The height of Mallows trees

Based on joint work with Benoit Corsini

McGill

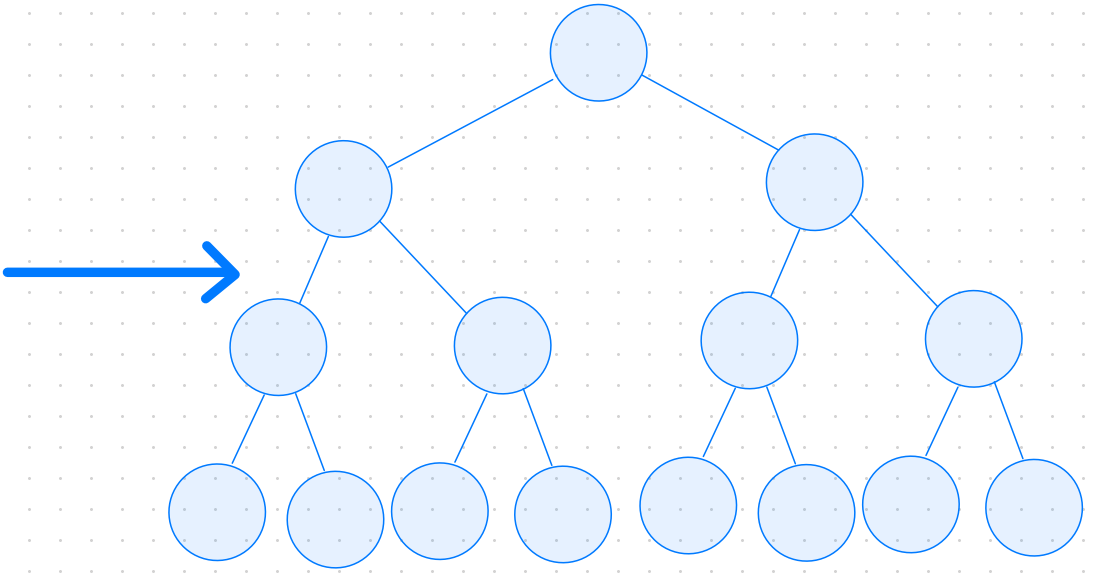
LOUIGI ADDARIO-BERRY

Giuseppe Penone

Binary search trees

1 2 3 4 5 6 7 8

(3, 5, 1, 7, 8, 4, 6, 2)

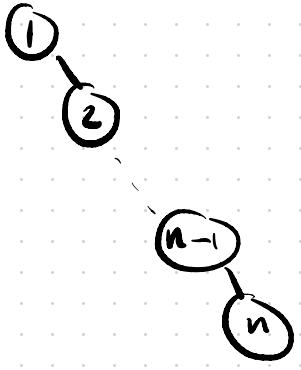


Permutation ⟶ Rooted tree (root at top)

Insert values one-at-a-time at top, routing down through tree until empty space is reached.

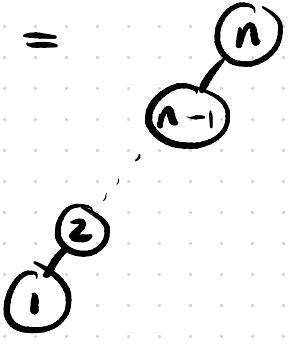
Binary search trees

$\sigma = id \in S_n$

$B(\sigma) =$ 

$\sigma \circ r \qquad r_i = n+1-i$

$\sigma = (n, n-1, \dots, 1) \in S_n$

$B(\sigma) =$ 

$\sigma \in_u S_n$

$B(\sigma) =$ Random Binary Search Tree (RBST)

Devroye's Theorem (1986)

Let B_n be an RBST with n nodes.

Then $\frac{ht(B_n)}{\log n} \xrightarrow{P} c^*$

and $\mathbb{E}\, ht(B_n) = (1+o(1))c^* \log n$

where c^* is the unique solution in $[2,\infty)$ of

$c \log\left(\frac{2}{c}\right) + c = 1.$

Reed's Theorem (2003)

Let B_n be an RBST with n nodes.

