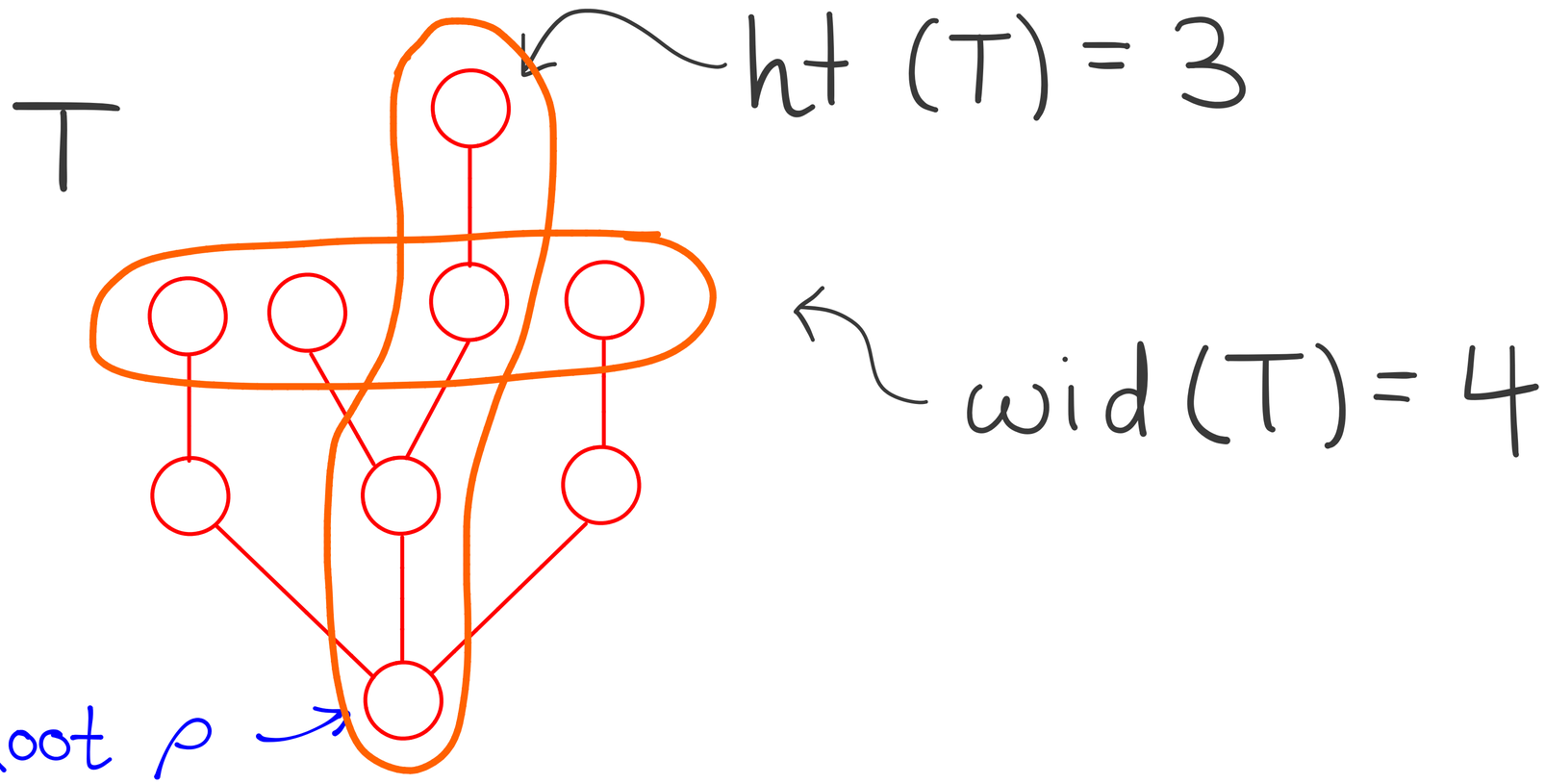
Most trees are short and fat.

