Squarish k-d trees

Luc Devroye, Jean Jabbour and Carlos Zamora-Cura
School of Computer Science
McGill University
Montreal, Canada H3A 2K6
{luc, jabbour, czamora}@cs.mcgill.ca

Abstract. We modify the k-d tree on $[0,1]^d$ by always cutting the longest edge instead of rotating through the coordinates. This modification makes the expected time behavior of lower-dimensional partial match queries behave as for perfectly balanced complete k-d trees on n nodes. This is in contrast to a result of Flajolet and Puech, who proved that for (standard) random k-d trees with cuts that rotate among the coordinate axes, the expected time behavior was much worse than for balanced complete k-d trees. We also provide results for range searching and nearest neighbor search for our trees.

Keywords and phrases. k-d trees, partial match query, range search, expected time, probabilistic analysis of algorithms, data structures.

Computing Reviews categories: 3.74, 5.25, 5.5.

AMS Subject Classifications: 68P05, 68Q25, 60C05.

Preferred short title: squarish k-d trees

Contact: Luc Devroye, School of Computer Science, McGill University, Montreal, Canada H3A 2K6. Tel 9514) 398-3738. FAX: (514) 398-3883. Email: luc@cs.mcgill.ca.

Research of the authors was sponsored by NSERC Grant A3456 and by FCAR Grant 90-ER-0291. The second author was on leave from the Department of Mathematics, Université de Versailles, France. Address: School of Computer Science, McGill University, 3480 University Street, Montreal, Canada H3A 2K6. The third author received a DGAPA-UNAM Scholarship.

§1. Introduction

The k-d tree, or k-dimensional binary search tree, was proposed by Bentley (1975). In this paper, we propose a modification, the squarish k-d tree, and analyze its expected time performance for partial match queries, orthogonal range searching, and nearest neighbor search under the standard random model for the input (n points independently and uniformly distributed on the unit hypercube). We point out its superiority over the standard k-d tree for this model.

Bentley's k-d tree is a binary search tree that generalizes the 1-d tree or ordinary binary search tree to $\mathbb{R}^k$. A partition of space into hyper-rectangles is obtained by splitting alternating coordinate axes by hyper-planes through data points. Figure 1 shows the partition and the corresponding k-d tree. Insertion and search are implemented as for the standard binary search tree algorithms. These trees are used for a variety of other operations, including orthogonal range searching (report all points within a given rectangle), partial match queries (report all points whose values match a given k-dimensional vector with possibly a number of wildcards, e.g., we may search for all points with values $(a_1, *, *, a_4, a_5, *)$, where $*$ denotes a wild card). Additionally, nearest neighbor searching is greatly facilitated by k-d trees. For orthogonal range searching, a host of particular data structures have been developed, such as the range tree and variations or improvements of it (for surveys, see Bentley and Friedman, 1979; Bentley, 1979; Yao, 1990; Samet, 1990a, 1990b; and Agarwal, 1997). However, the k-d tree offers several advantages–it takes $O(kn)$ space for n data points, it is easily updated and maintained, it is simple to implement and comprehend, and it is useful for other operations besides orthogonal range search.