Then $\mathbb{E}\, ht(B_n) = c^* \log n - \beta^* \log\log n + O(1)$

$\mathrm{Var}\, ht(B_n) = O(1)$

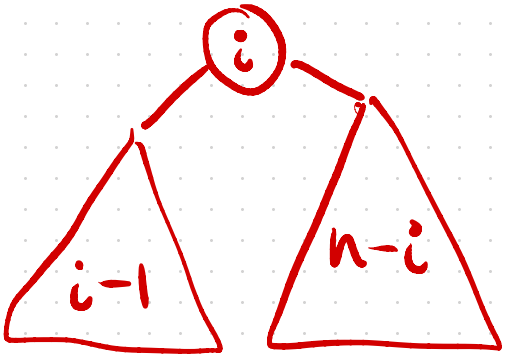
where $\beta^* = \frac{3}{2} \frac{1}{\log(c^*/2)}$

Proof idea (Devroye's Theorem)

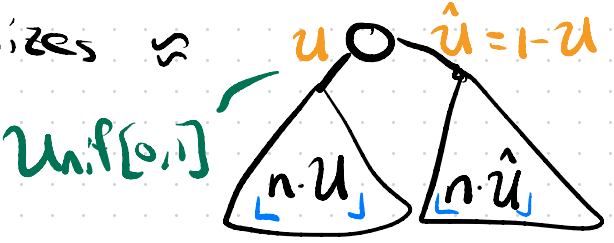
$\mathcal{B}_n \leftarrow \sigma \in_u S_n$

- Root value $\sigma(1)$ is Unif on $\{1, 2, \ldots, n\}$

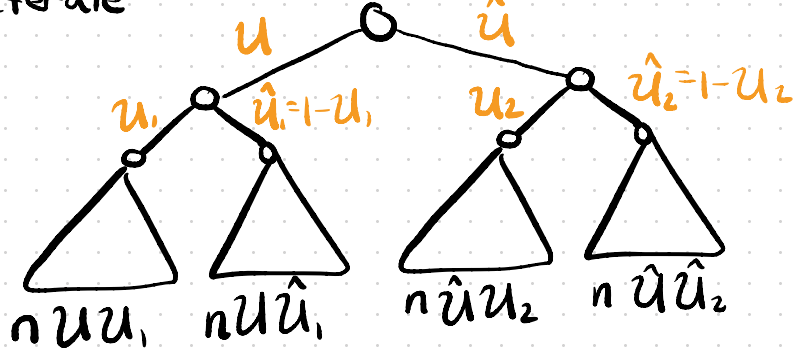
$\sigma(1) = i$

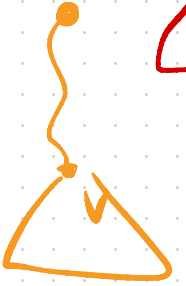


- So subtree sizes $\approx$



- Iterate





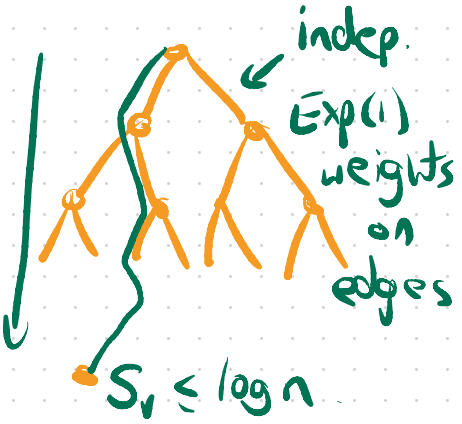
size $\stackrel{d}{=} \lfloor \lfloor n \cdot U_1 \rfloor \cdot U_2 \rfloor$

- Size of subtree rooted at $v \approx n \cdot \prod_{e \le v} U_e$

where terms in product are indep Unif[0,1]

- Height $\approx \max \{ht(v) : \prod_{e \le v} U_e \ge \frac{1}{n}\}$.

- Height $\approx \max\{ht(v) : \underbrace{\sum_{e \ni v}(-\log U_e) \le \log n}_{= S_v \leftarrow \text{Random walk with Exp(1) steps.}}\}$



Write $M_k = \min\{S_v : ht(v) = k\}$

Then
Height $\approx \min\{k : M_k \ge \log n\}$.

Law of large #s for minima in branching random walks
(Hammersley-Kingman-Biggins 1970's)

$\Rightarrow \frac{M_k}{k} \xrightarrow{a.s.} \frac{1}{C^*}$ $\Rightarrow$ Height $\approx C^* \log n$.

□

Proof idea (Reed's Theorem)

Write $M_k = \min\{S_v : ht(v) = k\}$

Then

Height $\approx \min\{k : M_k \geq \log n\}$

Reed used "truncated second moment method" to prove that

$$M_k = \frac{1}{c^*}k + \frac{1}{\beta^*}\log k + O(1) \quad \text{and} \quad Var(M_k) = O(1)$$

then used this to control the height.

