THE BINARY SEARCH TREE EQUATION

Michael Drmota *

Inst. of Discrete Mathematics and Geometry

Vienna University of Technology, A 1040 Wien, Austria

michael.drmota@tuwien.ac.at

www.dmg.tuwien.ac.at/drmota/

Partly joint work with **Brigitte Chauvin** (Université de Versailles)

* supported by the Austrian Science Founcation FWF, Project S9604

Workshop on Branching Random Walks and Searching in Trees

BIRS Banff, February 1–5, 2010

Outline of the Talk

- The “Binary Search Tree Equation”
- 3 Motivations
- Left/Right-most Particle in Branching Random Walks
- Height of Binary Search Trees
- Profile of Binary Search Trees
- Intersection Property

The "Binary Search Tree Equation"

$$\boxed{\boxed{\Phi'(u) = -\frac{1}{\alpha^2}\Phi\left(\frac{u}{\alpha}\right)^2}}$$

$$\alpha > 0,\ u > 0$$

Laplace transform: $\Phi(u) = \int_0^\infty \Psi(y)e^{-uy}\,dy$:

$$\boxed{y\,\Psi(y/\alpha) = \int_0^y \Psi(w)\Psi(y-w)\,dw}$$

$$\boxed{\Psi(y/\alpha) = \int_0^1 \Psi(yt)\Psi(y(1-t))\,dt = \mathbb{E}\left[\Psi(yU)\Psi(y(1-U))\right]}$$

Additive version: $w(x) = \Psi(e^x)$, $\gamma = \log\alpha$, $X_1 = \log\frac{1}{U}$, $X_2 = \log\frac{1}{1-U}$:

$$\boxed{w(x-\gamma) = \mathbb{E}\left[w(x-X_1)\,w(x-X_2)\right]}$$

The "Binary Search Tree Equation"

Trivial Solutions

- $\Phi(u) = \dfrac{1}{u}$ (for all $\alpha > 0$), $\Psi(y) = 1$

- $\Phi(u) = \dfrac{1}{1+u}$ (for $\alpha = 1$), $\Psi(y) = e^{-y}$

Non-trivial solution

- $\Phi(u) = \dfrac{1+u^{1/4}}{u} e^{-u^{1/4}}$ (for $\alpha = 16$), $\Psi(y) = e^{-y/4}$

The "Binary Search Tree Equation"

First attempt for a solution

$$\Phi(u) = \sum_{n\ge 1} c_n u^n$$

$$c_{n+1} = -\frac{\alpha^{-n-2}}{n+1}\sum_{k=0}^{n} c_k c_{n-k}, \qquad c_0 = \Phi(0) = 1$$

This provides a (unique) and **entire** solution for $\boxed{\alpha > 1}$.

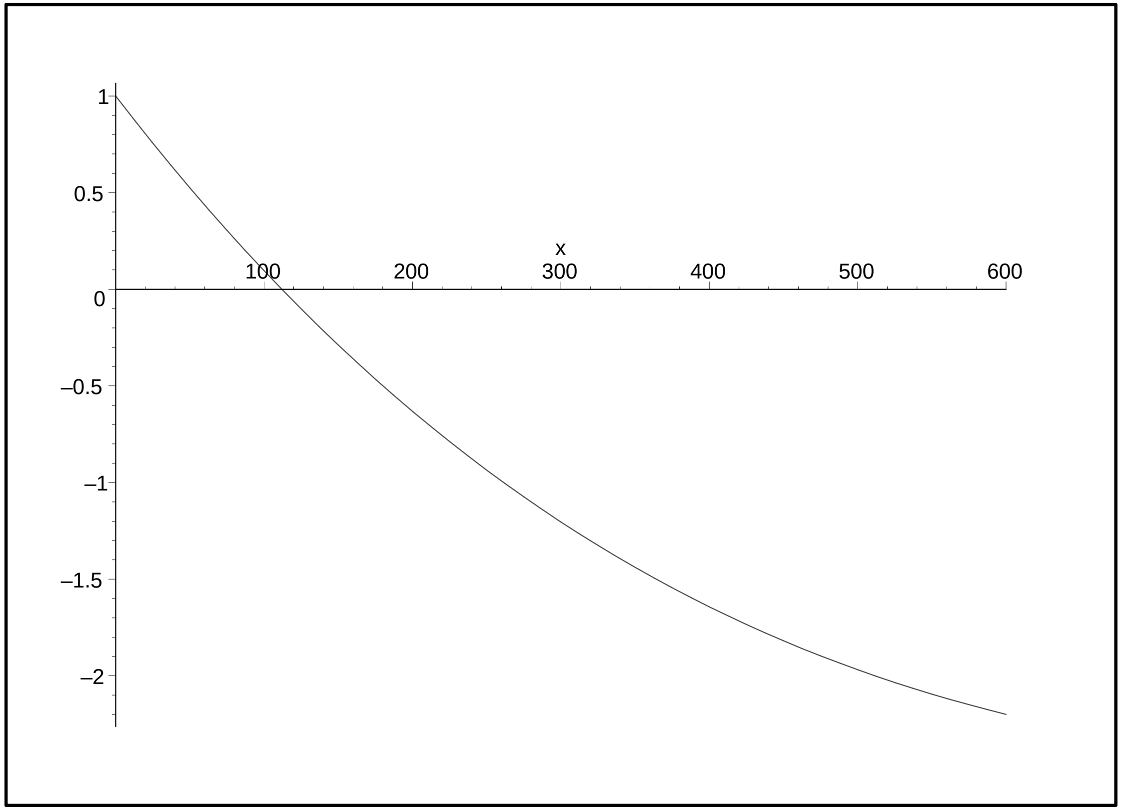
Remark. ($c = 4.31107\ldots$ and $c' = 0.3733\ldots$ satisfy $c\log\left(\frac{2e}{c}\right) = 1$)

- $\alpha \in (0, e^{1/c}] = (0, 1.26\ldots]$: $\Phi(u) \sim \dfrac{1}{u} \quad (u \to \infty)$

- $\alpha \in [e^{1/c'}, \infty) = [14.56\ldots, \infty)$: $\Phi(u) \sim \dfrac{1}{u} \quad (u \to 0)$

The “Binary Search Tree Equation”

Out of Range: e.g. $\alpha = 10$: $\Phi(u) \not\sim 1/u$
(no Laplace transform of a (tail) distribution function $\Psi(y)$)