Louigi Addario-Berry
McGill

Journees de combinatoire
de BORDEAUX

January 25, 2017

Trees:

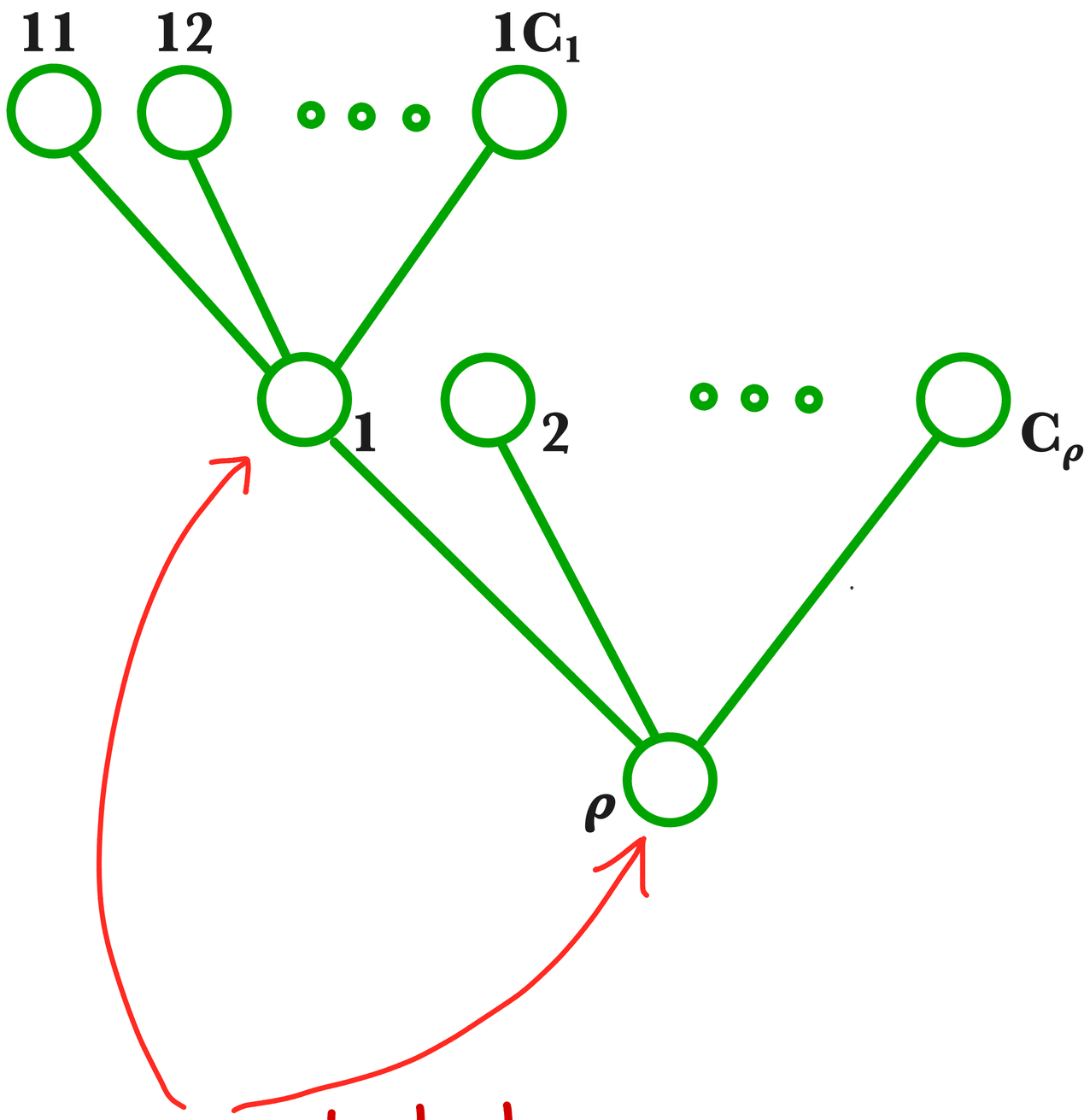


Height: Greatest distance from any node to the root } ht(T)

Width: Greatest # nodes on a single level. } wid(T)

Main Results Fix any r.v. C with $\sum_{k\geq 0}\mathbb{P}(C=k)=1$, write $p_k=\mathbb{P}(C=k)$.

