The Height of a Random Tree

Tomasz Łuczak
Adam Mickiewicz University, Poznan, Poland

March 29, 1993

[summary by Wojtek Szpankowski]

1. Introduction

Let T_n be a random labelled rooted tree on the vertex set $[n] = \{1, 2, \ldots, n\}$ with the root $v_0 \in [n]$ (here and below we assume that a root is always the vertex number 1). The limit distribution of the height of $\tilde{H} = \tilde{H}(n)$ of T_n, was found by Rényi and Szekeres [3] who proved the following result.

THEOREM 1. *For every constant* $\beta > 0$

$$\begin{aligned}\lim_{n\to\infty}(\tilde{H} = \lfloor\sqrt{2n}/\beta\rfloor) &= 2\sqrt{\frac{2\pi}{n}}\beta^2\sum_{i=1}^{\infty}\left(2i^4\pi^4\beta - 3i^2\pi^2\right)\exp(-\beta\pi^2 i^2)\\ &= \sqrt{\frac{8}{n\beta}}\sum_{i=1}^{\infty}\left(\frac{2i^4}{\beta} - 3i^2\right)\exp\left(-\frac{i^2}{\beta}\right),\end{aligned}$$

where the convergence is uniform for $\beta \in (c, C)$ *for every constants* $0 < c < C < \infty$.

Furthermore, they proved that the s-th moment of random variable $h(T_n)/\sqrt{2n}$ tends to $2\Gamma(s/2+1)(s-1)\zeta(s)$. In particular, for the expectation and the variance of $h(T_n)$, one obtains

$$\begin{aligned}\lim_{n\to\infty}\frac{\mathrm{E}\,h(T_n)}{\sqrt{n}} &= \sqrt{2\pi} = 2.50663\ldots\\ \lim_{n\to\infty}\frac{\mathrm{Var}\,h(T_n)}{n} &= \frac{2\pi(\pi-3)}{3} = 0.29655\ldots.\end{aligned}$$

(See also [1] for a generalization of this result to other simply generated families of trees.)

Consider now the following greedy algorithm. For a tree T with the root v_0 let $\mathcal{F}(T)$ be the forest of rooted trees obtained from T by removing the root, where as the root of a tree $T' \in \mathcal{F}(T)$ we choose the vertex adjacent to v_0 in T. The height of a tree can be estimated by using the following simple greedy algorithm, which finds in a tree a path starting from the root. The algorithm starts with a tree $T^{(0)} = T$ on n vertices, removes its root v_0, chooses the largest tree $T^{(1)}$ from $\mathcal{F}(T^{(0)})$ (if there are more than one of them it picks the one with the lexicographically first root), appends its root to a path, and repeats this procedure until for some h tree $T^{(h)}$ consists only of one vertex.

This talk concerns the study of the height $H = H(n)$ found in a random tree by the above greedy algorithm. The limiting distribution of H is found and it is shown that the expected value of $H/\sqrt{n}$ tends to an absolute constant C, where

$$C = \frac{\sqrt{2\pi}}{2\sqrt{2} - \ln(3 + 2\sqrt{2})} = 2.353139\ldots.$$

Thus, on average, the algorithm finds a path whose length is roughly 93% of the expected height of the tree.

2. Main Results

We need some definitions. Let us define recursively two sequences of random variables $\{\hat{H}_i\}$ and $\{W_i\}$ by setting $\hat{H}_0 = \min_j\{|T_n{}^{(j)}| \le n/2\}$, $W_0 = |T_n^{(\hat{H}_0)}|$ whereas for $i \ge 1$ let

$$\hat{H}_i = \min_j\{|T_n^{(j)}| \le W_{i-1}/2\}$$

and $W_i = |T_n^{(\hat{H}_i)}|$. Furthermore, set $H_0 = \hat{H}_0$ and $H_i = \hat{H}_i - \hat{H}_{i-1}$ for $1 \le r \le n-1$. Thus, W_i denotes the size of the tree $T_n^{(k)}$ when it first drops under $W_{i-1}/2$ and H_i is the number of steps of the algorithm between two such moments. Note that for every $i \ge 0$ we have $W_i \le 2^{-i-1}n$.

