THE HEIGHT OF RAWA TREES

András Antos

Technical University of Budapest

and

Luc Devroye

McGill University

Definitions

Binary Search Tree (BST): a binary tree built from a sequence of real numbers $s = (s_0, \ldots, s_n)$ in the following way:

Put s_0 in the root and build the left subtree from the subsequence of s containing the elements less than s_0 and the right subtree from the subsequence of s containing the elements greater than s_0, recursively.

BST is a common data structure for storing information.

Depth of a node: the path distance from the node to the root.

Height of the tree (H_n): the maximal depth.

Times of searching, inserting, deleting are related to the height (Knuth'73).

Random models

Random Binary Tree: a binary tree chosen randomly in a certain way from all the binary trees with a given number of nodes.

Note: Depths and height are random variables.

Examples:

1. Every binary tree with n nodes has equal probability (Devroye and Kruszewski'95).

2. **Random Binary Search Tree:** a binary search tree built from a sequence of real random variables $S = (S_0, \dots, S_n)$.

The model is determined by the distribution of the sequence S.

Background

Studied model: $S_0, \ldots, S_n$ are i.i.d.

Results: for continuous distribution

$H_n \sim 4.33 \ldots \log n$ a.s.,

$\mathbf{Var}\, H_n = O(\log \log n)$,

$\mathbf{E}|H_n - \mathbf{E}H_n| < 7$ i.o.

(Robson'79, Devroye'86, '87).

An open question: Is $\mathbf{Var}\, H_n$ bounded ?

Another model: $S_0 = 0$ and

$$S_i = X_1 + \cdots + X_i,$$

where $X_1, \ldots, X_n$ (and X) are i.i.d.

Leads to a random walk $\rightarrow$ RAWA trees.

Another possibility: $S_i = \frac{X_1 + \cdots + X_i}{\sqrt{i}}$

Conditions

If $\mathbf{E}X \neq 0$, the random walk drifts off and the shape of the tree is uninteresting with $\mathbf{E}H_n = \Theta(n)$.

Interesting: $\mathbf{E}X = 0$. We need some more technical conditions.

Define X to be a *nice* r.v., if

A. X has zero mean and finite variance σ^2

B. X has a bounded density f, symmetric about zero, with $\liminf_{x \to 0} f(x) > 0$ and $\mathbf{E}\{|X|^5\} < \infty$.

Main result

While the shape of the tree depends heavily on the distribution of X, for all nice X the limit behavior for the height is identical. $S_0, \ldots, S_n$ are dependent and the tree will be higher than in the i.i.d. case. The height has order of $\sqrt{n}$ by the following limit law:

Theorem 1 *Let X be a nice random variable. Then*

$$\frac{H_n}{\sqrt{2n}} \xrightarrow{\mathcal{D}} \mathcal{L}$$

where $\mathcal{L}$ is the Erdős-Kac-Rényi distribution function

$$\mathcal{L}(x) = \begin{cases} \frac{4}{\pi} \sum_{n=0}^{\infty} \frac{(-1)^n}{2n+1} e^{-\frac{(2n+1)^2 \pi^2}{8x^2}} & \text{if } x > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Main idea

Observation: The height is not much more than the depth of the "roof", that is the sides of the tree.

Let H'_n be the maximum of the depths of the two roofs. The proof of Theorem 1 is based on the following two lemmas:

Lemma 1 *Let X be a nice random variable. Then*

$$\frac{H'_n}{\sqrt{2n}} \xrightarrow{\mathcal{D}} \mathcal{L}$$

where $\mathcal{L}$ is the Erdős-Kac-Rényi distribution function.

Lemma 2 *Let X be a nice random variable. Then*

$$\frac{H_n - H'_n}{\sqrt{n}} \to 0.$$

in probability.

Connection with records

Observation: The elements of the right "roof" correspond to those elements of S which are greater than every previous element. These are the **up-records** in S. Analogously, **down-records**.

