Distributional Analysis of Recursive Algorithms by the Contraction Method

Ralph Neininger
Universität Freiburg

INRIA–Rocquencourt, November 1999

Contents

Partial Match Query

retrieve multiattribute records belonging to some k-dimensional domain

$$D = D_1 \times \ldots \times D_k$$

Given:

file $F \subset D$

query $q = (q_1, \ldots, q_k)$,

$$q \in (D_1 \cup \{*\}) \times \ldots \times (D_k \cup \{*\})$$

Problem:

Find all records in the file F which *satisfy* the query q, i.e. find all $r = (r_1, \ldots, r_k) \in F$ with

$$r_j = q_j \quad \text{for all} 1 \leq j \leq k \quad \text{with } q_j \neq *.$$

The *specification pattern* of a query q is the word $u = (u_1, \ldots, u_k) \in \{S, *\}^k$ with

$$u_j = S \quad \text{if} \quad q_j \in D_j \quad \text{(specified)}$$
$$u_j = * \quad \text{if} \quad q_j = * \quad \text{(unspecified)}$$

The model

For the attributes' domains assume

$$D_i = [0, 1] \quad \text{for all} \quad 1 \leq i \leq k.$$

The uniform probabilistic model assumes all attributes (=components, coordinates) in the files and queries to be independent and uniformly distributed over the unit intervall.

Appropriate for comparison based algorithms.

Comparison based structures

Quadtree (Finkel, Bentley '74)

K-d trees

- "classical" K-d tree (Bentley '75)

- K-d-t tree (locally balanced version)
. (Cunto, Lau, Flajolet '89)

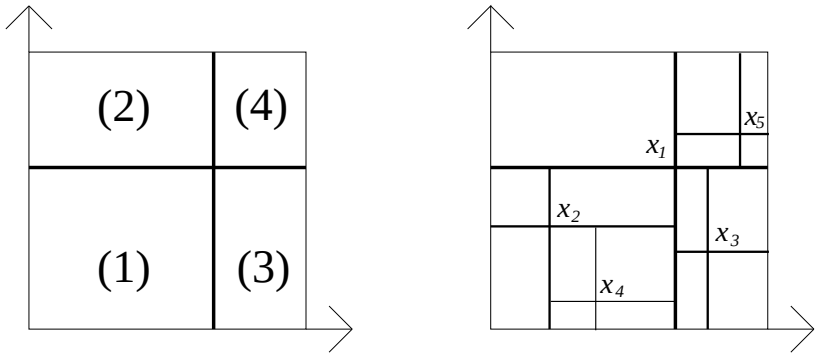
- random relaxed K-d tree (randomized version)
. (Duch, Estivill-Castro, Martínez '98)

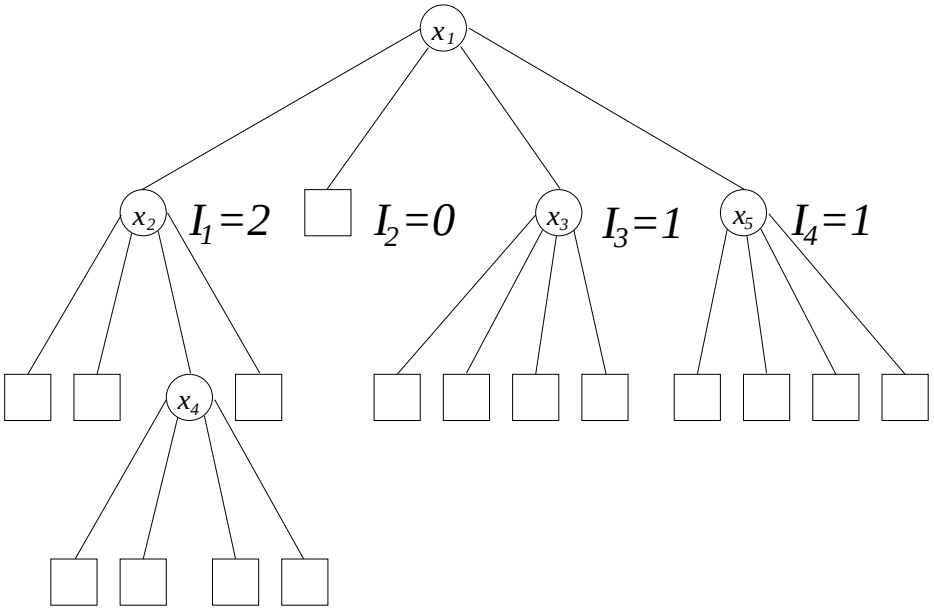
- squarish K-d tree
. (Devroye, Jabbour, Zamora-Cura '99)

A digital structure: The 2-d trie
. (Schachinger '99)

The Quadtree

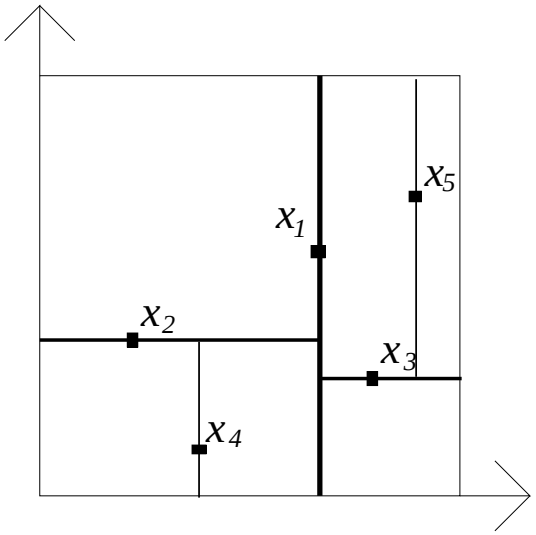
The data partition the unit-cube recursively into quadrants. The quadtree corresponds to this partitioning.

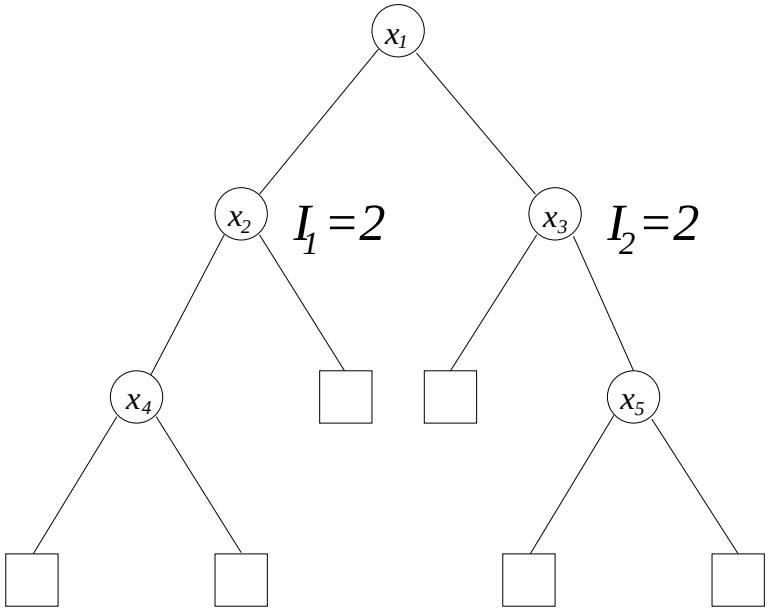




The K-d-Tree

The components of the data are used cyclically as discriminators.





