THE VARIANCE OF THE HEIGHT OF BINARY 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 binary search trees of size n is bounded.

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

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 [6]. The height of a binary tree is the maximal distance of a leaf (external node) to the root.)

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 $c = 4.31107\ldots$ is the solution greater than 2 of the equation $\left(\frac{2e}{c}\right)^c = e$. (For related work and extensions see Devroye [2], Devroye and Reed [3], Pittel [7], and Robson [9, 10, 11].)

Based on numerical data Robson conjectured that the variance $\mathbf{V}H_n$ is bounded. Quite recently, B. Reed [8] gave a first proof of Robson's conjecture. He used an approach related to that of [3], where he could also show that

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

In what follows we present a second (short) proof of Robson's conjecture.[1] It was found independently from that of Reed (and a little bit later) and is based on a completely different approach.

In order to formulate the result we introduce the polynomials[2]

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

Since the numbers $a_{n,h}$ satisfy the recurrent relation

$$a_{n,h+1} = \sum_{k=0}^{n-1} \binom{n-1}{k} a_{k,h} a_{n-1-k,h} \tag{2}$$

Date: October 19, 2000.

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

[1]This note is a short version of a part of [5] which also contains tighter bounds, some (possible) extensions concering the saturation level, m-ary search trees and distribution, and a review on the history of this problem and its solution.

[2]This point of view to the height of binary search trees was suggested to me by Philippe Flajolet.

with initial conditions $a_{h,0} = 1$ for $h \ge 0$ and $a_{n,0} = 0$ for $n > 0$, these polynomials are recursively given by $y_0(x) \equiv 1$ and by

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

(Alternatively we can characterize them by $y'_{h+1}(x) = y_h(x)^2$ and $y_h(0) = 1$.)

Theorem. *The expected value of the height H_n of binary search trees of size n is given by*

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

and the variance is bounded:

$$\mathbf{V}H_n = \mathbf{E}(H_n - \mathbf{E}H_n)^2 = \mathcal{O}(1) \qquad (n \to \infty). \tag{5}$$

Note that we also prove an asymptotic relation for the expected value $\mathbf{E}H_n$ which is, however, implicit and *does not* reprove the explicit asymptotic relation $\mathbf{E}H_n \sim c\log n$. In view of the following proof it seems to be *easier* just to prove the boundedness of $\mathbf{V}H_n$. (Nevertheless, by an refined analysis of the tools presented below one can show $\mathbf{E}H_n = c\log + \mathcal{O}(\log\log n)$, but at the moment not more, see [4]).

2. Proof of the Theorem

For convenience we will use the abbreviation $\alpha = e^{1/c} = 1.26\ldots$ (where $c = 4.31107\ldots$ is the solution greater than 2 of the equation $\left(\frac{2e}{c}\right)^c = e$).

We will first show that a specific (retarded) differential equation has a unique solution.

Lemma 1. *There exists a unique entire solution of the differential equation*

$$\Phi'(u) = -\frac{1}{\alpha^2}\Phi(u/\alpha)^2 \tag{6}$$

with the initial condition $\Phi(0) = 1$. Furthermore, this solution has the following properties for real $u > 0$:

1. *$0 < \Phi(u) < 1/u$.*
2. *$\Phi(u)$ is decreasing, $u\Phi(u)$ is increasing.*
3. *$u\Phi(u) = 1 + \mathcal{O}(u^{-1})$ as $u \to \infty$.*

Proof. Starting with a formal power series $\Phi(u) = \sum_{n\ge0} c_n u^n$ with $c_0 = 1$ and inserting it into (6) 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 $|c_n| \le 1$ for all $n \ge 0$. Hence, $\sum_{n\ge0} c_n u^n$ represents an analytic function for $|u| < 1$ and thus provides an analytic solution of the functional equation (6). Now (6) ensures that $\Phi'(u)$ (and consequently $\Phi(u)$) has an analytic continuation to the disc $|u| < \alpha$ which satisfies (6). 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.

Next, it is clear that $\Phi(u)$ is strictly decreasing for $u \ge 0$. However, in order to prove the other properties it seems to be necessary to introduce the auxiliary function

$$f(x) := e^{(c-1)x}\left(1 - e^x\Phi(e^x)\right). \tag{7}$$

First of all the differential equation (6) translates to (note that $c = \frac{2e}{\alpha}$)

$$f'(x) = c\big(f(x) - f(x - c^{-1})\big) + (c/2)^2 e^{-(c-1)x} f(x - c^{-1})^2. \tag{8}$$

Since $\Phi(u) = 1 - \alpha^{-2}u + \mathcal{O}(u^2)$ and $\Phi'(u) = -\alpha^{-2} + \mathcal{O}(u)$ as $u \to 0$ it follows that there exists x_0 such that $f(x)$ is increasing for $x \le x_0$. Assume that $x_1 = \sup\{x : f'(x) > 0\}$ is finite. By continuity of f' we have $f'(x_1) = 0$. Furthermore, by definition $f'(x) > 0$ for $x < x_1$. However, with help of (8) this implies $f'(x_1) > 0$ which contradics our assumption. Thus, $f(x)$ is increasing and positive for all x, and by definition (7) this implies that $u\Phi(u) < 1$ for all $u > 0$.