Motivation 1: Branching Random Walk

Random point measure

$$\boxed{Z = \delta_{X_1} + \delta_{X_2}}$$

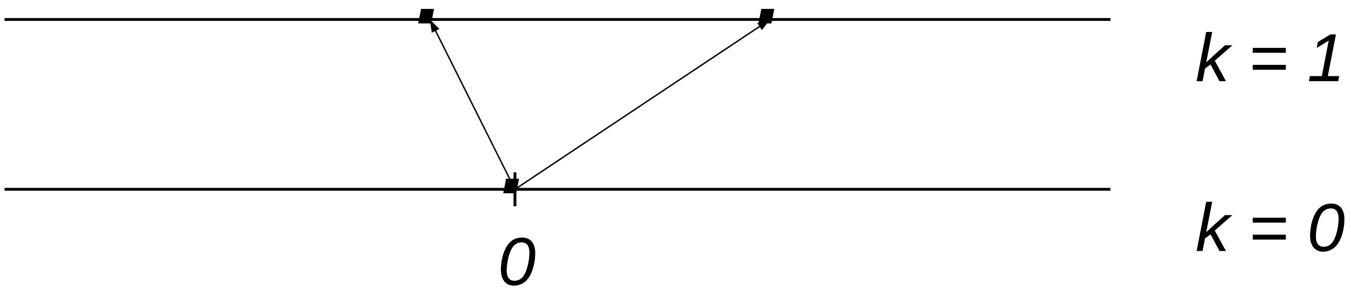
For example: $X_1 = \log(1/U)$, $X_2 = \log(1/(1-U))$.

Branching Random Walk: Sequence Z_k of random point measures:

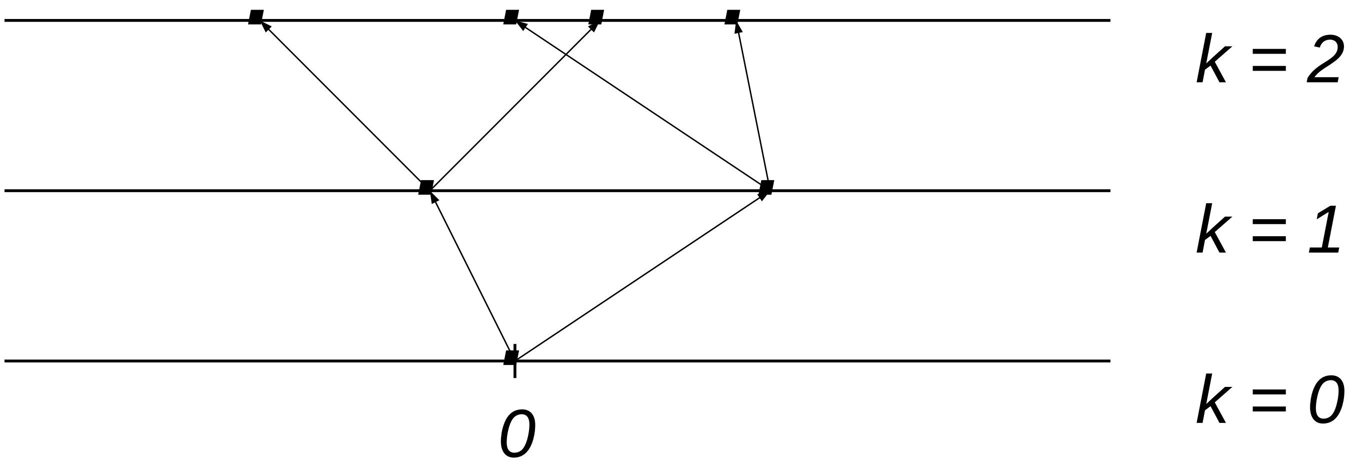
- $Z_0 = \delta_0$.
- Z_{k+1} is induced by Z_k by adding independent copies of Z to all points of Z_k.

Motivation 1: Branching Random Walk

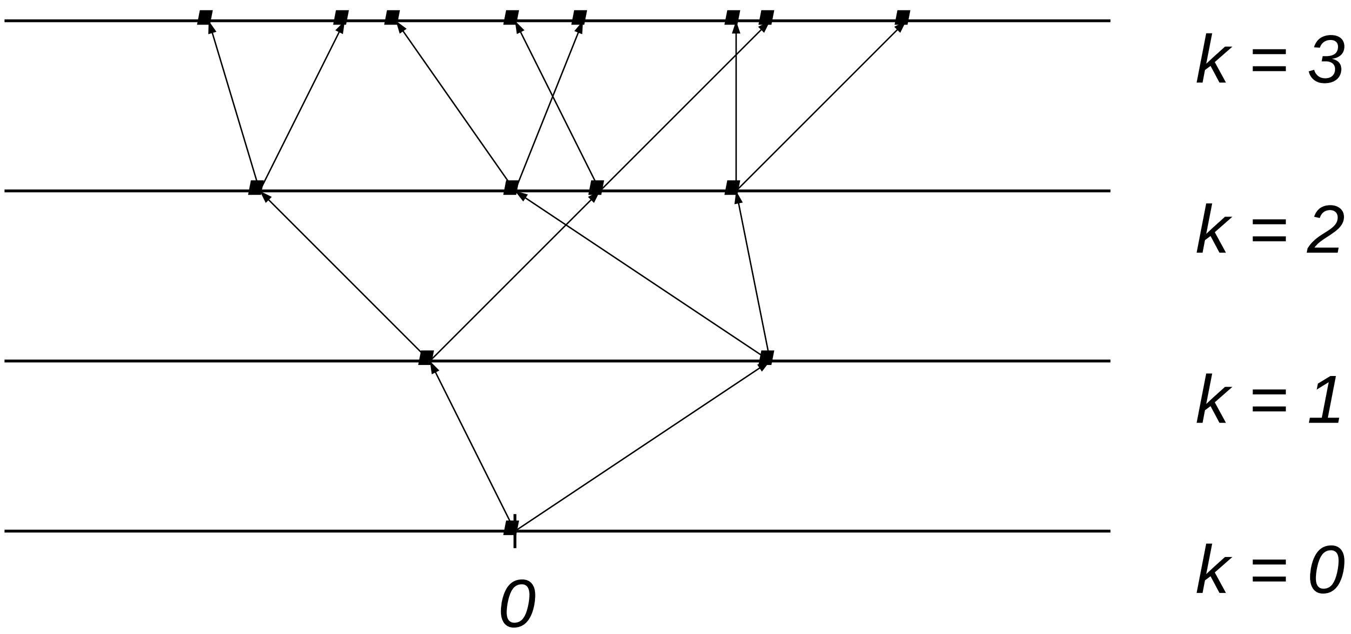
Motivation 1: Branching Random Walk



Motivation 1: Branching Random Walk



Motivation 1: Branching Random Walk



Motivation 1: Branching Random Walk

L_k ... Position of the **leftmost** point (after k steps)