Let T be $GW(C)$ distributed.



children is distributed as C, independently at each node.

Main Results

Fix any r.v. C with $\sum_{k\geq 0}\mathbb{P}(C=k)=1$, write $p_k=\mathbb{P}(C=k)$.

Let T be $GW(C)$ distributed.

Theorem ("Most trees are short & fat")

There is a universal constant $\delta>0$ s.t.

$$\mathbb{P}\left(ht(T)\geq \frac{k}{1-p_1}\cdot wid(T)\right)\leq \exp(-\delta k).$$

Remark: If $\mathbb{E}C>1$ then $\mathbb{P}(\sigma=\infty)>0$, and

$$\mathbb{P}(ht(T)=wid(T)=\infty \mid \sigma=\infty)=1.$$

Also, given that $\sigma<\infty$, the cond. dist. of T is $GW(\hat{C})$ where $\mathbb{P}(\hat{C}=1)=p_1$,

$\mathbb{E}(\hat{C})<1$ so can assume $\mathbb{E}C\leq 1$.

Heuristic: GW trees satisfy $wid(T)\cdot ht(T)\cong vol(T):=\sigma$

Implies "$ht>C^2\cdot wid$" $\cong$ "$ht^2\geq C^2\cdot vol$" so $\mathbb{P}\left(ht(T)>\frac{k}{\sqrt{1-p_1}}\sqrt{vol(T)}\right)\leq \exp(-\delta k^2)$

Theorem: $\mathbb{P}\left(ht(T)\geq \frac{k}{\sqrt{1-p_1}}\, vol(T)^{1/2}\right)\leq \exp\left(-\delta k^2\right)$

Galton-Watson Trees

- Each node has random # of children
- Nodes reproduce independently

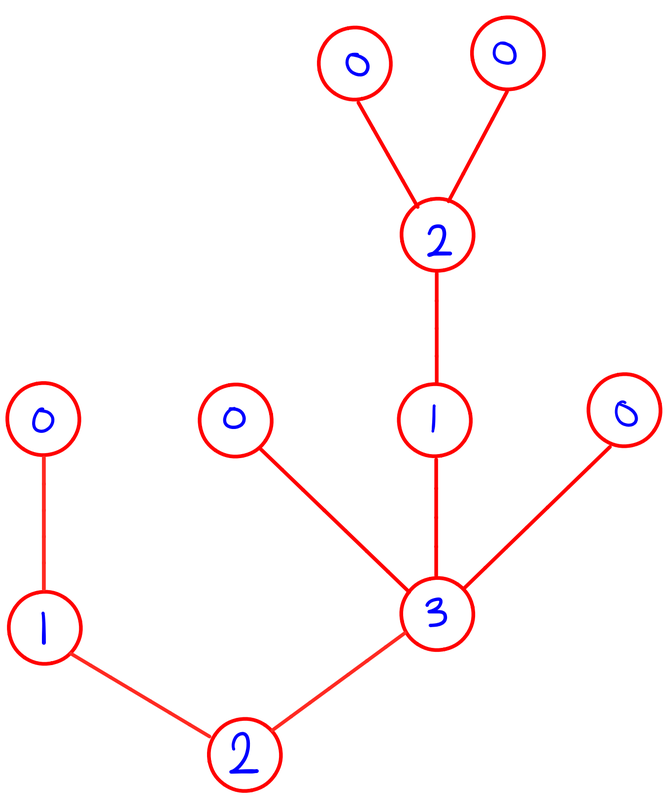
Construction

- $(C_i, i \geq 1)$ independent copies of a random variable C with $\sum_{k \geq 0} \mathbb{P}(C = k) = 1$.

