Algorithms Seminar 2001–2002,
F. Chyzak (ed.), INRIA, (2003), pp. 63–68.

Available online at the URL
`http://algo.inria.fr/seminars/`.

Traveling Waves and the Height of Binary Search Trees

Michael Drmota
Institut für Geometrie, Technische Universität Wien (Austria)

September 24, 2001

Summary by Brigitte Chauvin

1. Introduction

Binary search trees are widely used to store (totally ordered) data, and many parameters have been discussed in the literature (the monograph of Mahmoud [6] gives a very good overview of the state of the art). Starting from a permutation of $\{1, 2, \ldots, n\}$ we get a binary tree T_n with n internal nodes such that the keys of the left subtree of any given node x are smaller than the key of x and the keys of the right subtree are larger than the key of x. Usually it is assumed that every permutation of $\{1, 2, \ldots, n\}$ is equally likely and hence any parameter of binary search trees may be considered as a random variable.

Here we consider the height H_n which is the largest distance of an internal node from the root. In 1986, Devroye [2] proved that the expected value $\mathbf{E}\,H_n$ satisfies the asymptotic relation

$$\mathbf{E}\,H_n \sim c \log n, \tag{1}$$

and it is also proved [1] that

$$\frac{H_n}{c \log n} \to 1 \quad a.s., \tag{2}$$

(as $n \to \infty$), where $c = 4.31107\ldots$ is the (largest real) solution of the equation

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

Better bounds for the expected value were given by two completely different methods by Devroye and Reed [3] and by Drmota [4]. Finally Drmota [5] and Reed [8, 9] proved the so-called Robson conjecture

$$\mathbf{V}\,H_n = \mathcal{O}(1). \tag{4}$$

Reed [8, 9] was also able to obtain a very precise bound for the expected value:

$$\mathbf{E}\,H_n = c \log n - \frac{3c}{2(c-1)} \log\log n + \mathcal{O}(1). \tag{5}$$

Notice that properties analogous to (1) and (2) hold for the (dual) saturation level H'_n with constant c replaced by the other real solution of Equation (3) [1, 5, 6].

Here, the purpose is to obtain more precise information on the asymptotic behaviour of the distribution of the height H_n. This will also lead to a perspective of improving (4) and (5). To this end, we first need to understand the two main ideas. They are:

1. an analytic approach, due to Drmota, of the generating function
$$Y_k(z) = \sum_{n\geq 0} \mathbf{P}(H_n \leq k) z^n$$
2. Devroye's connection between Binary Search Trees (bst) and Branching Random Walks (brw), which allows to use the above analytic approach to a "close" model (brw), easier to deal with. Moreover, the analytic approach is applied to the Random Bisection Problem, considered as a brw with a continuous parameter.

This seminar is devoted to connect such methods to some facts and results. Very precise estimates are shown to be consequences of rather natural conjectures.

2. **Results and Conjectures**

Following the analytic approach, the generating function
$$Y_k(z) = \sum_{n\geq 0} \mathbf{P}(H_n \leq k) z^n$$
is a solution of the difference equation

$$\begin{cases} Y_0(z) = 1 \\ Y'_{k+1}(z) = Y_k(z)^2, \qquad Y_k(0) = 1. \end{cases} \tag{6}$$

For
$$x_k := Y_k(1) = \sum_{n\geq 0} \mathbf{P}(H_n \leq k),$$
it is shown in [4, 5] that x_k is related to $\mathbf{E}\, H_n$ by the following result.

Fact 1.
$$\mathbf{E}\, H_n = \max\{\, k \mid x_k \leq n \,\} + \mathcal{O}(1).$$

We also already noticed the following result by Reed [8, 9].

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

Together, Facts 1 and 2 give the following bounds:
$$c_2 \alpha^k k^\beta \leq x_k \leq c_1 \alpha^k k^\beta$$
where $\alpha = e^{1/c}$ and $\beta = \frac{3}{2(c-1)}$.

It follows that the following conjectures are quite natural.

Conjecture 1.
$$x_k \sim \gamma \alpha^k k^\beta \qquad (k \to +\infty).$$

Conjecture 2.
$$\lim_{k\to+\infty} \frac{x_{k+1}}{x_k} \textit{ exists.}$$

Assume for a while that Conjecture 2 is true,[1] then the following theorem holds.

[1] Recently Conjecture 2 could be verified so that Theorem 1 is now an unconditioned result.

Theorem 1. *There exists some distribution function $F(x)$ such that*

$$\mathbf{P}(H_n \leq k) = F(\log n - \log x_k) + o(1) \tag{7}$$

uniformly in k as $n \to +\infty$.

Let us point out here that, if Conjecture 1 is true, there exists some distribution function $F(x)$ such that

$$\mathbf{P}(H_n \leq k) = F\left(\log n - \frac{1}{c}k - \beta \log k\right) + o(1) \tag{8}$$

uniformly in k as $n \to +\infty$. The limit distribution F which appears in (7) and (8) can be understood as a traveling wave.

