RAWA TREES

Andras Antos
Informatics Laboratory Research Division
Computer and Automation Research Institute
of the Hungarian Academy of Sciences
H–1111 Lágymányos u 11
Budapest, Hungary
antos@szit.bme.hu

Luc Devroye
School of Computer Science
McGill University
i
3480 University Street
Montreal, Canada H3A 2K6
luc@cs.mcgill.ca

Abstract. A rawa tree is a binary search tree for an ordinary **ra**ndom **wa**lk $0, S_1, S_2, S_3, \ldots$, where $S_n = \sum_{i=1}^n X_i$ and the X_i's are i.i.d. distributed as X. We study the height H_n of the rawa tree, and show that if X is absolutely continuous with bounded symmetric density, if X has finite variance, and if the density of X is bounded away from zero near the origin, then $H_n/\sqrt{2n}$ tends to the Erdős-Kac-Rényi distribution.

Keywords and phrases. Random binary search tree, probabilistic analysis, random walk, Catalan constant, limit distributions.

Introduction

Binary search trees are the most common data structures for storing information. Given a sequence S of real numbers and a real number x, let $L(S, x)$ denote the subsequence of S consisting of all numbers less than x and let $R(S, x)$ be the subsequence consisting of numbers greater than x. In particular, if $S = (S_0, S_1, \ldots, S_n)$ are real numbers then the binary search tree for these data is recursively defined as follows: it is a binary tree with root S_0, with left subtree the binary search tree for $L(S, S_0)$ and with right subtree the binary search tree for $R(S, S_0)$. Binary search trees permit searching in time bounded by the height of the tree, where the height H_n is the maximal path distance from any node to the root (Knuth, 1973; Cormen, Leiserson and Rivest, 1990).

The standard random binary search tree is based on an i.i.d. sequence

$$(S_0, S_1, \ldots, S_n) \ .$$

In that case, it is well-known (Robson, 1979, Devroye, 1986, 1987) that if the common distribution is absolutely continuous, then

$$H_n \sim 4.33107 \ldots \log n$$

almost surely as $n \to \infty$. Rather little is known about H_n when the defining sequence is not i.i.d. We define a running average model for a generic random variable X and an averaging parameter $p \in [0, \infty)$ as follows: let $X_1, \ldots, X_n$ be i.i.d., distributed as X, and define $S_0 = 0$ and

$$S_n = \frac{X_1 + \cdots + X_n}{n^p} \ , n \geq 1 .$$

The running average random binary search tree is based on $(S_0, S_1, \ldots, S_n)$. Within this model, there are only two choices of practical interest, $p = 1/2$ and $p = 0$: for $p = 0$, S_n is just a partial sum of all X_i's with $i \leq n$, and we obtain a binary search for an ordinary random walk. We define a <u>rawa tree</u> $\mathcal{T}_n$ (or $\mathcal{T}_n(X)$) as a binary search tree based on $(S_0, S_1, \ldots, S_n)$ where $S_0 = 0$, $S_n = \sum_{i=1}^n X_i$ and $X_1, \ldots, X_n$ are i.i.d. distributed as X. We study H_n for a large class of rawa trees. If X has nonzero mean, the random walk will drift off, and the binary search tree has an uninteresting shape and height $H_n = \Theta(n)$. For this reason, and technical reasons that will be encountered later, we restrict ourselves to <u>nice</u> random variables, which are random variables with the following properties:

A. X has a density, is symmetric about zero, and has a finite variance σ^2.

B. The density f of X is bounded: $\|f\|_\infty < \infty$.

C. f is bounded away from 0 in a neighborhood of the origin: $\liminf_{x \downarrow 0} f(x) > 0$.

Random variables satisfying A only will be called <u>simple</u>.

While the shape of $\mathcal{T}_n(X)$ indeed depends heavily on the distribution of X, it is quite interesting that for all nice random variables, the limit behavior for H_n is essentially identical.

THEOREM. *Let X be a nice random variable, and let $\mathcal{T}_n$ be a rawa tree. Then, for all $x > 0$,*

$$\lim_{n\to\infty} \mathbf{P}\left\{\frac{H_n}{\sqrt{2n}} < x\right\} = \mathcal{L}(x) \ ,$$

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

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

Note that the limit law does not depend upon the distribution of X, yet the shape of the rawa tree depends upon the distribution in a substantial manner. Also, the Theorem does not exclude the possibility that there are symmetric random variables with different limit laws. Other questions of a more universal nature may be asked: for example, is $\mathbf{E}H_n \geq \Omega(\sqrt{n})$ for all X with a density? Is $\mathbf{E}H_n = O(\sqrt{n})$ for all X with a symmetric density? Binary search trees extract a lot of fine detail from the underlying sequence in terms of permutations and other global properties. They thus help in the understanding of the behavior of sequences.

