The distribution of the height of random BSTs

Marko R. Riedel

May 21, 2001

1 Introduction

Binary search trees (BSTs) are an important and commonly used data structure whose applications and characteristics have been and are the subject of extensive study. The reader should consult [MP97] or [FGM96] for historical reviews and recent results that concern binary search tree statistics. Random binary search trees (RBSTs) are BSTs obtained by performing n random insertions into an initially empty tree. We may also view RBSTs as a probability distribution on BSTs, where the distribution is the one produced by selecting a size n permutation uniformly at random and inserting its elements into a BST.

The height of a RBST is the maximum path length from the root to a node. Its asymptotics (mean, i.e. first moment) have been studied by Devroye, whose work is discussed in e.g. [VF90] and [SF96]. Owing to the fact that many tree statistics have straightforward recursive definitions, they are usually studied by means of generating functions that reflect the recursion via functional equations. It is worth pointing out that height is a statistic that yields generating functions that are somewhat more involved than those associated to, e.g., enumeration problems or internal path length. In particular, little is known about the distribution of the height of RBSTs, although it is generally conjectured that it is highly concentrated around its mean.

This paper's contribution is a tail inequality for said probability distribution. In consideration of the difficulties encountered by the standard methodology, i.e. the use of generating functions, we shift focus and construct an RBST height martingale in order to apply Azuma's inequality to get the tail bound. The procedure is straightforward. There are three steps:

- Recall the isomorphism that maps permutations to inversion tables and some of its basic properties.
- Express the RBST height statistic as a functional on the universe of permutations and verify that it fulfills the Lipschitz condition.
- Apply Azuma's inequality, making use of the fact that the Lipschitz condition implies the bounded difference property for martingales.

We start by reviewing permutations, RBSTs and martingales.

2 Preliminaries and notation

2.1 Permutations

We use the shorthand $[n]$ to denote the interval $[1, n] \subset \mathbb{N}$. A permutation π of size $|\pi| = n$ is a function $\pi : [n] \mapsto [n]$ where $|\pi([n])| = n$. The set Π_n is the set of permutations of size n.

We denote permutations in two ways, i.e. by lists and inversion tables. Consider $\pi \in \Pi_n$. The string $\pi(1)\pi(2)\ldots\pi(n) \in [n]^n$ identifies π. We use σ_i to refer to $\pi(i)$ throughout this paper. Hence $\pi = \sigma_1\sigma_2\ldots\sigma_n$. The inverse permutation π^{-1} of π maps σ_i to i and its elements are referenced by σ_i^{-1}, so that $\pi(\sigma_i^{-1}) = \sigma_i$. Note that σ_i^{-1} is the position of i in $\sigma_1\sigma_2\ldots\sigma_n$.

Fix π. The inversion table $\{y_j\}_{j=1}^{n}$ of the permutation π is given by $y_j = \left|\left\{ k \mid \sigma_k^{-1} < \sigma_j^{-1} \wedge k > j \right\}\right|$, i.e. the number of left inversions of j (elements that are larger than and to the left of j) in $\sigma_1\sigma_2\ldots\sigma_n$. We have the isomorphism

$$\Pi_n = [0..n-1] \times [0..n-2] \times \ldots \times [0..1] \times [0].$$

Consult [SF96] or [VF90] for background readings.

The image $\pi([a, b])$ of $[a, b]$ under π is denoted $\pi(a\ldots b)$.

The restriction $\pi_{[a,b]}$ of the permutation π to the interval $[a, b]$ is defined by the inversion table $\{y_j\}_{j=1}^{b-a+1}$, where

$$y_{j-a+1} = \left|\left\{ k \mid \sigma_k^{-1} < \sigma_j^{-1} \wedge k > j \wedge k \in [a, b]\right\}\right|$$

and is an element of Π_{b-a+1}. We write $\pi_{\geq a} \in \Pi_{n-a+1}$ and $\pi_{\leq b} \in \Pi_b$ to denote the restrictions $\pi_{[a,n]}$ and $\pi_{[1,b]}$, and further define $\pi_{<a}$, $\pi_{>b}$ and $\pi_{\neq c}$.

2.2 Random BSTs

The set $\mathcal{B}$ of binary trees is the set of binary trees that are either leaves $\square$ or binary nodes $(\cdot, \circ, \cdot)$ whose two children are binary trees, i.e.

$$\mathcal{B} = \square + \mathcal{B} \times \{\circ\} \times \mathcal{B}.$$

These trees are also known as Catalan trees. The size $|b|$ of a binary tree $b \in \mathcal{B}$ is the number of nodes, i.e. $|\square| = 0$ and $|(b_1, \circ, b_2)| = |b_1| + 1 + |b_2|$. Introduce $\mathcal{B}_n$, the set of Catalan trees of size n, and let $b_n = |\mathcal{B}_n|$.

We may define a set $\mathcal{I}$ of freely labelled binary trees, as follows.

$$\mathcal{I} = \square + \mathcal{I} \times \mathbb{N} \times \mathcal{I}.$$

The set of labels $l(u)$ of $u \in \mathcal{I}$ is $\emptyset$ if $u = \square$ and $l(u_1) \cup \{v\} \cup l(u_2)$ if $u = (u_1, v, u_2)$. The Catalan tree $B(u)$ that corresponds to $u \in \mathcal{I}$ is given by $B(\square) = \square$ and

$$B((u_1, v, u_2)) = (B(u_1), \circ, B(u_2)).$$

The subtree relation $v \preceq u$ holds iff $u = v$ or $u = (u_1, \cdot, u_2)$ and $v \preceq u_1$ or $v \preceq u_2$. The relation $v \prec u$ excludes the case $u = v$.

The set $\mathcal{T}$ of labelled binary trees is the set

$$\{ u \mid u \in \mathcal{I} \wedge l(u) = [|u|] \} .$$

Introduce the set $\mathcal{T}_n$ of size n labelled binary trees. Note that $|\mathcal{T}_n| = b_n n!$.

