DIE MATHEMATIK VON QUICKSORT UND VERWANDTE FRAGESTELLUNGEN

Michael Drmota

Institut für Diskrete Mathematik und Geometrie,

TU Wien

michael.drmota@tuwien.ac.at

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

TU Wien, 2. Juni 2006

Wechselwirkung

Anwendungsorientierte — Mathematik
Fragestellungen

Algorithmen für Datenstrukturen	—	• probabilistische Grenzwertsätze • diophantische Approximation

Inhalt

- Quicksort – Binäre Suchbäume
- Profil
- Höhe
- Präfix-Codes
- Diophantische Approximation
- Tunstall-Codes (Varn-Codes)

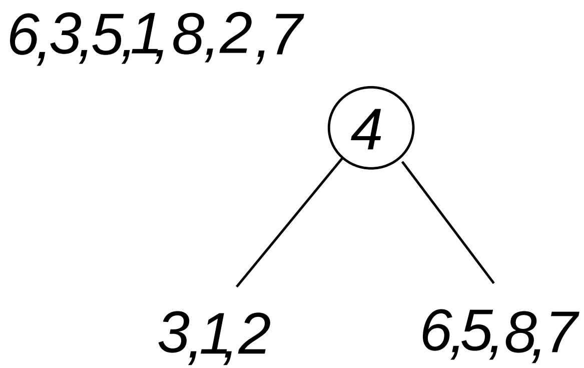
Quicksort

Sortieren von Daten

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

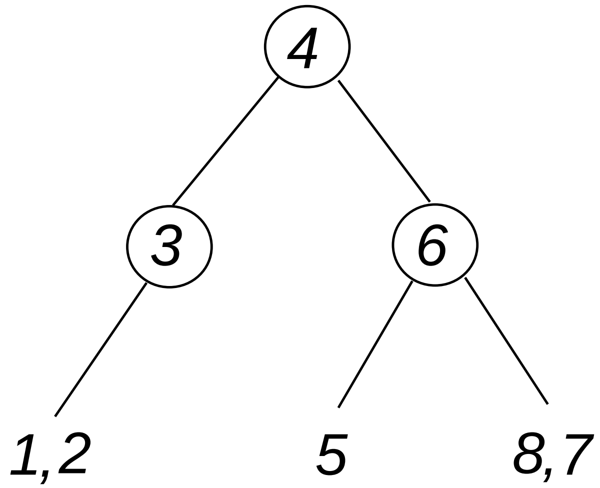
Quicksort

Sortieren von Daten



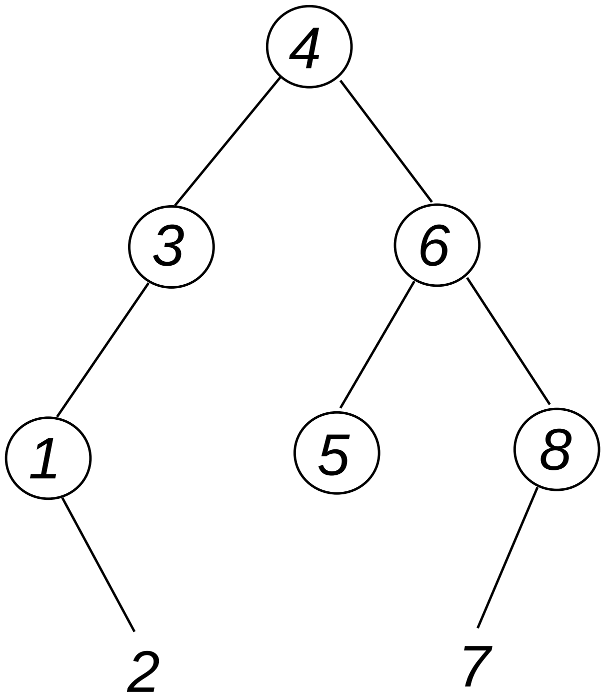
Quicksort

Sortieren von Daten



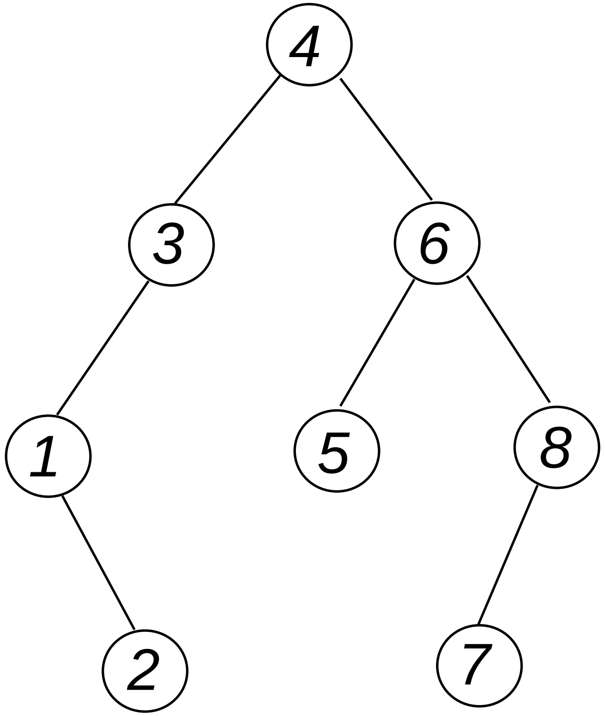
Quicksort

Sortieren von Daten



Quicksort

Sortieren von Daten



Binäre Suchbäume

Speichern von Daten

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

Binäre Suchbäume

Speichern von Daten

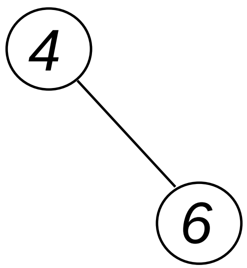
6,3,5,1,8,2,7

4

Binäre Suchbäume

Speichern von Daten

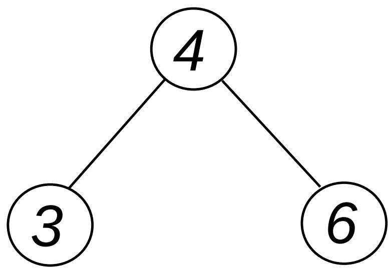
3,5,1,8,2,7



Binäre Suchbäume

Speichern von Daten

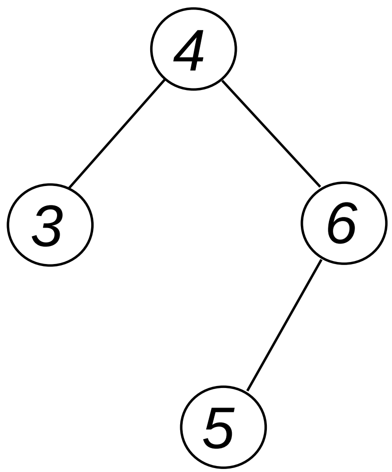
5,1,8,2,7



Binäre Suchbäume

Speichern von Daten

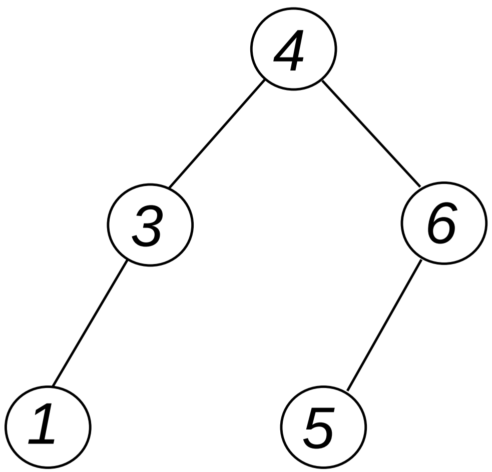
1,8,2,7



Binäre Suchbäume

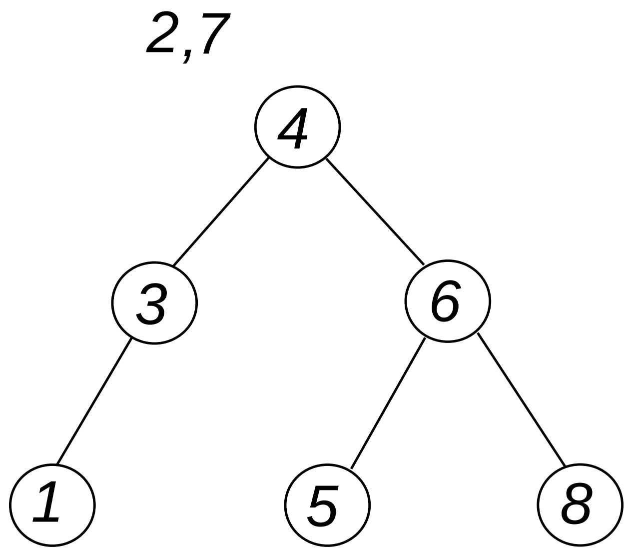
Speichern von Daten

8,2,7



Binäre Suchbäume

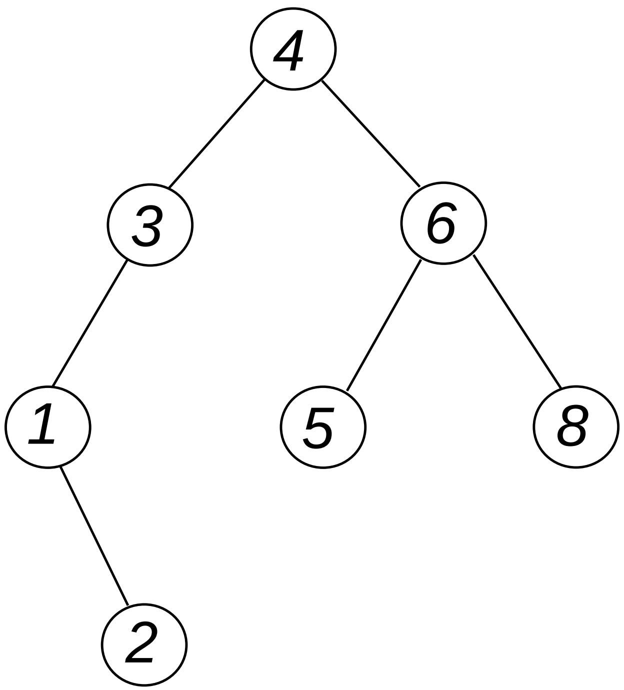
Speichern von Daten



Binäre Suchbäume

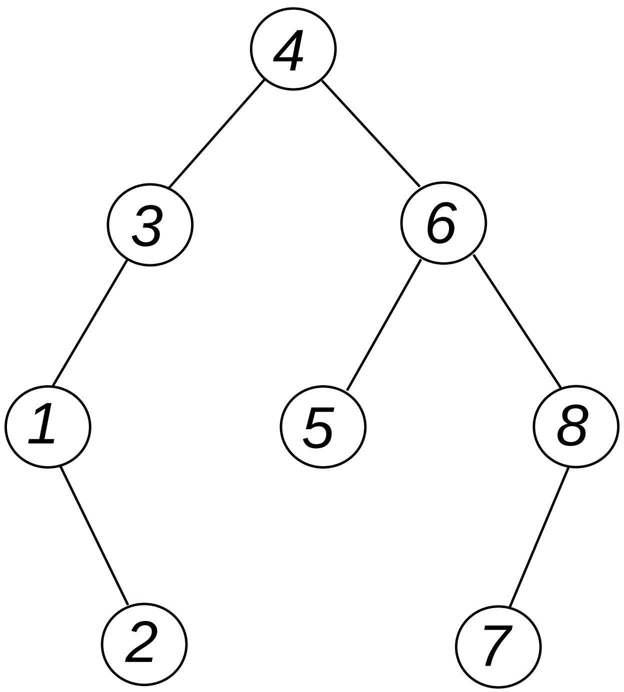
Speichern von Daten

7



Binäre Suchbäume

Speichern von Daten



Quicksort

Median of 3 – Variante