In the figure below, we show six rawa trees to indicate possible pathways for the proof.

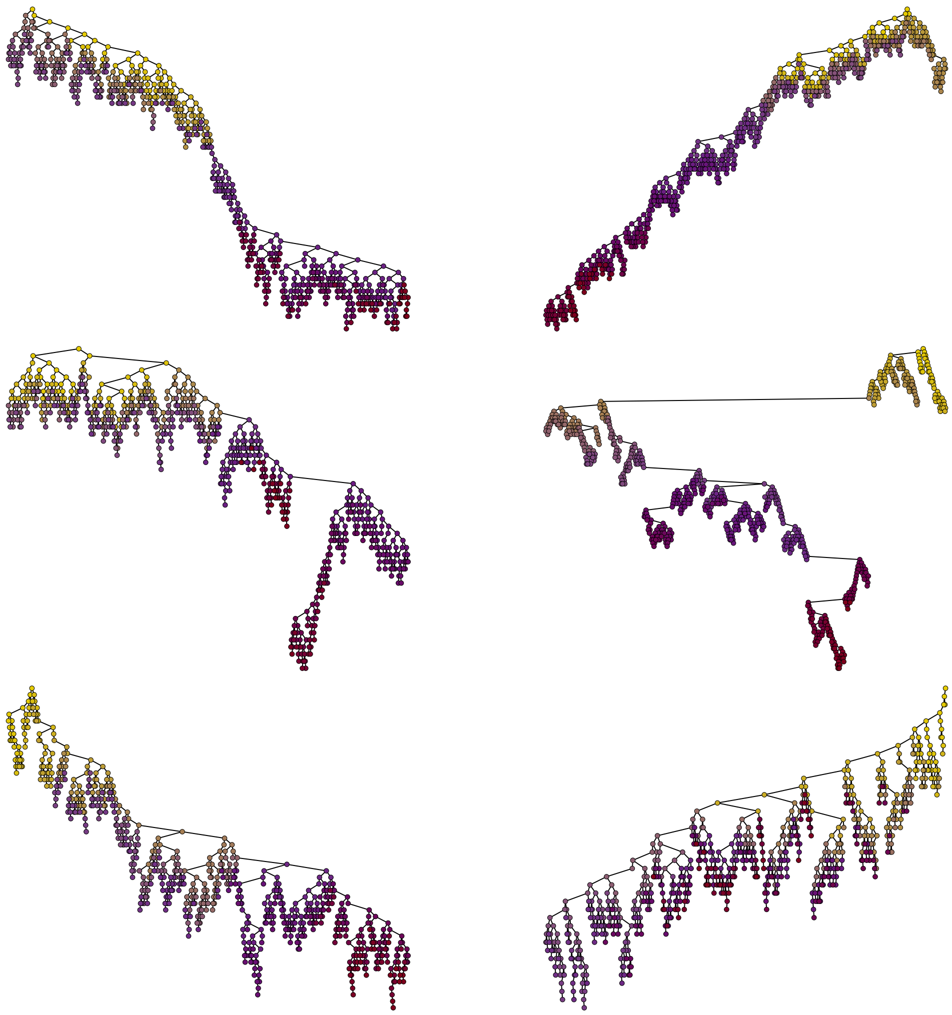

Figure 1. Six rawa trees are shown, from left to right, top to bottom: the random variables X are normal, Laplace, $s/\sqrt{|U_2|}$, $1/U_1$, $U_1U_2U_3U_4$ and $s(|U_1| + 5Z)$, where s is a random sign, Z is Bernoulli $(1/3)$ and the U_i's are uniform $[-1, 1]$.

Relationship between height and record sequences

The proof uses records in sequences. In a sequence $s = (s_0, s_1, s_2, \ldots)$, we say that s_n is an up-record if $s_n = \max\{s_0, s_1, \ldots, s_n\}$ and that it is a down-record if $s_n = \min\{s_0, s_1, \ldots, s_n\}$. Let $U(s)$ and $D(s)$ be the number of up-records and down-records in the sequence s. It is well-known (Devroye, 1988) that the path distance from the node for S_n to the root in a binary search tree is

$$D_n = U(L(S_{<n}, S_n), S_n) + D(R(S_{<n}, S_n), S_n) - 2$$

where $S_{<n} = (S_0, S_1, \ldots, S_{n-1})$. Thus, the height of the binary search tree is

$$H_n = \max_{1 \le i \le n} D_i = \max_{1 \le i \le n} \left(U(L(S_{<i}, S_i), S_i) + D(R(S_{<i}, S_i), S_i)\right) - 2 \ .$$

It is easy to see that

$$H_n \ge H'_n \stackrel{\text{def}}{=} \max\left(U(S_{<n+1}), D(S_{<n+1})\right) - 1 \ .$$

The proof is based on the following two lemmas, proved in the remainder of the paper.