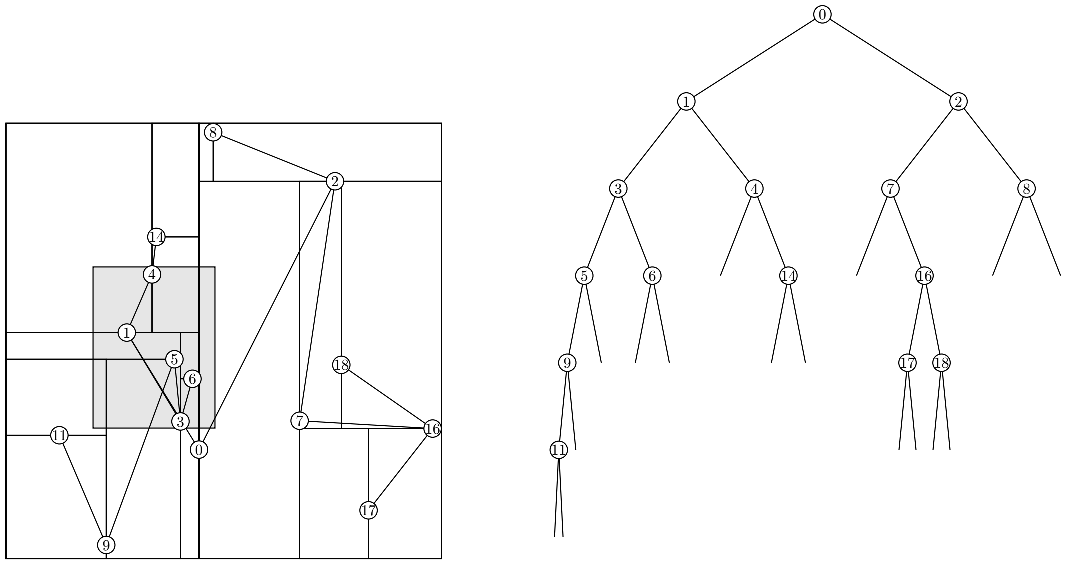

FIGURE 1. The k-d tree and its partition of the plane.
The query rectangle is shaded.

Bentley's orthogonal range search algorithm simply visits recursively all subtrees of the root that have a nonempty intersection with the query rectangle. In figure 1 for example, the left and right subtrees of the root are visited. Note that each node in the tree represents both a point of the data and a rectangle in the partition, namely the rectangle split by that point. Leaf regions thus have no points strictly in their interior. The query time for orthogonal search depends upon many factors, such as the location of the query rectangle, and the distribution of the points. One may construct a median k-d tree off-line by splitting each time about the median, thus obtaining a perfectly balanced binary tree, in which ordinary point search takes $\Theta(\log n)$ worst-case time, and a partial match query with s coordinates specified takes worst-case time $O(n^{1-s/k} + N)$, where N is the number of points returned (see, for example, Lee and Wong 1977).

For on-line insertion, balancing is notoriously difficult. If we assume that the data are independent and have a common distribution, then the expected query time is clearly of interest. For standard random binary search trees, it is known (Knuth, 1997; Pittel, 1984; Devroye, 1986, 1987; Mahmoud, 1992) that most properties of balanced search trees are inherited: the expected depth of a randomly selected node is about $2 \log n$ and the expected height is $O(\log n)$. One would hope that the random k-d tree, constructed by consecutive insertion of n data points, would also have a performance close to that of the median (off-line) k-d tree. Assuming that the data points are drawn from the uniform distribution on the unit k-cube, Flajolet and Puech (1986) showed that a random partial match query (carried out with s values also drawn uniformly and independently on $[0, 1]$, so, there are $k - s$ wild-cards) has expected time performance $\Theta(n^{1-s/k+\theta(s/k)})$, where $\theta(u)$ is a strictly positive function of $u \in (0, 1)$, with maximum not exceeding 0.07. Thus, random k-d trees behave a bit worse than their balanced counterparts, the median k-d trees. Surveys of related known probabilistic results are provided by Vitter and Flajolet (1990), and Gonnet and Baeza-Yates (1991).

FIGURE 2. Two random k-d tree partitions clearly show the elongated rectangles.

We propose a minor modification of the insertion procedure, namely, each time a rectangle is split by a newly inserted leaf point, the longest side of its rectangle is cut, that is, the cut is a $(k-1)$-dimensional hyper-plane through the new point perpendicular to the longest edge of the rectangle. It was shown by Chanzy, Devroye, and Zamora-Cura (1999) that elongated rectangles explain the poor performance of random k-d trees. In this paper, we show to what extent the proposed k-d trees have more squarish-looking rectangles, and will therefore call these trees **squarish k-d trees**. For the probabilistic model of Flajolet and Puech, it will be shown that the expected time for a partial match query is $\Theta(n^{1-s/k})$, just as for random median k-d trees. Furthermore, the expected complexity of any orthogonal range search in median k-d trees is asymptotically equivalent to that for the simple squarish k-d trees proposed here.

