Median and Hybrid Median K-Dimensional Trees




Amalia Duch Conrado Martínez Mercè Pons Salvador Roura

Univ. Politècnica de Catalunya, Spain

Analytic and Probabilistic Combinatorics BIRS Workshop
November 14–18, 2022
Banff, Canada

Median and Hybrid Median K-Dimensional Trees

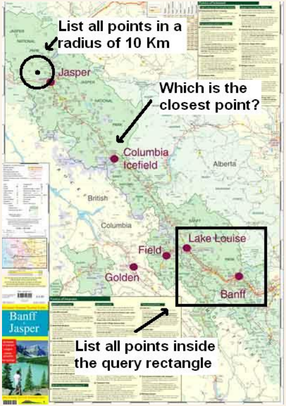


Amalia Duch   Conrado Martínez & Mercè Pons   Salvador Roura

Univ. Politècnica de Catalunya, Spain

Analytic and Probabilistic Combinatorics BIRS Workshop
November 14–18, 2022
Banff, Canada

The problem



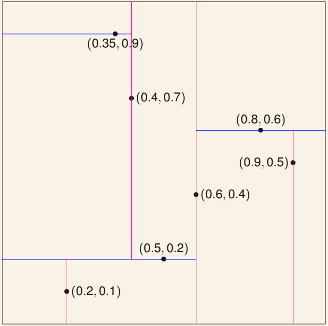
- **INPUT:** A set of *n* multidimensional data points + an associative query
- **OUTPUT:** Data points matching the query

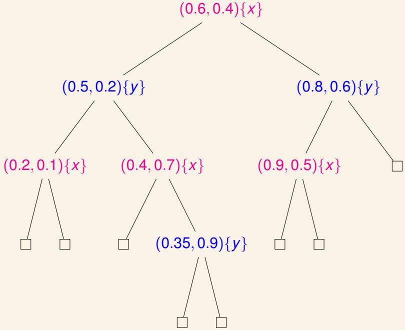
In this talk

- Two variants of multidimensional trees: median K-d trees and hybrid median K-d trees
- Analysis of the expected internal path length and the expected cost of partial match queries
- Trees are randomly built from n points where each coordinate x_i of a data point $\mathbf{x}$ is independently and uniformly drawn from $[0, 1]$

Standard K-d trees (Bentley, 1975)

Jon L. Bentley





Internal Path Length and Partial Match

K-d trees provide efficient (on expectation) support for dynamic insertions, exact searches and several associative queries

We focus here on:

- Internal path length (IPL)
  - cost of building the tree
  - cost of a successful search $= 1 + \frac{\text{IPL}}{n}$
- Partial match (PM) queries
  - most basic associative query: find all points matching a query with non-specified coordinates
  - a fundamental block for the analysis of other associative queries (orthogonal range, nearest neighbour queries, ...)

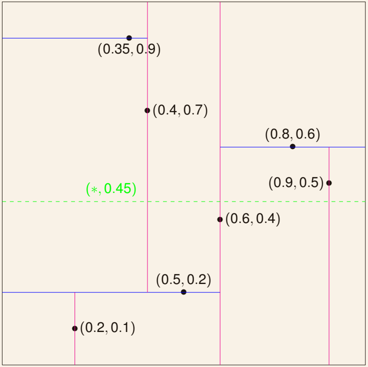
Partial match queries

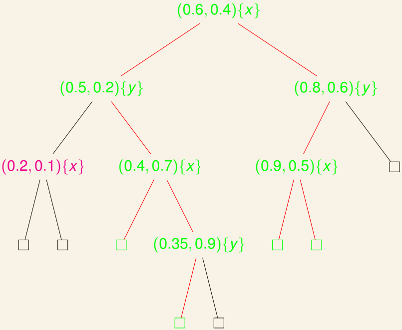
Definition

A random partial match query (RPM) is a K-dimensional tuple $\mathbf{q} = (q_0, q_1, \ldots, q_{K-1})\rangle$ where each $q_i \in [0,1] \cup \{*\}$