Partial match query in a quadtree

Dimension d=2

$W = (U, V)$ the first key to be inserted (the root)

$q = (Y, *)$ the query

Y, U, V independent, uniformly distr. on $[0, 1]$

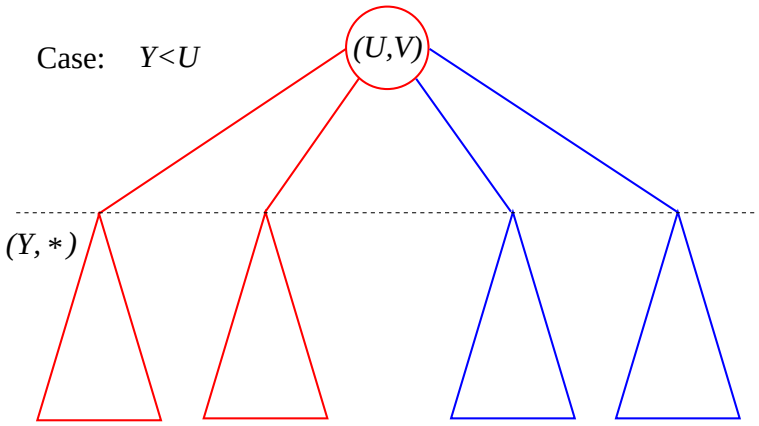
$I^{(n)}$ vector of the cardinalities of the subtrees of the root

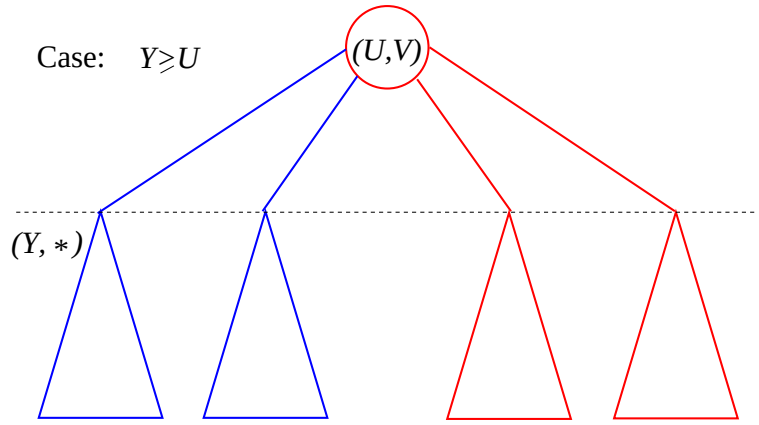
$\langle W \rangle = (UV, U(1-V), (1-U)V, (1-U)(1-V))$ volumes of the generated quadrants

C_n cost of a partial match query in a tree of size n, measured as the number of nodes traversed during the search

Search for query $(Y, *)$

Search for $(Y,*)$ with root (U,V)





Distributional recursive equation for the cost C_n

$$C_n \stackrel{\mathcal{D}}{=} \mathbf{1}_{\{Y<U\}} \left(C^{(1)}_{I_1^{(n)}} + C^{(2)}_{I_2^{(n)}} \right) + \mathbf{1}_{\{Y\geq U\}} \left(C^{(3)}_{I_3^{(n)}} + C^{(4)}_{I_4^{(n)}} \right) + 1$$

with

$Y, U, V, (C_i^{(1)})_{i\in\mathbb{N}}, \ldots, (C_i^{(4)})_{i\in\mathbb{N}}$ independent

Y, U, V uniformly distributed on $[0, 1]$

$C_i^{(k)} \stackrel{\mathcal{D}}{=} C_i, \quad 1 \leq k \leq 4$

$I^{(n)}$ given $(U, V) = w$ multinomial $M(n-1, \langle w \rangle)$ distributed

Moments of the cost C_n

$$\mathbb{E}C_n \sim \gamma n^{\alpha-1},$$

$$\alpha = \frac{\sqrt{17}-1}{2}, \quad \gamma = \frac{\Gamma(2\alpha)}{2\Gamma^3(\alpha)}$$

(Flajolet, Gonnet, Puech, Robson '93)

$$\mathrm{Var}(C_n) \sim \beta n^{2\alpha-2},$$

$$\beta = \frac{(2\alpha-1)\Gamma(2\alpha)}{3\alpha(\alpha-1)\Gamma^4(\alpha)} - \frac{\Gamma^2(2\alpha)}{4\Gamma^6(\alpha)}$$

(Martínez, Panholzer, Prodinger '98)

Normalization:

$$X_n = \frac{C_n - \mathbb{E}C_n}{n^{\alpha-1}}$$

The Limiting equation

$$X_n \stackrel{\mathcal{D}}{=} \mathbf{1}_{\{Y<U\}} \sum_{k=1}^{2} \left(\frac{I_k^{(n)}}{n}\right)^{\alpha-1} \left(X^{(k)}_{I_k^{(n)}} + \gamma\right)$$
$$+\mathbf{1}_{\{Y\geq U\}} \sum_{k=3}^{4} \left(\frac{I_k^{(n)}}{n}\right)^{\alpha-1} \left(X^{(k)}_{I_k^{(n)}} + \gamma\right)$$
$$-\gamma + o(1)$$

Recall

$$\frac{I^{(n)}}{n} \xrightarrow{\mathbb{P}} \langle W \rangle$$
$$= (UV, U(1-V), (1-U)V, (1-U)(1-V))$$

Limiting equation:

$$X \stackrel{\mathcal{D}}{=} \mathbf{1}_{\{Y<U\}} \sum_{k=1}^{2} \langle W \rangle_k^{\alpha-1} (X^{(k)} + \gamma)$$
$$+\mathbf{1}_{\{Y\geq U\}} \sum_{k=3}^{4} \langle W \rangle_k^{\alpha-1} (X^{(k)} + \gamma) - \gamma$$

The limiting operator

$$T : M^1(\mathbb{R}^1, \mathcal{B}^1) \longrightarrow M^1(\mathbb{R}^1, \mathcal{B}^1)$$

$$T(\mu) \stackrel{\mathcal{D}}{=} \mathbf{1}_{\{Y<U\}} \sum_{k=1}^{2} \langle W \rangle_k^{\alpha-1} (Z^{(k)} + \gamma)$$
$$+\mathbf{1}_{\{Y \geq U\}} \sum_{k=3}^{4} \langle W \rangle_k^{\alpha-1} (Z^{(k)} + \gamma) - \gamma$$

$Y, U, V, Z^{(1)}, \ldots, Z^{(4)}$ are independent

$W = (U, V)$

Y, U, V uniformly distr. on $[0, 1]$

$Z^{(1)}, \ldots, Z^{(4)} \stackrel{\mathcal{D}}{=} \mu$

The Wasserstein-metric

$$l_2(\mu,\nu) := \inf\{(\mathbb{E}(X-Y)^2)^{1/2} : X \stackrel{\mathcal{D}}{=} \mu, Y \stackrel{\mathcal{D}}{=} \nu\}$$

$$M_{0,2} := \{\mu \in M^1(\mathbb{R}^1, \mathcal{B}^1) : \mathbb{E}\mu = 0, \mathrm{Var}(\mu) < \infty\}$$

The infimum is attained for "optimal couplings".

$(M_{0,2}, l_2)$ is a complete metric space.

$$l_2(\mu_n, \mu) \to 0 \quad \Longleftrightarrow \quad \mu_n \stackrel{\mathcal{D}}{\to} \mu \text{ plus conv. of seconds moments}$$

Contraction property

Lemma: *The limiting operator* $T : M_{0,2} \to M_{0,2}$ *is a contraction w.r.t.* l_2*:*

$$l_2(T(\mu), T(\nu)) \leq \xi \, l_2(\mu, \nu) \quad \forall \, \mu, \nu \in M_{0,2},$$

$$\xi = \frac{2}{\sqrt{19 - 3\sqrt{17}}} = 0.776\ldots$$

Banach's fixed point theorem:
T has a unique fixed point ρ in $M_{0,2}$ and

$$l_2(T^n(\mu), \rho) \to 0$$

exponentially fast for any $\mu \in M_{0,2}$.