In the last part of the paper, we deal with orthogonal range search in general when the query rectangles may have dimensions that depend upon n in an arbitrary fashion. The proofs are probabilistic, rather than analytical, and do not offer explicit constants for expected times but only $\Theta(\cdot)$ results. However, they are short and explain many of the phenomena at work. Interestingly, very little probability beyond Hölder's inequality is needed. We conclude the paper by showing that a natural nearest neighbor search (with a randomly selected probe point) takes $O(\log n \log\log n)$ expected time in any dimension.

We should note that there are indeed more sophisticated data structures for some of the subproblems dealt with here. For example, if one is just interested in partial match queries, then one could just make j-d trees for each of the $2^k - 1$ nonempty subsets of size j of the k coordinates separately, so that search in the proper tree is just a point search, taking expected worst-case time $O(\log n)$, while the space used is still $O(n2^k)$. However, these would not be helpful for general orthogonal, simplex, or convex range searches. For an analysis of range search based on multiattribute trees see Gardy, Flajolet, and Puech (1989).

§2. The random processes

In this section, we will try to explain the differences between alternating cuts and longest-edge cuts in sequences of randomly cut rectangles. To explain the processes at work, we consider the following simplification of our problem: in $\mathbb{R}^2$, start with a rectangle with one vertex permanently pegged at the origin and the opposite one at $(1,1)$, and let (U_n, V_n) denote the coordinates of the top right vertex after n iterations, with $(U_0, V_0) = (1,1)$. The rectangle will be reduced in size, first by alternating uniform cuts, that is, if $Z_1, Z_2, \ldots$ are i.i.d. uniform $[0,1]$ random variables, then we set

$$(U_n, V_n) = \begin{cases} (Z_n U_{n-1}, V_{n-1}) & \text{if } n \text{ is odd;} \\ (U_{n-1}, Z_n V_{n-1}) & \text{if } n \text{ is even.} \end{cases}$$

If we denote by $\stackrel{\mathcal{L}}{=}$ equality in distribution, clearly, at time $2n$, we have

$$U_{2n} \stackrel{\mathcal{L}}{=} V_{2n} \stackrel{\mathcal{L}}{=} \prod_{i=1}^{n} Z_i \stackrel{\mathcal{L}}{=} e^{-\sum_{i=1}^{n} E_i} \stackrel{\mathcal{L}}{=} e^{-G_n}$$

where the E_i are independent exponential random variables, and G_n denotes a gamma random variable with parameter n. As U_k and V_j are independent of each other for all k, j, we see that the ratio

$$\frac{U_{2n}}{V_{2n}} \stackrel{\mathcal{L}}{=} e^{G_n - G'_n}$$

where G_n, G'_n are i.i.d. gamma random variables. By the central limit theorem, it is easy to see then that

$$\frac{1}{\sqrt{n}} \log\left(\frac{U_{2n}}{V_{2n}}\right) \xrightarrow{\mathcal{L}} \mathcal{N} - \mathcal{N}',$$

a difference of two independent standard normal random variables. Thus, the raw ratio behaves asymptotically like $\exp(\sqrt{n}(\mathcal{N} - \mathcal{N}'))$, and thus exhibits wide swings. In fact, if we stop at a large value for n, the rectangle will look very skinny indeed (see figure 2).

Since we would like to preserve squarish rectangles, we may opt instead to always cut the longest side of the rectangle. More formally, with notation as above, $(U_0, V_0) = (1, 1)$, we have

$$(U_n, V_n) = \begin{cases} (Z_n U_{n-1}, V_{n-1}) & \text{if } U_{n-1} > V_{n-1}; \\ (U_{n-1}, Z_n V_{n-1}) & \text{if } U_{n-1} < V_{n-1}. \end{cases}$$

In case of equality $U_{n-1} = V_{n-1}$, which only occurs at $n = 1$, we flip a perfect coin and pick an edge to cut at random.