Let $n \geq 1$ and define the set $\mathcal{A}_n$ as

$$\{ u \mid u \in \mathcal{T}_n \wedge \forall v = (v_1, v, v_2) \preceq u : \ (\forall l \in l(v_1) : \ l < v \wedge \forall l \in l(v_2) : \ l > v) \}$$

The set $\mathcal{A}$ of binary search trees (BSTs) is defined by

$$\mathcal{A} = \bigcup_{n \geq 1} \mathcal{A}_n.$$

We point out that $B : \ \mathcal{A} \mapsto \mathcal{B}$ is one-to-one (the cardinality of the left and right subtrees determines the label of the root), hence $|\mathcal{A}_n| = b_n$. Let $B^{-1} : \ \mathcal{B} \mapsto \mathcal{A}$ be the inverse of B on $B(\mathcal{A})$.

Given $f : \ \mathbb{N} \mapsto \mathbb{N}$, we may relabel $u \in \mathcal{I}$ to obtain $S(u, f)$, where $S(\Box, f) = \Box$ and $S((u_1, v, u_2), f) = (S(u_1, f), f(v), S(u_2, f))$. Introduce the shift functions $\triangle_k$, $k \in \mathbb{N}$, where $\triangle_k = \{(i, i+k)\}_{i \in \mathbb{N}}$.

Define $T(\pi)$, the binary search tree that corresponds to $\pi \in \Pi_n$, as follows.

$$T(\pi) = \begin{cases} (\Box, \sigma_1, \Box) & \text{if} \quad |\pi| = 1, \\ (\Box, 1, S(T_{>1}, \triangle_1)) & \text{if} \quad |\pi| > 1 \wedge \sigma_1 = 1, \\ (T_{<n}, n, \Box) & \text{if} \quad |\pi| > 1 \wedge \sigma_1 = n, \\ (T_{<\sigma_1}, \sigma_1, S(T_{>\sigma_1}, \triangle_{\sigma_1})) & \text{if} \quad |\pi| > 1 \wedge \sigma_1 \notin \{1, n\}. \end{cases}$$

Note that we may consider various probability distributions on the universe $\mathcal{B}_n$ of Catalan trees of size n, e.g. under the uniform distribution,

$$\mathrm{P}\,[b] = \frac{1}{b_n}$$

for all $b \in \mathcal{B}_n$. We obtain a second distribution if we insert every size n permutation into a binary tree and remove labels. Assume that we have a uniform distribution on the permutations in Π_n. The result is a multiset over $\mathcal{B}_n$, where the probability of $b \in \mathcal{B}_n$ is the instance count divided by the total cardinality, i.e.

$$\mathrm{P}\,[b] = \frac{|\{ \pi \mid \pi \in \Pi_n \wedge B(T(\pi)) = b\}|}{n!},$$

The set of random BSTs and their associated probabilities is precisely the set of trees obtained under this distribution. We have seen that we may represent its elements b by either one of $b \in \mathcal{B}$ and $B^{-1}(b) \in \mathcal{A}$.

We introduce

$$\Pi_n^{=b} = \{ \pi \mid \pi \in \Pi_n \wedge B(T(\pi)) = b\} \subseteq \Pi_n$$

to simplify notation, so that

$$\mathrm{P}\,[b] = \frac{\left|\Pi_n^{=b}\right|}{n!}.$$

We derive an alternate expression for $\mathrm{P}\,[b]$ under the second, so-called random permutation model. Let $b = (b_1, \circ, b_2)$. Note that $B(T(\pi)) = b$ iff $T(\pi) = B^{-1}(b)$. We restrict our attention to the last of the four cases that serve to define $T(\pi)$. Let $u = (u_1, v, u_2) = B^{-1}(b)$. Evidently $T(\pi) = B^{-1}(b)$ iff

$$T(\pi_{<v}) = u_1 \quad \text{and} \quad S(T(\pi_{>v}), \triangle_v) = u_2.$$

Let $\pi_1 \in \Pi_a$ and $\pi_2 \in \Pi_b$ and note that

$$|\{\pi \mid \pi \in \Pi_{a+b} \wedge \pi_{\leq a} = \pi_1 \wedge \pi_{>a} = \pi_2\}| = \binom{a+b}{a}.$$

We conclude that

$$\left|\Pi_n^{=b}\right| = \binom{n-1}{v-1} \left|\Pi_v^{=b_1}\right| \left|\Pi_{n-v}^{=b_2}\right|$$

and substitute back into $\mathrm{P}\,[b]$ to obtain

$$\mathrm{P}\,[b] = \frac{1}{n!}\;(v-1)!\mathrm{P}\,[b_1]\;(n-v)!\mathrm{P}\,[b_2]\;\binom{n-1}{v-1} = \frac{1}{n}\mathrm{P}\,[b_1]\,\mathrm{P}\,[b_2]\,.$$

Hence a tree b of size $|b| = n$ has probability

$$\mathrm{P}\,[b] = \prod_{a \preceq B^{-1}(b)} \frac{1}{|a|}$$

under this model, the product being extended over all subtrees $a \preceq B^{-1}(b)$ of the RBST $B^{-1}(b)$. (We need to distinguish multiple occurences of the same binary subtree and hence may not extend the product over the subtrees of b itself.) Fig. 1 illustrates the preceding remarks. Consult [FGM96] for more information.

The height $h(b)$ of a binary tree $b \in (\mathcal{B} - \{\square\})$ is the number of edges on the longest path from the root to a node whose children are leaves, i.e.

$$h:\ (\mathcal{B} - \{\square\}) \mapsto \mathbb{N}$$

according to

$$h((\square, \circ, \square)) = 0,$$
$$h((\square, \circ, b) = h((b, \circ, \square) = 1 + h(b) \quad \text{if} \quad b \neq \square$$

and

$$h((b_1, \circ, b_2)) = 1 + \max h(\{b_1, b_2\}) \quad \text{if} \quad \square \notin \{b_1, b_2\}.$$

Define the expectation A_n of h for a RBST of size n as

$$A_n = \sum_{a \in \mathcal{A}_n} \mathrm{P}\,[B(a)]\, h(B(a)) = \sum_{b \in \mathcal{B}_n} \mathrm{P}\,[b]\, h(b).$$

This quantity has been studied by Devroye; consult [VF90, p. 485] for an overview of the generating functions concerned. Devroye shows that A_n is asymptotic to $c \log n$, where $c = 4.311070\ldots$ is the root ≥ 2 of $(2e/c)^c = e$.