Exponential rate $\longrightarrow$ quick approximation of the fixed point by iterating the limiting operator

The Limit Law

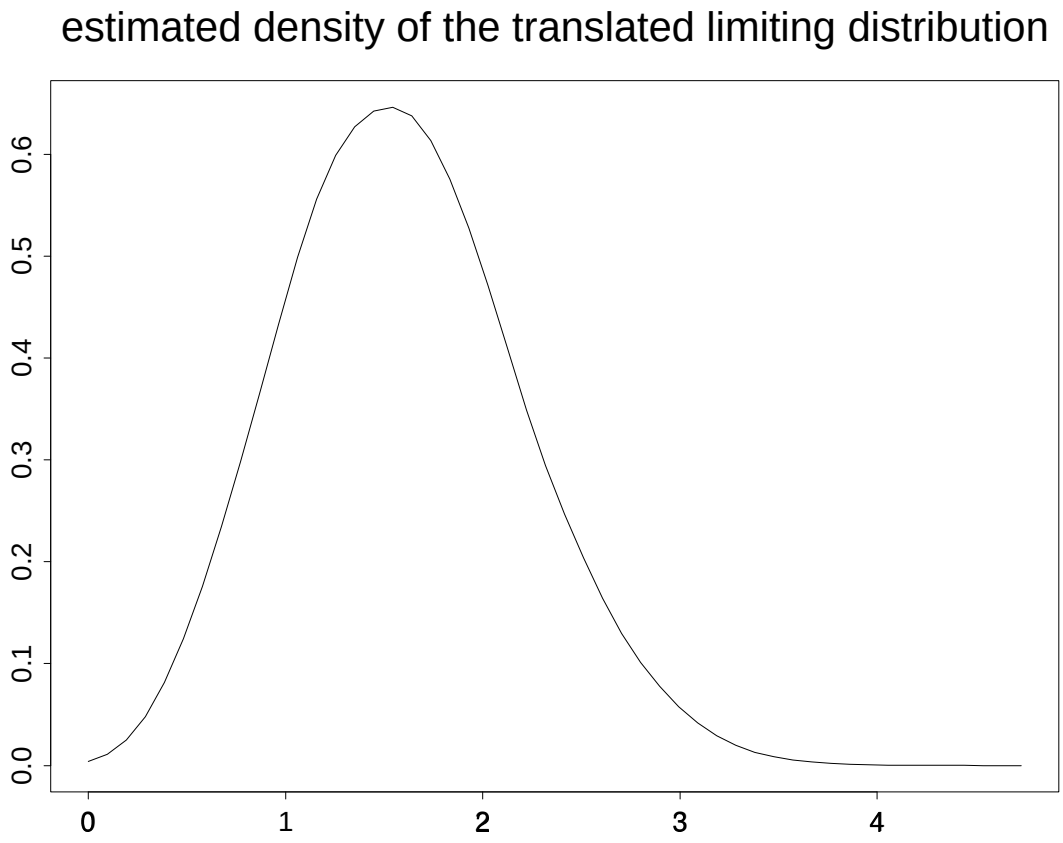
Theorem: (Limit law for PMQ in 2-dim. quadtrees) *Let X_n be the normalized number of traversed nodes and X the fixed point of the limiting operator (i.e. $T(\mathcal{L}(X)) = \mathcal{L}(X)$), then:*

$$l_2(X_n, X) \to 0.$$

The translated limit $\tilde{X} := X + \gamma$ is the unique solution in $M_{\gamma,2}$ of the equation

$$Z \stackrel{\mathcal{D}}{=} U^{\frac{\alpha-1}{2}} \left(V^{\alpha-1} Z^{(1)} + (1-V)^{\alpha-1} Z^{(2)} \right).$$

Approximation of the limiting distribution



Iterate 10 times

$$Z \stackrel{\mathcal{D}}{=} U^{\frac{\alpha-1}{2}} \left(V^{\alpha-1} Z^{(1)} + (1-V)^{\alpha-1} Z^{(2)} \right).$$

starting with δ_γ. From this distribution 15 000 samples were produced. Then a S-Plus smoothing-routine was applied to the histogramm of the data.

Multidimensional Quadtree

Dimension: d

$U = (U_1, \ldots, U_d)$ the root

$q = (Y_1, \ldots, Y_s, *, \ldots, *)$ the query

$1 \leq s \leq d - 1$ number of specified components

$\langle U \rangle = (\langle U \rangle_0, \ldots, \langle U \rangle_{2^d - 1})$
volumes of generated quadrants

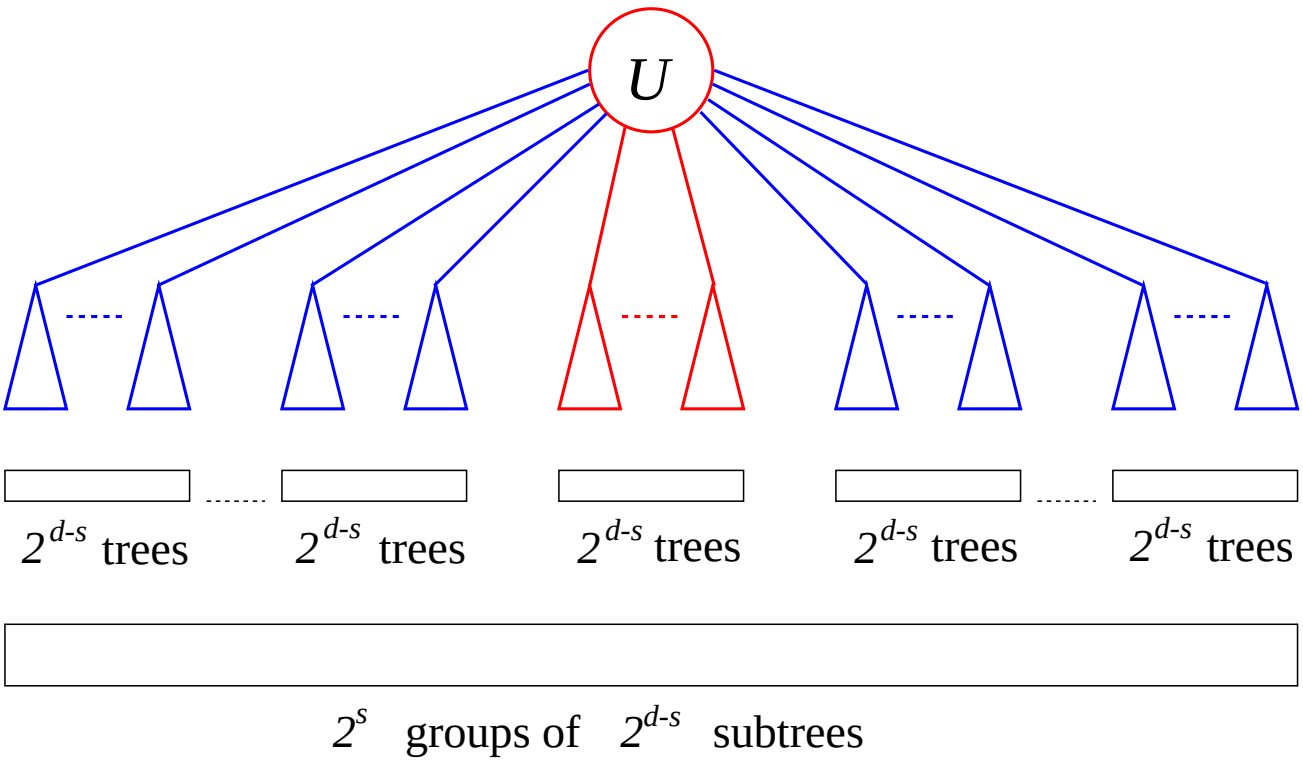
$I^{(n)}$ vector of the subtree-sizes

PMQ in the multidim. Quadtree

$U = (U_1, \ldots, U_d)$ the root

$q = (Y_1, \ldots, Y_s, *, \ldots, *)$ the query

Decision in s components



The cost C_n

$$\mathbf{1}_{j_1,\dots,j_s}(Y,U) := \prod_{\substack{1\le i\le s\\ j_i=0}} \mathbf{1}_{\{Y_i<U_i\}} \prod_{\substack{1\le i\le s\\ j_i=1}} \mathbf{1}_{\{Y_i\ge U_i\}}$$

$$(j_1,\dots,j_s\in\{0,1\})$$

Distributional equation for C_n:

$$C_n \stackrel{\mathcal{D}}{=} \sum_{j_1,\dots,j_s=0,1} \mathbf{1}_{j_1,\dots,j_s}(Y,U) \sum_{j_{s+1},\dots,j_d=0,1} C^{(j)}_{I^{(n)}_j} + 1$$

The mean: (Flajolet, Gonnet, Puech, Robson '93)

$$\mathbb{E}C_n \sim \gamma_{s,d} n^{\alpha-1}$$

$\gamma_{s,d} > 0$ (unknown)

$\alpha \in (1,2)$, satifying the indicial equation

$$\alpha^{d-s}(\alpha+1)^s = 2^d$$

Scaling

Var(C_n): unknown

Assume: Var(C_n) $\sim \beta_{s,d} n^{2\alpha-2}$ (later proved)