Lemma 1. *With the longest-edge cutting method, the sequence U_n/V_n, $n \geq 1$, is identically distributed. The common distribution is that of Z_1/Z_2, the ratio of two independent uniform $[0, 1]$ random variables.*

PROOF. Clearly, U_1/V_1 is distributed as Z_1 with probability $1/2$ and as $1/Z_1$ otherwise. It is easy to verify that this has the required density $1/(2\max(z, 1))^2$, $z > 0$. By induction, we need to show that if Z_1, Z_2, Z are i.i.d. uniform $[0, 1]$ random variables, then the random variable $ZZ_1/Z_2 I_{Z_1 > Z_2} + Z_1/(ZZ_2) I_{Z_1 < Z_2}$ is in turn distributed as Z_1/Z_2. This can be done by standard calculations, or even the method of characteristic functions. However, by far the quickest way to see this is by embedding. We note that Z_1/Z_2 is distributed as the random variable Z_4^S where $S = 1$ and $S = -1$ with equal probability, and Z_4 is another uniform $[0, 1]$ random variable. The case $Z_1 > Z_2$ corresponds to $S = -1$, and thus, we see that $ZZ_1/Z_2 I_{Z_1 > Z_2} + Z_1/(ZZ_2) I_{Z_1 < Z_2}$ is distributed as $(Z_4/Z)^S$, which was to be shown, as S is independent of Z and Z_4. □

Lemma 1 shows that cutting the longest edge is extremely stabilizing. Nevertheless, as U_n/V_n has Cauchy-like tails, its mean does not exist, and we will often see skinny rectangles, although by and large, the rectangles will be rather squarish (see figures 3 and 4). The above observations explain why the squarish k-d trees are useful. Our analysis is of course more involved, as rectangles participate in an evolving collection of rectangles, with very intricate dependencies. As soon as a rectangle becomes too small, it is unlikely to be picked again soon, and thus, the ratio of the sides of the rectangles must be considered in conjunction with the sizes. For this, we introduce a few new analysis methods.

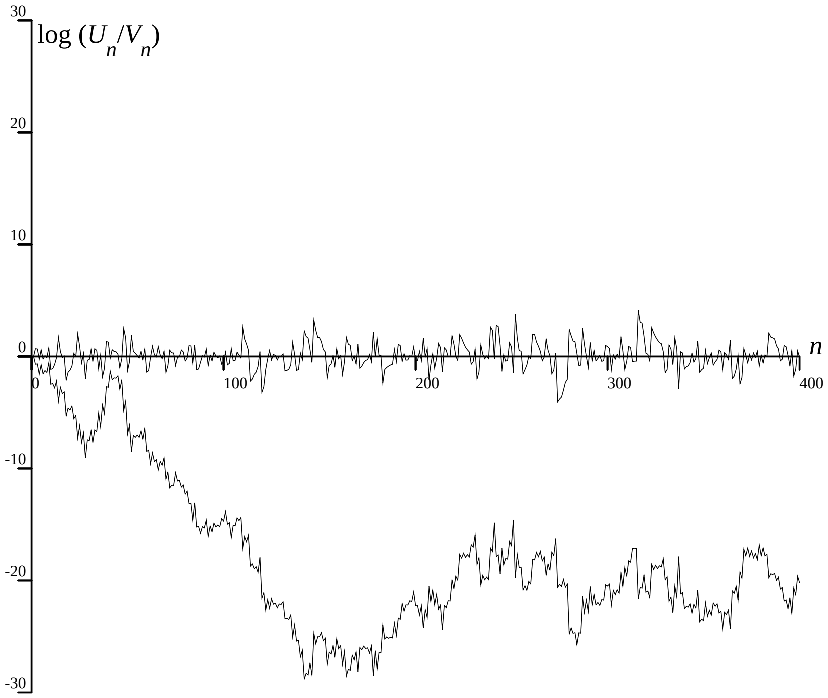

FIGURE 3. For the two processes above, $\log(U_n/V_n)$ is plotted versus n. The alternating cuts process wanders off just as a random walk. The largest edge cut strategy induces a sequence U_n/V_n that hovers near one and remains stationary.