Similar results independently proved by Bramson, much earlier (late 1970's) for branching Brownian motion.

□

Mallows trees

Any metric d on S_n defines a one-parameter family of distributions $(\varphi_q, q \geq 0)$ on the permutations of S_n by

$$\varphi_q(\sigma) \propto q^{d(\sigma, id)} \qquad \varphi_q(\sigma) = q^{d(\sigma, id)} / \sum_{\pi \in S_n} q^{d(\pi, id)}$$

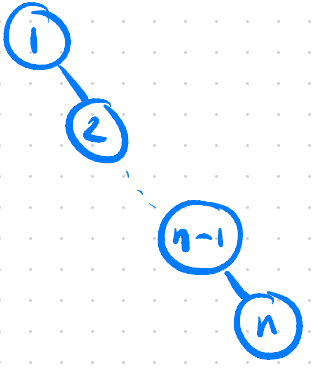
Well-used in statistics with various distances

> "Suppose that a set of k objects is to ranked independently by each of a panel of n judges, and that each ranking proceeds according to some criterion that presumes a true underlying ordering of the objects. Mallows (1957) introduces two one-parameter models to describe such a ranking process, the parameter reflecting the variability of the rankings about the true or modal ordering. The models are referred to as Mallows' Θ and Φ models, which are based on the correlation coefficients of Spearman (1904) and Kendall (1938), respectively. The Φ model is further investigated in Feigin and Cohen (1978), and Diaconis (1982) suggests use of the Mallows' models with other distances."
>
> Fligner and Verducci, JRSS B (1986)

Mallows trees

$\varphi_q(\sigma) \propto q^{d(\sigma, id)}$ dist. on S_n

$q \approx 0 \to \sigma \approx id \to B(\sigma) \approx$



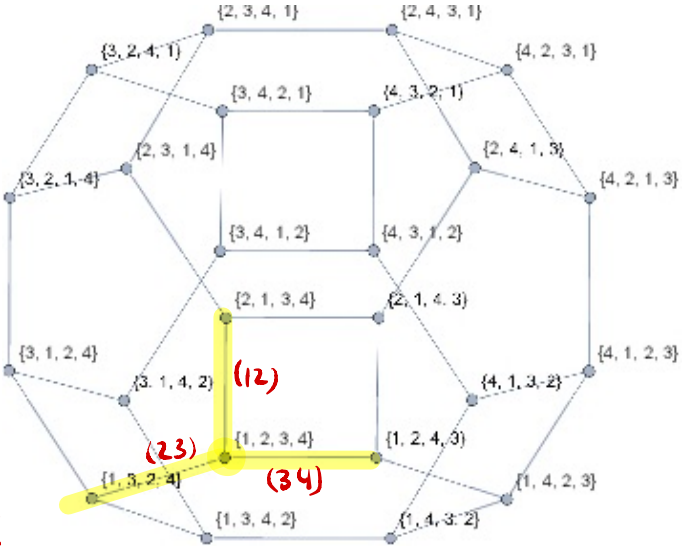
$q \approx 1 \to \sigma \approx$ unif. dist on S_n $\to B(\sigma) \approx$ RBST

$q \approx \infty \to \sigma \approx$ unif. dist. on perms. π with max dist. to id.

Law of $B(\sigma)$ depends on d.

Mallows' ϕ model:

$d(\pi, \sigma)$ = distance in adjacent transposition graph



For this model

$$d(\sigma, id) = inv(\sigma)$$

$$:= \#\{i<j : \sigma(i) > \sigma(j)\}$$

The corresponding Mallows distribution on S_n is

$$\phi_q(\sigma) = \frac{q^{inv(\sigma)}}{Z_{n,q}} \qquad Z_{n,q} = \prod_{i=1}^{n} \frac{1-q^i}{1-q}$$

$= i$ if $q = 1$

Mallows (n, q) dist.

Mallows Φ-model: previous research

Let $\pi \sim \text{Mallows}(n, q_n)$, $0 \le q_n \le 1$.