Normalisation:

$$X_n := \frac{C_n - \mathbb{E}C_n}{n^{\alpha-1}}$$

$\longrightarrow$ modified recursion + $I^{(n)}/n \to \langle U \rangle$
$\longrightarrow$ guess for limiting equation
$\longrightarrow$ limiting operator

$$T : M_1(\mathbb{R}, \mathcal{B}) \to M_1(\mathbb{R}, \mathcal{B})$$

$$T(\mu) \stackrel{\mathcal{D}}{=} \sum_{j_1,\ldots,j_s=0,1} \mathbf{1}_{j_1,\ldots,j_s}(Y, U) \times \sum_{j_{s+1},\ldots,j_d=0,1} \langle U \rangle^{\alpha-1}_{j_1,\ldots,j_d} (Z^{(j)} - \gamma_{s,d}) + \gamma_{s,d}$$

Contraction property

Lemma:

$T : M_{0,2} \to M_{0,2}$ is a contraction w.r.t. l_2:

$$l_2(T(\mu), T(\nu)) \leq \xi \, l_2(\mu, \nu) \quad \forall \mu, \nu \in M_{0,2},$$

$$\xi = \frac{1}{\sqrt{\alpha^s (\alpha - 1/2)^{d-s}}} < 1.$$

Remark to the proof:

First check that T is well-defined:

$$\mathbb{E}T(\mu) = \ldots = \gamma_{s,d} \left(\frac{2^d}{(\alpha+1)^s \alpha^{d-s}} - 1 \right) \stackrel{!}{=} 0$$

$\longrightarrow$ indicial equation

Limit law

Theorem: (Limit Law for PMQ in Quadtrees)
X_n the normalized cost for PMQ
X the fixed point of the limiting operator, then

$$l_2(X_n, X) \to 0.$$

The translated limit $\tilde{X} := X + \gamma_{s,d}$ is the unique solution in $M_{\gamma_{s,d},2}$ of the equation

$$\tilde{X} \stackrel{\mathcal{D}}{=} \prod_{i=1}^{s} U_i^{\frac{\alpha-1}{2}} \sum_{j_{s+1},\ldots,j_d=0,1} \bar{U}_{j_{s+1},\ldots,j_d}^{\alpha-1} \tilde{X}^{(j_{s+1},\ldots,j_d)}$$

with

$$\bar{U}_{j_{s+1},\ldots,j_d} := \prod_{\substack{s+1\le i\le d \\ j_i=0}} U_i \prod_{\substack{s+1\le i\le d \\ j_i=1}} (1-U_i),$$

the volumes of generated quadrants in the subspace of the last $d-s$ components.

Asymptotic for the variance

Corollary: (Variance of PMQ in Quadtrees)
The variance of the cost C_n for a partial match query in a d-dimensional quadtree with $1 \leq s \leq d-1$ components specified satisfies:

$$\mathsf{Var}(C_n) \sim \beta_{s,d} n^{2\alpha-2},$$

$$\beta_{s,d} = \left[\frac{((2\alpha-1)B(\alpha,\alpha)+1)^{d-s}-1}{\alpha^s(\alpha-1/2)^{d-s}-1} - 1 \right] \gamma_{s,d}.$$

Proof now is easy:

$$\begin{aligned} \mathsf{Var}(C_n) &= \mathsf{Var}(n^{\alpha-1}X_n) = (\mathsf{Var}(X)+o(1))n^{2\alpha-2} \\ &\sim \beta_{s,d} n^{2\alpha-2}. \end{aligned}$$

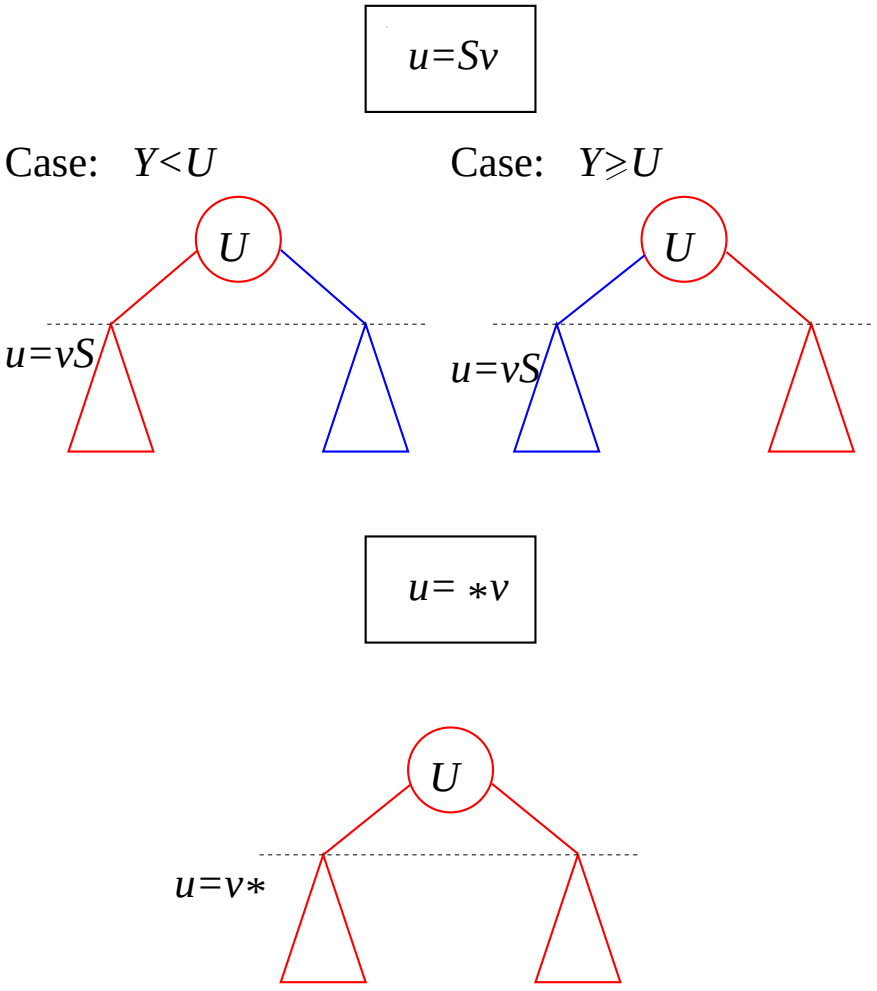
Calculate $\mathsf{Var}(X) = \beta_{s,d}$: Take square and expectation in

$$\tilde{X} \stackrel{\mathcal{D}}{=} \prod_{i=1}^{s} U_i^{\frac{\alpha-1}{2}} \sum_{j_{s+1},\ldots,j_d=0,1} \bar{U}^{\alpha-1}_{j_{s+1},\ldots,j_d} \tilde{X}^{(j_{s+1},\ldots,j_d)}.$$