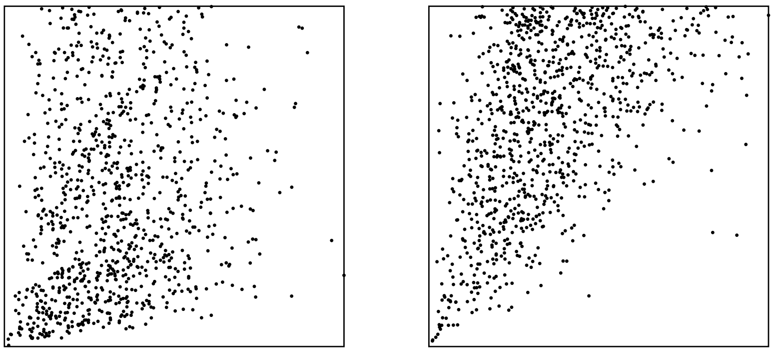

FIGURE 4. For the two processes above, let L_n and S_n be the long and short dimensions of a leaf rectangle in the 2-d partition. The values $\sqrt{L_n/S_n}$ are plotted versus $\sqrt{L_nS_n}$ (normalized so that the largest value is one) for the two k-d trees. For squarish k-d trees (on the right), there are many more rectangles in which L_n and S_n are close. And nearly all big rectangles are squarish. For the ordinary random 2-d tree, on the left, most rectangles have very small edge ratios.

§3. Notation and preliminaries

In k-d trees, nodes represent rectangular regions. Bentley's algorithm for orthogonal range search and partial match queries starts at the root of a k-d tree and recursively visits all subtrees that have a nonempty overlap with the rectangular regions of the children, and reports all points that fall in the search region. Let $u_1, u_2, \ldots, u_n$, $n \geq 1$, denote the nodes in the k-d tree, and let $U_1, \ldots, U_n$ denote the data points, which are i.i.d. and uniformly distributed on $[0,1]^k$. Thus, U_i is the data point corresponding to u_i. The rectangle split by u_i is R_i. Thus, $R_1 = [0,1]^k$. Let $|R_i|$ denote the volume of rectangle i. The $n+1$ leaf rectangles (the dangling edges in figure 1) are also denoted R_i, with the index i now running from $n+1$ to $2n+1$. The collection of rectangles is denoted by $\mathcal{R}_n$. The collection of the indices of the $n+1$ final rectangles is $\mathcal{F}_n$. We will denote by T the k-d tree constructed by inserting successively $u_1, u_2, \ldots, u_n$ into an initially empty k-d tree. Given a node u in T, we will denote by T_u the subtree of T rooted at u. With rotating coordinate cuts, such a tree is called a **random k-d tree**. With our method of cutting the largest edges, it will be called a **random squarish k-d tree**.

The dimensions of rectangle R_i are $X_{ij}, 1 \leq j \leq k$. For 2-d trees, we will use the lighter notation X_i, Y_i for the x and y dimensions of R_i. The query rectangle Q is $Z + [-m_1, m_1] \times \cdots \times [-m_k, m_k]$, $m_i \geq 0$ for all i, where the m_i's are fixed (that is, they may depend upon n only) and Z is uniformly distributed on $[0,1]^k$ and independent of $(U_1, \ldots, U_n)$. Bentley's range search applied to Q is called a **random orthogonal range search**. Note that a node u_i is visited by the range search algorithm if and only if the query rectangle Q intersects R_i. Any rectangle R_i is visited if and only if it intersects Q. Let N_n be the time complexity of Bentley's orthogonal range search. Then,

$$N_n = \sum_{i=1}^{2n+1} \mathbb{1}_{[R_i \cap Q \neq \emptyset]},$$

where $\mathbb{1}_{[A]}$ is the characteristic function of the event A. This quantity will be analyzed further on for random squarish k-d trees.

In a **random partial match query**, we specify a subset of s dimensions, $j_1, \ldots, j_s$, and perform an orthogonal range query with the i-th interval in the rectangle either $\{Z_i\}$ (a uniform random number on $[0,1]$) if $i \in \{j_1, \ldots, j_s\}$, or $(-\infty, \infty)$ otherwise. It is assumed that the Z_i's are independent, and independent of $(U_1, \ldots, U_n)$. In this paper, we first study random partial match queries for random squarish k-d trees and obtain results that should be compared against the following result for random k-d trees:

Theorem 1 (Flajolet and Puech, 1986). *For a random k-d tree and a random partial match query, in which s of the k fields are specified with $k > s \geq 0$, let $N_n^{(s)}$ be the number of comparisons that Bentley's orthogonal range search performs. Define*

$$\alpha(u) = \max_{0 \leq t \leq 1} \left\{ t + 2\left(\frac{1-t}{1-u}\right)^{1-u} \left(\frac{t}{u}\right)^{u} - 2 \right\} \;, \; 0 < u < 1 \;,$$

and note in particular that α is decreasing on $(0,1)$, $\alpha(0)=1$, and that $1-u<\alpha(u)<1.07-u$, $0<u<1$. Then, the expected value of $N_n^{(s)}$ is such that

$$\mathbf{E}\left\{N_n^{(s)}\right\}=(c+o(1))n^{\alpha(s/k)}\ ,$$

where c is a constant depending on the indices of the s fixed coordinates.