Bhatnagar-Peled (2015) If $q_n \to 1$ and $n(1-q_n) \to \infty$ then

Length of longest increasing subseq, LIS_n, satisfies $\dfrac{LIS_n}{n\sqrt{1-q_n}} \xrightarrow[\text{prob}]{L^1} 1$

Gladkich-Peled (2018) $0 \le q_n \le 1$

List cycles in decreasing order of length as $\mathcal{C}_1^{\downarrow}, \mathcal{C}_2^{\downarrow}, \ldots$

- If $n(1-q_n)^2 \to 0$ then $\frac{1}{n}(|\mathcal{C}_i|, i \ge 1) \xrightarrow{\text{dist}} \text{Poisson-Dirichlet}(1)$
- $\mathbb{E}[\#\text{ cycles in } \pi] \approx (1-q_n) \cdot n + O(1)$.

Other results by Diaconis+Ram; Mueller & Starr; Basu & Bhatnagar; Mukherjee; Gnedin & Olshanski; Evans, Gr übel + Wakolbinger; Angel, Holroyd, Hutchcroft & Levy; ...

Theorem (A-B, Corsini). Fix non-negative reals $(q_n, n \geq 1)$.

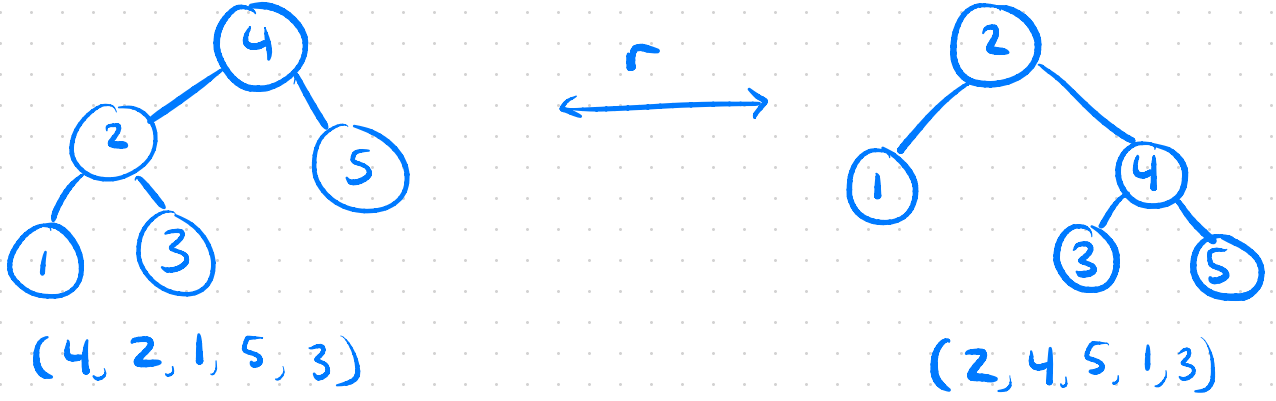
let σ_n be a Mallows(n, q_n) perm., let $T_n = B(\sigma_n)$.

↳ Mallows(n, q_n)-tree.

Then $$\frac{\text{height}(T_n)}{n \cdot \max(1-q_n, 1-\frac{1}{q_n}) + c^* \log n} \longrightarrow 1 \text{ in prob. and in } L_p \text{ for all } p.$$

$n(1-q_n) + c^* \log n$.

Remark: If σ is Mallows(n,q) and r has $r_i = n+1-i$ then $\sigma \circ r$ is Mallows$(n, \frac{1}{q})$. So can assume $q_n \leq 1 \ \forall n$.



Theorem (A-B, Corsini). Fix reals $(q_n, n \geq 1)$ in $[0,1]$, let σ_n be a Mallows(n, q_n) perm., let $T_n = B(\sigma_n)$.

If $nq_n \to \lambda \in [0, \infty)$ then $n-1-\text{height}(T_n) \xrightarrow{d} \text{Poisson}(\lambda)$

If $nq_n \to \infty$ and $\frac{n(1-q_n)}{\log n} \to \infty$ then

$$\frac{\text{height}(T_n) - n(1-q_n) - c^* \log\left(\frac{1}{1-q_n}\right)}{\sqrt{nq_n(1-q_n)}} \xrightarrow{d} N(0,1)$$

Behaviour when $\frac{n(1-q_n)}{\log n} = O(1)$ an open question.

When $q_n \equiv 1$, the RBST case, neither a Poisson nor a Normal limit obtains.

Key proof input: insertion algorithm for Mallows trees

Fix $q \in [0,1)$.

Let $(G_i, i \geq 1)$ be IID Geometric$(1-q)$ $\quad \mathbb{P}(G_i = k) = q^{k-1}(1-q)$

value $\approx \frac{1}{1-q}$.