Rule

The sequence $(C_i, i \geq 1)$ gives # children of nodes, in breadth-first search order

Example: $(2, 1, 3, 0, 0, 1, 0, 2, 0, 0, 1, 4, 0, \ldots)$



Halting Condition

Trees with n nodes have $n-1$ edges.

For $n \geq 0$

#nodes discovered by time n (n steps of BFS).

$$= 1 + \sum_{i=1}^{n} C_i$$

vertices explored by time n

$$= n$$

Total # vertices $= \sigma$

$$= \inf\left\{n : 1 + \sum_{i=1}^{n} C_i = n\right\}.$$

Setup

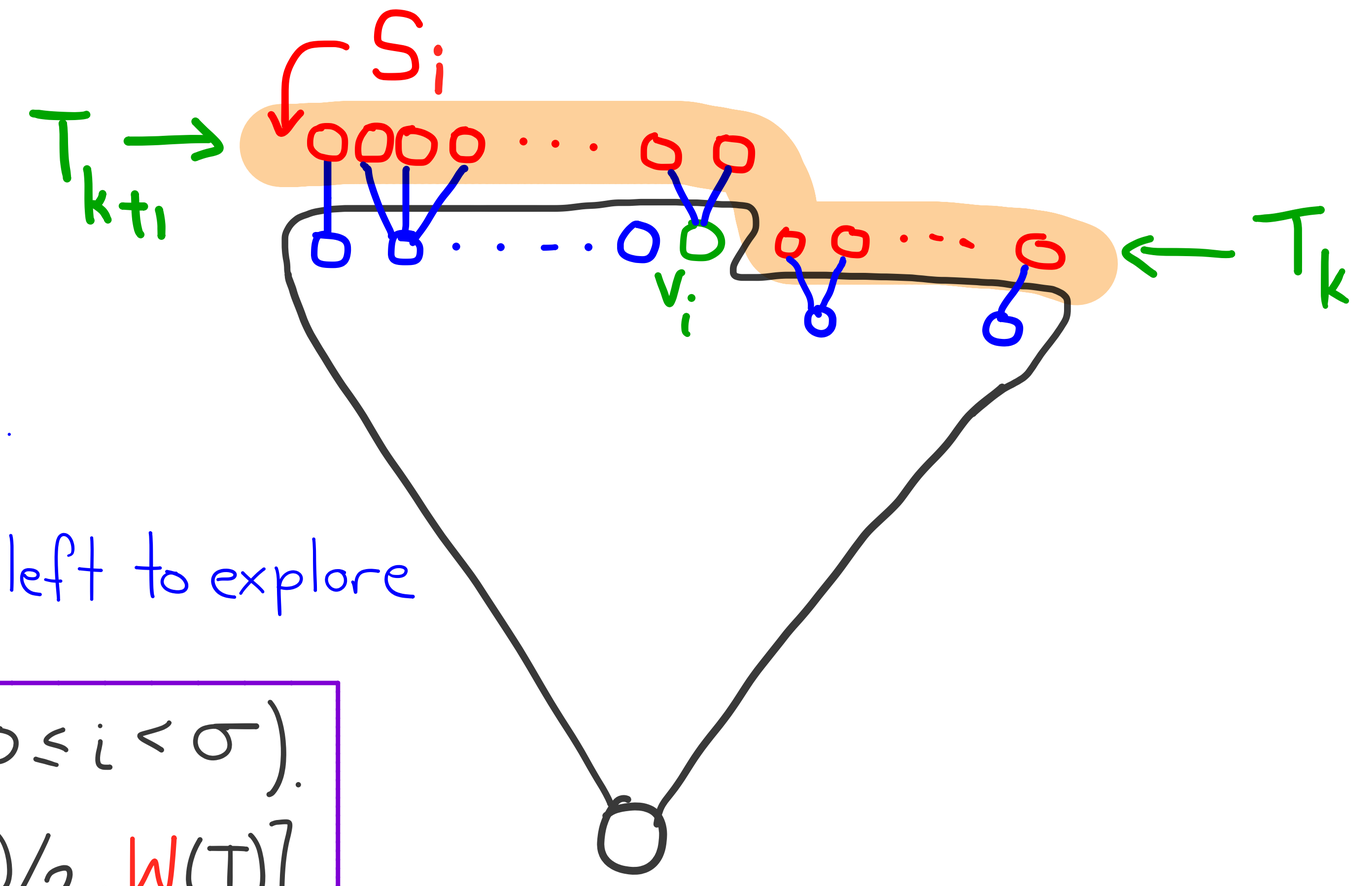
$1 + \sum_{j=1}^{i} C_j$ = # nodes discovered by time i.