$$w_k(x) = \mathbb{P}\{L_k > x\}$$

R_k ... Position of the **rightmost** point (after k steps)

$$\overline{w}_k(x) = \mathbb{P}\{L_k \leq x\}$$

$$w_{k+1}(x) = \mathbb{E}\left[w_k(x - X_1)\, w_k(x - X_2)\right]$$

$$\overline{w}_{k+1}(x) = \mathbb{E}\left[\overline{w}_k(x - X_1)\, \overline{w}_k(x - X_2)\right]$$

Travelling wave: $w_k(x) = w(x - k\gamma)$:

$$\boxed{w(x - \gamma) = \mathbb{E}\left[w(x - X_1)\, w(x - X_2)\right]}$$

Motivation 1: Branching Random Walk

Special case: $X_1 = \log(1/U)$, $X_2 = \log(1/(1-U))$

- **Iteration**

$$Y_k(u) := \int_0^\infty w_k(\log y) e^{-uy}\, dy$$

$$\boxed{Y'_{k+1}(u) = Y_k(u)^2}$$

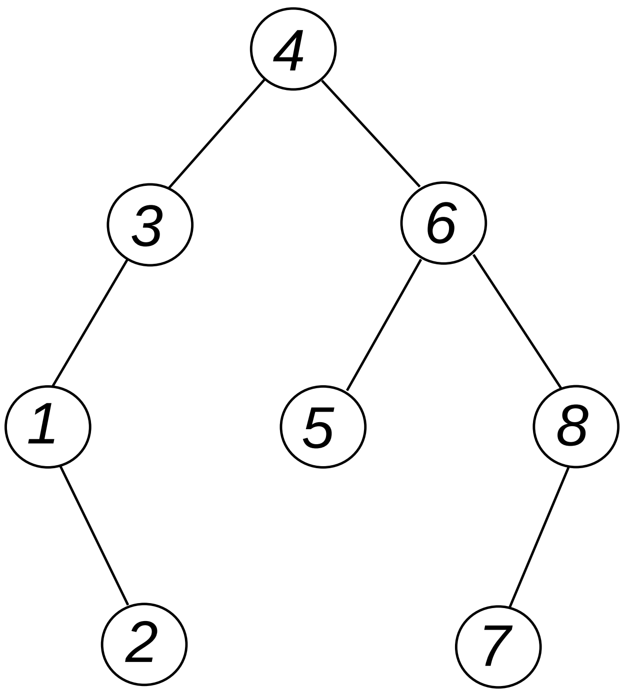
- **Travelling wave:** $w_k(x) = w(x - k\gamma)$, $\alpha = e^\gamma$

$$\Phi(u) = \int_0^\infty w(\log y) e^{-uy}\, dy$$

$$\boxed{\Phi'(u) = -\alpha^{-2} \Phi(u/\alpha)^2}$$

Motivation 2: Binary Search Trees

Vertex labelled binary tree:



Motivation 2: Binary Search Trees

Storing Data:

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

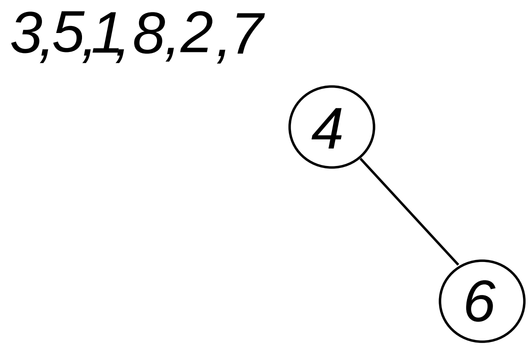
Motivation 2: Binary Search Trees

Storing Data:

6,3,5,1,8,2,7

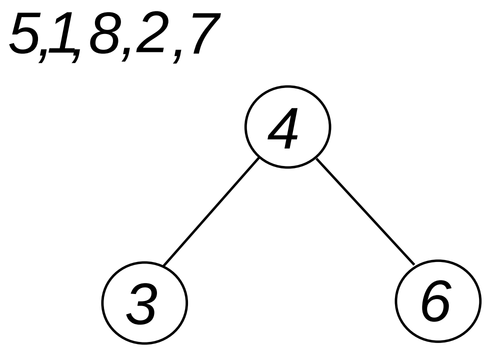
Motivation 2: Binary Search Trees

Storing Data:



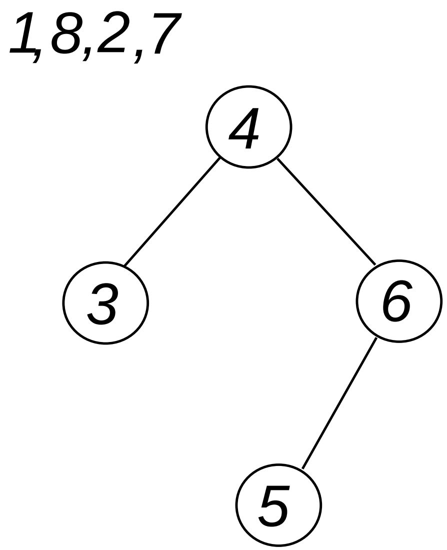
Motivation 2: Binary Search Trees

Storing Data:



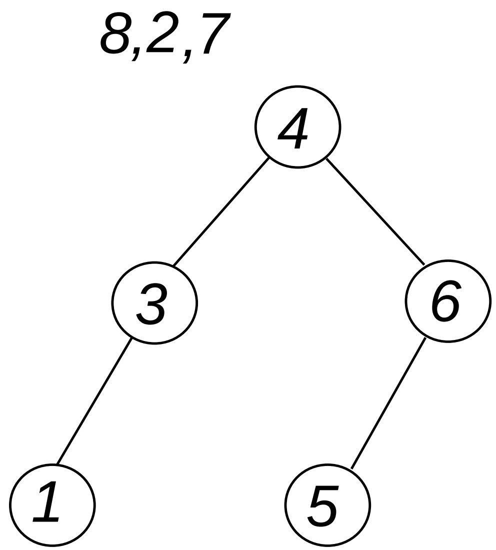
Motivation 2: Binary Search Trees

Storing Data:



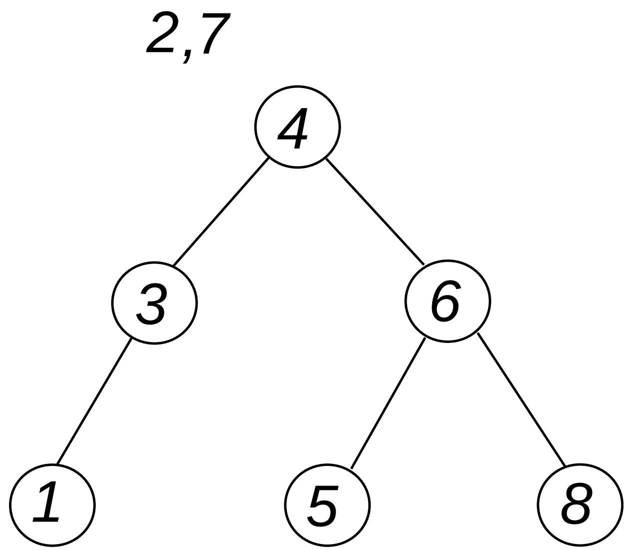
Motivation 2: Binary Search Trees

Storing Data:



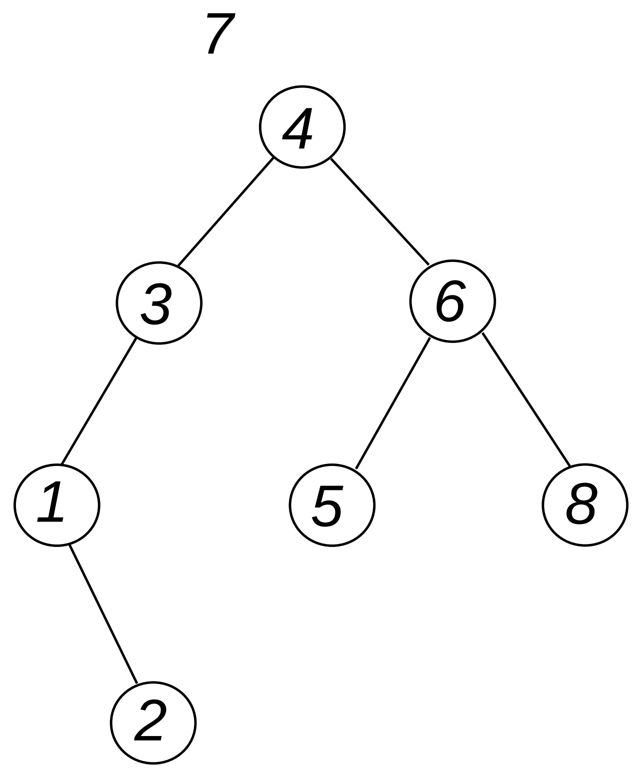
Motivation 2: Binary Search Trees

Storing Data:



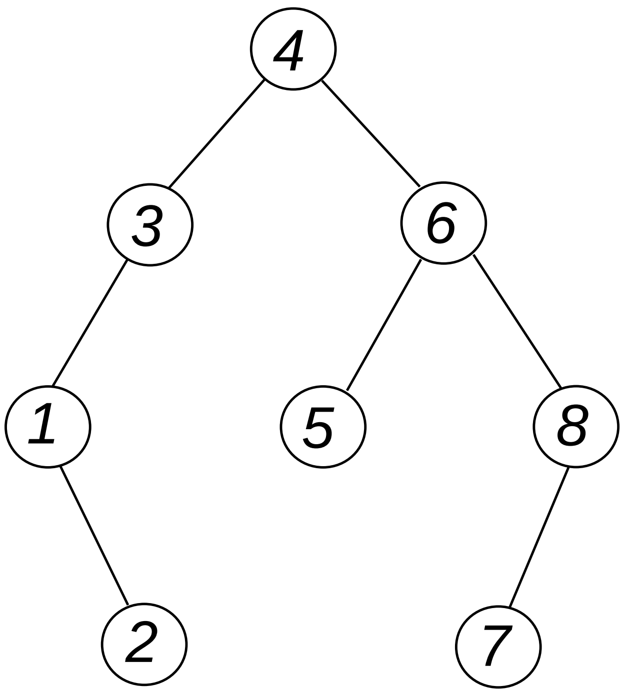
Motivation 2: Binary Search Trees

Storing Data:



Motivation 2: Binary Search Trees

Storing Data:



Motivation 2: Binary Search Trees

Probabilistic Model:

Every permutation of $\{1, 2, \ldots, n\}$ is equally likely.

$\longrightarrow$ probability distribution on binary trees of size n

$\longrightarrow$ every parameter on trees is a **random variable**

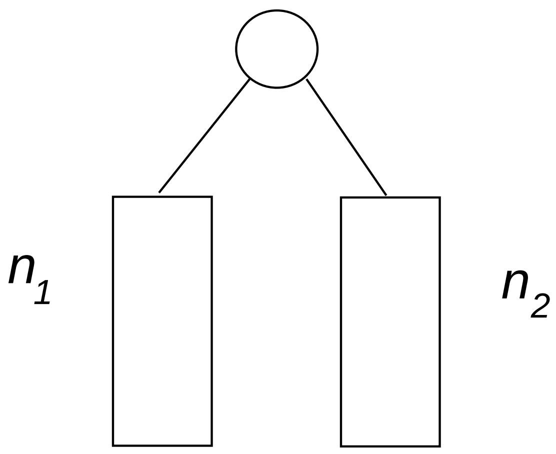
Notation

H_n ... **height** of trees (of size n)

Motivation 2: Binary Search Trees

Observation: Subtrees of the root are also binary search trees, the splitting probabilities are $\frac{1}{n}$.

$$\boxed{\mathbb{P}\{H_{n+1} \leq k+1\} = \frac{1}{n} \sum_{n_1+n_2=n} \mathbb{P}\{H_{n_1} \leq k\} \cdot \mathbb{P}\{H_{n_2} \leq k\}}$$



Motivation 2: Binary Search Trees

Generating Functions:

$$\boxed{y_k(x) = \sum_{n \ge 0} \mathbb{P}\{H_n \le k\} \cdot x^n}$$

$$\boxed{\boxed{y'_{k+1}(x) = y_k(x)^2}}$$

with initial conditions $y_1(x) = 1, \ y_k(0) = 1$.