Let $s_1 = G_1$, $\quad \mathbb{N}_1 = \mathbb{N} \setminus \{G_1\}$

Inductively let

$s_k = k$'th smallest element of $\mathbb{N}_{k-1}$, $\quad \mathbb{N}_k = \mathbb{N}_{k-1} \setminus \{s_k\} = \mathbb{N} \setminus \{s_1, \ldots, s_k\}$

Then the rank permutation of $(s_1, \ldots, s_n)$ is Mallows(n,q)-dist

$(2,1,3,1,5,2) \longrightarrow (2,1,5,3,9,6) \longrightarrow (2,1,4,3,6,5)$

$(G_1, G_2, \ldots, G_6) \quad (s_1, s_2, \ldots, s_6) \quad (\sigma_1, \sigma_2, \ldots, \sigma_6)$

Key proof input: insertion algorithm for Mallows trees

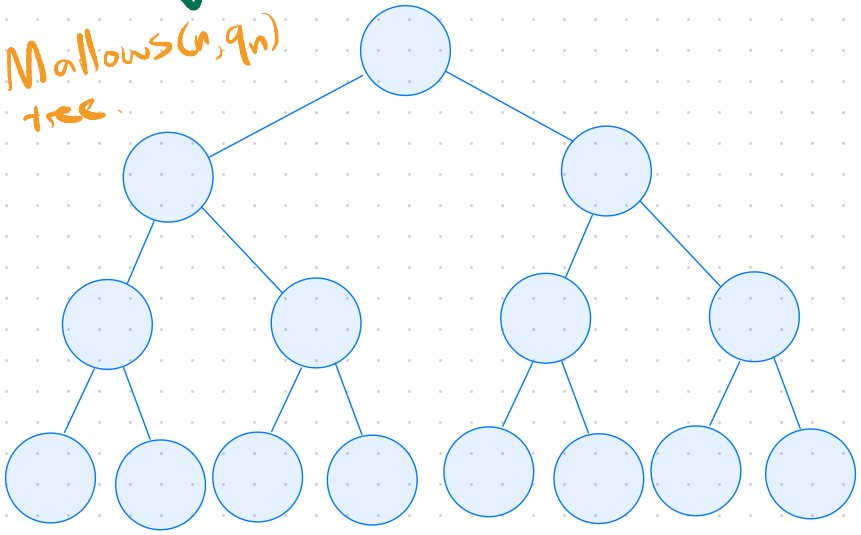
$(2,1,3,1,5,2) \rightarrow (2,1,5,3,9,6) \rightarrow (2,1,4,3,6,5)$

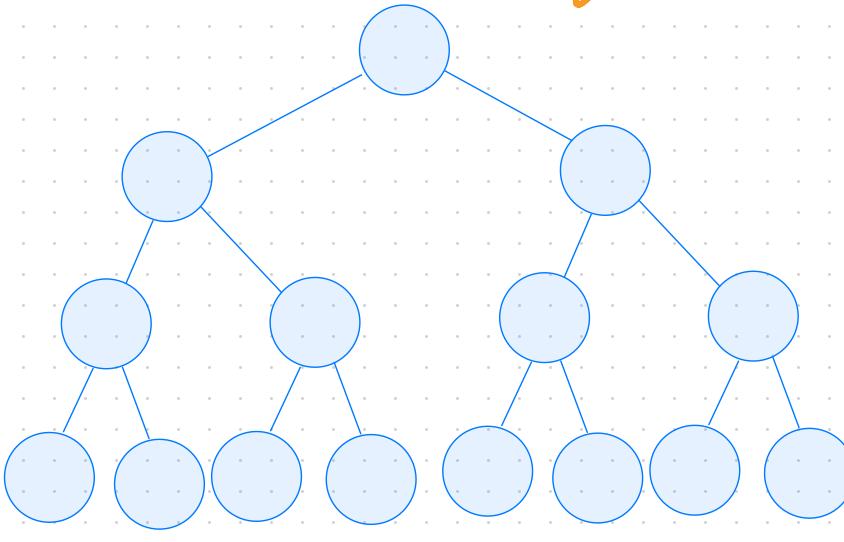
$(G_1, G_2, \ldots, G_6)$ $(S_1, S_2, \ldots, S_6)$ $(\sigma_1, \sigma_2, \ldots, \sigma_6)$

Insert the values $(S_1 \ldots S_n)$ using the BST rule

Insert the values $(\sigma_1 \ldots \sigma_n)$ using the BST rule

Mallows(n, q_n) tree.



