e^{-1/2} + \mathcal{O}(n_k^{-1/2}).$$

Thus, for sufficiently large k we have

$$P_k(n_k/\sqrt{2+\varepsilon}) \ge P_k(n_{k+1}/2) = e^{-1/2}$$

and consequently

$$n_k/\sqrt{2+\varepsilon} \le n_{k+1}/2.$$

This completes the proof of Proposition 1. □

5. Tries

In this section we want to show that the height of tries can be handled in the same fashion as above to obtain the same result (which is well known, compare with [10, 4, 2, 11, 12]).

Let $H_n = H_n^{TR}$ denote the random variable describing the height of tries of size n. Then it is known that and set the generating functions

$$G_k(x) := \sum_{n\ge 0} \mathbf{P}[H_n \le h]\,\frac{x^n}{n!} \tag{18}$$

satisfy the recurrent relation

$$G_{k+1}(x) = G_k(x/2)^2 \tag{19}$$

with $G_0(x) = 1 + x$. Thus,

$$G_k(x) = \left(1 + \frac{x}{2^k}\right)^{2^k} \tag{20}$$

or asymptotically

$$G_k(x) = e^{x-x^2/2^{k+1}+\mathcal{O}(x^3/2^{2k})}.$$

Hence, if we define n_k by $G(n_k) = e^{n_k-1}$ we get

$$n_k = 2^{(k+1)/2} + \mathcal{O}(1).$$

Especially, we have

$$\lim_{n\to\infty} \frac{n_{k+1}}{n_k} = \sqrt{2} > 1.$$

If we define

$$\tilde{G}(x) := e^{(1-1/n_k)x}$$

then it follows that

$$G_k(x) \ge \tilde{G}(x) \qquad (\text{for } x \le n_k),$$
$$G_k(x) \le \tilde{G}(x) \qquad (\text{for } x \ge n_k).$$

We only have to oberve that the difference

$$D_l(x) = \tilde{G}(x) - G_l(x)$$

satisfy

$$D_{l+1}(x) = (\tilde{G}(x/2) + G_l(x/2))D_l(x/2).$$

Hence, it follows inductively that all functions $D_l(x)$ have a unique zero $x_l > 0$ satisfying $x_{l+1} = 2x_l$. Thus, we get as above

$$\mathbf{E}H_n = 2\log_2 N + \mathcal{O}(1) \qquad (n \to \infty). \tag{21}$$

and all centralized moments are bounded:

$$\mathbf{E}|H_n - \mathbf{E}H_n|^L = \mathcal{O}(1). \tag{22}$$

References

[1] D. Aldous and P. Shields, *A diffusion limit for a class of random-growing binary trees*, Prob. Th. Rel. Fields **79** (1988), 509–542.

[2] L. Devroye, *A probabilistic analysis of the height of tries and of the complexity of triesort*, Acta Inf. **21** (1984), 229–237.

[3] M. Drmota, *The variance of the height of binary search trees*, Theoret. Comput. Science, to appear..

[4] Ph. Flajolet, *On the performance evaluation of extendible hashing and trie searching*, Acta Inf. **20** (1983), 345–369.

[5] C. Knessl and W. Szpankowski, *Limit laws for heigths in PATRICIA tries*, LATIN'2000, Punta del Este, Lecture Notes in Computer Science, No. 1776, 298-307, 2000.

[6] C. Knessl and W. Szpankowski *Asymptotic behavior of the height in a digital search tree and the longest phrase of the Lempel-Ziv scheme*, SIAM J. Computing **30** (2000), 923–964.

[7] P. Jacquet and W. Szpankowski *Analytical depoissonization and its applications*, Theoret. Comput. Science **201** (1998), 1–62.

[8] D. E. Knuth, *The Art of Computer Programming*, Vol. 3, 2nd ed., Addison-Wesley, 1998.

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

[10] H. Mendelson *Analysis of extendible hashing*, IEEE Trans. Software Eng. **SE-8** (1982), 611–619.

[11] B. Pittel, *Paths in a random digital tree: Limiting distributions*, Adv. Appl. Probab. **18** (1986), 139–155.

[12] W. Szpankowski, *On the height of digital trees and related problems*, Algorithmica **6** (1991), 256–277.