Let $S_i = 1 + \sum_{j=1}^{i} (C_j - 1)$

= # nodes in "BFS queue" at time i

$\mathbb{E}\, C \le 1 \Rightarrow \mathbb{E}\, S_n = 1 + n(\mathbb{E}\, C - 1) \le 1.$

$\sigma = \inf\{t : S_t = 0\}$ = first time no nodes left to explore



Prop: Let $W(T) = \max(S_i, 0 \le i < \sigma)$.
Then $\mathrm{wid}(T) \in (W(T)/2, W(T)]$

Proof: During BFS on level k, "exploration queue" $\subset T_k \cup T_{k+1}$; and $= T_k$ at start of level k. ∎

Idea: $\mathrm{ht}(T) = \sum_{k=1}^{\mathrm{ht}(T)} 1 = \sum_{k=1}^{\mathrm{ht}(T)} \sum_{v \in T_k} \frac{1}{|T_k|}$.

When $v_i \in T_k$ then $S_i \approx |T_k|$ so perhaps

$\mathrm{ht}(T) \approx \sum_{k=1}^{\mathrm{ht}(T)} \sum_{v_i \in T_k} \frac{1}{S_i} = \sum_{i=1}^{\sigma} \frac{1}{S_i} =: H(T)$?

[False; consider a star with n leaves. But...]

Prop:
$\mathrm{ht}(T) \le 3H(T)$.

Corollary Suffices to prove
$\mathbb{P}(H(T) \ge \frac{k}{1-p} W(T)) \le e^{-\delta k}$,
thm. follows.

$W(\sigma) = \max(S_i, 0 \le i < \sigma)$ $H(\sigma) = \sum_{i=1}^{\sigma} \frac{1}{S_i}$ Aim: $\mathbb{P}\left(H(\sigma) \ge \frac{k}{1-p_1} W(\sigma)\right) \le e^{-\delta k}$

Key Tool: Decomposition into scales.

When $S_j \approx 2^\ell$ ("scale ℓ") for $j \in \{i, \dots, i+2^\ell\}$, have

$$H(i+2^\ell) - H(i) = \sum_{j=i+1}^{i+2^\ell} \frac{1}{S_j} \simeq 2^\ell \cdot \frac{1}{2^\ell} = 1.$$

So bound (a) time to change scales,

(b) "# visits to scales" $= (M(\ell), \ell \ge 1)$

$W(\sigma)=\max(S_i, 0\le i<\sigma)$ $H(\sigma)=\sum_{i=1}^{\sigma}\frac{1}{S_i}$ **Aim**: $\mathbb{P}\left(H(\sigma)\ge \frac{k}{1-p_1}W(\sigma)\right)\le e^{-\delta k}$

Decomposition into scales.