PMQ in K-d-trees

Specification pattern specified in the components

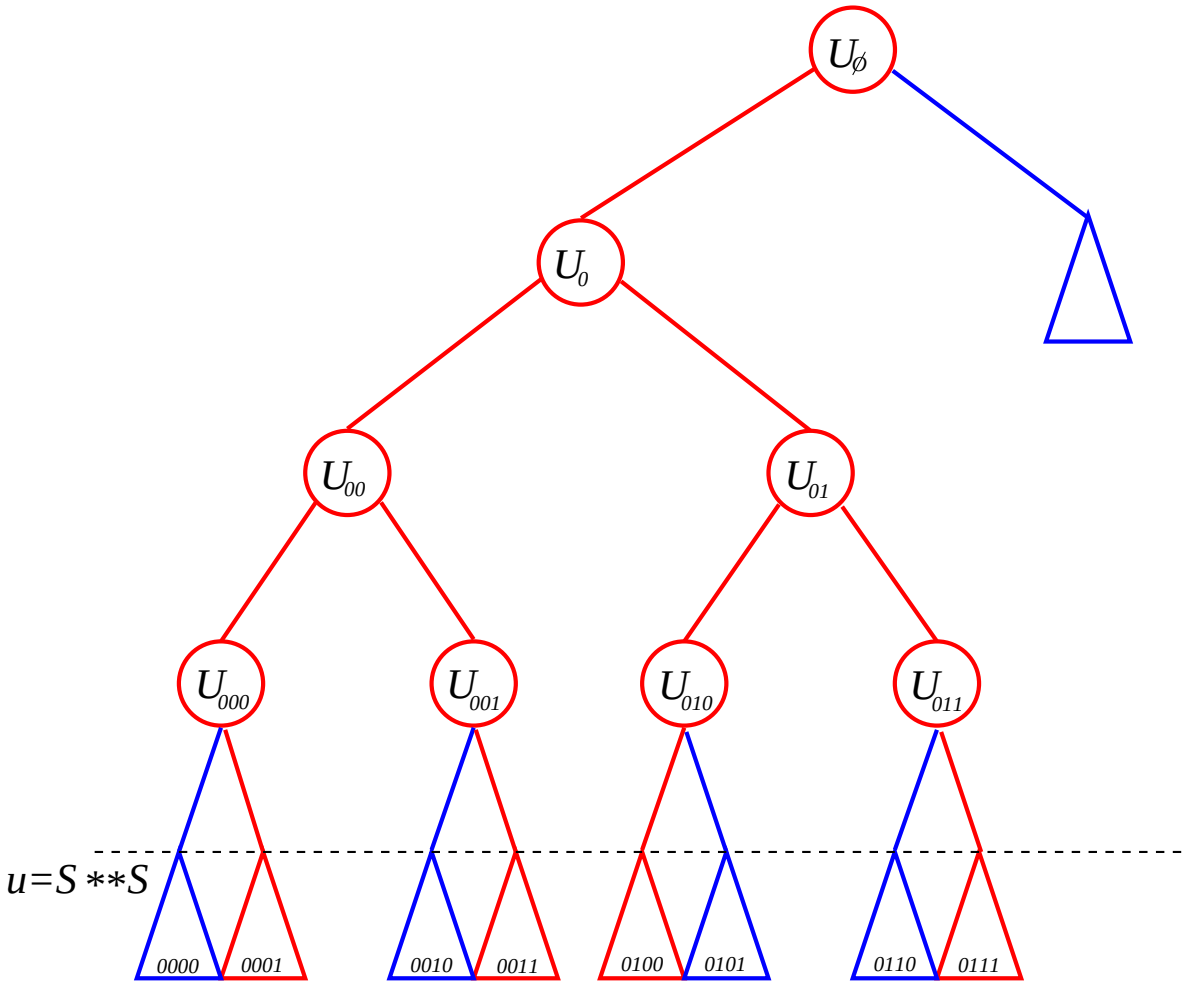
$$1 \leq r_1 < r_2 < \ldots < r_s \leq K, \quad 1 \leq s \leq K-1$$



$$C_n^{Sv} \stackrel{\mathcal{D}}{=} 1_{\{Y<U\}} C^{vS}_{I_1^{(n)}} + 1_{\{Y \geq U\}} \bar{C}^{vS}_{I_2^{(n)}} + 1$$

PMQ in K-d trees

Example: $K = 4,\ u = (S * *S)$



$D_n := \{0,1\}^n,\ D_0 := \{\emptyset\}$

$D^{(K)} := \bigcup_{n=0}^{K-1} D_n$

$\sigma|j := (\sigma_1, \ldots, \sigma_j) \in D_j$ for $\sigma \in D_n,\ 0 \le j \le n$

The recursion for the cost C_n

Distributional recursive equation for C_n:

$$C_n \stackrel{\mathcal{D}}{=} \sum_{\sigma \in D_K} \mathbf{1}_\sigma(\mathcal{U}, Y)\, C^{(\sigma)}_{I^{(n)}_\sigma} + N_n,$$

$Y, \mathcal{U}, (C_i^{(\sigma)})_{i\in\mathbb{N}}$ $(\sigma \in D_K)$ independent

$(C_i^{(\sigma)})_{i\in\mathbb{N}} \sim (C_i)_{i\in\mathbb{N}}, \quad \sigma \in D_K$

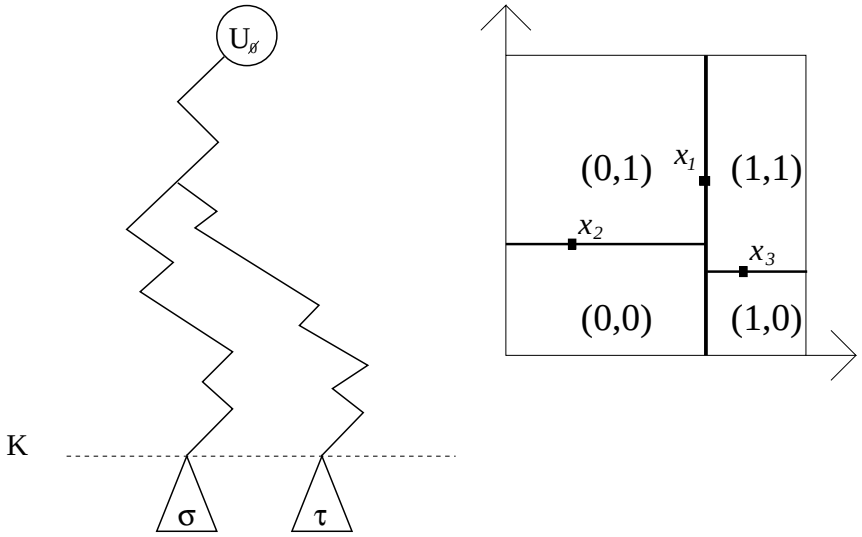
N_n number of nodes traversed on levels $0, \ldots, K-1$

$0 \le N_n \le 2^K$

$I^{(n)} = (I_\sigma^{(n)})_{\sigma\in D_K}$ vector of the cardinalities of the subtrees on level K

$$\mathbf{1}_\sigma(\mathcal{U}, Y) := \prod_{\substack{1\le j\le s\\ \sigma r_j = 0}} \mathbf{1}_{\{Y_j < U_{\sigma|r_j-1}\}} \prod_{\substack{1\le j\le s\\ \sigma r_j = 1}} \mathbf{1}_{\{Y_j \ge U_{\sigma|r_j-1}\}}$$

The subtree-sizes $I^{(n)}$



Volumes generated by the discriminators:

$$\langle \mathcal{U} \rangle_\sigma := \prod_{\substack{1 \le j \le K \\ \sigma_j = 0}} U_{\sigma|j-1} \prod_{\substack{1 \le j \le K \\ \sigma_j = 1}} (1 - U_{\sigma|j-1})$$

Given the levels $0, \ldots, K-1$ are full and given the discriminators $\mathcal{U}$, $I^{(n)}$ is multinomial distributed.

$$\frac{I^{(n)}}{n} \xrightarrow{\mathbb{P}} \langle \mathcal{U} \rangle = (\langle \mathcal{U} \rangle_\sigma)_{\sigma \in D_K}$$

Scaling

The mean: (Flajolet, Puech '86)

$$\mathbb{E}C_n \sim \gamma_u n^{\alpha-1}$$

$\gamma_u > 0$ depends on the specification pattern u,
$\alpha = \alpha(s, K) \in (1, 2)$ satisfying the indicial equation

$$\alpha^{K-s}(\alpha+1)^s = 2^K$$

Assume: $\mathrm{Var}(C_n) \sim \beta_u n^{2\alpha-2}$

Normalized version:

$$X_n := \frac{C_n - \mathbb{E}C_n}{n^{\alpha-1}}$$

Distributional recursion for the normalized cost:

$$X_n \stackrel{\mathcal{D}}{=} \sum_{\sigma \in D_k} \mathbf{1}_\sigma(\mathcal{U}, Y) \left(\frac{I_\sigma^{(n)}}{n}\right)^{\alpha-1} \left(X^{(\sigma)}_{I_\sigma^{(n)}} + \gamma_u\right)$$
$$-\gamma_u + o(1)$$

The limiting operator

Limiting equation:

$$X \stackrel{\mathcal{D}}{=} \sum_{\sigma \in D_k} \mathbf{1}_\sigma(\mathcal{U}, Y) \langle \mathcal{U} \rangle_\sigma^{\alpha-1} (X^{(\sigma)} + \gamma_u) - \gamma_u$$

Limiting operator:

$$T : M_1(\mathbb{R}, \mathcal{B}) \to M_1(\mathbb{R}, \mathcal{B})$$

$$T(\mu) \stackrel{\mathcal{D}}{=} \sum_{\sigma \in D_K} \mathbf{1}_\sigma(\mathcal{U}, Y) \langle \mathcal{U} \rangle_\sigma^{\alpha-1} (Z^{(\sigma)} + \gamma_u) - \gamma_u$$

$\mathcal{U}, Y, Z^{(\sigma)}$, $\sigma \in D_K$ independent, $Z^{(\sigma)} \sim \mu$.