Denote the number of up-records by R_n, the number of down-records by R'_n. Then $H'_n = \max(R_n, R'_n)$.

The increase between two consecutive up-records (or down-records) is called **ladder height**.

Proposition 1 *Let X be a nice random variable. Then the ladder heights form an i.i.d. sequence with mean $\sigma/\sqrt{2}$.*

Idea: the number of up-records may be related to M_n, the maximum of S. Analogously M'_n is the absolute value of the minimum of S.

Results on records and maxima

Define the distribution function

$$\mathcal{G}(x) = \begin{cases} 2\Phi(x) - 1 & \text{if } x > 0 \\ 0 & \text{otherwise,} \end{cases}$$

where Φ is the Gaussian distribution function. The following is known (Erdős and Kac'46, Feller'71, Rényi):

Proposition 2 *Let X be a nice random variable. Then*

$$\frac{M_n}{\sigma\sqrt{n}} \xrightarrow{\mathcal{D}} \mathcal{G}\,, \ \frac{M'_n}{\sigma\sqrt{n}} \xrightarrow{\mathcal{D}} \mathcal{G} \ \textit{and} \ \frac{\max(M_n, M'_n)}{\sigma\sqrt{n}} \xrightarrow{\mathcal{D}} \mathcal{L}.$$

Furthermore,

$$\frac{R_n}{\sqrt{2n}} \xrightarrow{\mathcal{D}} \mathcal{G} \ \textit{and} \ \frac{R'_n}{\sqrt{2n}} \xrightarrow{\mathcal{D}} \mathcal{G}.$$

Note: Lemma 1 is just the missing link for $\max(R_n, R'_n)$.

Proof of Lemma 1

$$\frac{\max(R_n, R'_n)}{\sqrt{2n}} \xrightarrow{\mathcal{D}} \mathcal{L}$$

The proof is based on Proposition 2 and

Proposition 3 *Let X be a nice random variable. Then, for all x,*

$$\mathbf{P}\left\{\frac{\max(R_n, R'_n)}{\sqrt{2n}} < x\right\} - \mathbf{P}\left\{\frac{\max(M_n, M'_n)}{\sigma\sqrt{n}} < x\right\}$$

tends to zero.

Idea: M_n/R_n is the average of the ladder heights.

The proof uses Propositions 1 and 2, the fact that $\lim_{n\to\infty} R_n = \lim_{n\to\infty} R'_n = \infty$ a.s. and the weak law of large numbers for the sequence of the ladder heights.

The law of large numbers

We generalize the weak law of large numbers for **random number of terms**:

Lemma 3 *Let* $\mathcal{H}_1, \mathcal{H}_2, \ldots$ *be i.i.d. nonnegative random variables with expectation* m *and* $N_1, N_2, \ldots$ *be nonnegative integer-valued random variables such that* $\liminf_{n\to\infty} N_n > 0$ *a.s.,* $N_n/a_n \xrightarrow{\mathcal{D}} K$ *for some function* K *on* $[0,\infty)$ *where* $a_n > 0$, $\lim_{n\to\infty} a_n = \infty$ *and* $\lim_{x\to 0} K(x) = 0$. *Then for all* $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbf{P}\left\{ \left| \frac{\sum_{i=1}^{N_n} \mathcal{H}_i}{N_n} - m \right| > \epsilon \right\} = 0.$$

Proof of Lemma 2

$$\frac{H_n - \max(R_n, R'_n)}{\sqrt{n}} \to 0.$$

The idea is splitting the real line into intervals and counting the maximal number of **records restricted to an interval**, that is, records in the subsequence falling in the interval. This turns out to be not much more than $n^{1/4} \log n$. (The central limit theorem suggests that the number of points falling in an interval is at most $\sqrt{n}$.)

The difference $H_n - H'_n$ can be related to this maximum, and this gives the result.