Next, introduce the function $\Psi(u) := u\Phi(u)$, which satisfies the differential equation $u\Psi'(u) = \Psi(u) - \Psi(u/\alpha)^2$. Since $\lim_{u\to 0} \Psi'(u) = 1$ it follows that $\Psi(u) = u\Phi(u)$ is monotonically increasing at least for sufficiently small (real) $u \in [0, u_0]$. However, if $[0, u_1]$ denotes a (maximal) interval, where $\Psi(u)$ is increasing we have by $\Psi(u) < 1$ that $u_1\Psi'(u_1) = \Psi(u_1) - \Psi(u_1/\alpha)^2 > 0$ which ensures that $\Psi(u)$ is increasing for all $u > 0$. This also shows that $\Phi(u) > 0$ for all $u \ge 0$.

Now, by the monoticity of $u\Phi(u)$ we have $u\Phi(u) \ge (u/\alpha)\Phi(u/\alpha)$ and consequently by (6)

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

for all $u \ge 0$. Hence,

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

or $\Phi(u) \ge 1/(1+u)$. Since $u\Phi(u) < 1$ we thus get $u\Phi(u) = 1 + \mathcal{O}(u^{-1})$ as $u \to \infty$. □

The most important tool for the proof of the Theorem is a set of auxiliary functions

$$\tilde{y}_h(x) := \alpha^h \Phi(\alpha^h(1-x)) \qquad (h \ge 0). \tag{9}$$

The properties of $\Phi(u)$ easily translate to corresponding properties for $\tilde{y}_h(x)$.

Lemma 2. *The functions $\tilde{y}_h(x)$ defined by (9) satisfy the following properties:*

1. $0 < \tilde{y}_h(0) < 1$.
2. $1 - \tilde{y}_h(0) = \mathcal{O}(\alpha^{-h})$ *($h \to \infty$).*
3. $\tilde{y}_h(1) = \alpha^h$.
4. $\tilde{y}_{h+r}(x) \ge \tilde{y}_h(x)$ *for all $x \ge 0$ and $r \ge 0$.*
5. $\tilde{y}'_{h+1}(x) = \tilde{y}_h(x)^2$.
6. *For every integer $h \ge 0$ and every real D the difference $y_h(x) - \tilde{y}_{h+D}(x)$ has exactly one zero $x_{h,D}$ on the positive real line. They satisfy $x_{h+1,D} > x_{h,D}$.*

Proof. The first four properties are obvious in view of Lemma 1.

Even the fifth property is easy to check. By using (6) 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 last property can be proved by induction. Set $\delta_{h,D}(x) := y_h(x) - \tilde{y}_{h+D}(x)$. We show that there exists $x_{h,D} > 0$ such that $\delta_{h,D}(x) > 0$ for $0 \le x < x_{h,D}$ and $\delta_{h,D}(x) < 0$ for $x > x_{h,D}$. Since $\tilde{y}_D(x)$ has a power series expansions at $x_0 = 0$ with positive coefficients (note that $\operatorname{sgn} \Phi^{(k)}(u) = (-1)^k$ for $u \ge 0$) it follows that $\tilde{y}_D(x)$ is strictly increasing and grows faster than any polynomial. Together with $0 < \tilde{y}_D(0) < 1$, this implies that the assertion is surely true for $h = 0$. Now suppose that the assertion is true for some $h \ge 0$. Then we have

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

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

The last property of Lemma 2 has in interesting consequence which we will use for the proof of the theorem.

Lemma 3. *For every $h \ge 0$ we have*

$$\frac{y_{h+1}(1)}{y_h(1)} \ge \alpha. \tag{10}$$

Proof. Set $e_h := c \log y_h(1)$. Then by definition $\tilde{y}_{e_h}(1) = \alpha^{e_h} = y_h(1)$, i.e. $x_{h,e_h-h} = 1$. Thus, $x_{h+1,e_h+1-(h+1)} = x_{h+1,e_h-h} > x_{h,e_h-h} = 1$ implies

$$\alpha^{e_{h+1}} = y_{h+1}(1) > \tilde{y}_{e_h+1}(1) = \alpha^{e_h+1} = \alpha\, y_h(1),$$

which proves the lemma. □

Note that (10) is equivalent to $e_{h+1} \ge e_h + 1$.

Now we directly enter the proof of the Theorem.