Motivation 2: Binary Search Trees

Generating Functions:

Special solution of the recurrence $y'_{k+1}(x) = y_k(x)^2$**:**

$$\boxed{y_k(x) = \alpha^k \Phi(\alpha^k(1-x))}$$

where $\Phi'(u) = -\alpha^{-2}\Phi(u/\alpha)^2$.

Analogue of the **travelling wave solution** in BRW's.

Motivation 2: Binary Search Trees

Profile of Binary Search Trees

$X_{n,k}$ ... number of nodes at level k (in a BST with n vertices)

$X_k(x,u) = \sum_{n \ge 0} \mathbb{P}\{X_{n,k} = \ell\}\, x^n u^\ell$:

$$\boxed{\frac{\partial}{\partial x} X_{k+1}(x,u) = X_k(x,u)^2.}$$

Motivation 3: Stochastic Fixed Point Equations

$$\boxed{Y \equiv V_1 Y^{(1)} + V_2 Y^{(2)}}$$

$Y^{(1)}, Y^{(2)}$ copies of Y, $((V_1, V_2), Y^{(1)}, Y^{(2)})$ independent.

$$G(x) = \mathbb{E}\, e^{-xY}$$

$$\boxed{G(x) = \mathbb{E}\,[G(xV_1)\, G(xV_2)]}$$

Motivation 3: Stochastic Fixed Point Equations

Special case: $z > 0, z \neq \frac{1}{2}$

$$\boxed{Y \equiv zU^{2z-1}Y^{(1)} + z(1-U)^{2z-1}Y^{(2)}}$$

$$G(x) = \mathbb{E}\, e^{-xY}$$

$$G(x) = \mathbb{E}\left[G(xzU^{2z-1})\, G(xz(1-U)^{2z-1})\right]$$

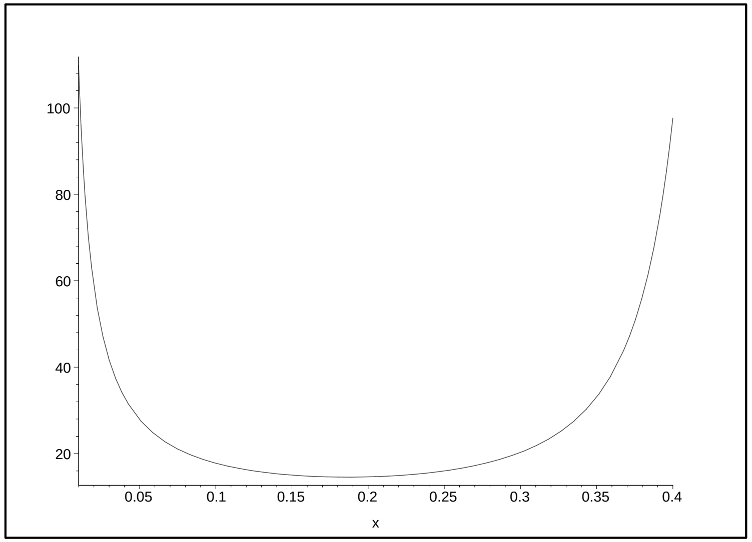
$$\Psi(y) = G(y^{2z-1}) = \mathbb{E}\left(e^{-y^{2z-1}Y}\right)$$

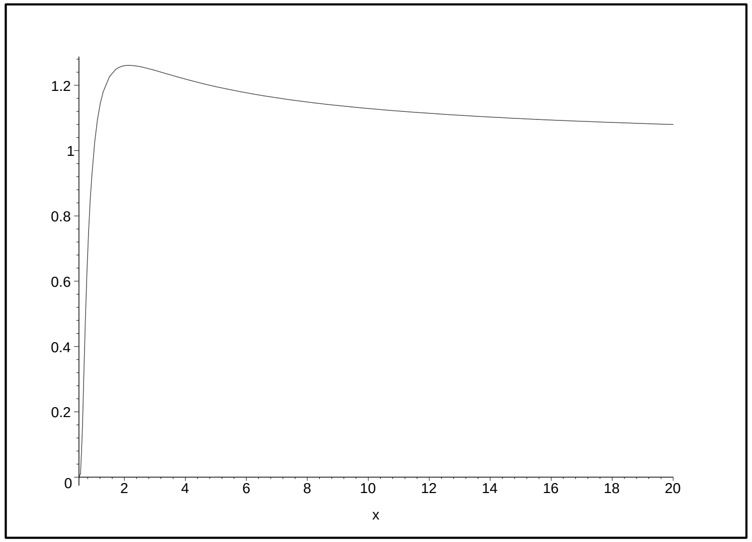
$$\boxed{\Psi\left(y/z^{\frac{1}{2z-1}}\right) = \mathbb{E}\left[\Psi(yU)\,\Psi(y(1-U))\right]}$$

$$\alpha = \alpha(z) = z^{\frac{1}{2z-1}}$$

Motivation 3: Stochastic Fixed Point Equations

Behaviour of $\alpha(z) = z^{\frac{1}{2z-1}}$**:**





$$0 < z < \frac{1}{2} \qquad \frac{1}{2} < z < \infty$$

$$\alpha(z) \geq e^{1/c'} = 14.56\ldots \qquad 0 \leq \alpha(z) \leq e^{1/c} = 1.26\ldots$$

Motivation 3:
Stochastic Fixed Point Equations

Existence of solutions: [Biggins + Kyprianou, Liu]

$$G(x) = \mathbb{E}\left[\prod_j G(xV_j)\right]$$

$$v(\gamma) = \log\left(\mathbb{E}\left[\sum_j V_j^\gamma\right]\right), \quad v(0) > 0, \quad v(1 \pm \varepsilon) < \infty$$

- $v(1) = 0$, $v'(1) = 0$: $\quad \dfrac{1 - G(x)}{-x \log x} \to c_1 \quad (x \to 0)$