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

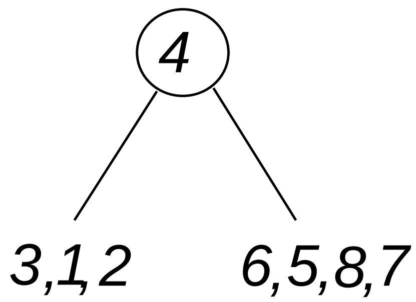
Quicksort

Median of 3 – Variante

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

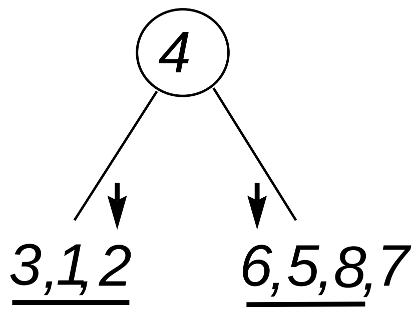
Quicksort

Median of 3 – Variante



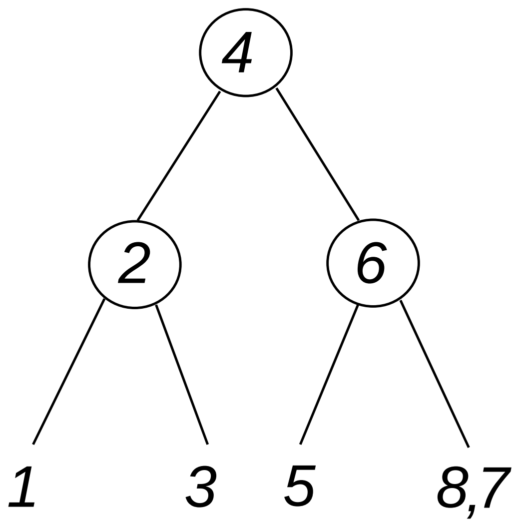
Quicksort

Median of 3 – Variante



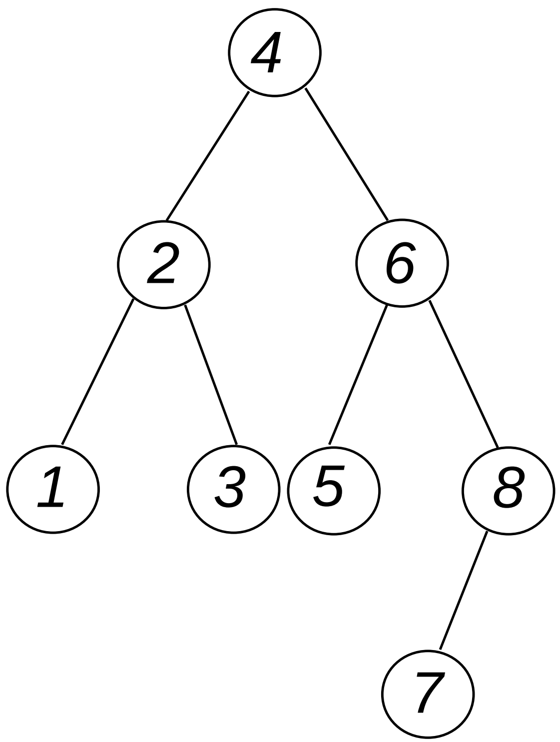
Quicksort

Median of 3 – Variante



Quicksort

Median of 3 – Variante



Binäre Suchbäume

$$\boxed{\textbf{Analyse von Quicksort} = \textbf{Analyse von binären Suchbäumen}}$$

Wahrscheinlichkeitsmodell:

Jede Permutation von $\{1, 2, \ldots, n\}$ ist gleichwahrscheinlich

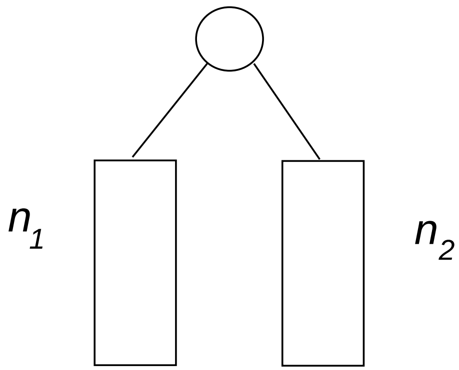
$\longrightarrow$ Wahrscheinlichkeitsverteilung auf den Binärbäumen der Größe n

$\longrightarrow$ jeder Baumparameter wird zu einer **Zufallsvariablen**

Binäre Suchbäume

Rekursiver Aufbau

Die Unterbäume der Wurzel tragen wieder dieselbe Struktur:
$(n = n_1 + n_2 + 1)$.



Verzweigungswahrscheinlichkeiten: p_{n_1,n_2}

Quicksort: $\boxed{p_{n_1,n_2} = \frac{1}{n}}$ Median of 3: $\boxed{p_{n_1,n_2} = \frac{n_1 n_2}{\binom{n}{3}}}$

Binäre Suchbäume

Parameter:

- **Tiefe** eines zufällig gewählten Knoten: D_n
- **Interne Pfadlänge**: I_n (Summe aller Distanzen zur Wurzel)
- **Höhe** H_n
- **Profil** $X_{n,k}$ (Anzahl der Knoten mit Tiefe k)

Bemerkung:

Anzahl der Vergleichsopertationen in Quicksort = int, Pfadlänge I_n

Binäre Suchbäume

Bedeutung des Profils $X_{n,k}$:

- $\mathbf{Pr}\{D_n = k\} = \dfrac{1}{n}\,\mathbf{E}\,X_{n,k}$

- $I_n = \sum_{k\geq 0} k X_{n,k}$

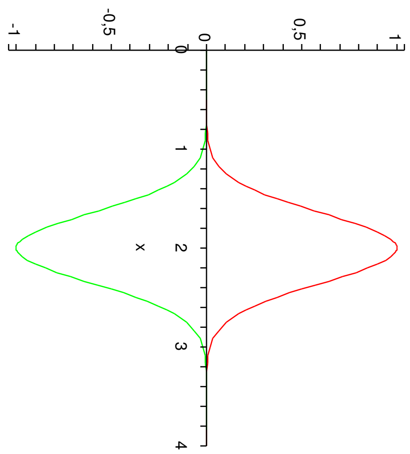
- $H_n = \max\{k \geq 0 : X_{n,k} > 0\}$

- Das Profil beschreibt die *Gestalt* des Baums.