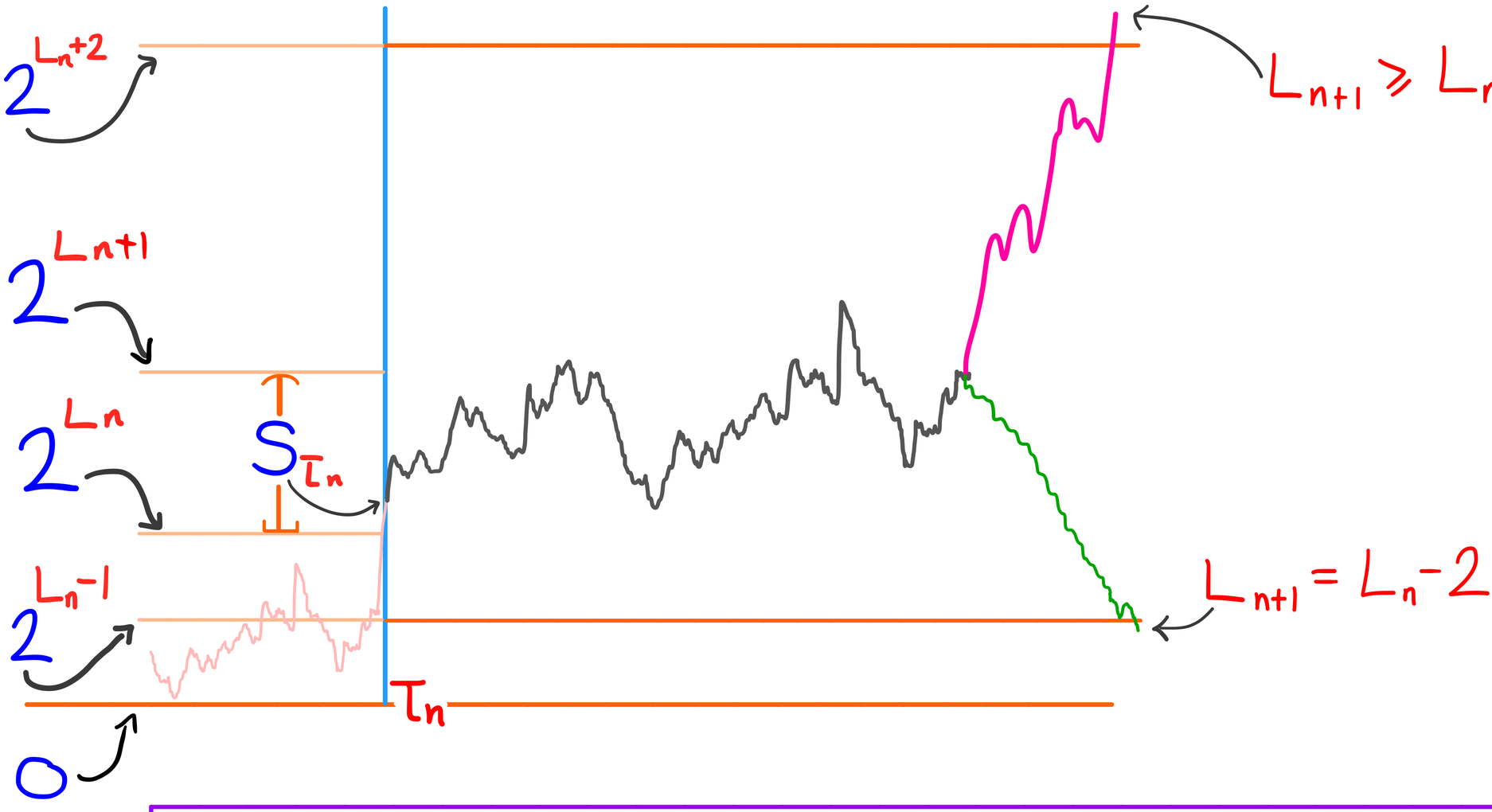
Def

$\tau_0=0$ = initial time

$L_0=0=\log_2 S_0$ = initial scale

$\tau_{n+1}=\min\{t\ge\tau_n : S_t\notin[2^{L_n-1}, 2^{L_n+2})\}$

$L_{n+1}=\sup\{k : 2^k\le S_{\tau_{n+1}}\}$



NB: $S_{\tau_n}<2^{L_{n+1}}$ so $\mathbb{P}(L_{n+1}<L_n)=\mathbb{P}(\text{Hit } 2^{L_n-1} \text{ before } 2^{L_n+2})>\frac{1}{2}$.