Lemma 1. *Let X be a simple random variable. Then, for all $x > 0$,*

$$\lim_{n \to \infty} \mathbf{P}\left\{ \frac{H'_n}{\sqrt{2n}} < x \right\} = \mathcal{L}(x) \ ,$$

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

Lemma 2. *Let X be a nice random variable. Then for all $\epsilon > 0$,*

$$\lim_{n \to \infty} \mathbf{P}\{H_n - H'_n > \epsilon\sqrt{n}\} = 0 \ .$$

Records and ladder heights for random walks

We first consider the increasing ladder epochs and ladder heights for

$$S_0, S_1, \ldots, S_n,$$

where $S_0 = 0$. The (increasing) ladder epochs are at $0 = T_0 < T_1 < T_2 < \cdots$, where

$$T_i = \inf\{j > T_{i-1} : S_j > S_{T_{i-1}}\} \ .$$

We say that S_{T_i} is an up-record and call $S_{T_i} - S_{T_{i-1}} = \mathcal{H}_{T_i}$ an ascending ladder height. Similarly, we have decreasing ladder epochs for consecutive minima. The epochs are denoted by $0 = T'_0 < T'_1 < T'_2 < \ldots$. We say that $S_{T'_i}$ is a down-record and call $S_{T'_i} - S_{T'_{i-1}}$ a descending ladder height. Define $M_n = \max_{i \le n} S_i$, $M'_n = -\min_{i \le n} S_i$. Let $R_n = \max\{k : T_k \le n\}$ and $R'_n = \max\{k : T'_k \le n\}$. The R_n ascending ladder heights are denoted by $\mathcal{H}_1, \ldots, \mathcal{H}_{R_n}$. The absolute values of the R'_n descending ladder heights are denoted by $\mathcal{H}'_1, \ldots, \mathcal{H}'_{R'_n}$. Note that $R_n = U(S_0, S_1, \ldots, S_n) - 1$, $R'_n = D(S_0, S_1, \ldots, S_n) - 1$ and $H'_n = \max(R_n, R'_n)$. We 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 standard normal distribution function. Thus, $\mathcal{G}$ is the distribution function of the absolute value of a standard normal random variable. Most of the properties of this section are available in standard references such as Feller (1971, chapter 12) or Spitzer (1976). The first half of proposition 2 is due to Erdős and Kac (1946). For more on $\mathcal{L}$, see Rényi (1963). Feller (1971) and Spitzer (1976) may be consulted for proposition 1 and for more properties of ladder heights and ladder epochs in random walks.

Proposition 1. *Let X be a simple random variable. Then $\mathcal{H}_1, \mathcal{H}_2, \ldots$ are independent identically distributed with mean $\sigma/\sqrt{2}$. The same is true for $\mathcal{H}'_1, \mathcal{H}'_2, \ldots$.*

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

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

Furthermore,

$$\frac{R_n}{\sqrt{2n}} \xrightarrow{\mathcal{L}} \mathcal{G} \text{ and } \frac{R'_n}{\sqrt{2n}} \xrightarrow{\mathcal{L}} \mathcal{G} \; .$$

Remark: Expected values. As $\mathcal{G}$ has expected value $\sqrt{2/\pi}$, and $\mathcal{L}$ has expected value $\kappa \stackrel{\text{def}}{=} \gamma\sqrt{32/\pi^3} = 0.930527\ldots$, where $\gamma = 0.915965\ldots$ is Catalan's constant ($\sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)^2}$), it is possible to show that for simple random variables,

$$\mathbf{E}M_n \sim \sigma\sqrt{2n/\pi} \, , \, \mathbf{E}M'_n \sim \sigma\sqrt{2n/\pi}$$

and

$$\mathbf{E}\{\max(M_n, M'_n)\} \sim \kappa\sigma\sqrt{n} \; .$$

PROPOSITION 3. *Let X be a simple random variable.* $\lim_{n\to\infty} R_n = \lim_{n\to\infty} R'_n = \infty$ *almost surely and* $\mathbf{E}R_n = \Theta(\sqrt{n})$, $\mathbf{E}R'_n = \Theta(\sqrt{n})$.

The proof of the following routine fact is omitted.