Profil

Das durchschnittliche Profil:

$$\mathbf{E}\,X_{n,k} = \frac{n}{\sqrt{4\pi \log n}}\left(e^{-\frac{(k-2\log n)^2}{4\log n}} + \mathcal{O}\left(\frac{1}{\sqrt{\log n}}\right)\right).$$



Profil

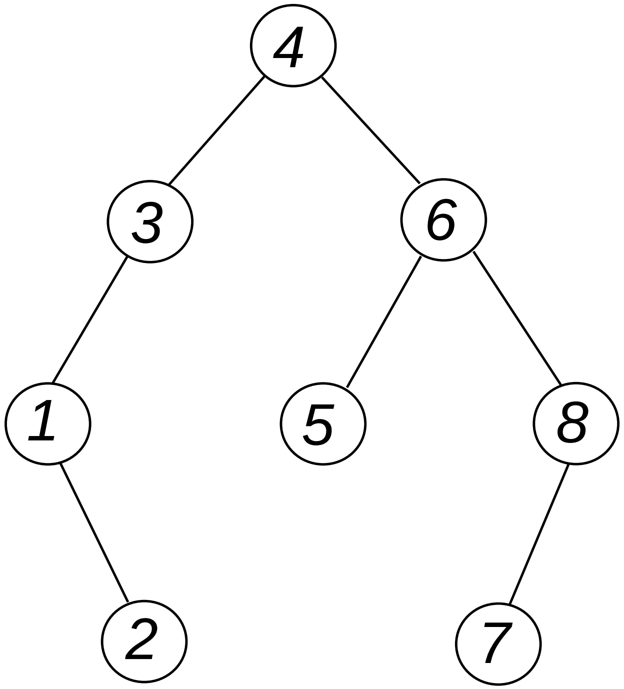
Zentraler Grenzwertsatz für die Tiefe

$$\boxed{\frac{D_n - 2\log n}{\sqrt{2\log n}} \to N(0,1)}$$

$$\mathbf{E}\, D_n = 2\log n + O(1), \qquad \mathbf{Var}\, D_n = 2\log n + O(1).$$

Profil

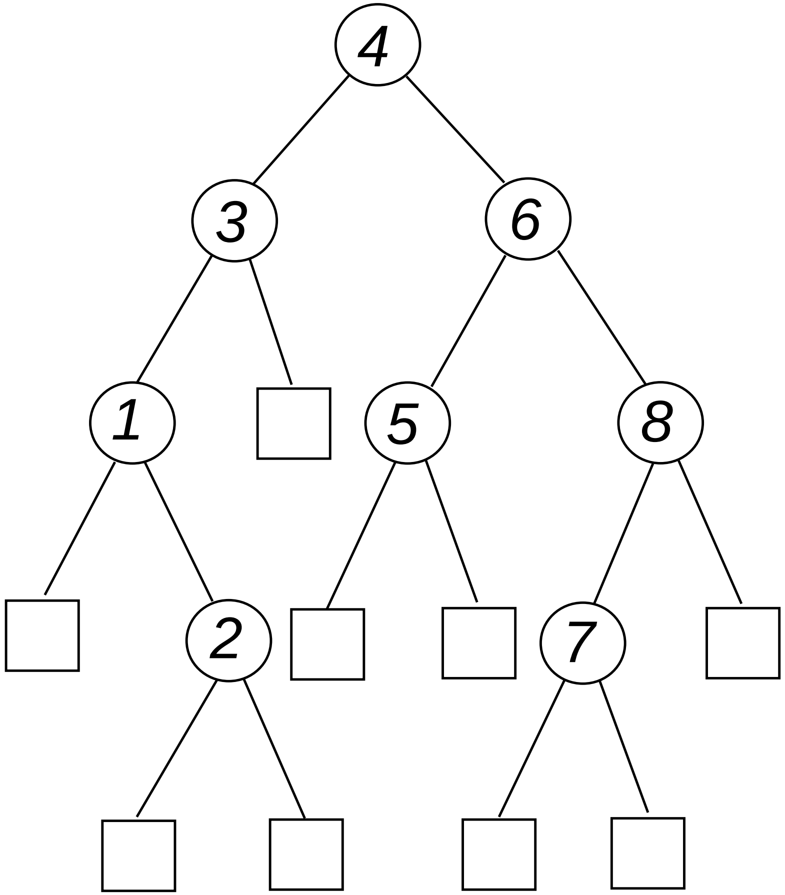
Das externe Profil:



Einfügen freier Plätze

Profil

Das externe Profil:



□ ... freie Plätze

Profil

Das externe Profil:

$Y_{n,k}$ = Anzahl der freien (= externen) Knoten der Tiefe k.

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

Profil

Das Profilpolynom

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

Lemma. Die normierten Profilpolynome

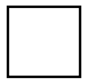
$$M_n(z) = \frac{W_n(z)}{\mathbf{E}\, W_n(z)}$$

bilden bezüglich der natürlichen Filtrierung (die von der Folge der Bäume $(T_n)_{n\geq 0}$ induziert wird) ein **Martingal**.