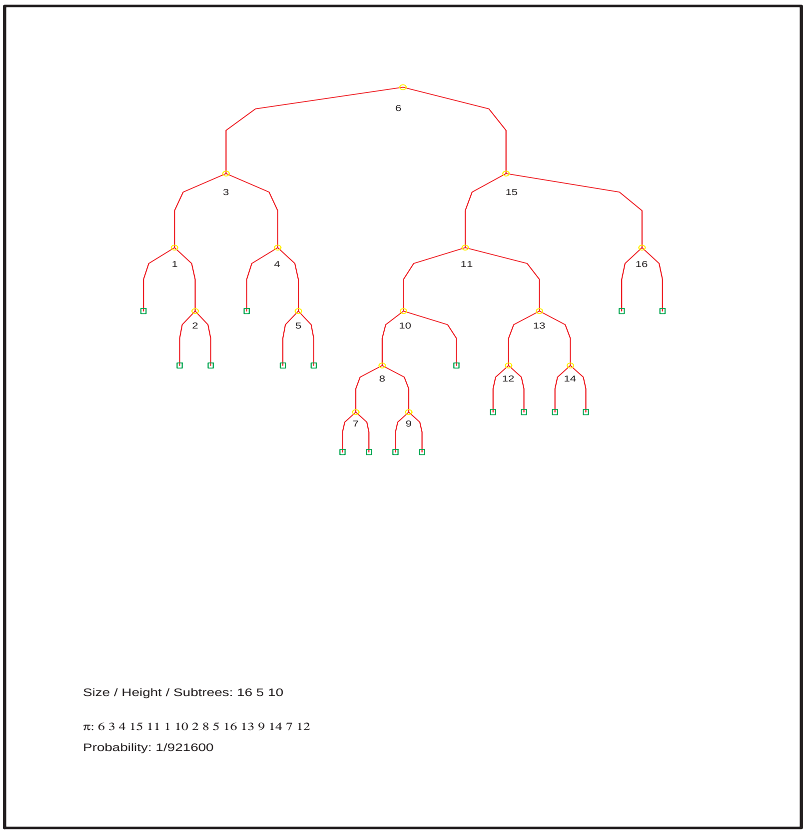

Figure 1: A random element π of Π_{16} and the BST $T(\pi)$ that corresponds to π.

2.3 Martingales

We use the framework of [AS92, p. 89]. The purpose of the following series of definitions is to facilitate the statement of Azuma's inequality.

Consider a set (universe) $\Omega = A^V$ of functions $g:\ V \mapsto A$, i.e. we obtain the elements g of Ω by setting every one of $m = |V|$ variables V, where

$$V = \{v_1, v_2, \ldots v_m\},$$

to one of $|A|$ possible values from the range A. Note that we may identify g by a vector $\mathbf{a} \in A^m$, so that

$$(g:\ V \mapsto A) = \{\, (v_i, a_i) \mid i \in [m]\}\,.$$

We assume that the variables are mutually independent and introduce a probability measure

$$\mathrm{P}\left[g(v) = a\right] = p_{a,v} \quad \text{where} \quad \sum_{a \in A} \mathrm{P}\left[g(v) = a\right] = 1.$$

This measure induces a probability distribution on Ω. An element g has probability

$$\mathrm{P}\left[g\right] = \prod_{i \in [m]} \mathrm{P}\left[g(v_i) = a_i\right]$$

The reader may verify that indeed

$$\sum_{g \in \Omega} \mathrm{P}\left[g\right] = 1.$$

Let $h \in \Omega$, $C \subseteq V$, and fix a gradation $\{V_i\}_{i=0}^{l}$ of the set V, i.e.

$$\emptyset = V_0 \subset V_1 \subset V_2 \subset \ldots V_{l-1} \subset V_l = V$$

and define the sets

$$\Omega(V_i, h) = \{\, g \mid v \in V_i \Rightarrow g(v) = h(v)\}$$

and

$$\Delta(C, h) = \bigcup_{V - C \subseteq D \subseteq V} \Omega(D, h).$$

The set $\Omega(V_i, h)$ is the set of g that agree with h on the set V_i of variables, while $\Delta(C, h)$ is the set of g that differ from h on all of the variables in C, or some subset thereof. (I.e. the g that agree with h on all of the variables not in C, or some superset thereof.)

Introduce a functional (random variable)

$$L:\ \Omega \mapsto \mathbb{R}$$

on $\Omega = A^V$ (L maps functions to real numbers) and define the expectation of L over a subset $H \subseteq \Omega$, as follows.

$$E[L \mid H] = \sum_{g \in H} L(g) \mathrm{P}[g \mid H].$$

Define the sequence $\{X_i\}_{i=0}^{l}$ of random variables $X_i : \Omega \mapsto \mathbb{R}$ by setting

$$X_i(h) = E[L \mid \Omega(V_i, h)].$$

The random variable $X_i(h)$ is the expectation of L over those g that agree with h on the variables in V_i and the sequence $\{X_i\}_{i=0}^{l}$ is a martingale. Note that X_0 is the constant function $X_0 : \Omega \mapsto \mu$ that maps every g to the expectation $E[L] = \mu$ of L over Ω, and that $X_l = L$.

We say that the functional L satisfies the Lipschitz condition if there exists a sequence $\{c_i\}_{i=1}^{l}$ of constants, possibly dependent on i, such that

$$\forall i \in [0, l-1]\ h_1 \in \Omega :\ h_2 \in \Delta(V_{i+1} - V_i, h_1) \Rightarrow |L(h_2) - L(h_1)| \leq c_{i+1}.$$

If L satisfies the Lipschitz condition, then

$$\forall h \in \Omega\ \forall i \in [0, l-1] :\ |X_{i+1}(h) - X_i(h)| \leq c_{i+1}.$$

This inequality is known as the bounded difference property.