Limit law

Theorem: (Limit law for PMQ in K-d trees)
X_n the normalized cost for PMQ
X the fixed point of the limiting operator, then

$$l_2(X_n, X) \to 0.$$

Corollary: (Variance of PMQ in K-d trees)

$$\mathrm{Var}(C_n) \sim \beta_u n^{2\alpha-2}$$

with

$$\beta_u = \left[c_\alpha \sum_{l \in \mathbf{U}} \left(\frac{2(2\alpha-1)}{\alpha^2} \right)^{K-l} \eta_\alpha^{s-s_l} - 1 \right] \gamma_u^2,$$

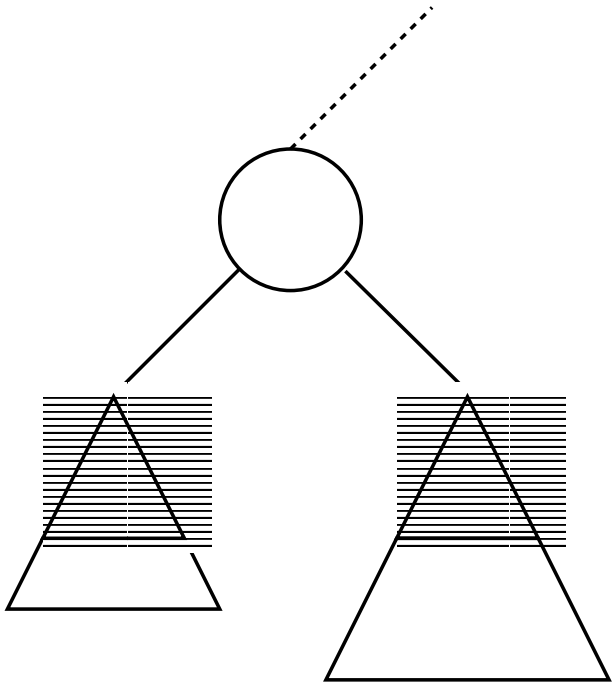
$$c_\alpha = \frac{(2\alpha-1)B(\alpha,\alpha)}{\alpha^s(\alpha-1/2)^{K-s}-1},$$

$$\eta_\alpha = \frac{\alpha(8\alpha^2-2\alpha-2-\alpha(\alpha+1)B(\alpha,\alpha))}{2(\alpha+1)(2\alpha-1)(2\alpha+1)}.$$

K-d-t trees

Locally balanced version of K-d trees introduced in Cunto, Lau and Flajolet ('89).

Subtrees of size greater than $2t$ have at least t nodes on each side. ($t = 0$: k-d trees)



$$\mathbb{E}C_n^{(t)} \sim \gamma_{t,u} n^{1-s/K+\theta(s/K,t)}$$

$$\theta(s/K, t) \longrightarrow 0 \quad \text{for} \quad t \to \infty$$

Random relaxed K-d tree

Each internal node contains a key plus an associated discriminant $j \in \{1, \ldots, K\}$.

The discriminant determines, which component is used for the comparison.

randomized version: The discriminants are independent r.v., uniformly distributed on $\{1, \ldots, K\}$.

Duch, Estivill-Castro, Martínez ('98)
Martínez, Panholzer, Prodinger ('98)

$1 \leq s \leq K - 1$ components specified

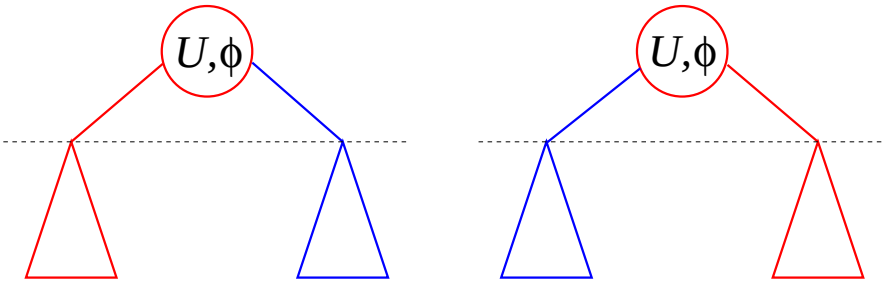
$\rho := \frac{s}{K}$

PMQ in random-r-K-d-trees

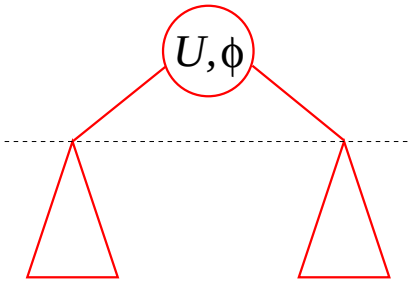
ϕ the random discriminant associated to the root

ϕ -th component specified

Case: $Y<U$ Case: $Y\geqslant U$



ϕ-th component unspecified



$$C_n \stackrel{\mathcal{D}}{=} B_\rho \left(\mathbf{1}_{\{Y<U\}} C^{(1)}_{I_1^{(n)}} + \mathbf{1}_{\{Y\geq U\}} C^{(2)}_{I_2^{(n)}} \right)$$
$$(1 - B_\rho) \left(C^{(1)}_{I_1^{(n)}} + C^{(2)}_{I_2^{(n)}} \right) + 1$$

The moments

$$\mathbb{E}C_n = \gamma n^{\alpha-1} + \mathcal{O}(1)$$

$$\alpha = \frac{1+\sqrt{9-8\rho}}{2}, \qquad \gamma = \frac{\Gamma(2\alpha-1)}{(1-\rho)\alpha\Gamma^3(\alpha)}$$

$$\mathsf{Var}(C_n) \sim \left(\frac{8\Gamma(2\alpha)}{\alpha^2(\alpha-1)^2(2\alpha-1)(3\alpha-2)\Gamma^4(\alpha)} - \frac{4\Gamma^2(2\alpha)}{\alpha^4(\alpha-1)^2(2\alpha-1)^2\Gamma^6(\alpha)} \right) n^{2\alpha-2}$$

$$X_n = \frac{C_n - \mathbb{E}C_n}{n^{\alpha-1}}$$

Limiting operator

$$T : M_1(\mathbb{R}, \mathcal{B}) \to M_1(\mathbb{R}, \mathcal{B})$$

$$\begin{aligned} T(\mu) \stackrel{\mathcal{D}}{=} B_\rho \Big(& \mathbf{1}_{\{Y<U\}} U^{\alpha-1}(X^{(1)} + \gamma) \\ & + \mathbf{1}_{\{Y \geq U\}} (1-U)^{\alpha-1}(X^{(2)} + \gamma) \Big) \\ + (1 - B_\rho) \Big(& U^{\alpha-1}(X^{(1)} + \gamma) \\ & + (1-U)^{\alpha-1}(X^{(2)} + \gamma) \Big) - \gamma \end{aligned}$$

Lemma:

$T : M_{0,2} \to M_{0,2}$ is a contraction w.r.t. l_2

$$l_2(T(\mu), T(\nu)) \leq \xi \, l_2(\mu, \nu) \quad \forall \, \mu, \nu \in M_{0,2},$$

$$\xi = \left[2 \frac{1 - \rho + \sqrt{9 - 8\rho}}{9 - 8\rho + \sqrt{9 - 8\rho}} \right]^{1/2} < 1$$