LEMMA 3. *If $X_1, X_2, \ldots$ are i.i.d. random variables with finite mean m, and if $R_n \to \infty$ almost surely, then $\sum_{i=1}^{R_n} X_i/R_n \to m$ almost surely. In particular, with M_n and R_n as in proposition 2, we have $M_n/R_n \to \sigma/\sqrt{2}$ almost surely.*

PROPOSITION 4. *Let X be a simple random variable. Then, for all x,*

$$\mathbf{P}\left\{\frac{R_n}{\sqrt{2n}} < x\right\} - \mathbf{P}\left\{\frac{M_n}{\sigma\sqrt{n}} < x\right\} \to 0 \ ,$$

$$\mathbf{P}\left\{\frac{R'_n}{\sqrt{2n}} < x\right\} - \mathbf{P}\left\{\frac{M'_n}{\sigma\sqrt{n}} < x\right\} \to 0 \ ,$$

and

$$\mathbf{P}\left\{\frac{\max(R_n, R'_n)}{\sqrt{2n}} < x\right\} - \mathbf{P}\left\{\frac{\max(M_n, M'_n)}{\sigma\sqrt{n}} < x\right\} \to 0 \ .$$

PROOF. We only show the proof for the last statement. We may assume $x > 0$. For arbitrary small $\epsilon > 0$

$$\begin{aligned}
&\mathbf{P}\{\max(R_n, R'_n) < \sqrt{2n}x\} \\
&= \mathbf{P}\{\max(M_n, M'_n) < (\sigma+\epsilon)\sqrt{n}x, \max(R_n, R'_n) < \sqrt{2n}x\} + \\
&\quad + \mathbf{P}\{\max(M_n, M'_n) \geq (\sigma+\epsilon)\sqrt{n}x, \max(R_n, R'_n) < \sqrt{2n}x\} \\
&\leq \mathbf{P}\{\max(M_n, M'_n) < (\sigma+\epsilon)\sqrt{n}x\} + \\
&\quad + \mathbf{P}\left\{\frac{\mathcal{H}_1 + \ldots + \mathcal{H}_{R_n}}{R_n} > \frac{\sigma+\epsilon}{\sqrt{2}}\right\} + \mathbf{P}\left\{\frac{\mathcal{H}'_1 + \ldots + \mathcal{H}'_{R'_n}}{R'_n} > \frac{\sigma+\epsilon}{\sqrt{2}}\right\} \ .
\end{aligned}$$

By Lemma 3, the last two terms tend to zero, so

$$\limsup_{n\to\infty} \mathbf{P}\{\max(R_n, R'_n) < \sqrt{2n}x\} \leq \limsup_{n\to\infty} \mathbf{P}\{\max(M_n, M'_n) < (\sigma+\epsilon)\sqrt{n}x\},$$

and by the continuity of the limit distribution of $\max(M_n, M'_n)$ in Proposition 2

$$\limsup_{n\to\infty} \mathbf{P}\{\max(R_n, R'_n) < \sqrt{2n}x\} \leq \limsup_{n\to\infty} \mathbf{P}\{\max(M_n, M'_n) < \sigma\sqrt{n}x\} = \mathcal{L}(x).$$

On the other hand for arbitrary small $\epsilon > 0$

$$
\begin{aligned}
&\mathbf{P}\{\max(M_n, M_n') < (\sigma-\epsilon)\sqrt{n}x\} \\
&= \mathbf{P}\{\max(R_n, R_n') < \sqrt{2n}x, \max(M_n, M_n') < (\sigma-\epsilon)\sqrt{n}x\} + \\
&\quad + \mathbf{P}\{\max(R_n, R_n') \geq \sqrt{2n}x, \max(M_n, M_n') < (\sigma-\epsilon)\sqrt{n}x\} \\
&\leq \mathbf{P}\{\max(R_n, R_n') < \sqrt{2n}x\} + \\
&\quad + \mathbf{P}\left\{\frac{\mathcal{H}_1 + \ldots + \mathcal{H}_{R_n}}{R_n} < \frac{\sigma-\epsilon}{\sqrt{2}}\right\} + \mathbf{P}\left\{\frac{\mathcal{H}_1' + \ldots + \mathcal{H}'_{R_n'}}{R_n'} < \frac{\sigma-\epsilon}{\sqrt{2}}\right\}
\end{aligned}
$$

and by Lemma 3 the last two terms tend to zero, so

$$\liminf_{n\to\infty} \mathbf{P}\{\max(R_n, R_n') < \sqrt{2n}x\} \geq \liminf_{n\to\infty} \mathbf{P}\{\max(M_n, M_n') < (\sigma-\epsilon)\sqrt{n}x\} \ .$$

By the continuity of $\mathcal{L}$,

$$\liminf_{n\to\infty} \mathbf{P}\{\max(R_n, R_n') < \sqrt{2n}x\} \geq \liminf_{n\to\infty} \mathbf{P}\{\max(M_n, M_n') < \sigma\sqrt{n}x\} = \mathcal{L}(x)$$

gives the result. □

Corollary. Let X be a simple random variable. For all $x > 0$

$$\lim_{n\to\infty} \mathbf{P}\{\max(R_n, R_n')/\sqrt{2n} < x\} = \mathcal{L}(x) \ .$$

Note that this implies Lemma 1.