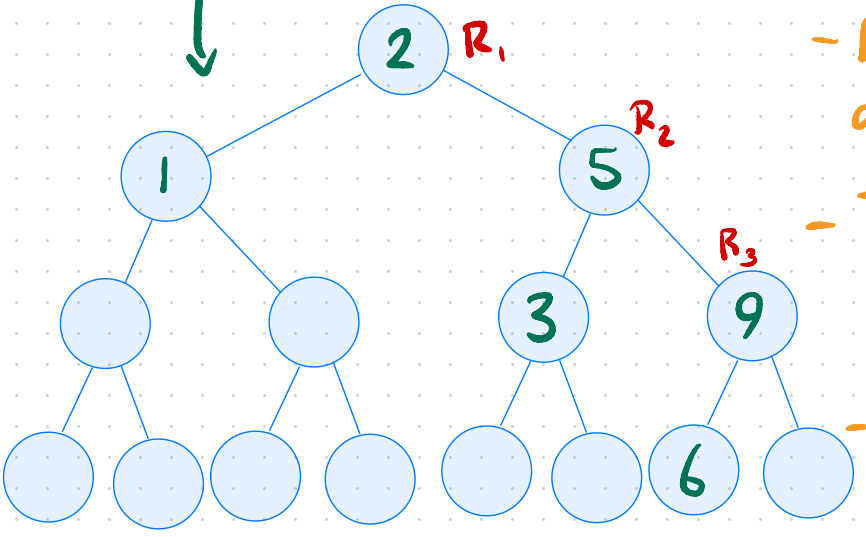


The structure of Mallows trees

$(2,1,3,1,5,2) \rightarrow (2,1,5,3,9,6) \rightarrow (2,1,4,3,6,5)$

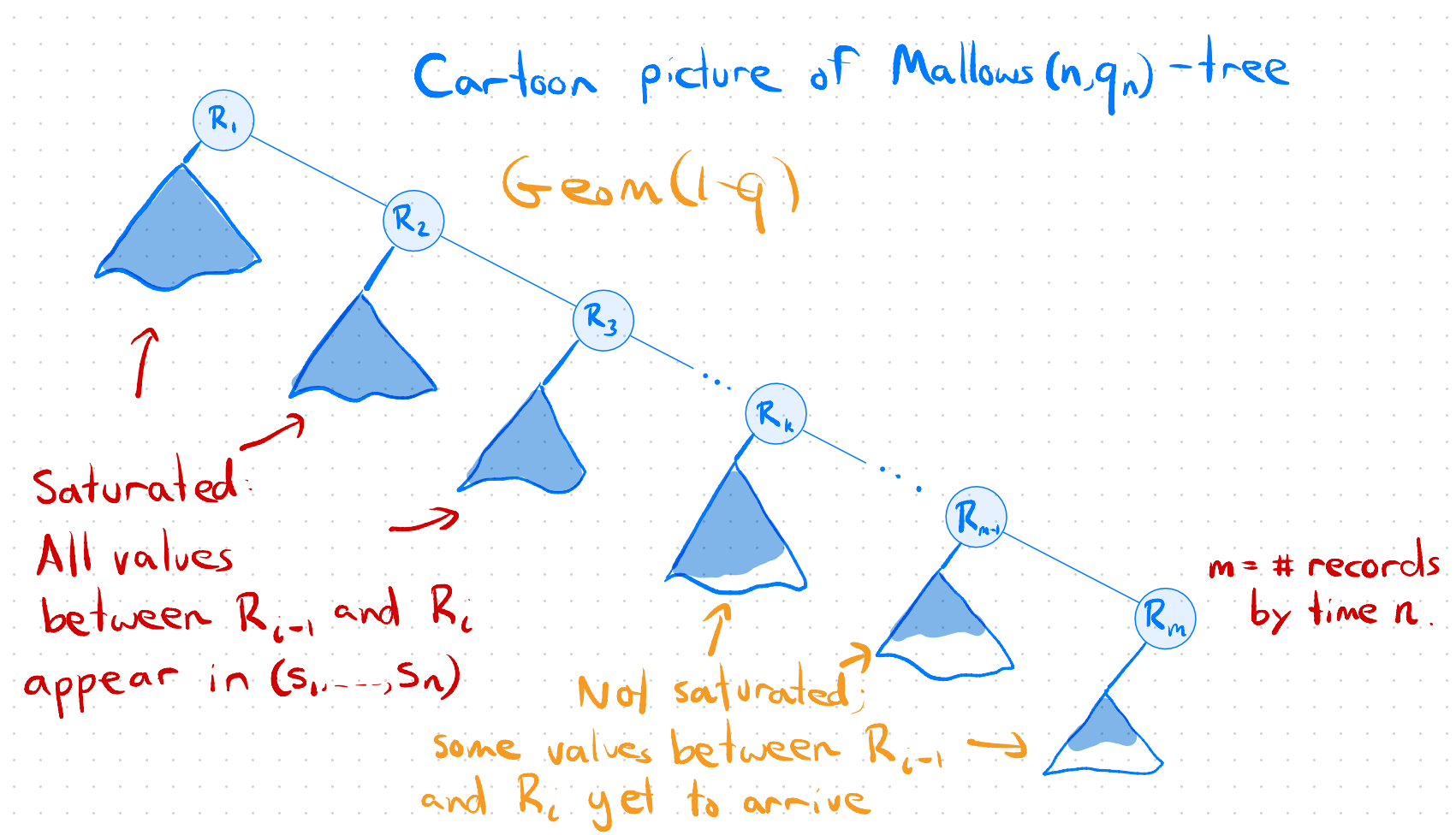
$(G_1, G_2, \ldots, G_6)$ $(S_1, S_2, \ldots, S_6)$ $(\sigma_1, \sigma_2, \ldots, \sigma_6)$