- $v(1) = 0$, $v'(1) < 0$: $\quad \dfrac{1 - G(x)}{x} \to c_2 \quad (x \to 0)$

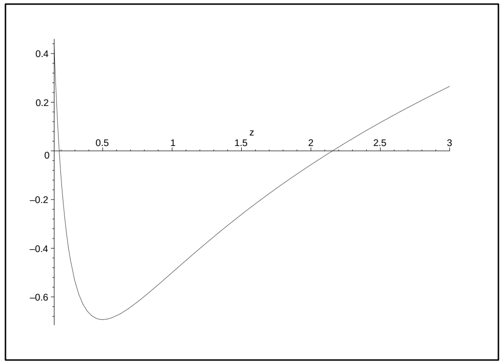
Motivation 3:
Stochastic Fixed Point Equations

Special case: $V_1 = zU^{2z-1}$, $V_2 = z(1-U)^{2z-1}$

$$v(\gamma) = \log \frac{2z^\gamma}{(2z-1)\gamma + 1}$$

$$v(1) = 0, \qquad v'(1) = \log z + \frac{1}{2z} - 1$$

$$v'(1) \le 0 \qquad \Longleftrightarrow \qquad \boxed{\frac{c'}{2} \le z \le \frac{c}{2}}$$



Motivation 3: Stochastic Fixed Point Equations

$$\Psi(y/\alpha) = \mathbb{E}\left[\Psi(yU)\,\Psi(y(1-U))\right]$$

Case 1. $0 < \alpha \le e^{1/c} = 1.26\ldots$, $2\alpha^\beta = \beta + 1$

$$1 - \Psi(y) \sim c_1 y^\beta \quad (y \to 0) \quad \text{for } 0 < \alpha < e^{1/c}$$

$$1 - \Psi(y) \sim c_1 y^{c-1} \log y \quad (y \to 0) \quad \text{for } \alpha = e^{1/c}$$

$\Psi(u)$ is monotonely decreasing (tail distribution function).

Case 2. $e^{1/c'} \le \alpha < \infty$, $2\alpha^\beta = \beta + 1$

$$1 - \Psi(y) \sim c_1 y^\beta \quad (y \to \infty) \quad \text{for } e^{1/c'} < \alpha < \infty$$

$$1 - \Psi(y) \sim c_1 y^{c'-1} \log y \quad (y \to \infty) \quad \text{for } \alpha = e^{1/c'}$$

$\Psi(u)$ is monotonely increasing (distribution function).

Motivation 3: Stochastic Fixed Point Equations

Notation

- $\alpha = e^{1/c} = 1.26\ldots$ (or $z = c/2$):

$$\Psi_c(y), \quad w_c(x) = \Psi_c(e^x)$$

- $\alpha = e^{1/c'} = 14.56\ldots$ (or $z = c'/2$):

$$\Psi_{c'}(y), \quad w_{c'}(x) = \Psi_{c'}(e^x)$$

Left/Right-most Point in BRW's

Theorem [Chauvin + D.]

Z_k ... BRW with $Z_0 = \delta_0$ and increments $X_1 = \log(1/U)$, $X_2 = \log(1/(1-U))$.

L_k, R_k ... position of the left/right-most particle (after k steps)
$m_1(k)$, $m_2(k)$ ... median of the distributions of L_k, R_k, resp.

$$\boxed{\mathbb{P}\{L_k > x\} = w_c(x - m_1(k)) + o(1)}$$

$$\boxed{\mathbb{P}\{R_k \le x\} = w_{c'}(x - m_2(k)) + o(1)}$$

$$m_1(k) = \frac{1}{c}k + \Theta(\log k) \qquad m_2(k) = \frac{1}{c'}k + \Theta(\log k) \qquad (k \to \infty),$$

$$\mathbb{P}\{|L_k - m_1(k)| > x\} \le Ce^{-\eta x}, \qquad \mathbb{P}\{|R_k - m_2(k)| > x\} \le Ce^{-\eta x}.$$

Left/Right-most Point in BRW's

Extensions

$m \geq 2$, $(V_1, \ldots, V_m)$ r.v.'s with $V_1 + \cdots + V_m = 1$ and density

$$f(x_1, \ldots, x_m) = \frac{(m(t+1)-1)!}{(t!)^m}(x_1 x_2 \cdots x_m)^t$$

on the simplex $x_1 + \cdots + x_m = 1$, $0 \leq x_j \leq 1$
($t \geq 0$ is a integer parameter.)

Z_k BRW with increments $X_j = \log(1/V_j)$ $(1 \leq j \leq m)$.

Then there exist functions $w_1(x)$ and $w_2(x)$ such that

$$\boxed{\mathbb{P}\{L_k > x\} = w_1(x - m_1(k)) + o(1)}$$

$$\boxed{\mathbb{P}\{R_k \leq x\} = w_2(x - m_2(k)) + o(1)}$$

with medians

$$m_1(k) = k \log \rho_1 + \Theta(\log k), \qquad m_2(k) = k \log \rho_2 + \Theta(\log k) \qquad (k \to \infty).$$

Height of Binary Search Trees

$y'_{k+1}(x) = y_k(x)^2$, $y_1(x) = 1$, $y_k(0) = 1$.

Theorem

H_n ... height of binary search trees with n nodes.

$$\boxed{\mathbb{P}\{H_n \le k\} = \Psi_c(n/y_k(1)) + o(1)}$$

$$\log y_k(1) = \frac{k}{c} + \frac{3c}{2(c-1)} \log k + O(1)$$

$$\mathbb{E}\, H_n = \max\{k \ge 0 : y_k(1) \le n\} + O(1) = c \log n - \frac{3c}{2(c-1)} \log\log n + O(1)$$

$$\mathbb{P}\{|H_n - \mathbb{E}\, H_n| > y\} = O(e^{-\eta y})$$

Remark. The function $\Psi_{c'}(y)$ describes the distribution of the saturation level (up to this level the tree is a complete binary trees).

Height of Binary Search Trees

Extensions

- m-ary search trees (also fringe-balanced versions)
- recursive trees
- plane oriented recursive trees
- m-ary recursive trees
- ...