Concentration results for random walks

In this section, we study upper tail bounds for an empirical concentration function for the random walk,

$$Q_n(\ell) = \sup_x \sum_{i=1}^n \mathbf{1}_{[S_i \in [x, x+\ell]]}$$

for a particular range of interval sizes ℓ roughly between $1/\sqrt{n}$ and $1/n^{1/3}$. Classical concentration inequalities for S_n are nicely described by Petrov (1995, section 2.4). For example, Petrov (1995, (2.71)) shows that there exists a positive constant A (referred to below as Petrov's constant) such that uniformly over all ℓ,

$$\sup_x \mathbf{P}\{S_n \in [x, x+\ell]\} \leq \frac{A \sup_x \mathbf{P}\{X \in [x, x+\ell]\}}{\sqrt{n(1 - \sup_x \mathbf{P}\{X \in [x, x+\ell]\})}} \ ,$$

where X is the generic summand in the random walk. In particular, if X has a density f bounded by $\|f\|_\infty$, and $\ell\|f\|_\infty \leq 3/4$, then

$$\sup_x \mathbf{P}\{S_n \in [x, x+\ell]\} \leq \frac{2A\ell\|f\|_\infty}{\sqrt{n}} \ .$$

We do not expect $Q_n(\ell)$ to be substantially larger than n times the latter bound, i.e., $Q_n(\ell)$ should be $O(\ell\sqrt{n})$. This result, even without a tight constant, will suffice for the present paper. Lemma 4 takes care of this.

LEMMA 4. *Let X have density f, and let f be symmetric and bounded by $\|f\|_\infty$. Let $\ell > 0$ be so small that $\ell\|f\|_\infty \le 3/4$, and let A be Petrov's constant. Then, if $Y_i = \mathbf{1}_{[S_i\in[x,x+\ell]]}$, we have*

$$\mathbf{E}\left\{\left(\sum_{i=1}^n Y_i\right)^4\right\} \le 543\max\left(16, \lambda^4 n^2\right)$$

where $\lambda = 2A\|f\|_\infty\ell$.

PROOF. Note that Y_i is increasing in ℓ. If $\lambda\sqrt{n} < 2$, then we increase ℓ to make $\lambda\sqrt{n} = 2$. Thus, without loss of generality, we can assume that $\lambda\sqrt{n} \ge 2$, and we need only prove the inequality with $543\lambda^4 n^2$ on the right-hand-side. We proceed by repeated use of Petrov's inequality. The following simple summation bounds are easy to verify:

$$\sum_{i=1}^n \frac{1}{\sqrt{i}} \le 2\sqrt{n}\ ,$$

$$\sum_{1\le i_1<i_2\le n} \frac{1}{\sqrt{i_1}\sqrt{i_2-i_1}} \le (2\sqrt{n})^2\ ,$$

$$\sum_{1\le i_1<i_2<i_3\le n} \frac{1}{\sqrt{i_1}\sqrt{i_2-i_1}\sqrt{i_3-i_2}} \le (2\sqrt{n})^3\ ,$$

$$\sum_{1\le i_1<i_2<i_3<i_4\le n} \frac{1}{\sqrt{i_1}\sqrt{i_2-i_1}\sqrt{i_3-i_2}\sqrt{i_4-i_3}} \le (2\sqrt{n})^4\ .$$

Thus,

$$\begin{aligned}
&\mathbf{E}\left\{\left(\sum_{i=1}^n Y_i\right)^4\right\} \\
&= \sum_{1\le i_1,i_2,i_3,i_4\le n} \mathbf{E}\left\{Y_{i_1}Y_{i_2}Y_{i_3}Y_{i_4}\right\} \\
&= 24 \sum_{1\le i_1<i_2<i_3<i_4\le n} \mathbf{E}\left\{Y_{i_1}Y_{i_2}Y_{i_3}Y_{i_4}\right\} + 36 \sum_{1\le i_1<i_2<i_3\le n} \mathbf{E}\left\{Y_{i_1}Y_{i_2}Y_{i_3}\right\} \\
&\quad + 14 \sum_{1\le i_1<i_2\le n} \mathbf{E}\left\{Y_{i_1}Y_{i_2}\right\} + \sum_{1\le i_1\le n} \mathbf{E}\left\{Y_{i_1}\right\}
\end{aligned}$$

$$
\begin{aligned}
&\le 24 \sum_{1\le i_1<i_2<i_3<i_4\le n} \frac{\lambda^4}{\sqrt{i_1(i_2-i_1)(i_3-i_2)(i_4-i_3)}} \\
&\quad + 36 \sum_{1\le i_1<i_2<i_3\le n} \frac{\lambda^3}{\sqrt{i_1(i_2-i_1)(i_3-i_2)}} \\
&\quad + 14 \sum_{1\le i_1<i_2\le n} \frac{\lambda^2}{\sqrt{i_1(i_2-i_1)}} + \sum_{1\le i_1\le n} \frac{\lambda}{\sqrt{i_1}} \\
&\le 24(2\lambda\sqrt{n})^4 + 36(2\lambda\sqrt{n})^3 + 14(2\lambda\sqrt{n})^2 + (2\lambda\sqrt{n}) \\
&\le (2\lambda\sqrt{n})^4(24 + 36/4 + 14/16 + 1/64) \\
&\le 543(\lambda\sqrt{n})^4 \ . \square
\end{aligned}
$$