Bemerkung. $\mathbf{E}\, W_n(z) = \binom{2z+n-1}{n}$

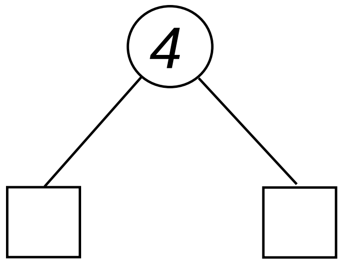
Binäre Suchbäume

Wachstumsprozess



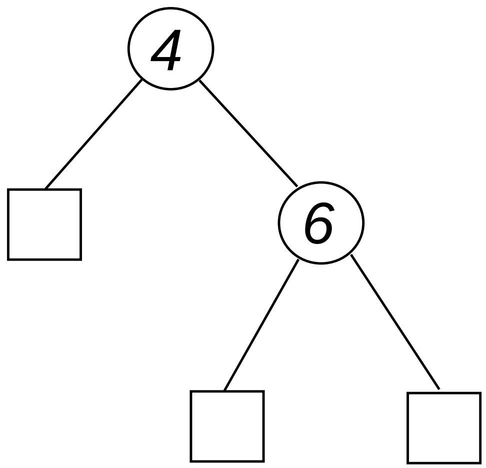
Binäre Suchbäume

Wachstumsprozess



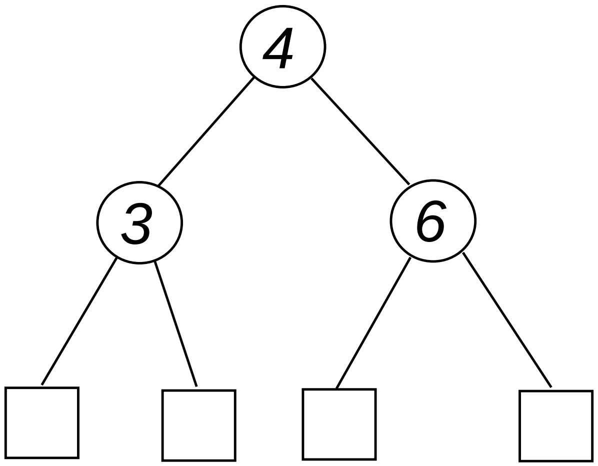
Binäre Suchbäume

Wachstumsprozess



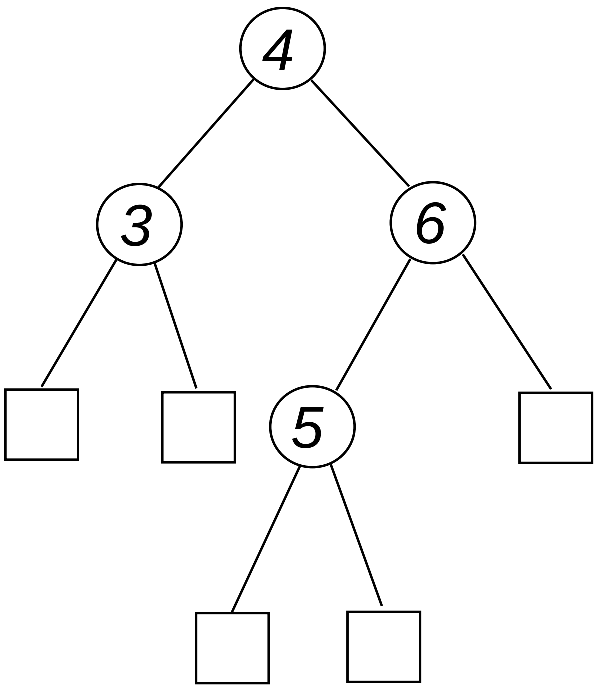
Binäre Suchbäume

Wachstumsprozess



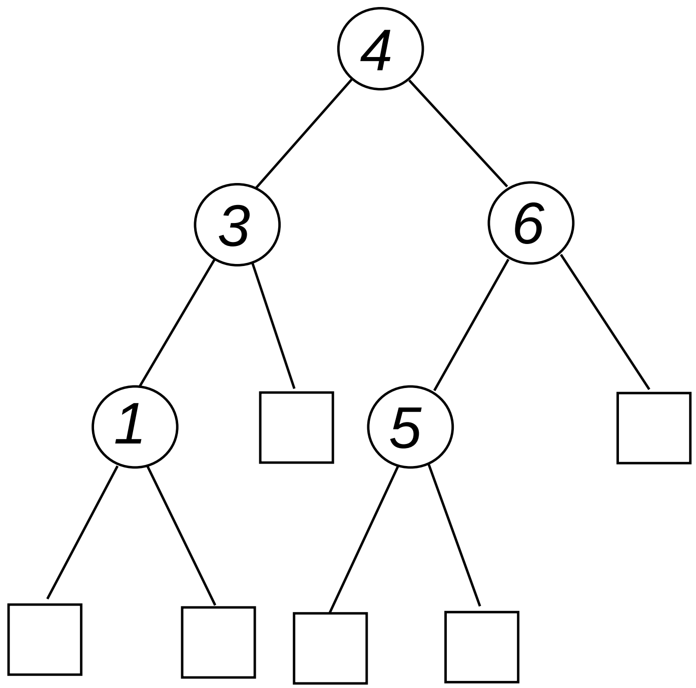
Binäre Suchbäume

Wachstumsprozess



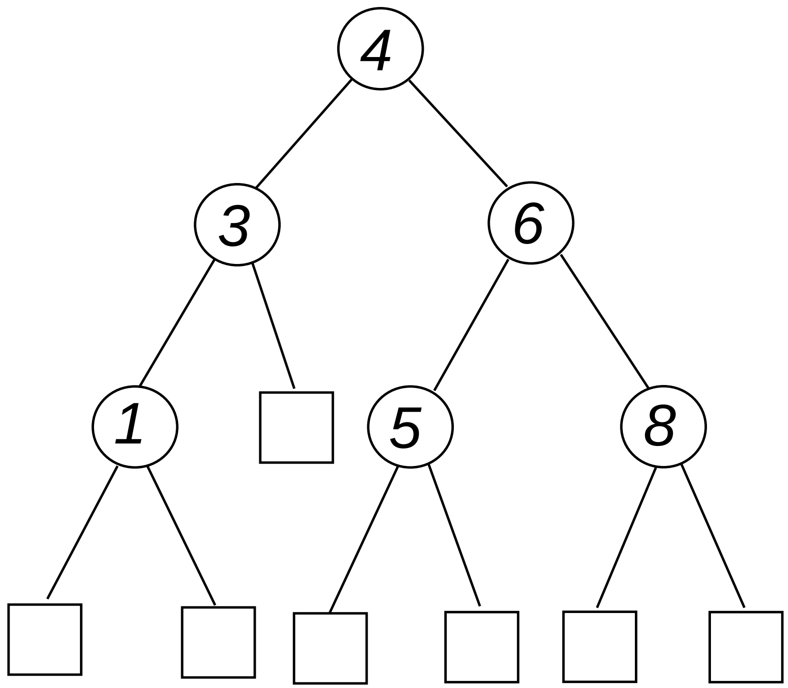
Binäre Suchbäume

Wachstumsprozess



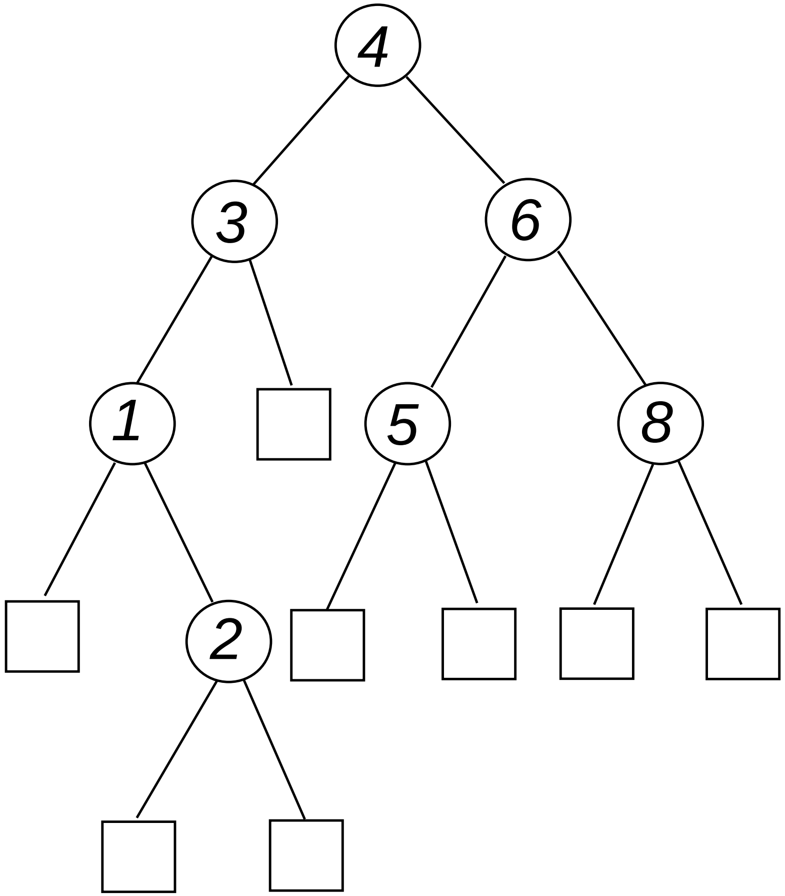
Binäre Suchbäume

Wachstumsprozess



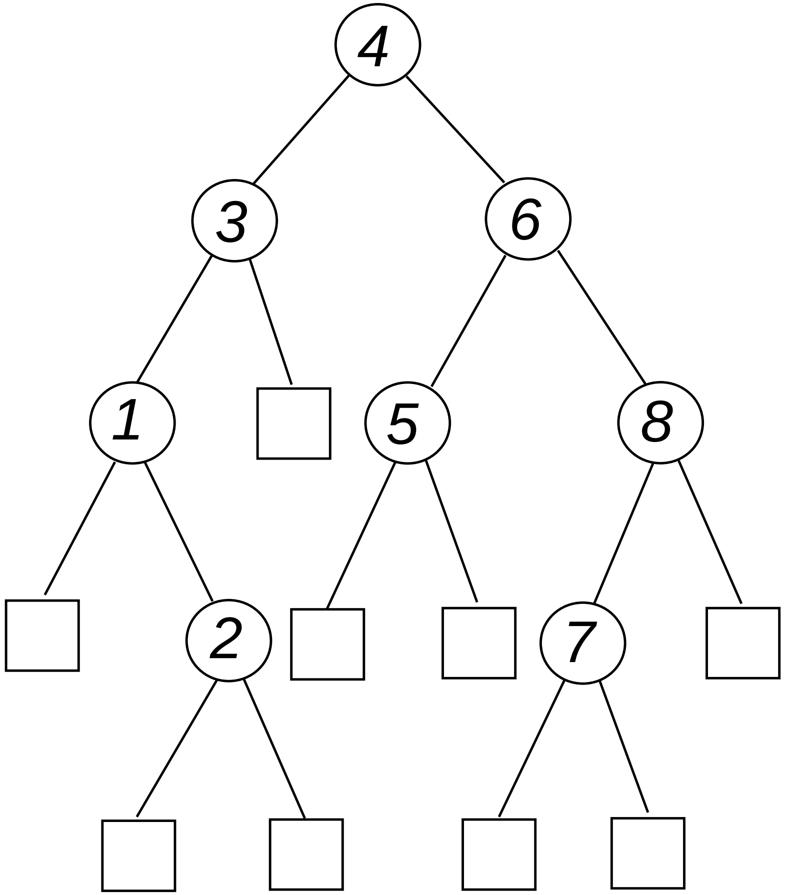
Binäre Suchbäume

Wachstumsprozess



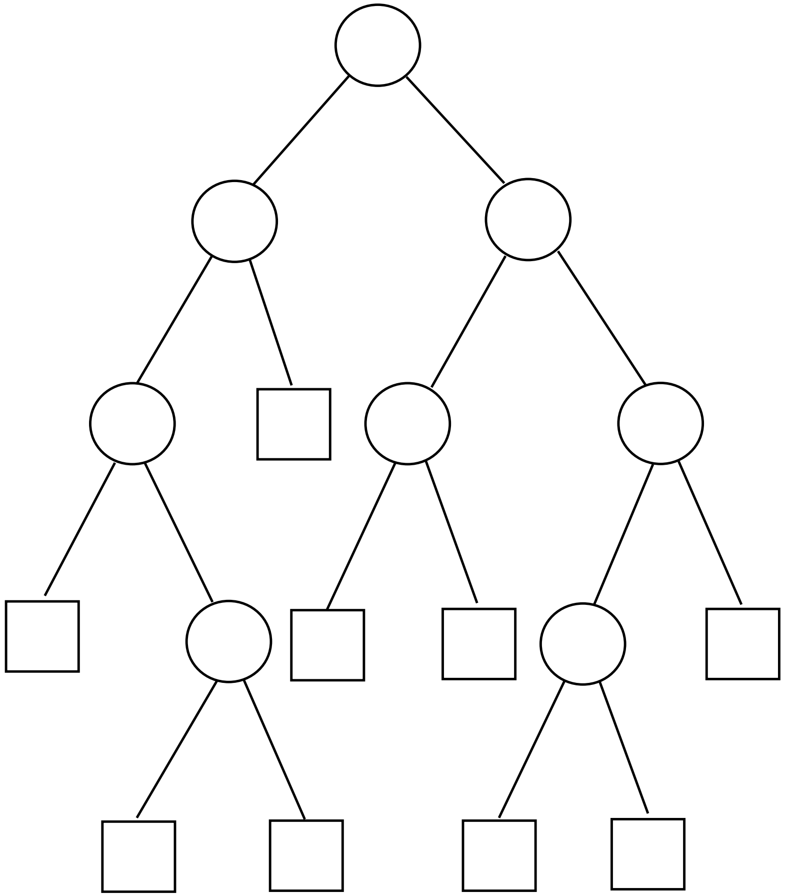
Binäre Suchbäume

Wachstumsprozess



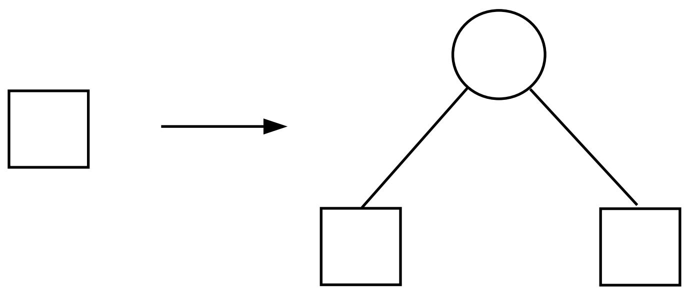
Binäre Suchbäume

Wachstumsprozess



Binäre Suchbäume

Wachstumsprozess



Profil

Grenzprozess des Profils [Chauvin + D. + Jabbour-Hattab]

Satz 1

$$\left(\frac{W_n(z)}{\mathbf{E}\,W_n(z)}, z \in B\right) \to (M(z), z \in B)$$

für einen geeigneten Bereich $B \subseteq \mathbb{C}$.

Satz 2

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

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

Profil

Bemerkungen:

- $(M(z), z \in B)$ stochastischer Prozess von **zufälligen analytischen Funktionen**.

- Fixpunktgleichung:

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

- **Interne Pfadlänge**

$$\boxed{M_n'(1) = \frac{I_n - \mathbf{E}\, I_n}{n+1} \to M'(1)}$$

$M'(1)$ ... **Quicksortverteilung**, $\mathbf{E}\, I_n \sim 2n \log n$.

Profil

Die "Quicksortgleichung":

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

$$\alpha > 0,\ \Phi(0) > 0$$

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

$$\boxed{\Psi(y/\alpha) = \frac{1}{y}\int_0^y \Psi(w)\Psi(y-w)\,dw}$$

Profil

Die “Quicksortgleichung”:

- Für $\alpha = z^{1/(2z-1)}$ gilt

$$\boxed{\mathbf{E}\, e^{-\frac{y^{2z-1}}{\Gamma(2z)} M(z)} = \Psi(y)}$$

- Genaue Eigenschaften der Verteilung von $M(z)$ sind unbekannt.

- Dieselbe “Quicksortgleichung” tritt auch bei der **Analyse der Höhe** auf.

Median of 3 – Variante

Grenzprozess des Profils [D. + Janson + Neininger]

Satz 1'

$$\left(\frac{W_n(z)}{\mathbf{E}\,W_n(z)}, z \in B\right) \to (M_3(z), z \in B)$$

für einen geeigneten Bereich $B \subseteq \mathbb{C}$.

Satz 2' $\beta(\alpha) := (2\alpha^2 + \sqrt{4\alpha^4 + \alpha^2})/12$.