Proof. We start by observing that

$$\mathbf{P}[H_{n+1} \le h] \le \mathbf{P}[H_n \le h], \tag{11}$$

i.e. the coefficients of $y_h(x)$ are decreasing. Since $\mathbf{P}[H_n \le h] = a_{n,h}/n!$ we equivalently have

$$a_{n+1,h} \le (n+1)a_{n,h}, \tag{12}$$

which we will prove by induction. Obvioulsly, (12) it is true for $h = 0$. Now suppose that it is true for some $h \ge 0$. Then, we have for $n \ge 0$

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

Since $a_{0,h+1} = 1$ and $a_{1,h+1} \le 1$ we also have $a_{0,h+1} \le a_{1,h+1}$. This completes the proof of (12) and consequently that of (11).

The next step is to prove upper bounds for $\mathbf{P}[H_n \le h]$ if $n \ge y_h(1)$ and lower bounds for $\mathbf{P}[H_n \le h]$ if $n \le y_h(1)$. As above set $e_h = c \log y_h(1)$, i.e. $y_h(1) = \tilde{y}_{e_h}(1)$.

First we suppose that $x \ge 1$. By 6. of Lemma 2 and (11) we get

$$\tilde{y}_{e_h}(x) \ge y_h(x) \ge \sum_{k=0}^{n} \mathbf{P}[H_k \le h]\, x^k \ge \mathbf{P}[H_n \le h]\, \frac{x^{n+1}-1}{x-1}.$$

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

$$\mathbf{P}[H_n \le h] \le \frac{1}{(1+\alpha^{-e_h})^{n+1}-1} \Phi(-1) \ll \frac{1}{n\alpha^{-e_h}} = \alpha^{-(c\log n - e_h)} \tag{13}$$

which is significant for $c \log n \ge e_h$, resp. for $n \ge y_h(1)$.

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

$$\begin{aligned} \frac{1}{1-x} - \tilde{y}_{e_h}(x) \ge \frac{1}{1-x} - y_h(x) &\ge \sum_{k=n}^{\infty} \left(1 - \mathbf{P}[H_k \le h]\right) x^k \\ &\ge \left(1 - \mathbf{P}[H_n \le h]\right) \frac{x^n}{1-x}. \end{aligned}$$

Setting $x = 1 - 1/n$ we directly get

$$1 - \mathbf{P}[H_n \le h] \ll 1 - \alpha^{e_h - c\log n}\Phi(\alpha^{e_h - c\log n}) \ll \alpha^{-(e_h - c\log n)} \tag{14}$$

if $e_h \ge c\log n$, resp. if $n \le y_h(1)$. Now recall that $e_{h+1} \ge e_h + 1$. Hence we have

$$\begin{aligned}\sum_{h:y_h(1)\le n} \mathbf{P}[H_n \le h] &\ll \sum_{h:e_h \le c\log n} \alpha^{-(c\log n - e_h)} \\ &\ll \sum_{j\ge 0} \alpha^{-j} = \mathcal{O}(1) \qquad (n\to\infty)\end{aligned}$$

and

$$\begin{aligned}\sum_{h:y_h(1)> n} (1 - \mathbf{P}[H_n \le h]) &\ll \sum_{h:e_h > c\log n} \alpha^{-(e_h - c\log n)} \\ &\ll \sum_{j\ge 0} \alpha^{-j} = \mathcal{O}(1) \qquad (n\to\infty).\end{aligned}$$

These estimates are sufficient to prove the proposed expansion (4) for the expected value $\mathbf{E}H_n$:

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

In view of this estimate (and the property $e_{h+1} - e_h \ge 1$) we have for $h \le \mathbf{E}H_n$

$$c\log n - e_h \ge (\mathbf{E}H_n - h) + C_1$$

and for $h \ge \mathbf{E}H_n$

$$e_h - c\log n \ge (h - \mathbf{E}H_n) + C_2$$

for suitably chosen constants C_1, C_2. Thus, (13) and (14) yield

$$\mathbf{P}[|H_n - \mathbf{E}H_n| \ge j] \ll \alpha^{-j}$$

which directly implies that

$$\mathbf{E}(H_n - \mathbf{E}H_n)^2 \ll \sum_{j\ge 0} j\,\mathbf{P}[|H_n - \mathbf{E}H_n| \ge j] = \mathcal{O}(1)$$

as $n \to \infty$. This completes the proof of the Theorem. □

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, *An Analytic Approach to the Height of Binary Search Trees*, Algorithmica, to appear.

[5] M. Drmota, *An Analytic Approach to the Height of Binary Search Trees. II*, J. Assoc. Comput. Mach., submitted.

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

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

[8] B. Reed, *The height of a random binary search tree*, J. Assoc. Comput. Mach., submitted.

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

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

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