Let $\lambda \in \mathbb{R}^+, h \in \Omega$, and introduce $\{\chi_i\}_{i=1}^{l}$ where

$$\chi_i = \sum_{j=1}^{i} c_j^2.$$

Then $\forall i \in [1, l]$

$$\mathrm{P}[X_i(h) > \mu + \lambda] < e^{-\lambda^2/(2\chi_i)} \quad \text{and} \quad \mathrm{P}[X_i(h) < \mu - \lambda] < e^{-\lambda^2/(2\chi_i)}.$$

This is Azuma's inequality.

3 The RBST height martingale

We connect the height statistic on RBSTs to the martingale framework. The universe Ω is the set Π_n of length n permutations. We use inversion tables to identify the permutations $g \in \Omega$, i.e. we view g as a function

$$g :\ \{y_j\}_{j=1}^{n-1} \mapsto [0, n-1] \subset \mathbb{N},$$

(The set V of variables is the sequence $\{y_j\}_{j=1}^{n-1}$ and the range A is the set $\{0, 1, 2, \ldots n-1\}$. The variable y_n is not included because it is zero, a constant.)

The next step is to define the probabilities $\mathrm{P}[g(y_j) = a]$. We have a uniform distribution on the set $\Pi_n = \Omega$, hence

$$\mathrm{P}[g(y_j) = a] = \begin{cases} 1/(n-j+1) & \text{if} \quad a \in [0, n-j] \\ 0 & \text{otherwise.} \end{cases}$$

The gradation $\{V_i\}_{i=0}^{n-1}$ is the obvious one. We have $l = n-1$ and

$$\emptyset = V_0 \subset \{y_1\} = V_1 \subset \{y_1, y_2\} = V_2 \subset \ldots \{y_j\}_{j=1}^{n-2} = V_{n-2} \subset V_{n-1} = V,$$

i.e. we fill the inversion table from left to right.

The choice of gradation also defines the sets $\Omega(V_i, h)$ and $\Delta(V_{i+1} - V_i, h)$. The first is the set of permutations whose inversion table agrees with h on the first i entries, the second, the set of permutations that differ from h by at most one entry, namely y_{i+1}. It can be enumerated by fixing all other entries and letting the value of y_{i+1} run through $[0, n-(i+1)]$.

It remains to choose the functional $L:\ \Omega \mapsto \mathbb{R}$. We let

$$(L:\ \Pi_n \mapsto \mathbb{N}) = h \circ B \circ T.$$

This yields the martingale $\{X_i\}_{i=0}^{n-1}$, where $X_i(h)$ is the expected height of a RBST that corresponds to a permutation that agrees with h on the first i entries of its inversion table.

We wish ultimately to apply Azuma's inequality. Hence we must show that $\{X_i\}$ has the bounded difference property, i.e. we must verify that L fulfills the Lipschitz condition. *Ergo* we ask the following question. Choose a permutation $h \in \Pi_n$ and consider the set $\Delta(j, h)$ of permutations whose inversion table agrees with h except possibly at position j, where $j \in [1, n-1]$. What is the range of L on $\Delta(j, h)$?

3.1 RBSTs and interleaved permutations

Claim. Let $\pi_1 \in \Pi_a$ and $\pi_2 \in \Pi_b$, where $a, b \geq 1$. If $\pi \in \Pi_{a+b}$ and $\pi_{\leq a} = \pi_1$, $\pi_{>a} = \pi_2$, then

$$\max\{L(\pi_1), L(\pi_2)\} \leq L(\pi) \leq L(\pi_1) + L(\pi_2) + 1.$$

Proof. Note that π is one of $\binom{a+b}{a}$ possible interleavings of π_1 and $\triangle_a \circ \pi_2$.

The left fringe $p_l(u) \in [n]^+$ of a RBST $u \in \mathcal{A}_n$ is the unique path from the root to $\min l(u)$, the right fringe $p_r(u) \in [n]^+$ the path from the root to $\max l(u)$, i.e. $p_l(\Box) = \lambda$, $p_l((u_1, v, \cdot)) = v\ p_l(u_1)$ and $p_r(\Box) = \lambda$, $p_r((\cdot, v, u_1)) = v\ p_r(u_1)$.

Let $t_1 = T(\pi_1)$ and $t_2 = S(T(\pi_2), \triangle_a)$. The RBST t_1 consists of a path $p_r(t_1)$, and a sequence of possibly empty subtrees that are left children of the nodes along this path; t_2 consists of the nodes on $p_l(t_2)$ and their right subtrees.

Observe that $\max l(t_1) < \min l(t_2)$, and hence the tree $t = T(\pi)$ may be decomposed into a central path that contains the nodes on $p_r(t_1)$ and $p_l(t_2)$, left subtrees that originate in t_1, and right subtrees that originate in t_2.

Start with the upper bound. Consider any path from the root of t to a leaf. The height of the central path is precisely $|p_r(t_1)| + |p_l(t_2)| - 1$. Left subtrees, originally from t_1 and now attached to the central path, are lowered by at most $|p_l(t_2)|$, while right subtrees, originally from t_2, are lowered by at most $|p_r(t_1)|$. Note that $|p_l(t_2)| \leq h(B(t_2)) + 1$ and $|p_r(t_1)| \leq h(B(t_1)) + 1$. We distinguish three cases.

- The central path is maximal. Then

$$h(B(t)) \leq (h(B(t_1)) + 1 + h(B(t_2)) + 1) - 1 = h(B(t_1)) + h(B(t_2)) + 1.$$

- There is a maximal path that terminates in a subtree $s \preceq t_1$. The height of such a path is at most

$$h(B(t_1)) + |p_l(t_2)| \leq h(B(t_1)) + h(B(t_2)) + 1.$$

- There is a maximal path that terminates in $s \preceq t_2$, and its height is at most
$$h(B(t_2)) + |p_r(t_1)| \leq h(B(t_2)) + h(B(t_1)) + 1.$$

These are the only possible cases and we have the upper bound. Fig. 2 illustrates the classification used in the proof.