Insert the values $(S_1 \ldots S_n)$ into a BST



- List the record values in $(S_n, n \geq 1)$ as $(R_i, i \geq 1)$, set $R_0 = 0$
- The record increments $(R_i - R_{i-1}, i \geq 0)$ are IID Geom(1-q)
- If $R_i - R_{i-1} = s$ then for n large, the left subtree hanging from the ith node along the right path will have size $s-1$.

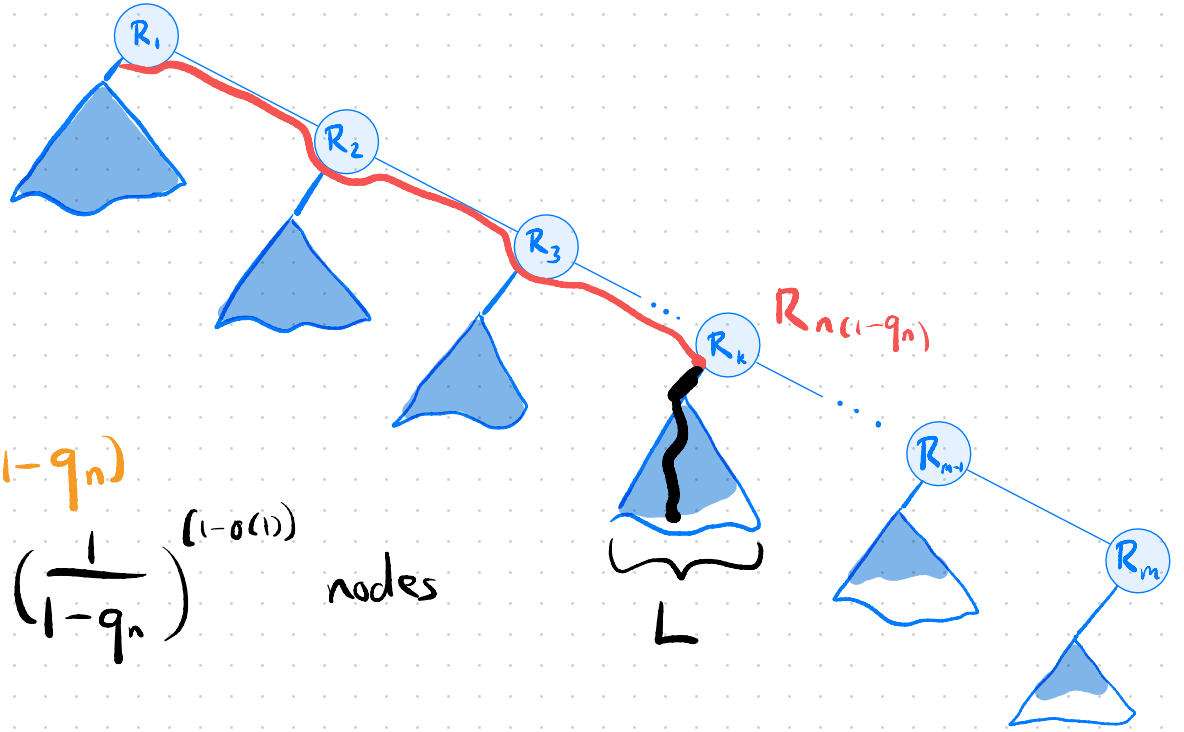
The structure of Mallows trees



Height of the Mallows tree

Two contributions

- Rightmost path
- Left subtree

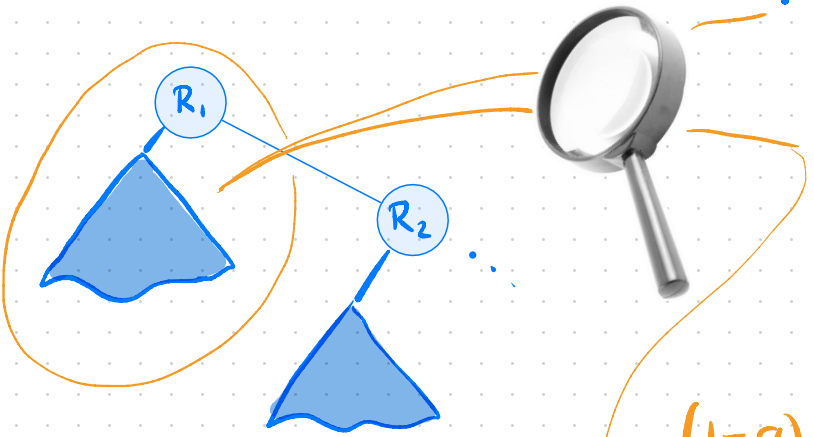


Rightmost path:
Stop at dist. $n(1-q_n)$

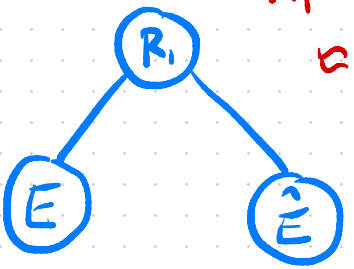
- Left subtree has $\left(\frac{1}{1-q_n}\right)^{(1-o(1))}$ nodes with high prob.

Prove left subtree has height $(1+o_p(1))\,c^* \log|L|$ like in Devroye's thm
(For CLT need finer control over fluctuations of left subtree heights)

The left subtrees (when $q = 1 - O(1)$)