LEMMA 5. *Let f be a symmetric density bounded by $\|f\|_\infty$, and assume that f has finite variance σ^2. Let ℓ be a sequence of interval sizes depending upon n such that $n\ell^3 = o(1)$. Then, for every $\epsilon > 0$,*

$$
\lim_{n\to\infty} \mathbf{P}\{Q_n(\ell) \ge \epsilon\sqrt{n}\} = 0 \ .
$$

PROOF. We may assume without loss of generality that $n\ell^2 \to \infty$. If not, we increase ℓ artificially, which by virtue of the monotonicity of $Q_n(\ell)$ with respect to ℓ is allowable. We partition $\mathbf{R}$ into intervals of length 2ℓ each and let N_i denote the number of S_j's that land in the i-th interval, $1 \le j \le n$. Clearly, $Q_n(\ell) \le 2\max_i N_i$. Let B denote the set of interval indices for intervals that intersect $[-n, n]$, and note that $|B| \le 2 + n/\ell$. Let N denote the number of S_j's that fall outside $[-n, n]$. Let n be so large that $2\ell\|f\|_\infty \le 3/4$. Define $k = \epsilon\sqrt{n}/2$. We have, if X denotes the generic summand with density f,

$$
\begin{aligned}
\mathbf{P}\{N \ge k\} \le \mathbf{E}\{N\}/k &= \sum_{j=1}^{n} \mathbf{P}\{|S_j| \ge n\}/k \\
&\le \frac{\sum_{j=1}^n \mathbf{E}\{S_j^2\}}{kn^2} = \frac{\sum_{j=1}^n j\mathbf{E}\{X^2\}}{kn^2} \\
&= \frac{(n+1)\sigma^2}{2kn} \le \frac{\sigma^2}{k} \ .
\end{aligned}
$$

Let A be Petrov's constant. By Lemma 4, if n is so large that $4A\ell\|f\|_\infty\sqrt{n} \ge 2$,

$$
\begin{aligned}
\mathbf{P}\{\max_{i\in B} N_i \ge k\} &\le |B| \max_i \mathbf{P}\{N_i \ge k\} \\
&\le \frac{|B|\max_i \mathbf{E}\{N_i^4\}}{k^4} \\
&\le \frac{543|B|(4A\|f\|_\infty\ell)^4 n^2}{k^4} \ .
\end{aligned}
$$

Thus,

$$\begin{aligned}\mathbf{P}\{Q_n(\ell) \geq 2k\} &\leq \mathbf{P}\{N \geq k\} + \mathbf{P}\{\max_{i \in B} N_i \geq k\} \\ &= O(1/k) + O\left(n\ell^3\right) \\ &= o(1) \ . \ \square\end{aligned}$$

The height of a rawa tree: proof of Lemma 2

Throughout this section, we set $\ell = 1/n^{4/10}$. Observe that the condition of Lemma 5, $n\ell^3 \to 0$, holds. We call the depth $D(x)$ of x the path distance from x to the root for the binary search tree defined by $S_0, S_1, \ldots, S_n, x$. Consider the collection $C = \{i\ell : i \text{ integer}\}$. Recalling the definition of $Q_n(\ell)$ from the previous section, we note that the height H_n of the rawa tree satisfies

$$H_n \leq Q_n(\ell) + \max_{x \in C} D(x) \ .$$

Fix $x > 0$. Let $1 \leq T_1 < T_2 < \cdots$ be the epochs at which the random walk reaches a maximum: so, T_{i+1} is the first index greater than T_i for which $S_{T_{i+1}} > S_{T_i}$. We define $A(x)$ and $B(x)$ as follows: let i be the unique index for which $S_{T_i} \leq x < S_{T_{i+1}}$, where $S_{T_{i+1}}$ is replaced by ∞ if $T_{i+1} > n$ (to take care of the rightmost interval); then set $A(x) = S_{T_i}$ and $B(x) = S_{T_{i+1}}$. A similar symmetric definition is used for $x < 0$ with possibly $A(x) = -\infty$. Observe that