Clearly, the length of the path found by the algorithm can now be written as a sum of H_i's, so

$$\begin{aligned}\Pr(H > k) &= \Pr(\sum_{i\ge 0} H_i > k) \\ &= \Pr(H_0 > k) + \sum_{j\ge 1}\Pr(\sum_{i=0}^{j} H_i > k \wedge \sum_{i=0}^{j-1} H_i \le k)\end{aligned}$$

In order to characterize the behaviour of the probabilities $\Pr(\sum_{i=0}^{j} H_i > k \wedge \sum_{i=0}^{j-1} \le k)$ let us define an integral operator A by setting

$$(Ag)(x) = \int_0^x \int_{1/4}^{1/2} f(z,y) g((x-z)/\sqrt{y}) dy\, dz\ ,$$

where f is a function defined as

$$f(x,y) = \frac{1}{2\pi}\int_{1/2-y}^{y} \frac{x}{t^{3/2}y^{3/2}(1-t-y)^{3/2}} \exp\left(-\frac{x^2}{2(1-y-t)}\right) dt.$$

Furthermore, let

$$g_0(x) = \int_x^\infty \int_{1/4}^{1/2} f(z,y) dy\, dz,$$

and for $j \ge 1$

$$g_j = Ag_{j-1} = A^j g_0.$$

The next result shows that functions g_j are closely related to our problem.

LEMMA 1. *For every $x > 0$ we have*

$$\Pr(H_0 > \lfloor x\sqrt{n}\rfloor) = (1 + o(1))g_0,$$

and for $j \ge 1$

$$\Pr(\sum_{i=0}^{j} H_i > \lfloor x\sqrt{n}\rfloor \wedge \sum_{i=0}^{j-1} H_i \le \lfloor x\sqrt{n}\rfloor) = (1 + o(1))g_i(x)\,,$$

where, for given constants c, C, the quantity $o(1)$ tends to 0 uniformly for every $x \in (c, C)$.

As a consequence of Lemma 1 one proves the limiting distribution of H.

THEOREM 2. *For every constant $x \ge 0$*

$$\lim_{n\to\infty} \Pr(H > x\sqrt{n}) = h(x),$$

where

$$h(x) = \sum_{j=0}^{\infty} g_j(x) = \sum_{j=0}^{\infty} (A^j g_0)(x).$$

In particular, the function h is the only continuous solution of the integral equation

$$\begin{aligned} h(x) &= g_0(x) + (Ah)(x) \\ &= \int_x^\infty \int_{1/4}^{1/2} f(z,y)dy\,dz + \int_0^x \int_{1/4}^{1/2} f(z,y)h((x-z)/\sqrt{y})dy\,dz\ , \end{aligned}$$

Having computed the distribution of H it is not hard to guess the value of its mean. Clearly, $\mathrm{E}\,H/\sqrt{n}$ should converge to the expected value of the random variable Z, where $P(Z > x) = h(x)$ and $h(x)$ is given by Theorem 2. But $xh(x) \to 0$ as $x \to \infty$ (as a matter of fact Theorem 1 says that the probability that the actual height of a random tree is larger than x decreases exponentially with x), so

$$\mu = \mathrm{E}\,Z = \int_0^\infty h(x)dx.$$

Integrating both sides of the formula on $h(x)$ in Theorem 2, after elementary calculations, one obtains

$$\mu = \int_0^\infty \int_{1/4}^{1/2} xf(x,y)dy\,dx + \mu \int_0^\infty \int_{1/4}^{1/2} \sqrt{y}f(x,y)dy\,dx.$$

Consequently,

$$\mu = \frac{\int_0^\infty \int_{1/4}^{1/2} xf(x,y)dy\,dx}{1 - \int_0^\infty \int_{1/4}^{1/2} \sqrt{y}f(x,y)dy\,dx}\ .$$

Finally, one proves the following result.

THEOREM 3. *The average height obtained by the greedy algorithm is*

$$\lim_{n\to\infty} \frac{\mathrm{E}\,H}{\sqrt{n}} = \mu\ ,$$

where

$$\mu == \frac{\int_0^\infty \int_{1/4}^{1/2} xf(x,y)dy\,dx}{1 - \int_0^\infty \int_{1/4}^{1/2} \sqrt{y}f(x,y)dy\,dx} = \frac{\sqrt{2\pi}}{2\sqrt{2} - \ln(3+2\sqrt{2})} = 2.353139\ldots.$$

This completes our presentation of main results of the talk. More details can be found in [2].

Bibliography

[1] Flajolet (Philippe), Gao (Zhicheng), Odlyzko (Andrew), and Richmond (Bruce). – *The Distribution of Heights of Binary Trees and Other Simple Trees.* – Research Report n° 1749, Institut National de Recherche en Informatique et en Automatique, September 1992. 12 pages. Accepted for Publication in *Combinatorics, Probability, and Computing.*

[2] Luczak (T.). – A greedy algorithm estimating the height of random trees. – Preprint, 1993.

[3] Rényi (A.) and Szekeres (G.). – On the height of trees. *Australian Journal of Mathematics*, vol. 7, 1967, pp. 497–507.