$W(\sigma) = \max(S_i, 0 \le i < \sigma)$ $\quad H(\sigma) = \sum_{i=1}^{\sigma} \frac{1}{S_i}$ $\quad$ Aim: $\mathbb{P}\left(H(\sigma) \ge \frac{k}{1-p_1} W(\sigma)\right) \le e^{-\delta k}$

Key Tool: Decomposition into scales.

When $S_j \approx 2^\ell$ ("scale ℓ") for $j \in \{i, \dots, i+2^\ell\}$, have

$$H(i+2^\ell) - H(i) = \sum_{j=i+1}^{i+2^\ell} \frac{1}{S_j} \simeq 2^\ell \cdot \frac{1}{2^\ell} = 1.$$

So bound (a) time to change scales,

(b) "# visits to scales" $= (M(\ell), \ell \ge 1)$

(a) Thm (Lévy; Doeblin; Kolmogorov; Rogozin; Le Cam; Esséen; Kesten):

With $p = \max p_i$, have $\boxed{\max_k \mathbb{P}(S_n = k) \le \frac{Cp}{\sqrt{n(1-p)}}}$ $C > 0$ universal.

"Any random walk spreads out over $\ge \sqrt{n}$ values by time n". Here $\sqrt{n} \simeq 2^\ell$.

Thus if $L_k = \ell$ then we expect $\tau_{k+1} - \tau_k \lesssim 4^\ell$ so

$$H(\tau_{k+1}) - H(\tau_k) = \sum_{j=\tau_k+1}^{\tau_{k+1}} \frac{1}{S_j} \lesssim 4^\ell \cdot \frac{1}{2^\ell} = 2^\ell \simeq 2^{L_k}.$$

More precisely, obtain $\mathbb{P}\left(H_{\tau_{k+1}} - H_{\tau_k} \ge \frac{C}{1-p} \ell \mid L_k = \ell\right) \le e^{-\delta C}$

$W(\sigma) = \max(S_i, 0 \le i < \sigma)$ $\quad H(\sigma) = \sum_{i=1}^{\sigma} \frac{1}{S_i}$ $\quad$ **Aim**: $\mathbb{P}\left(H(\sigma) \ge \frac{k}{1-p_1} W(\sigma)\right) \le e^{-\delta k}$

Key Tool: Decomposition into scales.

When $S_j \approx 2^\ell$ ("scale ℓ") for $j \in \{i, \dots, i+2^\ell\}$, have

$$H(i+2^\ell) - H(i) = \sum_{j=i+1}^{i+2^\ell} \frac{1}{S_j} \simeq 2^\ell \cdot \frac{1}{2^\ell} = 1.$$

So bound (a) time to change scales,

(b) "# visits to scales" $= (M(\ell), \ell \ge 1)$

(b) $M(\ell) =$ # visits to scale ℓ $= \#\{i : L_i = \ell\}$

$N(\ell, i) =$ Duration of i'th visit to scale ℓ

$N(\ell) = N(\ell,1) + \cdots + N(\ell, M(\ell)) =$ Total duration at scale ℓ.

Fact: Given that $M(\ell) \ne 0$, $M(\ell)$ dominated by sum of 2 Geom$(\frac{1}{2})$ r.v.s; $\Rightarrow \mathbb{P}(M(\ell) > k \,|\, M(\ell) > 0) \le 2^{-k/2}$.

Proof: visits to scale ℓ entail upcrossings of $[2^{\ell-1}, 2^{-\ell})$ or of $[2^{\ell+1}, 2^{\ell+2})$.

Both are hard since walk has non-positive drift. ∎

$W(\sigma) = \max(S_i, 0 \le i < \sigma)$ $H(\sigma) = \sum_{i=1}^{\sigma} \frac{1}{S_i}$ Aim: $\mathbb{P}\left(H(\sigma) \ge \frac{k}{1-p_1} W(\sigma)\right) \le e^{-\delta k}$

Key Tool: Decomposition into scales.