It remains to prove the lower bound. We say that the root k of $s \preceq t$ is a fringe node if $s = (\Box, k, \Box)$, i.e. if the children of k are leaves. The height of t is the maximum distance from the root to a fringe node s. Note that the fringe nodes of t are the fringe nodes of t_1 and t_2, except for one of the two fringe nodes that terminate $p_r(t_1)$ and $p_l(t_2)$, respectively. The central path of t contains an initial segment whose nodes coincide with the initial segment of either $p_r(t_1)$ or $p_l(t_2)$. All subtrees $s \prec t$ orginally from t_1 or t_2 are lowered except those attached to nodes on the initial segment. Hence no path in t_1 or t_2 to a fringe node not on $p_r(t_1)$ and $p_l(t_2)$ is shortened. The two fringe nodes (if any) at the ends of $p_r(t_1)$ and $p_l(t_2)$ in t_1 and t_2 are replaced by a single fringe node at distance $|p_r(t_1)| + |p_l(t_2)| - 1$, which is $> |p_r(t_1)| - 1$ and $> |p_l(t_2)| - 1$, because $a, b \geq 1$. Therefore none of the paths that contribute to $\max\{L(\pi_1), L(\pi_2)\}$ except for $p_r(t_1)$ and $p_l(t_2)$ are shortened, and the latter are replaced by a single path that is strictly larger than either; moreover, $L(\pi)$ is the maximal path length taken over this new set of paths. This gives the bound.

Remark. The two bounds are tight. (Consider the case $|t_1| = |p_r(t_1)|$ and $|t_2| = |p_l(t_2)|$ for the upper, and $|p_r(t_1)| = |p_l(t_2)| = 1$ for the lower bound.)

Claim. Let $\pi_1 \in \Pi_a$ and $\pi_2 \in \Pi_b$, where $a, b \geq 1$. If $\pi \in \Pi_{a+b+1}$ and $\pi_{\leq a} = \pi_1$, $\pi_{>a+1} = \pi_2$, then

$$\max\{L(\pi_1), L(\pi_2)\} \leq L(\pi) \leq L(\pi_1) + L(\pi_2) + 2.$$

Proof. Note that π is one of $(a+b+1)\binom{a+b}{a}$ possible interleavings of π_1 and $\triangle_{a+1} \circ \pi_2$, with the element $a+1$ being placed at any of $a+b+1$ possible positions.

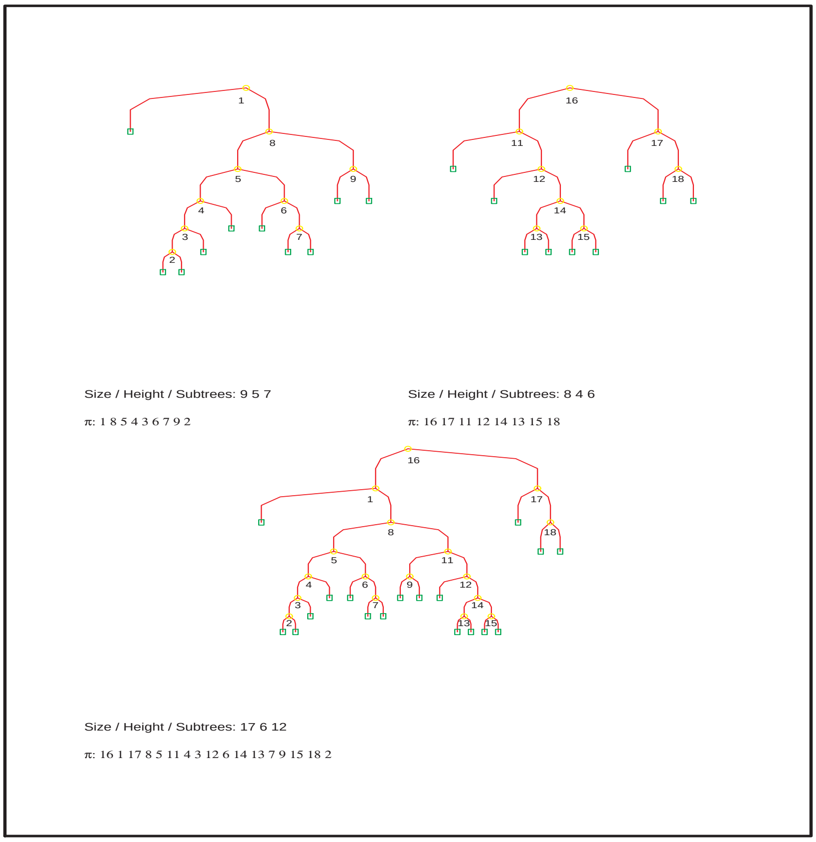

Figure 2: We interleave two permutations $\pi_1 \in \Pi_9$ and $\pi_2 \in \Pi_8$. The upper left shows $T(\pi_1)$, the upper right $S(T(\pi_2, \triangle_9))$. The lower half of the diagram shows one of $\binom{17}{8}$ possible interleavings of π_1 and π_2. The left fringe $p_l(t_2)$ of t_2 is the pair $(16, 11)$, while the right fringe $p_r(t_1)$ of t_1 is the triple $(1, 8, 9)$; the central path of t is $(16, 1, 8, 11, 9)$. The subtree rooted at 5 has been lowered by $1 \leq 2$, the one rooted at 17 by $0 \leq 3$, and the one at 12 by $2 \leq 3$.

Let $f = \{(i,i)\}_{i=1}^{a} \cup \{(i,i+1)\}_{i=a+1}^{a+b}$, i.e. f translates the interval $[a+1, a+b]$ to $[a+2, a+b+1]$. We prove the claim by comparing $t = T(\pi)$ to $u = S(T(\pi_{\neq a+1}), f)$. We know that

$$\max\{L(\pi_1), L(\pi_2)\} \leq L(\pi_{\neq a+1}) \leq L(\pi_1) + L(\pi_2) + 1.$$

Let $t_1 = T(\pi_1)$ and $t_2 = S(T(\pi_2), \triangle_{a+1})$. We may construct t from u by placing $a+1$ anywhere on the central path of u. (This is because $\max l(t_1) < a+1 < \min l(t_2)$.) If we do not place it at the very end, the remaining parts of $p_r(t_1)$ and $p_l(t_2)$ and the subtrees attached to them become the left and right subtrees of $a+1$ in t. (Note that the two subtrees of $a+1$, if any, are also subtrees of t_1 and t_2, respectively.)