$$\left(\frac{Y_{n,\lfloor \alpha \log n \rfloor}}{\mathbf{E}\,Y_{n,\lfloor \alpha \log n \rfloor}}, \alpha \in I\right) \to (M_3(\beta(\alpha)), \alpha \in I)\,.$$

$$\left(\frac{X_{n,\lfloor \alpha \log n \rfloor}}{\mathbf{E}\,X_{n,\lfloor \alpha \log n \rfloor}}, \alpha \in I\right) \to (M_3((\beta(\alpha)), \alpha \in I)\,.$$

Median of 3 − Variante

Bemerkungen.

- **Keine Martingalstruktur !!**

- Stochastische Fixpunktgleichung in Funktionenräumen analytischer Funktionen:

$$\boxed{M_3(z) \equiv zV^{\lambda(z)-1}M_3^{(1)}(z) + z(1-V)^{\lambda(z)-1}M_3^{(2)}(z) \quad (z \in B)}$$

  mit $\lambda(z) = (\sqrt{1+48z} - 3)/2$.
  V hat Dichte $v(x) = 6x(1-x)$ auf $[0,1]$.

- "Median-of-3-Gleichung"

$$\boxed{\Phi'''(u) = \frac{6}{\alpha^2}\Big(\Phi'(u/\alpha)\Big)^2}$$

Höhe

Satz 3. [Pittel, Robson, Devroye, Reed, D.]

$$\boxed{\mathbf{E}\, H_n = c \log n - \frac{3c}{2(c-1)} \log\log n + O(1)}$$

$c = 4.31107\ldots$ erfüllt die Gleichung $c \log\left(\frac{2e}{c}\right) = 1$.

$$\boxed{\mathbf{Var}\, H_n = O(1)}$$

Höhe

Satz 4. [D.]

$$\boxed{\mathbf{Pr}\{H_n \le h_n + r\} = W(r) + o(1)}$$

$W(x) = \Psi_\alpha(e^{-x/c})$ für $\alpha = e^{1/c} = 1.26107\ldots$,

$$\Psi(y/\alpha) = \frac{1}{y}\int_0^y \Psi(w)\Psi(y-w)\,dw$$

(h_n ist eine Folge mit $h_n = \mathbf{E}\,H_n + O(1)$.)

Höhe

Erzeugende Funktionen

$y_k(x) = \sum_{n\ge 0} \mathbf{Pr}\{H_n \le k\}\, x^n$:

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

mit $y_0(x) = 1$, $y_k(0) = 1$.

Ersatzfunktionen: $\boxed{\tilde{y}_k(x) = \alpha^k \Phi(\alpha^k(1-x))}$

(mit $\Phi(u) = \int_0^\infty \Psi(y) e^{-uy}\, dy$ und $\alpha = e^{1/c}$)

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

Profil

Erzeugende Funktionen

$Y_k(x,u) = \sum_{n\ge 0} \mathbf{Pr}\{Y_{n,k} = \ell\}\, x^n u^\ell$:

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

Aus diesem Ansatz scheint es nicht möglich zu sein, die Verteilung von $Y_{n,k}$ (ausser mittels Momentenmethode) zu bekommen, insbesondere keinen funktionalen Grenzwertsatz.

Höhe

Diskrete "Branching Random Walks"

Zufälliges Punktmaß:

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

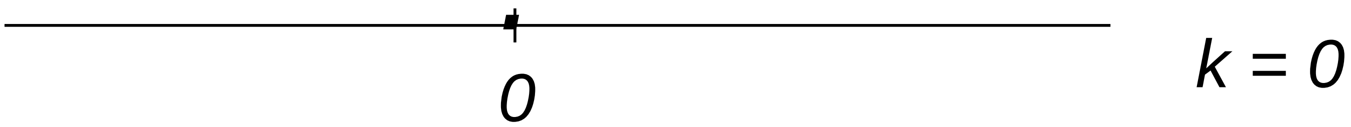
z.B.: $X_1 = \log(1/U)$, $X_2 = \log(1/(1-U))$.

"Branching Random Walk": Folge Z_k von zufälligen Punktmaßen:

- $Z_0 = \delta_0$.
- Z_{k+1} entsteht aus Z_k, indem jedes "Teilchen" von Z_k unabhängig voneinander gemäß der Verteilung (additiv) aufgeteilt wird.