As another consequences of Conjecture 1, precise estimates of the first and second moment of the height are:

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

and

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

where Δ_1 and Δ_2 are continuous, periodic functions with period 1.

There is an intimate relation of Random Binary Search Trees to Devroye's Tree Model, resp. a relation between a Binary Search Tree and a Branching Random Walk. Recall that in this connection, the considered Branching Random Walk is defined by an infinite binary tree with weights $\tilde{U}$, equal to U or $1-U$ on left and right edges respectively (U denotes a uniform random variable on $[0,1]$). In this model, each node v of the tree has a weight

$$l(v) = \prod_{e<v} \tilde{U}_e.$$

Let the tree $\bar{T}_n$ be defined by

$$\bar{T}_n := \left\{ v \;\middle|\; l(v) \geq \frac{1}{n} \right\},$$

and let $\bar{H}_n$ denotes the height of $\bar{T}_n$. Devroye has shown that the distribution of $\bar{H}_n$ is "very close" to that of H_n.

Let us see now why the the distribution of $\bar{H}_n$ is close to that of H_n. We work in terms of the Random Bisection Problem (which is a reformulation of $\bar{H}_n$): in that problem, an interval with length x is randomly cut into two intervals with length $x_1 := Ux$ and $x_2 := (1-U)x$, where U is uniformly distributed on $[0,1]$.

Let $P_k(x,l)$ be the probability that all segments are less than l after k steps, and let

$$\bar{P}_k\left(\frac{x}{l}\right) := P_k\left(\frac{x}{l}, 1\right) = P_k(x,l),$$

then $\bar{P}_k(x)$ looks like a wave, and is a solution of the following recursion:

$$\bar{P}_{k+1}(x) = \frac{1}{x}\int_0^x \bar{P}_k(y)\bar{P}_k(x-y)\,dy.$$

By definition of P_k, $\bar{H}_n$, $\bar{T}_n$,

$$\bar{P}_k(n) = P_k\left(1, \frac{1}{n}\right) = \mathbf{P}\left(\bar{H}_n \leq k\right)$$

so that the Random Bisection Problem appears as a generalized tree model with continuous parameter x:

$$\bar{T}_x = \left\{ v \;\middle|\; l(v) \geq \frac{1}{x} \right\}, \qquad \bar{H}_x = \text{height of } \bar{T}_x.$$

For this generalized tree model, the analytic approach is close to that for Binary Search Trees and it provides an analogy between H_n and $\bar{H}_n$: let

$$\bar{Y}_k(z) := \int_0^\infty \bar{P}_k(x) e^{(z-1)x}\, dx = \int_0^\infty P\left(\bar{H}_x \leq k\right) e^{(z-1)x}\, dx$$

then

$$\bar{Y}_0(z) = \frac{1}{z-1}(e^{z-1} - 1)$$

and

$$\bar{Y}'_{k+1}(z) = \bar{Y}_k(z)^2. \tag{9}$$

For

$$\bar{x}_k := \bar{Y}_k(1) = \int_0^\infty \bar{P}_k(x)\, dx = \int_0^\infty P\left(\bar{H}_x \leq k\right)\, dx$$

we have the following results.

Fact 1'.

$$\mathbf{E}\, \bar{H}_n = \max\{\, k \mid x_k \leq n \,\} + \mathcal{O}(1) \qquad (n \to \infty).$$

Fact 2'.

$$\begin{aligned} \mathbf{E}\, \bar{H}_n &= \mathbf{E}\, H_n + \mathcal{O}(1) \\ &= c \log n - \frac{3c}{2(c-1)} \log\log n + \mathcal{O}(1) \qquad (n \to \infty). \end{aligned}$$

Both results imply

$$\bar{c}_2 \alpha^k k^\beta \leq \bar{x}_k \leq \bar{c}_1 \alpha^k k^\beta$$

for the same constants α and β. Analogous conjectures are

Conjecture 1'.

$$\bar{x}_k \sim \bar{\gamma} \alpha^k k^\beta \qquad (k \to +\infty).$$

Conjecture 2'.

$$\lim_{k \to +\infty} \frac{\bar{x}_{k+1}}{\bar{x}_k} \text{ exists.}$$

Note that Conjectures 1 and 1' on the one hand, and Conjectures 2 and 2' on the other hand, are equivalent. Admitting these conjectures, the following theorem can be deduced as well:

Theorem 2. *If Conjecture 2' is true,*[2] *there exists some distribution function $\bar{F}(x)$ such that*

$$\mathbf{P}\left(\bar{H}_n \leq k\right) = \bar{P}_k(n) = \bar{F}(\log n - \log \bar{x}_k) + o(1) \tag{10}$$

uniformly in k as $n \to +\infty$.