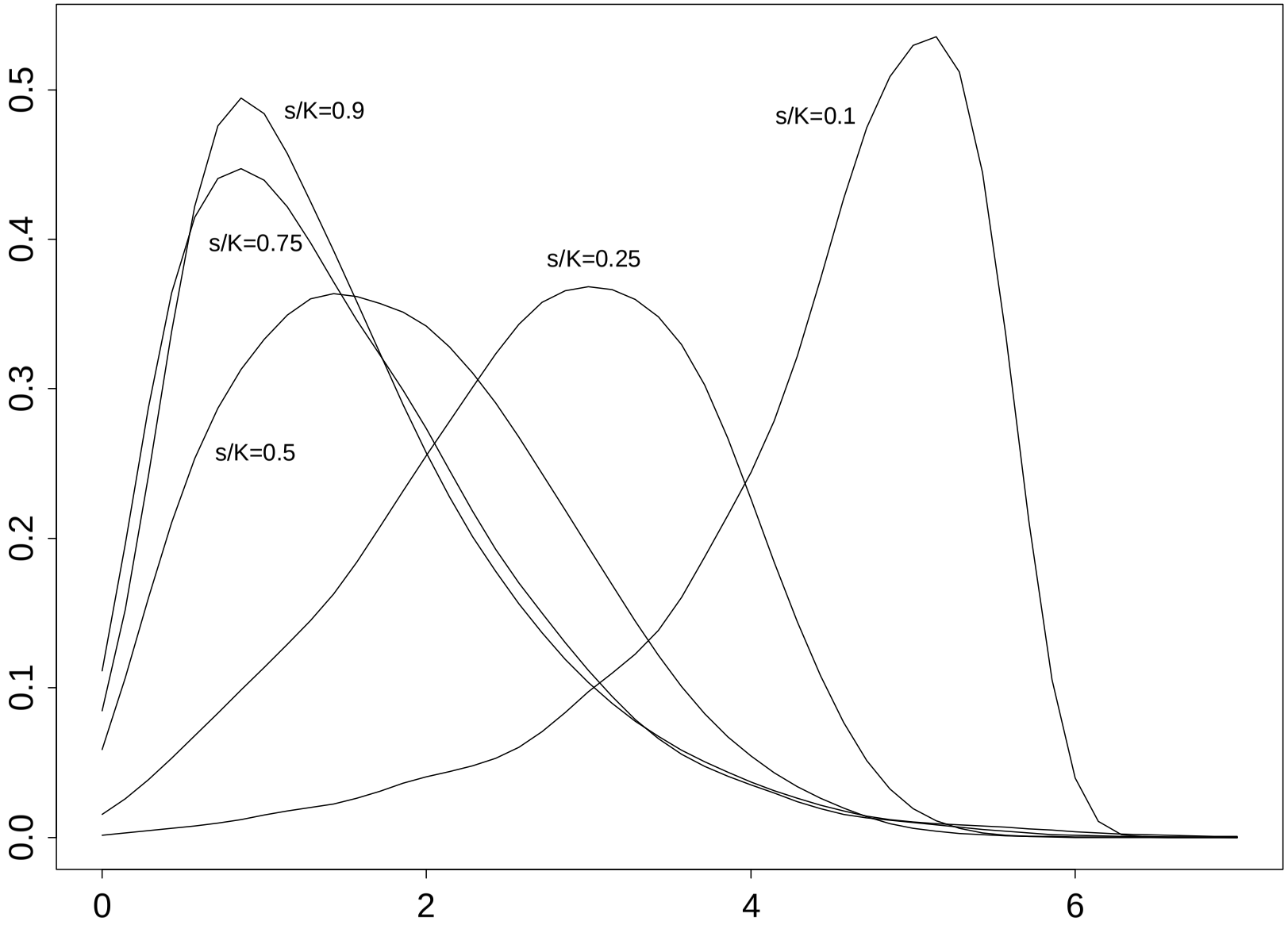
Limit law

Theorem: (Limit law for PMQ in random-r-K-d-trees)
X_n the normalized cost for PMQ
X the fixed point of the limiting operator, then

$$l_2(X_n, X) \to 0.$$

The translated limit $\tilde{X} := X + \gamma$ is the unique solution in $M_{\gamma,2}$ of the equation

$$Z \stackrel{\mathcal{D}}{=} B_\rho U^{\frac{\alpha-1}{2}} Z^{(1)} + (1-B_\rho)\left(U^{\alpha-1} Z^{(1)} + (1-U)^{\alpha-1} Z^{(2)}\right)$$


s/K=0.9
s/K=0.1
s/K=0.75
s/K=0.25
s/K=0.5
0.0
0.1
0.2
0.3
0.4
0.5
0
2
4
6

Laplace Transform

Theorem: (Laplace transforms)

$$\mathbb{E}\exp(\lambda X) < \infty \quad \textit{for all} \quad \lambda \in (-\lambda_0, \lambda_0).$$

Assume

$$0 < \frac{s}{K} \leq \frac{\ln(4/3)}{\ln(5/3)} = 0.563\ldots \quad \textit{for the } K\textit{-d tree},$$

$$0 < \frac{s}{K} \leq \frac{\ln\left(\frac{4+2t}{3+2t}\right)^{\overline{t+1}}}{\ln\left(\frac{5+2t}{3+2t}\right)^{\overline{t+1}}} \quad \textit{for the } K\textit{-d-}t \textit{ tree},$$

$$0 < \frac{s}{K} \leq 0.625 \quad \textit{for the random relaxed } K\textit{-d tree},$$

$$0 < \frac{s}{d} \leq \frac{\ln(4/3)}{\ln(5/3)} = 0.563\ldots \quad \textit{for the quadtree}.$$

Then

$$\mathbb{E}\exp(\lambda X_n) \to \mathbb{E}\exp(\lambda X) < \infty \quad \textit{for all} \quad \lambda \in \mathbb{R}.$$

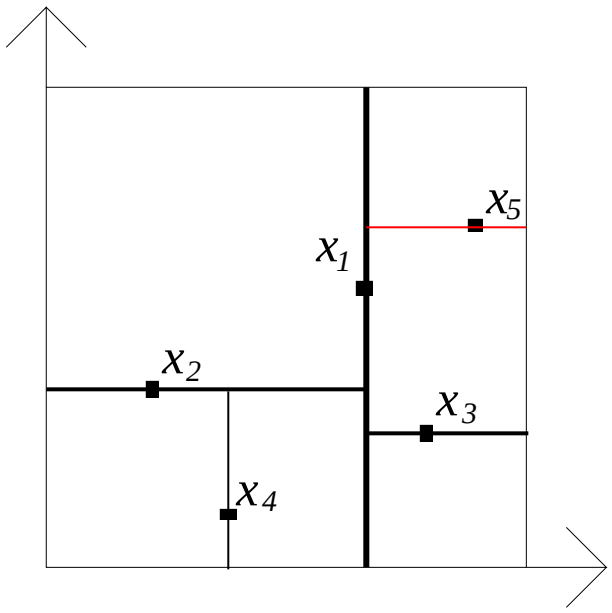
$$\mathbb{P}(C_n \geq a_n) \leq c_\lambda \exp\left(-\lambda \frac{a_n}{n^{\alpha-1}}\right) \quad \textit{for all} \quad \lambda > 0.$$

The squarish K-d tree

Devroye, Jabbour, Zamora-Cura '99

- relaxed K-d tree

- choose the discriminant at a node, so that the corresponding quadrant is cut perpendicular to the longest edge



$$\mathbb{E}C_n = \Theta(n^{1-s/K})$$

Problem: Prove

$$\mathbb{E}C_n \sim \gamma n^{1-s/K}.$$

Internal path length in random trees

Internal path length (IPL): Sum of the depths of the nodes in a tree.

- Cost (=key comparisons) for building up the tree.

- BST: Internal path length is distributed as the running time of the Quicksort algorithm.

The Quadtree

Y_n: IPL of the random d-dimensional quadtree,
$Y_1 = 1,\ Y_2 = 3, \ldots$

Distributional recursive equation for Y_n:

$$Y_n \stackrel{\mathcal{D}}{=} \sum_{k=0}^{2^d-1} Y^{(k)}_{I^{(n)}_k} + n$$

with

$I^{(n)}, (Y_i^{(0)}), \ldots, (Y_i^{(2^d-1)})$ independent

$Y_i^{(k)} \sim Y_i,\ 0 \leq k \leq 2^d - 1$

$I^{(n)}$ vector of the subtree sizes.

Moments of Y_n

$$\mathbb{E}Y_n = \frac{2}{d} n \ln n + \mu_d n + o(n)$$

(Flajolet, Labelle, Laforest, Salvy '95)

Var Y_n: unknown

Assume: Var $Y_n \sim v_d n^2$ (later proved)

Normalisation:

$$X_n := \frac{Y_n - \mathbb{E}Y_n}{n}.$$

The modified recursion

$$X_n \stackrel{\mathcal{D}}{=} \sum_{k=0}^{2^d-1} \frac{I_k^{(n)}}{n} X_{I_k^{(n)}}^{(k)} + C_n(I^{(n)})$$

with

$$C_n(i) = 1 + \frac{1}{n}\left(\sum_{k=0}^{2^d-1} \mathbb{E}Y_{i_k} - \mathbb{E}Y_n\right)$$

for $i = (i_0, \ldots, i_{2^d-1})$ with $\sum i_k = n - 1$.