$R_1 = \text{Geom}(1-q) \approx \frac{1}{1-q} \cdot \text{Exp}(1)$

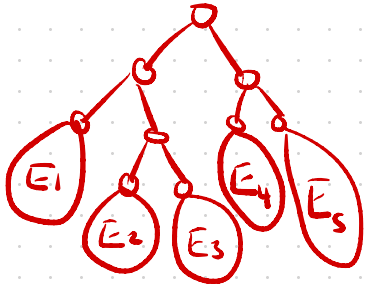


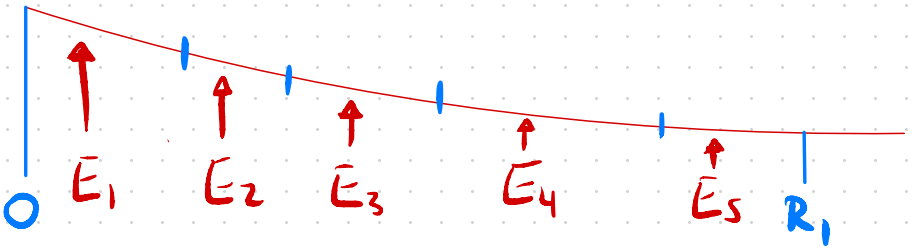
$(1-q)E \approx \text{Exp}(1)$

conditioned to be $\leq R_1$

$\hat{E} = R_1 - 1 - E$

After k-1 insertions:





Open questions & research directions

- Let $L(q)$ be a left subtree in a large Mallows(n,q) tree.
  Prove Reed's theorem for height$(L(q))$

$$\mathbb{E}\,\text{height}(L(q)) = c^* \log \frac{1}{1-q} + \beta^* \log\log \frac{1}{1-q} + O(1)$$

$$\text{Var}(\text{height}(L(q)) = O(1) \quad \text{uniformly in } 0 \le q < 1.$$

- Extend the range of the CLT, figure out when the CLT fails

- Carry out this research for other Mallows/Diaconis models.

Reference :

LAB+Benoît Corsini, The height of Mallows trees. Annals of Probability, to appear.
https://arxiv.org/pdf/2007.13728.pdf