If Conjecture 1' is true, there exists some distribution function $\bar{F}(x)$ such that

$$\mathbf{P}\left(\bar{H}_n \leq k\right) = \bar{P}_k(n) = \bar{F}\left(\log n - \frac{k}{c} - \beta \log k\right) + o(1) \tag{11}$$

uniformly in k as $n \to +\infty$. The limit distribution $\bar{F}$ which appears in (10) and (11) can be understood as a traveling wave.

[2]... which has been verified

Note that $F(x)$ of Theorem 1 and $\bar{F}(x)$ of Theorem 2 in fact coincide.

3. Sketch of Proof

To prove Theorem 1 (and similarly Theorem 2) it is necessary to get information on $\bar{Y}_k(x)$, the solution of Equation (6) (resp. of (9)). The method consists in considering an auxiliary function $\tilde{Y}_k(x)$, related to a solution of the Retarded Differential Equation with a parameter α:

$$\Phi'(u) = -\frac{1}{\alpha^2}\Phi\left(\frac{u}{\alpha}\right)^2, \qquad \Phi(0) = 1,$$

by

$$\tilde{Y}_k(x) := \alpha^k \Phi\left(\alpha^k(1-x)\right) \qquad (k \in \mathbb{R}).$$

The Retarded Differential Equation can be solved, because Φ is the Laplace transform of some function Ψ

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

solution of the integral equation

$$y\Psi\left(\frac{y}{\alpha}\right) = \int_0^y \Psi(z)\Psi(y-z)\,dz.$$

The existence and unicity of solutions of this integral equation, considered as a fixed-point equation, come from a contraction method which applies only for values of parameter α between 1 and a critical value $\alpha_0 = e^{1/c} = 1.26\ldots$

The relation between the auxiliary function $\tilde{Y}_k(x)$ and the true function $Y_k(x)$ relies on a scaling: define e_k by

$$\alpha^{e_k} = x_k,$$

then, locally around $x = 1$,

$$Y_k(z) \sim \tilde{Y}_{e_k}(x),$$

at least if Conjecture 2 is right!, i.e.,

$$\lim_{k\to\infty} \frac{x_{k+1}}{x_k} = \alpha.$$

Then, it remains to extract the coefficient with degree n in $Y_k(x)$

$$\mathbf{P}(H_n \le k) = [x_n]\,Y_k(x) = \Psi(n/x_k) + o(1)$$

to get by comparison with $\tilde{Y}_{e_k}(x)$, the asymptotics of Theorem 1:

$$\mathbf{P}(H_n \le k) \sim F(\log n - \log x_k)$$

with $F(x) = \Psi(\log x)$.

As a last remark, it is worth to connect the above objects, especially $\bar{x}_k$, to some heuristics in statistical physics literature (see for instance [7]), where quite similar traveling waves appear. There, $\bar{x}_k$ is the front position, it increases as $\alpha^k k^\beta$ (Conjecture 1') and parameter α of the Retarded Differential Equation is nothing but the velocity of the front wave.

Bibliography

[1] Biggins (J. D.). – How fast does a general branching random walk spread? In *Classical and modern branching processes (Minneapolis, MN, 1994)*, pp. 19–39. – Springer, New York, 1997.

[2] Devroye (Luc). – A note on the height of binary search trees. *Journal of the Association for Computing Machinery*, vol. 33, n° 3, 1986, pp. 489–498.

[3] Devroye (Luc) and Reed (Bruce). – On the variance of the height of random binary search trees. *SIAM Journal on Computing*, vol. 24, n° 6, 1995, pp. 1157–1162.

[4] Drmota (M.). – An analytic approach to the height of binary search trees. *Algorithmica*, vol. 29, n° 1-2, 2001, pp. 89–119. – Average-case analysis of algorithms (Princeton, NJ, 1998).

[5] Drmota (Michael). – The variance of the height of binary search trees. *Theoretical Computer Science*, vol. 270, n° 1-2, 2002, pp. 913–919.

[6] Mahmoud (Hosam M.). – *Evolution of random search trees.* – John Wiley & Sons, New York, 1992, *Wiley-Interscience Series in Discrete Mathematics and Optimization*, xii+324p.

[7] Majumdar (Satya N.) and Krapivsky (P. L.). – Traveling waves, front selection, and exact nontrivial exponents in a random fragmentation problem. *Physical Review Letters*, vol. 85, n° 26, 2000, pp. 5492–5495.

[8] Reed (Bruce). – How tall is a tree? In *Proceedings of the thirty-second annual ACM symposium on Theory of computing (Portland, Oregon, United States)*, pp. 479–483. – 1999. Proc. of STOC'00.

[9] Reed (Bruce). – The height of a random binary search tree. *Journal of the Association for Computing Machinery*, vol. 50, n° 3, 2003, pp. 306–332.