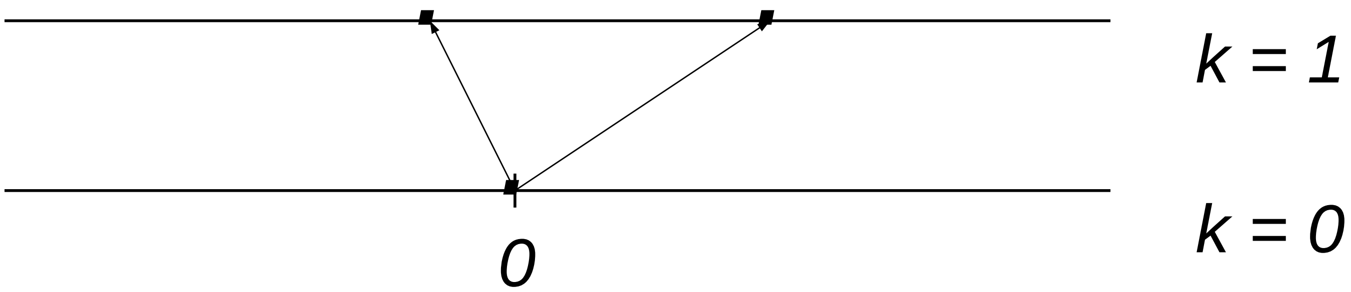
Höhe

Diskrete “Branching Random Walks”



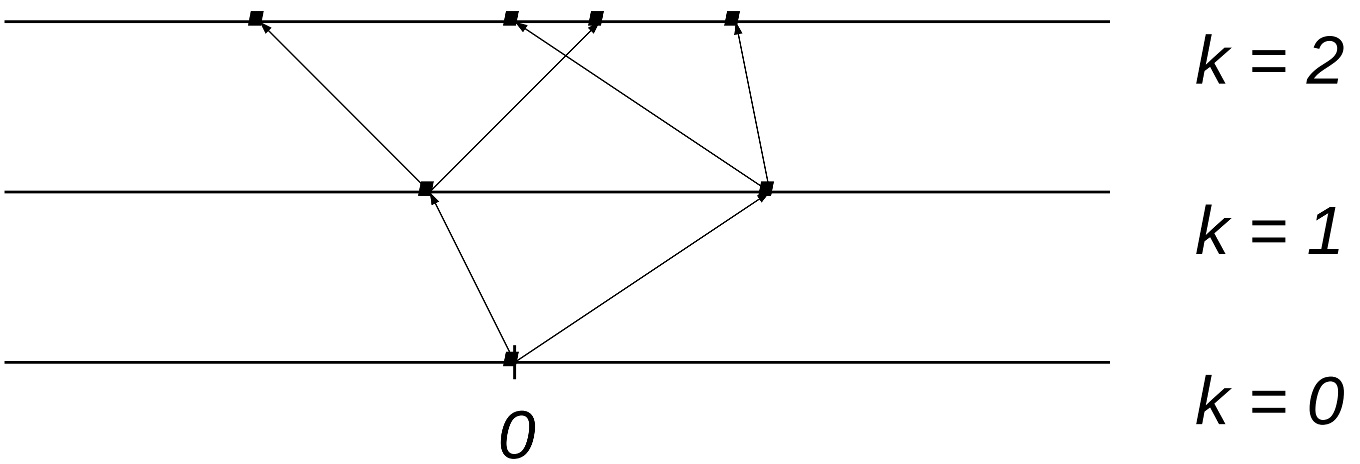
Höhe

Diskrete “Branching Random Walks”



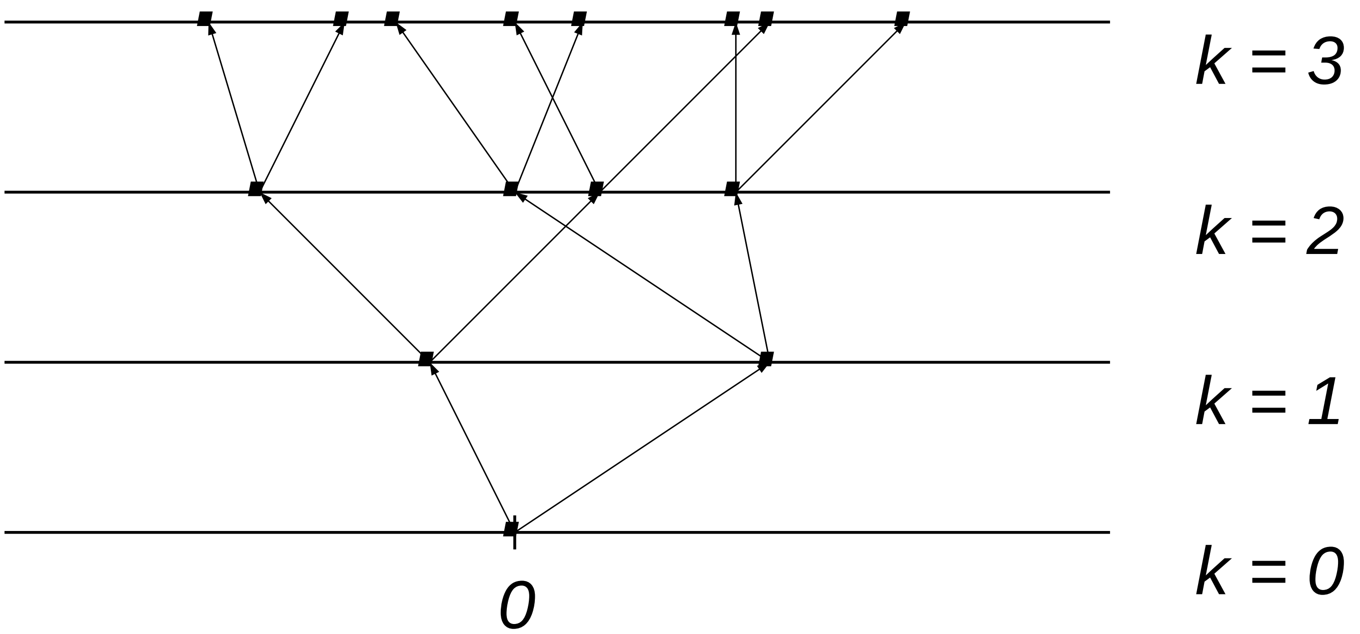
Höhe

Diskrete “Branching Random Walks”



Höhe

Diskrete "Branching Random Walks"



Höhe

Diskrete "Branching Random Walks"

L_k ... Position des am weitesten **links** liegenden Teilchens
(nach k Schritten)

R_k ... Position des am weitesten **rechts** liegenden Teilchens
(nach k Schritten)

$c = 4.31107\ldots$ und $c' = 0.373\ldots$ seien die 2 Lösungen der Gleichung

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

Höhe

Diskrete "Branching Random Walks"

Satz 5. $Z = \delta_{\log(1/U)} + \delta_{\log(1/(1-U))}$.

$$\begin{aligned}
\mathbf{Pr}\{L_k > x\} &= w_1(x - m_1(k)) + o(1), \\
\mathbf{Pr}\{R_k \le x\} &= w_2(x - m_2(k)) + o(1).
\end{aligned}$$

$m_1(k)$ ... Median von L_k, $m_2(k)$ ... Median von R_k,

$$\begin{aligned}
w_1(x) &= \Psi_\alpha(e^x) \quad \text{mit } \alpha = e^{1/c}, \\
w_2(x) &= \Psi_{\alpha'}(e^x) \quad \text{mit } \alpha' = e^{1/c'}.
\end{aligned}$$

Bemerkung: Diskretes Analogon zur "branching Brownian motion" (KPP-Gleichung etc.)

Präfix-Codes

Beispiel. $X = \{x_1, x_2, x_3, x_4\}$,

$$\mathbf{Pr}\{x_1\} = \tfrac{1}{2},\ \mathbf{Pr}\{x_2\} = \tfrac{1}{4},\ \mathbf{Pr}\{x_3\} = \tfrac{1}{8},\ \mathbf{Pr}\{x_4\} = \tfrac{1}{8},$$