We again distinguish several cases. If we place $a+1$ at the beginning of the central path of u, the tree t has height $1 + \max\{L(\pi_1), L(\pi_2)\}$. If $a+1$ is the last node on the central path of u, the height of t either equals that of u or exceeds it by one. In either case we have the claim, i.e. both the upper and the lower bound.

We now assume that $a+1$ is neither placed first nor last and proceed by fringe node i.e. path classification, as in the proof of the previous claim. The fringe nodes of t were either fringe nodes not on the right fringe of t_1, and proceding from t_1, fringe nodes not on the left fringe of t_2, and proceding from t_2, or the single fringe node at the end of the central path.

The new classification is as follows. There is no single fringe node at the end of the central path, because we assumed that $a+1$ is placed internally. There are two additional types of fringe nodes, i.e. those that procede from t_1 or t_2 and count $a+1$ among their ancestors.

Paths to fringe nodes from t_1 or t_2 that do not pass through $a+1$ are identical to their counterparts in u, except for relabelling. The central path case does not occur.

It remains to verify both bounds for the two new cases. Note that a fringe node from t_1 situated in the subtree at $a+1$ is lowered by at most the number of edges that it was lowered in u, because the placement of $a+1$ shortens or preserves the other fringe $p_l(t_2)$. The same argument but with respect to $p_r(t_1)$ applies to fringe nodes from t_2 that are children of $a+1$. This gives the upper bound. The lower bound follows because there is no central path case and neither of the two new path types is shortened with respect to its source in t_1 or t_2.

Claim. Let $\pi_1 \in \Pi_a$, where $a \geq 1$. If $\pi \in \Pi_{a+1}$ and $\pi_{>1} = \pi_1$, then

$$L(\pi_1) \leq L(\pi) \leq L(\pi_1) + 1.$$

Proof. We obtain $t = T(\pi)$ from $u = S(T(\pi_1), \triangle_1)$ by inserting 1 anywhere along $p_r(u)$. This insertion lowers the subtree that is attached at the insertion point by one. Hence the claim.

3.1.1 Application to L on $\Delta(j,h)$

We reformulate the results of the previous section in the terminology associated to martingales.

Let $h \in \Omega$ and $h_1 \in \Delta(j,h)$. If $j=1$, then $|L(h_1)-L(h)| \le 1$. If $1<j<n$, we let $\pi_1 = h_{<j}$ and $\pi_2 = h_{>j}$ to obtain $|L(h_1)-L(h)| \le \min\{L(\pi_1), L(\pi_2)\}+2$. Using $L(\pi_1) \le j-2$ and $L(\pi_2) \le n-j-1$, we reformulate this as

$$|L(h_1)-L(h)| \le \min\{j-2, n-j-1\}+2 = \min\{j, n-j+1\}.$$

3.2 A tail inequality for the height of RBSTs

We wish to study the deviation of L from its mean A_n. The functional L is identical to the last random variable X_{n-1} of the martingale sequence. (Recall that $X_{n-1}(h)$ is the expectation of L over the set of g that agree with h on all variables, i.e. over the singleton set $\{h\}$.)

Hence we have the tail inequality if we compute χ_{n-1}. Note that

$$\min\{j, n-j+1\}$$

is j iff $j \le (n+1)/2$ and $n-j+1$ otherwise. There are two cases: n odd or n even. Let $n=2k$, whence

$$\chi_{n-1} = 1+\sum_{j=2}^{k} j^2 + \sum_{j=k+1}^{n-1}(n-j+1)^2 = 1+2\sum_{j=2}^{k} j^2 = -1+\frac{1}{12}n(n+1)(n+2).$$

Now let $n=2k+1$.

$$\chi_{n-1} = 1+2\sum_{j=2}^{k} j^2 + (k+1)^2 = -1+\frac{1}{12}n(n+1)(n+2)+\frac{1}{4}(n+1).$$

We have

$$e^{-\lambda^2/(2\chi_{n-1})} < e^{-\frac{1}{2}\lambda^2/\binom{n+2}{3}}$$

in either case.

This yields a pair of tail inequalities. Let $\lambda \in \mathbb{R}^+$ and $a \in \mathcal{A}_n$ selected according to the probability measure

$$\mathrm{P}[a] = \prod_{u \preceq a} \frac{1}{|u|}.$$

Then

$$\mathrm{P}[h(a) > A_n + \lambda] < e^{-\frac{1}{2}\lambda^2/\binom{n+2}{3}} \quad \text{and} \quad \mathrm{P}[h(a) < A_n - \lambda] < e^{-\frac{1}{2}\lambda^2/\binom{n+2}{3}},$$

where $\lfloor \log_2 n \rfloor \le h(a) < n$, $A_n \sim c \log n$ and $c = 4.311070\ldots$ is the root ≥ 2 of $(2e/c)^c = e$.

Remark. This tail inequality supplies no new information. All useful values of λ are necessarily $\mathcal{O}(n)$. We have

$$\lim_{n\to\infty} e^{-\frac{1}{2}\lambda^2/\binom{n+2}{3}} = 1$$

for any choice of $\lambda = \lambda(n) \in \mathcal{O}(n)$. Therefore the inequality degenerates to the trivial bound (any probability is at most one) as n goes to infinity, no matter how we choose λ; Fig. 3 illustrates the problem.

4 A better RBST height martingale

The value of $\chi_l = \chi(l)$ is the critical parameter for any martingale that is used to study the deviation of a random variable from its mean. It must be a small as possible. Indeed if $r = \max L(\Omega) - \min L(\Omega) = r(l)$, no $\chi(l)$ that is not $o\left(r(l)^2\right)$ or at the very least $\theta\left(r(l)^2\right)$ will do.

There are several ways to lower χ_l, i.e. we might reduce the number of variables, or lower the c_i. We do the latter.