$$H_n \leq Q_n(\ell) + \max_{x \in C : |A(x)| + |B(x)| < \infty} D(x) \ .$$

For $x > 0$, there are no S_j's strictly in $(A(x), B(x))$ with $j \leq T_{i+1}$. We condition on the history of the random walk up to T_{i+1} and call it $\mathcal{F}$. Note that $D(x)$ is bounded by the sum of $i + 2$ and the number of local left records, plus the number of local right records. A local left record is a record value (maximum) among those S_j's that fall in $(A(x), x)$. Clearly, each value of j must exceed T_{i+1}. A local right record is a record value (minimum) among those S_j's that fall in $(x, B(x))$. Observe that $i + 1 \leq H_n'$ if $|A(x)| + |B(x)| < \infty$. Denoting by $L(x)$ and $R(x)$ the number of local left records and local right records for x respectively, we see that $D(x) \leq i + 2 + L(x) + R(x)$, and that

$$H_n - H_n' \leq Q_n(\ell) + \max_{x \in C : |A(x)| + |B(x)| < \infty} L(x) + \max_{x \in C : |A(x)| + |B(x)| < \infty} R(x) \ .$$

The proof is complete if we can show that for each $\epsilon > 0$,

$$\lim_{n \to \infty} \mathbf{P}\left\{Q_n(\ell) > \epsilon\sqrt{n}\right\} = 0 \ ,$$

$$\lim_{n \to \infty} \mathbf{P}\left\{\max_{x \in C : |A(x)| + |B(x)| < \infty} L(x) > \epsilon\sqrt{n}\right\} = 0 \ ,$$

and

$$\lim_{n\to\infty} \mathbf{P}\left\{ \max_{x\in C:|A(x)|+|B(x)|<\infty} R(x) > \epsilon\sqrt{n} \right\} = 0 \ .$$

The first part was shown in Lemma 5. We will show the second part involving $L(x)$, as the third part is similar.

Given $\mathcal{F}$, we may have large values for $L(x)$ if $B(x) - A(x)$ is large. So, let G be the event

$$\left[\max_{x\in C:|A(x)|+|B(x)|<\infty} |B(x) - A(x)| \le \delta\sqrt{n} \right] \ ,$$

where $\delta > 0$ is to be picked as a function of ϵ. Let G^c denote its complement. Clearly,

$$\begin{aligned} \mathbf{P}\{G^c\} &\le \mathbf{P}\left\{ \max_i |X_i| > \delta\sqrt{n} \right\} \\ &\le n\mathbf{P}\left\{ |X_1| > \delta\sqrt{n} \right\} \\ &\le n \frac{\mathbf{E}\left\{ X_1^2 \mathbf{1}_{[|X_1|>\delta\sqrt{n}]} \right\}}{\delta^2 n} \\ &= o(1) \end{aligned}$$

if $\mathbf{E}\left\{X_1^2\right\} < \infty$. Thus, by Lemma 5,

$$\begin{aligned} &\mathbf{P}\left\{ \max_{x\in C:|A(x)|+|B(x)|<\infty} L(x) > \epsilon\sqrt{n} \right\} \\ &\le \mathbf{P}\{G^c\} + \mathbf{P}\{Q_n(\ell) > (\epsilon/3)\sqrt{n}\} \\ &\quad + \mathbf{P}\left\{ G, Q_n(\ell) \le (\epsilon/3)\sqrt{n}, \max_{x\in C:|A(x)|+|B(x)|<\infty} L(x) > \epsilon\sqrt{n} \right\} \\ &\le o(1) \\ &\quad + \mathbf{P}\left\{ \bigcup_{\substack{x\in C: \\ |A(x)|+|B(x)|<\infty}} \left[Q_n(\ell) \le (\epsilon/3)\sqrt{n}, |B(x) - A(x)| \le \delta\sqrt{n}, L(x) > \epsilon\sqrt{n} \right] \right\} \\ &\le o(1) \\ &\quad + \mathbf{E}\left\{ \sum_{\substack{x\in C: \\ |A(x)|+|B(x)|<\infty}} \mathbf{1}_{[|B(x)-A(x)|\le\delta\sqrt{n}]} \mathbf{P}\left\{ Q_n(\ell) \le (\epsilon/3)\sqrt{n}, L(x) > \epsilon\sqrt{n} \middle| \mathcal{F} \right\} \right\} . \end{aligned}$$

We show that we can find $\delta > 0$ so that the last term is $o(1)$.

Using the chain of inequalities for N in the proof of Lemma 5, we see that the probability that there is one occurrence of $|S_j| > n^2$ is bounded by the expected number of such occurrences, which by Chebyshev's inequality does not exceed σ^2/n^2. Thus, with probability at least $1 - O(1/n^2)$, the number of $x \in C$ with