- The specified coordinates $q_i \in [0,1]$ are drawn from the same distribution as the coordinates of the data points
- $s =$ the number of specified coordinates in a query $\mathbf{q}$; we assume $0 < s < K$
- Goal: to report all data points $\mathbf{x} = (x_0, \ldots, x_{K-1})$ in the tree such that $x_i = q_i$ whenever $q_i \neq *$

Example of a random partial match query





Known results

- IPL $\sim c_K n \ln n$
- RPM $= \Theta(n^\alpha), \alpha = \alpha(s, K)$

Family	IPL (c_K)		RPM (α)	
	$K = 2$	$K \to \infty$	$s = 1$, $K = 2$	$s = K/2$, $K \to \infty$
Standard K-d trees	2	2	0.56155	0.56155
Relaxed K-d trees	2	2	0.618	0.618
Squarish K-d trees	2	2	0.5	0.5

Median K-d trees and hybrid median K-d trees

In median and hybrid median K-d trees we choose the discriminant of each node aiming at building more balanced trees

- Median K-d trees: choose as discriminant of each node the coordinate that is closest, after renormalization, to the center of the region associated to the node (bounding box)
- Hybrid median K-d trees: use the median rule but only with coordinates that haven't been used in the current path, until a full permutation of discriminants has been used

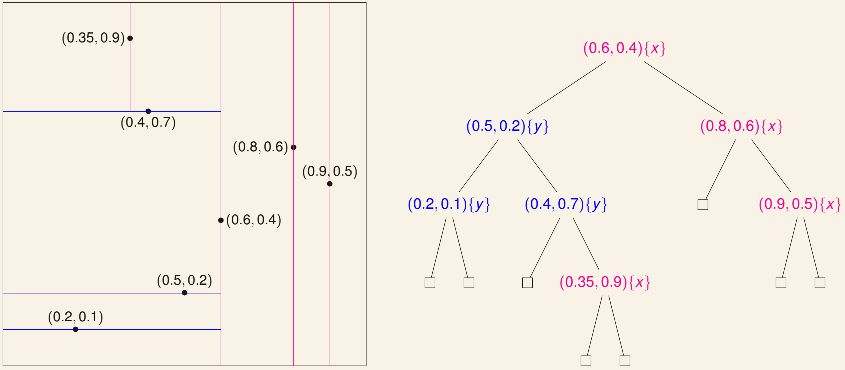
Median K-d trees

- Introduced in Pons's master thesis (2010)
- When a new data point $\mathbf{x} = (x_0, \ldots, x_{K-1})$ is inserted in the leaf associated to region $R = [\ell_0, u_0] \times \cdots [\ell_{K-1}, u_{K-1}]$ (bounding box) the discriminant j is chosen as follows

  $$j = \text{arg min}_{0 \leq i < K} \left\{ \left| \frac{x_i - \ell_i}{u_i - \ell_i} - \frac{1}{2} \right| \right\}$$

  that is, the coordinate such that x_j is closest, after renormalization, to the center

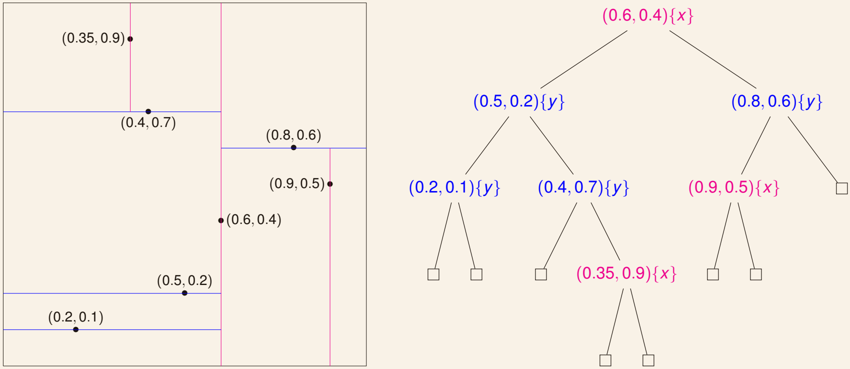
Example of a median K-d tree



Hybrid median K-d trees

- Hybrid median K-d trees also introduced by Pons in 2010,
- For an arbitrary dimension $K \geq 2$, the rule to assign the discriminants is the following
  1. Nodes at levels $\ell \equiv 0 \pmod{K}$ discriminate w.r.t. the median rule applied to all K coordinates
  2. Nodes at levels $\ell \equiv j \pmod{K}$, $0 < j < K$, discriminate w.r.t. the median rule applied to all the coordinates not used as discriminant by any of its $j - 1$ immediate ascendants
- Discriminants along any path from the root to a leaf form a sequence of permutations of order K, except perhaps for the last part of the path, which will contain only $< K$ distinct discriminants