4.1 Interleaved permutations revisited

Introduce the function

$$\Phi : \Pi^2 \mapsto \mathfrak{P}(\Pi) \quad \text{where} \quad \Pi_0 = \{\emptyset\} \quad \text{and} \quad \Pi = \bigcup_{n\geq 0} \Pi_n,$$

defined as follows.

$$\Phi(\pi_1, \pi_2) = \left\{ \pi \mid \pi \in \Pi_{|\pi_1|+|\pi_2|} \wedge \pi_{\leq |\pi_1|} = \pi_1 \wedge \pi_{>|\pi_1|} = \pi_2 \right\}$$

We pointed out earlier that

$$|\Phi(\pi_1, \pi_2)| = \binom{|\pi_1| + |\pi_2|}{|\pi_1|} \quad \text{and} \quad \Pi_{a+b} = \bigcup_{(\pi_1,\pi_2)\in \Pi_a\times\Pi_b} \Phi(\pi_1, \pi_2).$$

Note that the isomorphism between permutations and inversion tables is equivalent to the recurrence

$$\Pi_n = \bigcup_{(\pi_1,\pi_2)\in \Pi_1\times\Pi_{n-1}} \Phi(\pi_1, \pi_2) = \bigcup_{\pi_2\in \Pi_{n-1}} \Phi(\{(1,1)\}, \pi_2) \cong \left[\binom{n}{1}\right] \times \Pi_{n-1}.$$

Consider a different choice of recurrence.

$$\Pi_n = \bigcup_{(\pi_1,\pi_2)\in \Pi_{\lfloor n/2\rfloor}\times\Pi_{\lceil n/2\rceil}} \Phi(\pi_1, \pi_2) \cong \left[\binom{n}{\lfloor n/2\rfloor}\right] \times \Pi_{\lfloor n/2\rfloor} \times \Pi_{\lceil n/2\rceil}.$$

These two recurrences express the fact that a permutation of length n may be identified in various ways. E.g. we may interleave a permutation of length

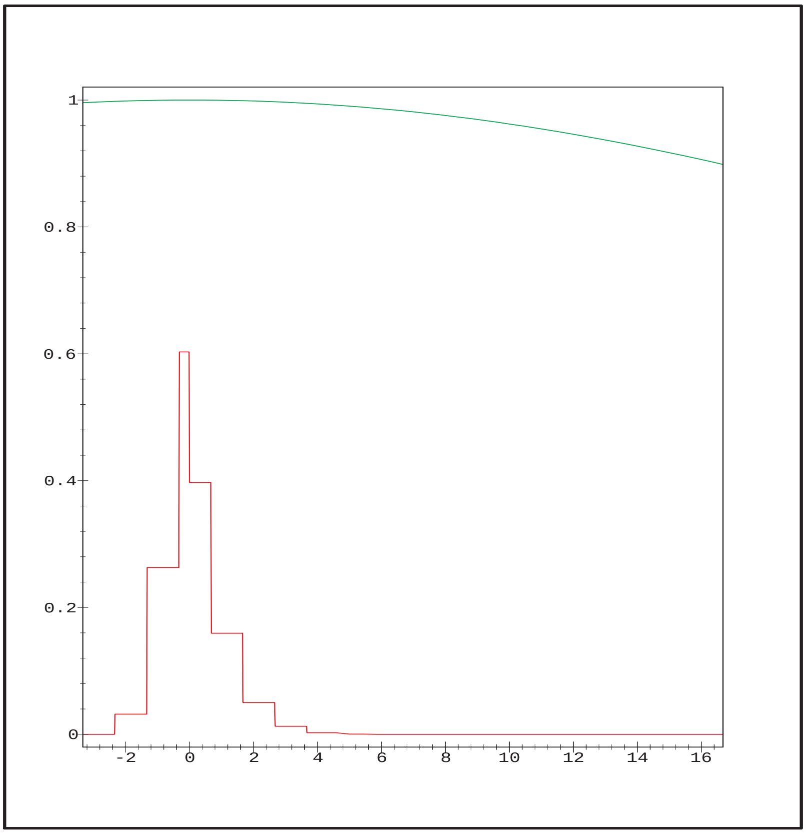

Figure 3: Tail probabilities for $h(a)$ when $n = 24$. The horizontal axis indicates the distance from the mean $A_{24} = 7.327725534\ldots \in \mathbb{Z}$. The upper bound is continuous. (The reader may wish to check his/her understanding of the paper by giving an algorithm to compute A_{24} and the exact tail probabilities. Note that $|\Pi_{24}| \sim 10^{23}$.)

$n-1$ and the singleton permutation in n possible ways or we may interleave two permutations of length $\lfloor n/2 \rfloor$ and $\lceil n/2 \rceil$, respectively, in $\binom{n}{\lfloor n/2 \rfloor}$ possible ways.

We define the three base cases of the new recurrence: $\Pi_1 \cong [1]$, $\Pi_2 \cong [2]$ and $\Pi_3 \cong [3] \times [2]$. This gives a complete decomposition of Π_n for every n. Examples include

$$\Pi_8 \cong \left[\binom{8}{4}\right] \times \left[\binom{4}{2}\right]^2 \times [2]^4$$

and

$$\begin{aligned} \Pi_9 &\cong \left[\binom{9}{4}\right] \times \left[\binom{5}{2}\right] \times [3] \times [2] \times [2] \times \left[\binom{4}{2}\right] \times [2]^2 \\ &= \left[\binom{9}{4}\right] \times \left[\binom{5}{2}\right] \times \left[\binom{4}{2}\right] \times [3] \times [2]^4 . \end{aligned}$$

The reader may verify that there are $n-1$ terms in the cartesian product when $n \geq 2$.

Introduce the function

$$m : \mathbb{N}^+ \mapsto \mathfrak{P}\left(\mathbb{N} \times \{[i]\}_{i \in \mathbb{N}}\right),$$

where

$$\begin{aligned} m(n) &= \left\{\left(n, \left[\binom{n}{\lfloor n/2 \rfloor}\right]\right)\right\} \\ &\cup \quad m\left(\lfloor n/2 \rfloor\right) \cup \{(a + \lfloor n/2 \rfloor - 1, r) \mid (a, r) \in m\left(\lceil n/2 \rceil\right)\}. \end{aligned}$$

if $n \geq 2$ and $m(1) = \emptyset$, e.g.

$$m(8) = \{(8, [70]), (7, [6]), (6, [2]), (5, [2]), (4, [6]), (3, [2]), (2, [2])\} .$$

The reader may wish to match $m(8)$ to the decomposition of Π_8 shown above.