$\mathcal{A} = \{0, 1\}$ ... (Binär)-Alphabet

Code: $\boxed{c : X \to \mathcal{A}^+}$

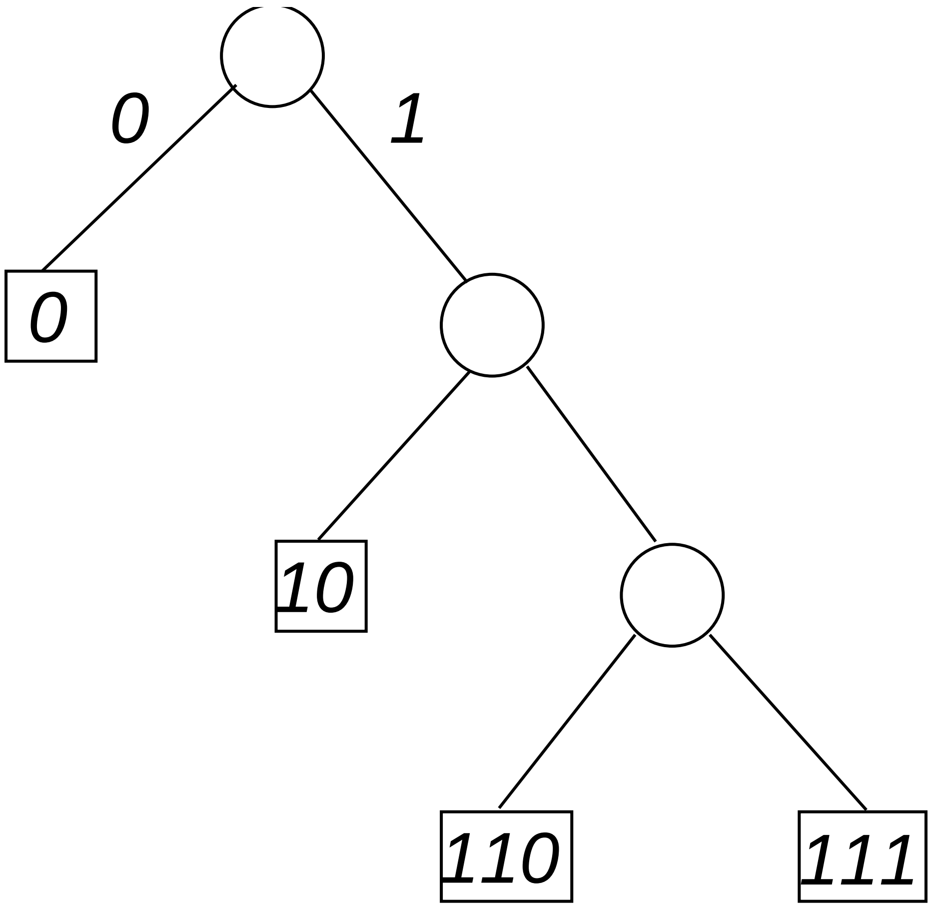
$$\begin{aligned} c(x_1) &= 0, \\ c(x_2) &= 10, \\ c(x_3) &= 110, \\ c(x_4) &= 111. \end{aligned}$$

Kein Codewort ist Präfix eines anderen (**Präfix-Code**).

$$c(x_1 x_1 x_4 x_2 x_3 x_1 \ldots) = 0|0|111|10|110|0|\ldots$$

Präfix-Codes

- Ein Präfix-Code entspricht den Blättern eines (Binär-)Baums



Präfix-Codes

- **Entropie:** $h_X = \sum_i \mathbf{Pr}\{x_i\} \log_2 (1/(\mathbf{Pr}\{x_i\}))$

- **Mittlere Codewortlänge:** $L = \sum_i \mathbf{Pr}\{x_i\} \cdot |c(x_i)|$

- **Redundanz:** $R = L - h_X$

$$\boxed{L \geq h_X} \Longrightarrow R \geq 0.$$

- **Kraftsche Ungleichung:**

$$\sum_i 2^{-|c(x_i)|} \leq 1.$$

Präfix-Codes

Shannon-Code

$X = \{x_1, x_2, \ldots, x_n\}$

$\ell_i = \lceil \log_2(1/(\mathbf{Pr}\{x_i\})) \rceil$ erfüellt die Kraftsche Ungleichung

$$\sum_{i=1}^{n} 2^{-\ell_i} \leq \sum_{i=1}^{n} 2^{-\log_2(1/(\mathbf{Pr}\{x_i\}))} = \sum_{i=1}^{n} \mathbf{Pr}\{x_i\} = 1.$$

Daher gibt es einen Code $c : X \to \mathcal{A}^+$ mit $|c(x_i)| = \ell_i$ und Redundanz

$$R = \sum_{i=1}^{n} \mathbf{Pr}\{x_i\} (\ell_i - \log_2(1/(\mathbf{Pr}\{x_i\}))) \leq \sum_{i=1}^{n} \mathbf{Pr}\{x_i\} = 1.$$

Der Unterschied zu einem optimalen Code ist maximal 1 Bit pro Quelltextzeichen.

Präfix-Codes

Verbesserungen

- $X \longrightarrow X^k$: Zusammenfassen von jeweils k Quelltextzeichen, darauf Shannon-Code $c_k : X^k \to \mathcal{A}^+$, $R_k \leq 1$

  Der Unterschied zu einem theoretisch optimalen Code ist maximal $1/k$ Bit pro Zeichen.

- Aufteilen des Quelltextes bezüglich einer präfixfreien Menge: z.B. $X = \{a, b\}$. $\mathcal{D} = \{a, ba, bba, bbb\}$.

  $$abaaabbbaabbbabababbba \ldots \longrightarrow a|ba|a|a|bbb|a|a|bbb|a|ba|ba|bbb|a| \ldots$$

  darauf Shannon-Code $c : \mathcal{D} \to \mathcal{A}^+$ (oder Varianten)

Präfix-Codes

Satz 6. [Khodak, D. + Szpankowski]

Quelle: Bernoulli oder Markoff-Prozess über einem endlichen Alphabet.

Es existieren präfixfreien Mengen $\mathcal{D}$ (über X) mit beliebig großer durchschnittlicher Länge $\boxed{D = \sum_{d \in \mathcal{D}} \mathbf{Pr}\{d\} \cdot |d|}$ und Prefix-Codes $c : \mathcal{D} \to \mathcal{A}^+$ mit Redundanz

$$R \le c' \, D^{-2/3}$$

(mit einer absoluten Konstante $c' > 0$). Weiters ist $\max_{d \in \mathcal{D}} |d| = O(D \log D)$.

Der Unterschied zu einem (theoretisch) optimalen Code ist daher maximal

$$\boxed{c' \cdot D^{-5/3} \text{ Bit pro Zeichen.}}$$

Präfix-Codes

Satz 7. [Bugeaud + D. + Szpankowski]

Für **fast alle** Bernoulli-Quellen über einem endlichen Alphabet X der Größe m gilt:

Es existieren präfixfreie Mengen $\mathcal{D}$ (über X) mit beliebig großer durchschnittlicher Länge D und Prefix-Codes $c : \mathcal{D} \to \mathcal{A}^+$ mit Redundanz

$$R \le D^{-(m+1)/3+\varepsilon}$$

(für beliebiges $\varepsilon > 0$ und $D \ge D_0(\varepsilon)$). Weiters ist $\max_{d\in\mathcal{D}} |d| = O(D \log D)$.

Der Unterschied zu einem theoretisch optimalen Code ist daher maximal

$$\boxed{D^{-(m+4)/3+\varepsilon} \text{ Bit pro Zeichen.}}$$

Umgekehrt gilt für alle präfixfreien Mengen $\mathcal{D}$

$$R \ge D^{-2m-\varepsilon}.$$

Diophantische Approximation

Linearformen:

$X = \{x_1, x_2, \ldots, x_m\}$, $p_i = \mathbf{Pr}\{x_i\}$.

$$\mathbf{Pr}\{x_{i_1} x_{i_2} \cdots x_{i_k}\} = p_{i_1} p_{i_2} \cdots p_{i_k} = p_1^{k_1} p_2^{k_2} \cdots p_m^{k_m}$$

$\boxed{\text{Redundanz ist klein}}$

$$\iff \quad p_1^{k_1} p_2^{k_2} \cdots p_n^{k_m} \approx 2^{-\ell} \text{ für alle } d \in \mathcal{D} \text{ (Blätter)}$$

$$\iff \quad k_1 \log_2 p_1 + k_2 \log_2 p_2 + \cdots k_m \log_2 p_m \approx -\ell \in$$

$$\iff \quad \boxed{k_1 \log_2 p_1 + k_2 \log_2 p_2 + \cdots k_m \log_2 p_m + \ell \approx 0}.$$

Diophantische Approximation

Problem:

Gegeben: $p_1 > 0, \ldots, p_m > 0$ mit $p_1 + \cdots + p_m = 1$.

Gesucht: m-ärer endlicher Baum mit

$$k_1 \log_2 p_1 + k_2 \log_2 p_2 + \cdots + k_m \log_2 p_m$$ **fast ganzzahig für alle Blätter.**

$(k_1, \ldots, k_m)$ ist der jeweilige *Typ* des Blatts.

Diophantische Approximation

Dispersion

$S \subset [0, 1)$:

$$\boxed{\delta(S) := \sup_{0 \le y < 1} \inf_{x \in S} |y - x|}$$

Maß für die Dichte von S in $[0, 1)$.

Diophantische Approximation

Satz 8.