Example of a hybrid median K-d tree



Median K-d trees: Expected IPL

Theorem (Pons, 2010)

The expected IPL of random median K-d tree of size n is

$$I_n = c_K^{[med]} n \ln n + o(n \log n)$$

where

$$c_K^{[med]} = \left(-K 2^K \left[A_K + \sum_{0 \le i < K} \binom{K-1}{i} (-1)^i B_{i+1} \right] \right)^{-1},$$

with $B_j = -(A_j + 1/(j+1)^2)$ *and*

$$A_j = \int_0^{1/2} z^j \ln z \, dz = -\frac{1 + (j+1) \ln 2}{2^{j+1} (j+1)^2}$$

Hybrid median K-d trees: Expected IPL

Theorem

The expected IPL of a random hybrid median K-d tree of size n is

$$I_n = c_K^{[hm]} n \ln n + o(n \log n)$$

where

$$c_K^{[hm]} = \frac{K}{\frac{1}{c_1^{[med]}} + \ldots + \frac{1}{c_K^{[med]}}}$$

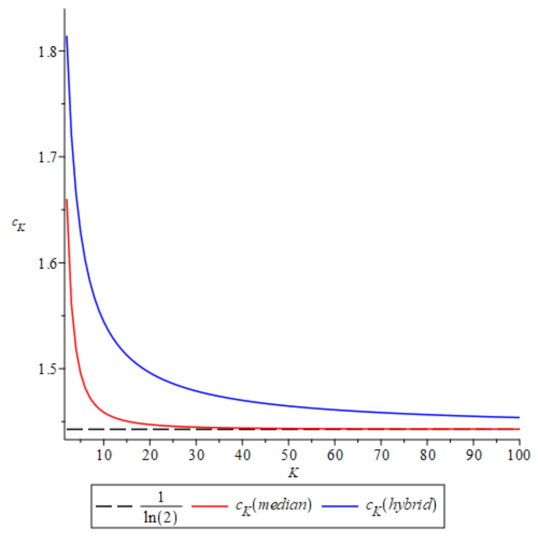
Expected IPL: the coefficients c_K

Proposition

For all $K \geq 2$,

1. $c_K^{[med]} \leq c_K^{[hm]} < 2 = c_K^{[rlx]} = c_K^{[sqr]} = c_K^{[std]}$
2. $c_K^{[med]} > c_{K+1}^{[med]}$ *and* $c_K^{[hm]} > c_{K+1}^{[hm]}$,
3. $$\lim_{K\to\infty} c_K^{[hm]} = \lim_{K\to\infty} c_K^{[med]} = \frac{1}{\ln 2} \leftarrow \textit{optimal}$$

Expected IPL: the coefficients c_K



Median K-d trees: Random partial matches

Theorem

The expected cost of a RPM query with s specified coordinates out of K, $0 < s < K$, in a random median K-d tree of size n is:

$$P_n = \Theta(n^\alpha),$$

where $\alpha \in [0, 1]$ is the unique real solution of:

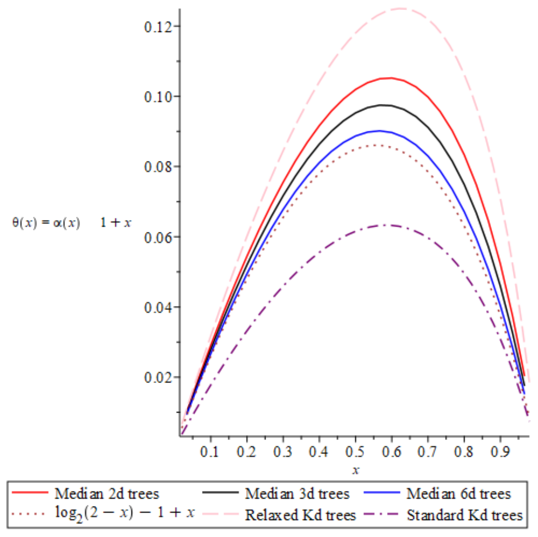
$$2^{-\alpha}\left(\frac{K(1-\rho)}{K+\alpha} + \frac{K\rho}{2(K+\alpha+1)}\right) + K2^K\Big\{\rho\, B(1/2; K+1, \alpha+1) + (1-\rho)\, B(1/2; K, \alpha+1)\Big\} = 1,$$

with $\rho = s/K$ and $B(z; a, b) = \int_0^z t^{a-1}(1-t)^{b-1}\, dt$ denoting the incomplete Beta function

Median K-d trees: Random partial matches

- Although it is not possible to give a closed form for α in terms of K and ρ it is possible to compute numerical approximations with any desired degree of accuracy
- It is possible also to find the value of α as K grows and $\rho = s/K$ is fixed. From known asymptotic expansions of the incomplete Beta function we get $\alpha \to \log_2(2 - \rho)$ as $K \to \infty$ and $\rho = s/K$ fixed.

Median K-d trees: Random partial matches



Hybrid median K-d trees: Random partial matches

Theorem

The expected cost of a RPM query with s specified coordinates out of K, $0 < s < K$, in a random hybrid median K-d tree of size n is

$$P_n^{(K,s)} = \Theta(n^{\alpha}),$$

where $\alpha \in [0, 1]$ is the unique real solution of

$$\det(\boldsymbol{I} - \boldsymbol{\Phi}(x)) = 0,$$

where $\boldsymbol{\Phi}(x) = \int_0^1 \boldsymbol{\Omega}(z)\, z^x\, dz$ and $\boldsymbol{\Omega}(z)$ is the shape matrix corresponding to a system of d divide-and-conquer recurrences, $d = (K - s + 1)(s + 1) - 1$

Hybrid median K-d trees: Random partial matches

	s				
K	1	2	3	4	5
2	0.546 (0.562)	-	-	-	-
3	0.697 (0.716)	0.368 (0.395)	-	-	-
4	0.771 (0.79)	0.53 (0.562)	0.275 (0.306)	-	-
5	0.815 (0.833)	0.624 (0.656)	0.425 (0.463)	0.218 (0.25)	-
6	0.845 (0.862)	0.685 (0.716)	0.522 (0.562)	0.354 (0.395)	0.181 (0.211)

In parentheses the values for standard K-d trees

A comparison of various K-d trees

Family	IPL (c_K)		Partial match (α)	
			$s = 1$,	$s = K/2$,
	$K = 2$	$K \to \infty$	$K = 2$	$K \to \infty$
Standard K-d trees	2	2	0.56155	0.56155
Relaxed K-d trees	2	2	0.618	0.618
Squarish K-d trees	2	2	0.5	0.5
Median K-d trees **[this paper]**	1.66	$\to$ 1.443	0.602	$\to$ 0.585
Hybrid median K-d trees **[this paper]**	1.814	$\to$ 1.443	0.546	$\to 0.5^*$

* conjectured

Sketch of the proofs

In order to prove previous theorems we follow these steps:

1. Set up recurrences for the expected IPL and expected cost of PM in median K-d trees
2. Solve the resulting divide-and-conquer recurrences by means of Roura's Continuous Master theorem (CMT)
3. For hybrid median K-d trees is more complicated since it requires considering systems of divide-and-conquer recurrences —not covered by CMT
4. We have generalized the CMT to solve systems of D&C recurrences such as those in the analysis of hybrid median K-d trees

The Continuous Master Theorem

CMT considers divide-and-conquer recurrences of the following type:

$$F_n = t_n + \sum_{0 \le j < n} \omega_{n,j} F_j, \qquad n \ge n_0$$

for some positive integer n_0, a function t_n, called the *toll function*, and a sequence of *weights* $\omega_{n,j} \ge 0$. The weights must satisfy two conditions:

1. $W_n = \sum_{0 \le j < n} \omega_{n,j} \ge 1$ (at least one recursive call).
2. $Z_n = \sum_{0 \le j < n} \frac{j}{n} \cdot \frac{\omega_{n,j}}{W_n} < 1$ (the size of the subinstances is a fraction of the size of the original instance).