We use the function m to define a martingale on $n-1$ variables, $n > 1$, where $\Omega = \Pi_n$ and $L = h \circ B \circ T$, as before. The permutations $g \in \Omega$ are functions

$$g = \cup_{j=1}^{n-1} \left(\{y_j\} \mapsto r(j+1)\right)$$

where $r = m(n)$ and the probability measure is

$$\mathrm{P}\left[g(y_j) = a\right] = 1/\left|r(j+1)\right| \quad \text{if} \quad a \in r(j+1),$$

i.e. the uniform distribution on Π_n.

We cannot verify the Lipschitz condition unless we know exactly how the value of $y_j \in V = \{y_j\}_{j=1}^{n-1}$ influences the permutation g. The variable y_j determines one choice among many of interleaving two permutations.

The function

$$\rho : \mathbb{N}^+ \mapsto \mathfrak{P}\left(\mathbb{N} \times \mathfrak{P}(\mathbb{N})^2\right)$$

where

$$\begin{array}{rcl}\rho(n) & = & \{(n,([1,\lfloor n/2\rfloor],[\lfloor n/2\rfloor+1,n]))\}\cup m(\lfloor n/2\rfloor) \\ & \cup & \{(a+\lfloor n/2\rfloor-1,(f(r_1),f(r_2)))\mid(a,(r_1,r_2))\in m(\lceil n/2\rceil)\}.\end{array}$$

and $f=\triangle_{\lfloor n/2\rfloor-1}$, if $n\geq 2$, and $\rho(1)=\emptyset$, helps to disambiguate this concept.

Let $(r_1,r_2)=\rho(j+1)$. We continue to use the gradation $V_0=\emptyset$ and $V_{j+1}-V_j=\{y_j\}$. The variable y_j denotes a uniform draw from

$$\Delta(\{y_j\},h)=\left\{g\mid\left(i\notin r_1\cup r_2\Rightarrow g^{-1}(i)=h^{-1}(i)\right)\wedge g_{r_1}=h_{r_1}\wedge g_{r_2}=h_{r_2}\right\},$$

i.e. from the set of permutations that do not reposition values not in r_1 or r_2 with respect to h and interleave the values in r_1 and r_2 such that the relative ordering of r_1 and r_2 in h is preserved.

E.g. let $r=\rho(8)$. This gives $r(8)=([1,4],[5,8])$, $r(7)=([5,6],[7,8])$, $r(4)=([1,2],[3,4])$, etc. Hence y_7 denotes a choice of interleaving the intervals $[1,4]$ and $[5,8]$, y_6 the intervals $[5,6]$ and $[7,8]$ etc. We hasten to point out that these variables are independent. E.g. whether we set y_7 to list the two intervals in order or to alternate between them does not affect, nor is affected by, our choice of listing $[5,8]$, determined by y_6.

Note that ρ generates only contiguous intervals.

The next step is to determine the range of L on $\Delta(\{y_j\},h)$.

This is as far as I was able to get. I will outline possible continuations.

Initially I considered the following two steps, which are based on the assumption that I can bound the range of L on $\Delta(\{y_j\},h)$ by a term that is linear in the minimum of the lengths of two permutations that are interleaved at y_j.

- Prove the linear bound on the range.
- Compute the trivial divide-and-conquer bound on the cost function χ_{n-1}. This bound is $\mathcal{O}(n^2)$, i.e. an improvement on the first martingale, but nonetheless insufficient because $n^2\in\theta(r(n-1)^2)$. If we choose a sublinear split e.g. $\lfloor\log_2 n\rfloor$, this can be lowered to $\mathcal{O}(n)$, which falls into the useful range.

Consideration of a sublinear split introduces the following paradox. The martingale that uses inversion tables does not have a linear cost function, but the split is constant, i.e. into 1 and $n-1$.

The above reasoning leads me to question my bound on the range of L on $\Delta(\{y_j\},h)$. Indeed my initial conjecture is false. It is not hard to find a counterexample. Consider a y_j embedded into a variable assignment that produces inwardly oriented fringes around the adjacent interval pair whose relative ordering is determined by y_j. This raises the cost function to at least the square of the fringe length of the two outer intervals, i.e. the total cost reverts to $\mathcal{O}(n^2)$ or worse.

I list two ways to continue this research.

- Further decomposition of the martingale to get constant bounds on the range of a single variable, with the number of variables an element of $o(n^2)$. E.g. the second martingale may be decomposed by replacing every variable by a sequence of variables whose length is the minimum of the respective lengths of the two permutations being interleaved. This subsequence of variables is a martingale on the universe of interleavings, i.e. fixed-cardinality subsets of a given set. (The reader may wish to construct this martingale and verify that the distribution on the set of interleavings is uniform, and that the number of variables is $\theta(n \log n)$.)

- Existence theorems on RBST height martingales, i.e. prove the existence of a martingale with a $o(n^2)$ cost function without actually specifying the mapping from variables to permutations.

References

[AS92] Noga Alon and Joel H. Spencer. *The Probabilistic Method.* John Wiley & Sons, Inc., 1992. 254 pages. (ISBN 0-471-53588-5). With an appendix by Paul Erdös.

[FGM96] Philippe Flajolet, Xavier Gourdon, and Conrado Martínez. Patterns in random binary search trees. Research Report 2997, Institut National de Recherche en Informatique et en Automatique, October 1996. 23 pages. Accepted for publication in *Random Structures and Algorithms.*

[MP97] Conrado Martínez and Helmut Prodinger. On the number of descendants and ascendants in random search trees. Research Report LSI-97-1-R, Departament de Llenguatges i sistemes informàtics. Universitat Politècnica de Catalunya, January 1997. 32 pages.

[SF96] Robert Sedgewick and Philippe Flajolet. *An Introduction to the Analysis of Algorithms.* Addison-Wesley Publishing Company, 1996. 512 pages. (ISBN 0-201-4009-X).

[VF90] Jeffrey Scott Vitter and Philippe Flajolet. Analysis of algorithms and data structures. In J. van Leeuwen, editor, *Handbook of Theoretical Computer Science*, volume A: Algorithms and Complexity, chapter 9, pages 431–524. North Holland, 1990.