$p_1 > 0, \ldots, p_m > 0$ mit $p_1 + \cdots + p_m = 1$ seien gegeben, die Menge

$$S = \{k_1 \log_2 p_1 + \cdots + k_m \log_2 p_m \bmod 1 : 0 \le k_j < N \ (1 \le j \le m)\},$$

habe Dispersion

$$\boxed{\delta(S) \le \frac{1}{N^\eta}.}$$

Dann existiert eine präfixfreien Menge $\mathcal{D}$ mit durchschnittlicher Länge $D \approx N^3$ und ein Prefix-Code $c : \mathcal{D} \to \mathcal{A}^+$ mit Redundanz

$$R \le c' \cdot D^{-(\eta+1)/3}$$

Der Unterschied zu einem optimalen Code ist daher maximal

$$\boxed{c' \cdot D^{-(\eta+4)/3} \text{ Bit pro Zeichen.}}$$

Diophantische Approximation

Lemma

Für fast alle $p_1 > 0, \ldots, p_m > 0$ mit $p_1 + \cdots + p_m = 1$ hat die Menge

$$S = \Big\{ k_1 \log_2 p_1 + \cdots + k_m \log_2 p_m \bmod 1 : 0 \le k_j < N \ (1 \le j \le m) \Big\}$$

Dispersion

$$\delta(S) \le \frac{1}{N^{m-\varepsilon}}$$

für $\varepsilon > 0$ und $N \ge N_0(\varepsilon)$.

Beweismethode: Metrische Diophantische Approximation auf Mannigfaltigkeiten.

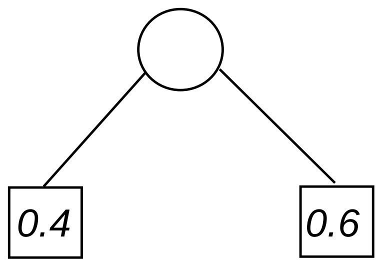
Tunstall-Codes

$p_1 = 0.4$, $p_2 = 0.6$.

$\boxed{1}$

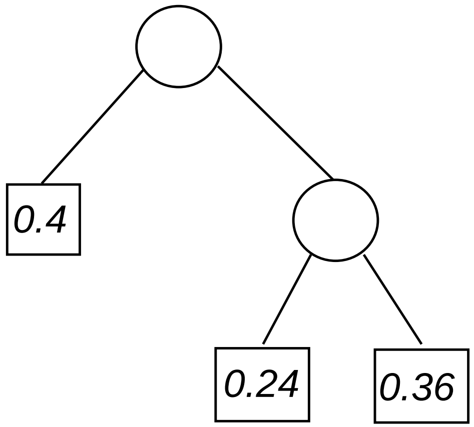
Tunstall-Codes

$p_1 = 0.4$, $p_2 = 0.6$.



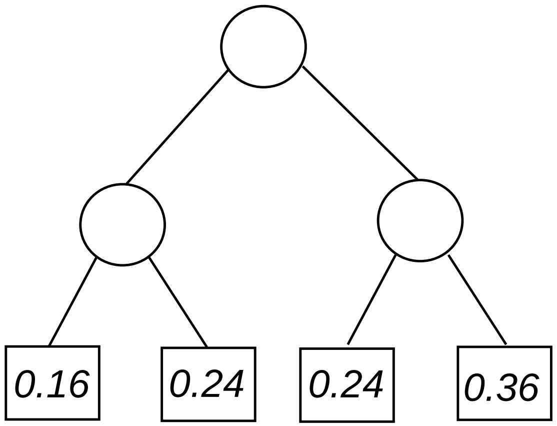
Tunstall-Codes

$p_1 = 0.4$, $p_2 = 0.6$.



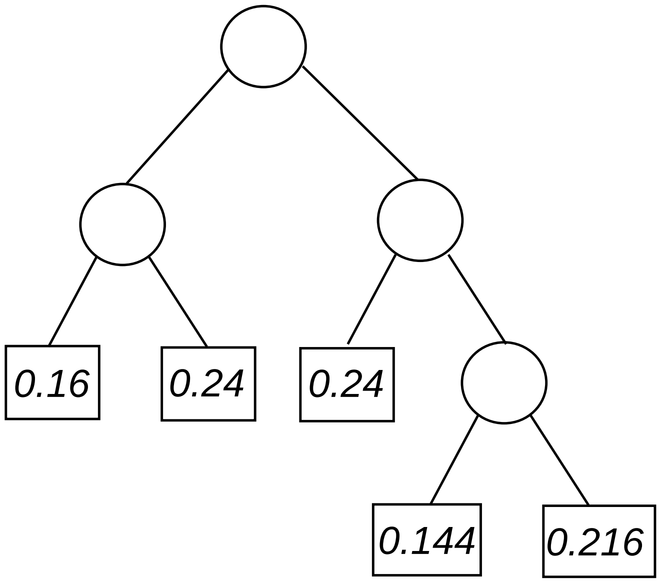
Tunstall-Codes

$p_1 = 0.4$, $p_2 = 0.6$.



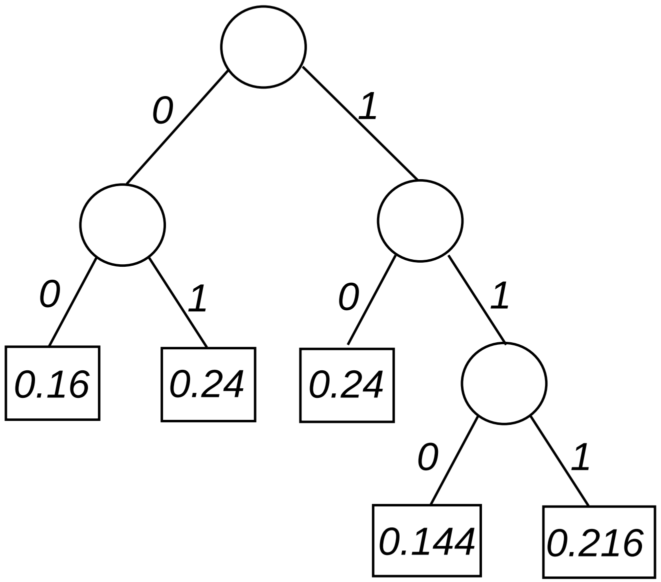
Tunstall-Codes

$p_1 = 0.4$, $p_2 = 0.6$.



Tunstall-Codes

$p_1 = 0.4$, $p_2 = 0.6$. $\mathcal{D}_5 = \{00, 01, 10, 110, 111\}$.



Tunstall-Codes

Satz 9. [D. + Reznik + Savari + Szpankowski]

$\mathcal{D}_M$ Tunstall-Code der Größe M

Ist $\log p_1/\log p_2$ $\boxed{\text{irrational}}$, so gilt

$$R(\mathcal{D}_M) = \frac{H}{\log M}\left(-\frac{H_2}{2H} - \log H\right) + o\left(\frac{1}{\log M}\right).$$

Ist $\log p_1/\log p_2$ $\boxed{\text{rational}}$ und $\Lambda > 0$ die größte reelle Zahl, sodass $\log(1/p_1)$ and $\log(1/p_2)$ ganzzahlige Vielfache von Λ sind, so gilt

$$R(\mathcal{D}_M) = \frac{H}{\log M}\left(-\frac{H_2}{2H} - \log H + \log \Lambda - \log(e^\Lambda - 1) + \frac{\Lambda}{2}\right) + O\left((\log M)^{-2}\right).$$

Tunstall-Codes

Satz 10. [D. + Reznik + Savari + Szpankowski]

$\mathcal{D}_M$ Tunstall-Code der Größe M, $p_1 \ne p_2$.

D_M Länge eines Wortes von $\mathcal{D}_M$ (gemäß der von p_1, p_2 auf $\mathcal{D}_M$ induzierten Wahrscheinlichkeitsverteilung). Dann gilt

$$\boxed{\frac{D_M - \frac{1}{H}\log M}{\left(\left(\frac{H_2}{H^3} - \frac{1}{H}\right)\log M\right)^{1/2}} \to N(0,1).}$$

$(H = p_1 \log(1/p_1) + p_2 \log(1/p_2),\ H_2 = p_1 \log^2(1/p_1) + p_2 \log^2(1/p_2))$

$$\mathbf{E}\, D_M = \frac{\log M}{H} + \frac{\log H}{H} + \frac{H_2}{2H^2} + \frac{-\log \Lambda + \log(1 - e^{-\Lambda}) + \frac{\Lambda}{2}}{H} + O((\log M)^{-1}),$$

$$\mathbf{Var}\, D_M = \left(\frac{H_2}{H^3} - \frac{1}{H}\right)\log M + O(1).$$

Tunstall-Codes

Inverse Mellin-Transformation

$$
\begin{aligned}
D_M(z) &= \sum_{d\in\mathcal{D}_M} \mathbf{Pr}\{d\}\cdot z^{|d|} \\
&= \frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty} \left(\frac{1-z}{s(1-zp_1^{1-s}-zp_2^{1-s})} - \frac{1}{s}\right) v^{-s}\, ds \\
&= v^{\frac{z-1}{H}-\left(\frac{1}{H}-\frac{H_2}{2H^3}\right)(z-1)^2+O(|z-1|^3)} \left(1+O\left(|z-1|^{1/2}\right)\right)
\end{aligned}
$$

$(v = v(M) \sim H\, M)$

Beweismethode: Quantifizierte Version des Taubersatzes von Wiener-Ikehara.

Die Nullstellen von $1-zp_1^{1-s}-zp_2^{1-s}=0$, die die *diophantische Struktur* von $\log p_1/\log p_2$ widerspiegeln, gehen in den Fehlerterm ein.