The next step is to find a *shape function* $\omega(z)$, a continuous function approximating the discrete weights $\omega_{n,j}$.

The Continuous Master Theorem

Definition

Given the sequence of weights $\omega_{n,j}$, $\omega(z)$ is a shape function for that set of weights if

1. $\int_0^1 \omega(z)\, dz \geq 1$
2. there exists a constant $\rho > 0$ such that

$$\sum_{0 \leq j < n} \left| \omega_{n,j} - \int_{j/n}^{(j+1)/n} \omega(z)\, dz \right| = \mathcal{O}(n^{-\rho})$$

A simple trick that works very often:

$$\omega(z) = \lim_{n \to \infty} n \cdot \omega_{n,z \cdot n}$$

The Continuous Master Theorem

Theorem (Roura, 1997)

Let F_n satisfy the recurrence

$$F_n = t_n + \sum_{0 \le j < n} \omega_{n,j} F_j,$$

with $t_n = \Theta(n^a (\log n)^b)$, for some constants $a \ge 0$ and $b > -1$, and let $\omega(z)$ be a shape function for the weights $\omega_{n,j}$. Let $\mathcal{H} = 1 - \int_0^1 \omega(z) z^a \, dz$ and $\mathcal{H}' = -(b+1) \int_0^1 \omega(z) z^a \ln z \, dz$. Then

$$F_n = \begin{cases} \frac{t_n}{\mathcal{H}} + o(t_n) & \text{if } \mathcal{H} > 0, \\ \frac{t_n}{\mathcal{H}'} \ln n + o(t_n \log n) & \text{if } \mathcal{H} = 0 \text{ and } \mathcal{H}' \neq 0, \\ \Theta(n^\alpha) & \text{if } \mathcal{H} < 0, \end{cases}$$

where $x = \alpha$ is the unique non-negative solution of the equation

$$1 - \int_0^1 \omega(z) z^x \, dz = 0.$$

Analyzing median K-d trees

Example:
$I_n =$ expected internal path lenght of a random median K-d tree

$$I_n = n - 1 + \sum_{0 \le j < n} \pi_{n,j} \cdot (I_j + I_n), I_0 = 0$$

where $\pi_{n,j}$ is the probability that the left subtree of a random median K-d tree of size n is of size j, $0 \le j < n$

$$\pi_{n,j} = \begin{cases} \frac{1}{n^K}\left[(2j+2)^K - (2j+1)^K\right] & \text{if } j < \lfloor n/2 \rfloor, \\ \frac{1}{n^K}\left[(2(n-j)-1)^K - (2(n-j)-2)^K\right] & \text{otherwise.} \end{cases}$$

CMT solves "easily" the complicated recurrence above with the shape function

$$\omega(z) = \begin{cases} K2^K z^{K-1} & \text{if } z \le 1/2, \\ K2^K (1-z)^{K-1} & \text{if } z \ge 1/2. \end{cases}$$

Analyzing hybrid median K-d trees

For hybrid median K-d trees you need to set up systems of divide-and-conquer recurrences.

Example:
$P_n^{(i,\ell)} =$ expected cost of a random PM in a random hybrid median K-d tree of size n such that there are only i ($1 \leq i \leq K$) possible choices for the discriminant at the root and ℓ of these i coordinates are specified in the query ($0 \leq \ell \leq s$)