Height of Binary Search Trees

History

- $\mathbf{Var}\, H_n = O(1)$??? [Robson 1979] (**Robson's conjecture**)
- $\mathbb{E}\, H_n \sim c \log n$ [Devroye 1986]
- $\mathbb{E}\, H_n = c \log n + O(\log\log n)$ [Devroye+Reed 1995]
- $\mathbb{E}\, H_n = c \log n - \frac{3c}{2(c-1)} \log\log n + O(1)$ [Reed 2003]
- $\mathbf{Var}\, H_n = O(1)$ [Reed 2003] [D. 2003]

Height of Binary Search Trees

More on the variance $\mathbf{Var}\, H_n$:

$$V(x) := \sum_{k\ge 0} (2k+1)\left(1-\Psi_c\left(\frac{x}{y_k(1)}\right)\right) - \left(\sum_{k\ge 0}\left(1-\Psi_c\left(\frac{x}{y_k(1)}\right)\right)\right)^2$$

$$V(e^{1/c}x) = V(x) + o(1) \qquad (x\to\infty).$$

$$\boxed{\mathbf{Var}\, H_n = V(n) + o(1) \qquad (n\to\infty)}$$

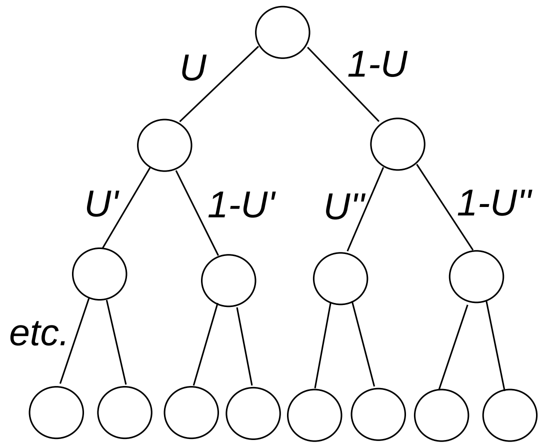
$$\boxed{\max_{n\ge n_1} |\mathbf{Var}\, H_n - v_0| \le 10^{-3}}$$

$$v_0 = c\int_0^\infty (E(u) + E(ue^{-1/c}))\Psi_c(u)\,\frac{du}{u} = 2.085687\ldots$$

$$E(u) := \sum_{k\ge 0}\left(1-\Psi_c(ue^{-k/c})\right).$$

Height of Binary Search Trees

Direct relation between BST's and BRW's [Devroye]



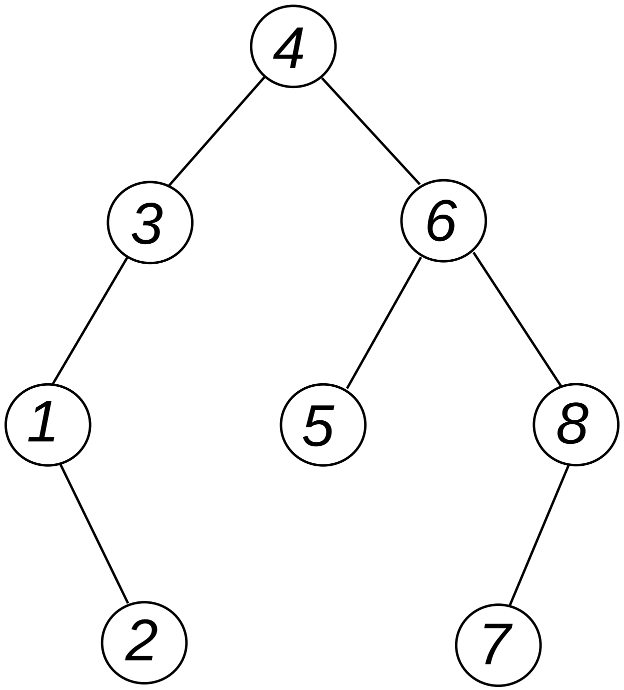
x ... vertex of (infinite) binary tree (at level k)
$U_1, U_2, \ldots, U_k$ ... r.v.'s on the path from the root to x

$h_n(x) = \lfloor U_k \lfloor \cdots \lfloor U_2 \lfloor U_1 n \rfloor \rfloor \cdots \rfloor \rfloor$

$$BST_n = \{x : h_n(x) \geq 1\}$$

Profile of Binary Search Trees

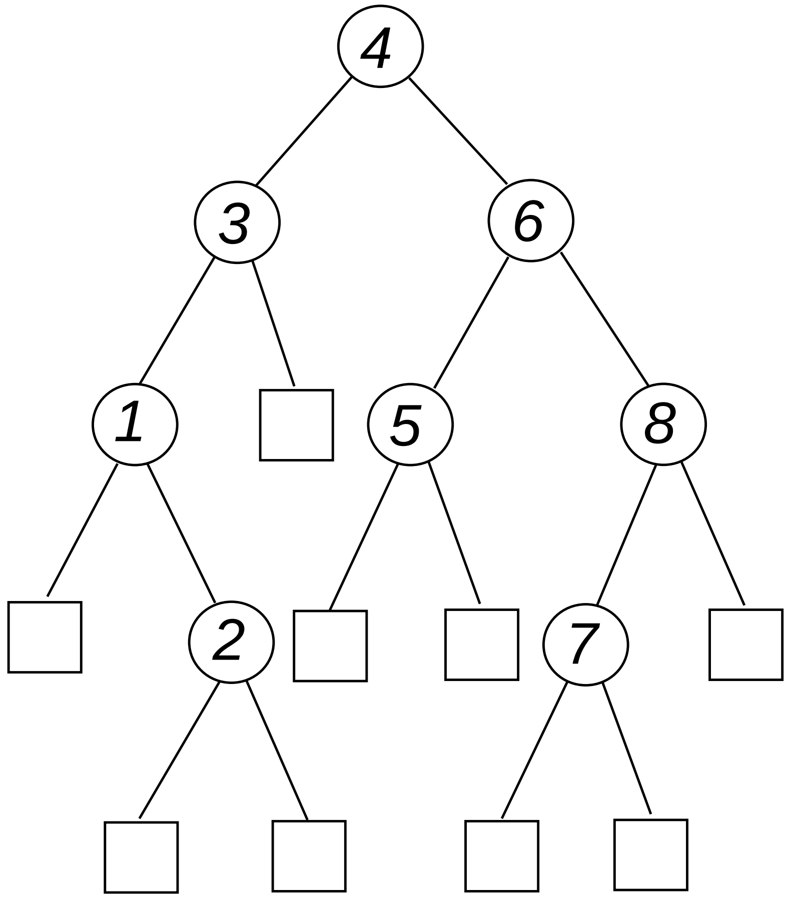
Internal and external profile:



Including “free” places

Profile of Binary Search Trees

Internal and external profile:



□ ... “free” place

Profile of Binary Search Trees

Internal and external profile:

$X_{n,k}$ ... number of **internal** vertices at level k

$Y_{n,k}$ ... number of **external** vertices at level k

$$\boxed{X_{n,k} = \sum_{j>k} 2^{k-j}\, Y_{n,j}}$$

Profile of Binary Search Trees

A stochastic process of analytic functions

$(M(z), z \in B)$ stochastic process of analytic functions that is defined by $\mathbb{E}\, M(z) = 1$ and the **stochastic fixed point equation**:

$$\boxed{M(z) \equiv zU^{2z-1}M^{(1)}(z) + z(1-U)^{2z-1}M^{(2)}(z)}$$

B ... domain in $\mathbb{C}$ with $B \cap \mathbb{R} = (\frac{c'}{2}, \frac{c}{2}) =: I$

Profile of Binary Search Trees

Theorem [Chauvin+D.+Jabbour, Chauvin+Klein+Marckert+Rouault]

$Y_{n,k}$ ... number of **external** vertices at level k

$$\left(\frac{Y_{n,\lfloor 2z \log n \rfloor}}{\mathbf{E}\, Y_{n,\lfloor 2z \log n \rfloor}}, z \in I\right) \to (M(z), z \in I)\,.$$

(almost surely!!)

$X_{n,k}$ ... number of **internal** vertices at level k

$$\left(\frac{X_{n,\lfloor 2z \log n \rfloor}}{\mathbf{E}\, X_{n,\lfloor 2z \log n \rfloor}}, z \in I'\right) \to \left(M(z), z \in I'\right).$$

$I' = (\frac{1}{2}, \frac{c}{2})$

Profile of Binary Search Trees

Extensions:

- m-ary search trees (also fringe balanced) [D.+Janson+Neininger]
- recursive trees, plane oriented recursive trees [Schopp]

Profile of Binary Search Trees

Profile polynomials

$$W_n(z) = \sum_{k \ge 0} Y_{n,k} z^k$$

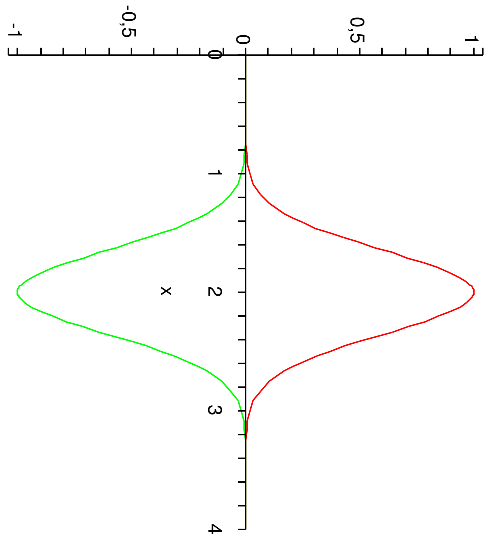
$$M_n(z) = \frac{W_n(z)}{\mathbb{E}\, W_n(z)} \quad \text{is a } \textbf{martingale}$$

$$\boxed{M_n(z) \to M(z)}$$

Profile of Binary Search Trees

Expected profile

$$\mathbf{E}\,Y_{n,k} \sim \frac{(2\log n)^k}{k!\,n\,\Gamma(k/\log n)}$$



Profile of Binary Search Trees

Fixed point equation

$$Y_{n,k+1} \equiv Y^{(1)}_{\lfloor Un \rfloor,k} + Y^{(2)}_{n-1-\lfloor Un \rfloor,k}$$

If the limit

$$\frac{Y_{n,\lfloor 2z \log n \rfloor}}{\mathbb{E}\, Y_{n,\lfloor 2z \log n \rfloor}} \to M(z)$$

exists then

$$M(z) \equiv zU^{2z-1} M^{(1)}(z) + z(1-U)^{2z-1} M^{(2)}(z).$$

Intersection Property

Point process:

$$Z = \sum_{j=1}^{N} \delta_{X_j},$$

Example: $N = 2$, $X_1 = \log(1/V)$, $X_2 = \log(1/(1-V))$.

Transform $\mathbf{T}$ (for distributions functions):

$$(\mathbf{T}G)(x) = \mathbf{E}\left(\prod_{j=1}^{N} G(x - X_j)\right).$$

Example: $G(x) = F(e^{-x})$: $F(x) = \mathbf{E}(F(xV)F(x(1-V)))$.

Intersection Property

Intersection property:

> *Suppose that $F(x)$ and $G(x)$ are continuous distribution functions such that the difference $F(x) - G(x)$ has exactly one zero. Then the difference $(\mathbf{T}\,F)(x) - (\mathbf{T}\,G)(x)$ has at most one zero.*

Intersection Property

Lemma.

Suppose that V is t-beta distributed and **T** is defined by $(\mathbf{T}F)(x) = \mathbf{E}(F(xV)F(x(1-V)))$.

Then the Laplace transforms $\Phi(u) = \int_0^\infty F(x)e^{-xu}\,dx$ satisfy an *intersection property*.

This property is the **key property** for the proof of the travelling wave property for the **left/right-most particle of BRW's** and also for the distribution of the **height of binary search trees**.

It is not clear whether this is also true on the level of distributions functions?

Intersection Property

Theorem

Let $G_0(x) = 0$ for $x < 0$ and $G_0(x) = 1$ for $x \geq 0$ and set $G_{k+1} = \mathbf{T}\, G_k$, that is,

$$G_{k+1}(x) = \mathbf{E}\left(\prod_{j=1}^{N} G_k(x - X_j)\right).$$

If $\mathbf{T}$ satisfies the *intersection property* then there exists $w(x)$ such that (uniformly for real x as $k \to \infty$)

$$\boxed{G_k(x) = w(x - m(k)) + o(1)},$$

where $m(k)$ is defined by $G_k(m(k)) = \frac{1}{2}$.

More precisely, we have

$$\boxed{m(k) = kc + o(k)}.$$

for some constant $c > 0$ and $w(x)$ satisfies

$$\boxed{w(x) = \mathbf{E}\left(\prod_{j=1}^{N} w(x + c - X_j)\right)}.$$

Thank You!