When $S_j \approx 2^\ell$ ("scale ℓ") for $j \in \{i, \dots, i+2^\ell\}$, have

$$H(i+2^\ell) - H(i) = \sum_{j=i+1}^{i+2^\ell} \frac{1}{S_j} \asymp 2^\ell \cdot \frac{1}{2^\ell} = 1.$$

So bound (a) time to change scales,

(b) "# visits to scales" $= (M(\ell), \ell \ge 1)$

(a) $\tau_{k+1} - \tau_k \sim (2^{L_k})^2$ so $\boxed{\mathbb{P}\left(H_{\tau_{k+1}} - H_{\tau_k} \ge \frac{c}{1-p} 2^\ell \mid L_k = \ell\right) \le e^{-\delta c}}$

(b) $M(\ell) =$ # visits to scale ℓ $\boxed{\mathbb{P}(M(\ell) > k \mid M(\ell) > 0) \le 2^{-k/2}}$

(a)+(b) $\Rightarrow$ Total contribution of level ℓ to height is

$\begin{cases} 0 & \text{if } M(\ell) = 0. \\ O(2^\ell) \text{ with exp. tails} & \text{if } M(\ell) > 0. \end{cases}$ $\Big\|$ So $H(\sigma) \lesssim 2^{\max(\ell : M(\ell) \ne 0)}$

$\le 4 \cdot W(\sigma)$.

Remarks

- Stronger results if add info. about tails of degrees.

Ex: • If $\mathbb{P}(C \geq k) = \Theta(t^{-\alpha})$, $\alpha \in (1,2)$,

then $\mathbb{P}(ht(T) > A \cdot m \cdot wid(T)^{\alpha-1}) \leq 2^{-\delta m}$

• ~~If $Var(C) = \infty$ then $\mathbb{P}(ht(T) > A \cdot m \cdot wid(T)) \leq e^{-m \cdot f(m)}$; $f(m) \xrightarrow{m \to \infty} \infty$~~

FALSE

e.g. if $p_0 > 0$, $p_2 > 0$ then

$$\mathbb{P}(ht(T) > m \cdot wid(T)) \geq \mathbb{P}\left(\text{path } 1, 2, \dots, m\right)$$

$$= p_2^{\,m} p_0^{\,m+1} > e^{-C \cdot m}.$$

But this is the only problem.

Thm: If $Var(C) = \infty$ then for any $\varepsilon(m) \to 0$ there is $f(m) \to \infty$ such that

$$\mathbb{P}(ht(T) > m \cdot wid(T),\ ht(T) < \varepsilon(m) \cdot vol(T)) \leq \exp(-m \cdot f(m))$$

Remarks

- Stronger results if add info. about tails of degrees.

Ex: • If $\mathbb{P}(C \geq k) = \Theta(t^{-\alpha})$, $\alpha \in (1,2)$,

then $\mathbb{P}(ht(T) > A \cdot m \cdot wid(T)^{\alpha-1}) \leq 2^{-\delta m}$

- Conjecture: All this works even conditional on size of tree: $\mathbb{P}(ht(T) > A \cdot m \cdot wid(T) \mid \sigma = n) \leq 2^{-\delta m}$

- Conjecture: Binary trees are the tallest.

Consider random trees $T_{\vec{n}}$ with a fixed degree seq $\vec{n} = (n_i, i \geq 1)$

Here n_i = # nodes of deg i. With $\sum n_i = n$, then $\sum i n_i = 2(n-1)$.

To stochastically maximize $ht(T_{\vec{n}})$ among sequences with $n_0 = k$, $n_1 = 0$, choose the seq. $(k, 0, k-1, 0, \ldots)$

I can prove binary trees stochastically maximize the depth of a random node, but not (yet) for the deepest node.


IMPACASTORINA
IMPA·NWL - NET
Thank you !

- Claimed theorem with dependence only on p_1 proved it with dependence on $p = \max p_i$.

Fix: requires more careful "disperson" bound for our setting.

(Idea: If subcritical then $p_{max} = p_0$ or p_1; if p_0 close to 1 then either very subcritical or make large jumps.)