Using the expectation formula it follows

$$C_n(i) = 1 + \frac{2}{d} \sum_{k=0}^{2^d-1} \frac{i_k}{n} \ln\left(\frac{i_k}{n}\right) + o(1)$$

Recall $\frac{I^{(n)}}{n} \xrightarrow{\mathbb{P}} \langle U \rangle$.
($\langle U \rangle$ volumes generated by the root.)

The limiting equation

Entropy functional

$$C : T_{2^d-1} \to \mathbb{R}, \;\; C(x) := 1 + \frac{2}{d} \sum_{k=0}^{2^d-1} x_k \ln x_k$$

defined on the simplex

$$T_{2^d-1} := \left\{ x \in [0,1]^{2^d} \;\; : \;\; \sum_{k=0}^{2^d-1} x_k = 1 \right\}.$$

Limiting operator:

$$T : M^1(\mathbb{R}^1, \mathcal{B}^1) \longrightarrow M^1(\mathbb{R}^1, \mathcal{B}^1)$$

$$T(\mu) \stackrel{\mathcal{D}}{=} \sum_{k=0}^{2^d-1} \langle U \rangle_k Z^{(k)} + C(\langle U \rangle)$$

$U, Z^{(0)}, \ldots, Z^{(2^d-1)}$ independent, U unif. on $[0,1]^d$ and $Z^{(k)} \sim \mu$.

The limit law

Theorem: (Limit law for IPL in quadtrees) *Let Y_n be the IPL in a d-dimensional quadtree, X_n be the normalized version and X the fixed point of the limiting operator, then:*

$$l_2(X_n, X) \to 0,$$

$$\mathsf{Var}(Y_n) \sim v_d n^2, \quad v_d = \frac{21 - 2\pi^2}{9d(1-(2/3)^d)},$$

$$\mathbb{E}\exp(\lambda X_n) \to \mathbb{E}\exp(\lambda X), \quad \lambda \in \mathbb{R},$$

$$\mathbb{P}(|Y_n - \mathbb{E}Y_n| \geq \epsilon \mathbb{E}Y_n) = O(n^{-k}) \quad \forall\, k \in \mathbb{N}.$$

General split trees

Other trees are of a similar type:
BST, m-ary search trees, median-of-$(2k+1)$ search trees,...

$\longrightarrow$ random split tree model (Devroye '98)
$\longrightarrow$ analysis uniformly valid for the depth and height

IPL:

$$Y_n \stackrel{\mathcal{D}}{=} \sum_{k=1}^{b} Y^{(k)}_{I_k^{(n)}} + n$$

$I^{(n)}$ has conditionally given a splitting vector $\mathcal{V} = (V_1, \ldots, V_b)$ a multinomial structure:

$$\mathbb{P}^{I^{(n)} \,|\, \mathcal{V}=v} = M(n-s, v),$$

$$\mathbb{P}^{I_1^{(n)}} = B(n-s, V_1)$$

$V \sim V_1$ is called the splitter .

Expectation of the IPL

$$\mathbb{E}Y_n = b \sum_{i=0}^{n-s} \mathbb{P}(I_1^{(n)} = i)\mathbb{E}Y_i + n$$

with $\mathbb{P}^{I_1^{(n)}} = B(n-s, V_1)$, i.e.

$$\mathbb{P}(I_1^{(n)} = i) = \int_0^1 \binom{n-s}{i} p^i (1-p)^{n-s-i} d\mathbb{P}^V(p)$$

BST: V uniformly on $[0,1]$ distributed

m-ary tree: V beta$(1, m-1)$ dist.
. (min. of m unif. on $[0,1]$ dist. r.v.)

quadtree: V product of d unif. dist. r.v.

median of $(2k+1)$ tree: V beta$(k+1, k+1)$ dist.
. (median of $2k+1$ unif. on $[0,1]$ dist. r.v.)

Problem

$$\mathbb{E}Y_n = b\sum_{i=0}^{n-s}\mathbb{P}(I_1^{(n)} = i)\mathbb{E}Y_i + n,$$

$$\mathbb{P}(I_1^{(n)} = i) = \int_0^1 \binom{n-s}{i} p^i (1-p)^{n-s-i} d\mathbb{P}^V(p)$$

Find conditions on V which imply

$$\mathbb{E}Y_n = \mu^{-1} n \ln n + dn + o(n)$$

with

$$\mu = b\mathbb{E}[V \ln(1/V)]$$

The formula is valid for the examples mentioned above!

The Find-algorithm

Search for an order statistic in a file of n keys using the partitioning procedure of quicksort.

- pick the pivot element at random
- partition the set of elements into the elements smaller resp. greater than the pivot
- continue recursively in the set containing the desired statistic

Model: Assume

- the keys permuted uniformly at random
- the order statistic M_n uniformly over $\{1, \ldots, n\}$
- the pivot Z_n chosen as the median of 3 independent, unif. over the keys distributed elements

Recursion for the cost C_n

C_n the number of key comparisons of Find applied to a set of n keys (without the comparisons for finding the median of three)

$$C_n \stackrel{\mathcal{D}}{=} \mathbf{1}_{\{Z_n > M_n\}} C^*_{Z_n - 1} + \mathbf{1}_{\{Z_n < M_n\}} C^{**}_{n - Z_n} + n - 1$$

$Z_n, M_n, (C_i^*), (C_i^{**})$ independent

$C_i^*, C_i^{**} \sim C_i$

M_n unif. dist. over $\{1, \ldots, n\}$

Z_n dist. as the median of three uniformly on $\{1, \ldots, n\}$ dist. r.v.

Scaling of C_n

$$\mathbb{E}C_n = \frac{5}{2}n + O(\ln n)$$

(Kirschenhofer, Martínez, Prodinger '97)

Assume: $\mathrm{Var}C_n \sim m_2 n^2$

$$Y_n := \frac{C_n - \mathbb{E}C_n}{n}$$

Modified recursion

$$\begin{aligned} Y_n \stackrel{\mathcal{D}}{=} & \quad 1_{\{Z_n > M_n\}} \frac{Z_n - 1}{n} \left(Y^*_{Z_n - 1} + \frac{5}{2} \right) \\ & + \quad 1_{\{Z_n < M_n\}} \frac{n - Z_n}{n} \left(Y^{**}_{n - Z_n} + \frac{5}{2} \right) \\ & - \quad \frac{5}{2} + \frac{n-1}{n} + o(1). \end{aligned}$$

Limiting equation/operator

$$\begin{aligned} Y \stackrel{\mathcal{D}}{=} & \quad \mathbf{1}_{\{T>U\}} T(Y^* + 5/2) \\ & \quad + \mathbf{1}_{\{T<U\}}(1-T)(Y^{**} + 5/2) \\ & \quad - 3/2 \\ \stackrel{\mathcal{D}}{=} & \quad X(Y + 5/2) - 3/2 \end{aligned}$$

with X, Y independent, X has density $t \mapsto 12t^2(1-t)$ for $t \in [0,1]$.

The translation $\overline{Y} := Y + 5/2$ is the solution of

$$\overline{Y} = X\overline{Y} + 1 \quad \text{with} \quad \mathbb{E}\overline{Y} = 5/2$$

and

$$\frac{C_n}{n} \xrightarrow{\mathcal{D}} \overline{Y}.$$

Moments

Existence and convergence of the Laplace transform implies convergence of all moments, i.e.

$$\mathbb{E}[(C_n)^k] \sim m_k n^k \quad \text{for} \quad n \to \infty$$

with $m_0 = 1, m_1 = 5/2$ and

$$m_k = \left(12 + \frac{144}{k^2 + 7k}\right) \sum_{j=0}^{k-1} \binom{k}{j} \frac{m_j}{(j+3)(j+4)}$$

for $k \geq 2$.

Fourier transform

$$\mathbb{E} \exp(it\overline{Y}) = e^{it}\phi(t)$$

where $\phi : \mathbb{R} \to \mathbb{C}$ is smooth and given by the linear homogeneous differential equation of second order

$$t^2\phi''(t) + 8t\phi'(t) - 12(e^{it} - 1)\phi(t) = 0$$

with the initial conditions $\phi(0) = 1$ and $\phi'(0) = 3i/2$.