Analyzing hybrid median K-d trees

If $i > 1$ and $0 < \ell < i$ then

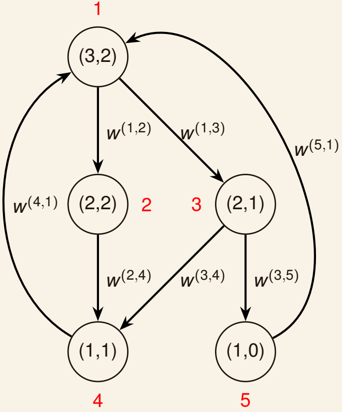
$$P_n^{(i,\ell)} = 1 + \frac{\ell}{i} \sum_{j=0}^{n-1} \left(\pi_{n,j}^{(i)} + \pi_{n,n-1-j}^{(i)}\right) \frac{j+1}{n+1} P_j^{(i-1,\ell-1)}$$
$$+ \frac{i-\ell}{i} \sum_{j=0}^{n-1} (\pi_{n,j}^{(i)} + \pi_{n,n-1-j}^{(i)}) P_j^{(i-1,\ell)},$$

with $\pi_{n,j}^{(i)}$ as in median K-d trees (but only i available coordinates, not K)

Other cases ($i = 1$, $i = \ell$, $\ell = 0$) are handled similarly

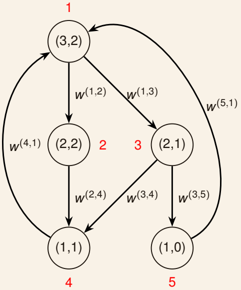
Analyzing hybrid median K-d trees

For example, with $K = 3$ and $s = 2$ we must set up a 5×5 system of linear recurrences and define an shape matrix $\mathbf{\Omega}$



$$\mathbf{\Omega} = \begin{array}{c c} & \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} & \begin{pmatrix} 0 & w^{(1,2)} & w^{(1,3)} & 0 & 0 \\ 0 & 0 & 0 & w^{(2,4)} & 0 \\ 0 & 0 & 0 & w^{(3,4)} & w^{(3,5)} \\ w^{(4,1)} & 0 & 0 & 0 & 0 \\ w^{(5,1)} & 0 & 0 & 0 & 0 \end{pmatrix} \end{array}$$

Analyzing hybrid median K-d trees



- $w^{(1,2)}$ = shape function for the weight $\frac{1}{3}(\pi^{(3)}_{n,j} + \pi^{(3)}_{n,n-1-j}) \to$ Algorithm #1 with cost $P^{(3,2)}_n$ calls recursively algorithm #2 with cost $P^{(2,2)}_j$
- $w^{(1,3)}$ = shape function for the weight $\frac{2}{3}\frac{j+1}{n+1}(\pi^{(3)}_{n,j} + \pi^{(3)}_{n,n-1-j}) \to$ Algorithm #1 ($P^{(3,2)}_n$) calls recursively algorithm #3 ($P^{(2,1)}_j$)
- ...

Conclusions and final remarks

- Both median and hybrid median K-d trees are simple and easy to implement, and neither requires significant extra space
- Both are more balanced than most other well known variants of K-d trees; their expected IPL is $\sim c_K n \ln n$ with $c_K < 2$ for all $K \geq 2$, and $c_K \to 1/\ln 2$ (optimal) as $K \to \infty$
- Their expected cost for PM is $\Theta(n^\alpha)$; for any s and $K \geq 2$ we have

$$1 - \frac{s}{K} \leq \alpha^{[\mathrm{hm}]} < \alpha^{[\mathrm{std}]} < \alpha^{[\mathrm{med}]} < \alpha^{[\mathrm{rlx}]} = \frac{1}{2}\left(\sqrt{9 - 8\frac{s}{K}} - 1\right)$$

Conclusions and final remarks

- Hybrid median K-d trees outperfom standard, median and relaxed K-d trees and we conjecture that they approach the optimal exponent $\alpha = 1 - s/K$ as K gets larger
- The special structure of the linear systems of recurrences for the IPL and RPM of hybrid median K-d trees can be exploited to find the constants c_K and the equations satisfied by the exponents $\alpha(s, K)$; we have developed a limited extension of the CMT to cope with these systems of recurrences
- This work is a new example of the power of the CMT as a fundamental tool in the analysis of algorithms, for example to analyze the expected cost of quicksort, quickselect, binary search trees, ... but it hasn't found its way into our algorithms textbooks yet 🤔

Please like ♥ and subscribe to my channel 🔔

just kidding... THANK YOU FOR YOUR ATTENTION!