$|A(x)| + |B(x)| < \infty$ is not more than $2 + 2n^2/\ell = O(n^{24/10})$. By trivial bounding then, it would suffice if we can show that for an appropriate choice of $\delta > 0$, and uniformly over all x, and all histories $\mathcal{F}$ with $|B(x) - A(x)| \leq \delta\sqrt{n}$,

$$\mathbf{P}\left\{Q_n(\ell) \leq (\epsilon/3)\sqrt{n}, L(x) > \epsilon\sqrt{n}\middle|\mathcal{F}\right\} \leq \exp\left(-cn^{1/10}\right)$$

for some $c > 0$. We drop the conditioning in the notation. By condition (C) on f (see the definition of nice random variables), we find a pair $b > 0$ and $r > 0$ such that $\inf_{|x| \leq r} f(x) \geq b > 0$. Any such pair will do. Then we partition $[A(x), x]$ into intervals of length $r/2$ starting from $A(x)$, taking care not to cover x (thus leaving an interval of length less than r that reaches x; call that interval I). Let J be the rightmost interval. Replace J by $J \cup I - [x - \ell, x]$ and I by $[x - \ell, x]$. Thus, $r - \ell \geq |J| \geq r/2 - \ell \geq r/3$ for n large enough. Consider the following process: a local left record S_j arrives in one of the intervals (I and J excluded). At that time, given that we are in an interval to the left of J, there is a probability of at least $(r/3)b$ that S_{j+1} hits one of the intervals to the right, J included. Picture this as a success. Let the number of intervals up to and including J be k, and note that $k \leq 1 + 2\delta\sqrt{n}/r$. Let N^* be the number of local left records in the intervals to the left of J. In what follows, we set $m = \lfloor(\epsilon/3)\sqrt{n}\rfloor$. $[N^* > m]$ implies that in m such trials we had less than k successes. In other words,

$$\mathbf{P}\{N^* > m\} \leq \mathbf{P}\{\text{binomial}(m, br/3) < k\} .$$

If we set $k < mbr/7$, then this probability is $\exp\left(-\Omega(\sqrt{n})\right)$, by Hoeffding's inequality (Hoeffding, 1963). This condition is satisfied for n large enough when $\delta = \epsilon br^2/43$, which is the choice we will adopt. Let N' be the number of local left records in J. Here, we move to I with probability at least $b\ell$, so that by the same argument,

$$\begin{aligned}\mathbf{P}\{N' > m\} &\leq \mathbf{P}\{\text{binomial}(m, b\ell) = 0\} = (1 - b\ell)^m \\ &\leq \exp\left(-mb\ell\right) = \exp\left(-\Omega\left(n^{1/10}\right)\right) .\end{aligned}$$

Finally, let N'' be the number of local left records in I. This is clearly bounded by $Q_n(\ell)$, uniformly over all x and all $A(x), B(x)$. As $L(x) \leq N^* + N' + Q_n(\ell)$, we see that

$$\begin{aligned}\mathbf{P}\{Q_n(\ell) \leq (\epsilon/3)\sqrt{n}, L(x) > \epsilon\sqrt{n}\} &\leq \mathbf{P}\{N^* > m\} + \mathbf{P}\{N' > m\} \\ &= \exp\left(-\Omega\left(n^{1/10}\right)\right) .\end{aligned}$$

This concludes the proof of Lemma 2. □

References

T. H. Cormen, C. E. Leiserson, and R. L. Rivest, *Introduction to Algorithms*, MIT Press, Boston, 1990.

L. Devroye, "A note on the height of binary search trees," *Journal of the ACM*, vol. 33, pp. 489–498, 1986b.

L. Devroye, "Branching processes in the analysis of the heights of trees," *Acta Informatica*, vol. 24, pp. 277–298, 1987.

L. Devroye, "Applications of the theory of records in the study of random trees," *Acta Informatica*, vol. 26, pp. 123–130, 1988.

P. Erdős and M. Kac, "On certain limit theorems of the theory of probability," *Bulletin of the American Mathematical Society*, vol. 52, pp. 292–302, 1946.

W. Feller, *An Introduction to Probability Theory and its Applications, Volume 2*, John Wiley, New York, 1971.

I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series and Products*, Academic Press, New York, 1980.

W. Hoeffding, "Probability inequalities for sums of bounded random variables," *Journal of the American Statistical Association*, vol. 58, pp. 13–30, 1963.

D. E. Knuth, *The Art of Computer Programming, Vol. 3 : Sorting and Searching*, Addison-Wesley, Reading, MA, 1973.

H. M. Mahmoud, *Evolution of Random Search Trees*, John Wiley, New York, 1992.

V. V. Petrov, *Limit Theorems of Probability Theory*, Oxford University Press, 1995.

J. M. Robson, "The height of binary search trees," *The Australian Computer Journal*, vol. 11, pp. 151–153, 1979.

A. Rényi, "On the distribution function $L(z)$," *Selected Translations in Mathematical Statistics and Probability*, vol. 4, pp. 219–224, 1963.

F. Spitzer, *Principles of Random Walk (2nd ed)*, Springer-Verlag, New York, 1976.

R. L. Wheeden and A. Zygmund, *Measure and Integral*, Marcel Dekker, New York, 1977.