The following proposition is useful in relating partial random partial match queries to the range search problem.

Proposition 1. *Given is a random k-d tree based on i.i.d. random variables $U_1,\ldots,U_n$, distributed uniformly on $[0,1]^k$. Consider a random partial match query, in which $s\geq 0$ of the k fields are specified. Let $N_n^{(s)}$ be the number of comparisons that Bentley's orthogonal range search performs. Let S be the set of specified coordinates. Then*

$$\mathbf{E}\left\{N_n^{(s)}\right\}=\mathbf{E}\left\{\sum_{i=1}^{2n+1}\prod_{j\in S}X_{ij}\right\},$$

where $X_{ij}, 1\leq j\leq k$, are the lengths of the sides of rectangle R_i in $\mathcal{R}_n$.

PROOF. Let Q be the query rectangle. Note that $\mathbf{P}\left\{Q\cap R_i\neq\emptyset \,|\, U_1,\ldots,U_n\right\}=\prod_{j\in S}X_{ij}$. Thus we have,

$$\mathbf{E}\left\{N_n^{(s)}\right\}=\mathbf{E}\left\{\sum_{i=1}^{2n+1}\mathbb{1}_{[Q\cap R_i\neq\emptyset]}\right\}=\sum_{i=1}^{2n+1}\mathbf{P}\left\{Q\cap R_i\neq\emptyset\right\}=\mathbf{E}\left\{\sum_{i=1}^{2n+1}\prod_{j\in S}X_{ij}\right\}.\ \square$$

The next observation is important. It follows immediately by considering the random growth of our k-d trees. And of course, it implies that the joint distribution of the ordered volumes of the $n+1$ leaf rectangles is identical for both random k-d trees considered here!

Lemma 3. *Consider a random k-d tree or a random squarish k-d tree. Then, the volumes of the rectangles in $\mathcal{F}_n$ are distributed as the set $\mathcal{V}_n$ of the consecutive spacings between the order statistics of n i.i.d. random variables, uniformly distributed on $[0,1]$.*

§4. Random partial match queries with squarish 2-d trees

In a vertical random partial match query on a 2-d tree, we take a uniformly distributed value Z, and visit all nodes in the tree whose rectangle cuts the vertical line at Z. Horizontal partial match queries on 2-d trees are defined in an analogous manner. The probability of hitting a rectangle with dimensions $X_i\times Y_i$ is of course X_i, so that the expected number of nodes visited, and hence, the expected time for a partial match query, is simply $\mathbf{E}\left\{\sum_{i=1}^{2n+1}X_i\right\}$, where the sum is taken over all $2n+1$ rectangles in the partition. A similar formula holds of course for horizontal partial match queries. In this section, we prove that a random partial match query in a random squarish 2-d tree takes expected time $\Theta(\sqrt{n})$ as opposed to $\Theta(n^{0.5616})$ for random 2-d trees (see Theorem 1).

Theorem 2. *For random squarish 2-d trees,*

$$\frac{\sqrt{\pi n}}{3} \le \mathbf{E}\left\{\sum_{i=1}^{2n+1} Y_i\right\} \le 180\sqrt{n}.$$

The same result holds for $\mathbf{E}\left\{\sum_{i=1}^{2n+1} X_i\right\}$. *Hence, the expected time for a random partial match query is* $\Theta(\sqrt{n})$.

Of course, no attempt was made to optimize the constants. A few technical results will be needed in the sequel. In particular, the next lemma is valid for squarish k-d trees in arbitrary dimension.

Lemma 4. *For random squarish k-d trees, let* $p \ge 0, n \ge 1$, *then*

$$\left(\frac{1}{1+p}\right)^{\lfloor p\rfloor+1}\frac{\Gamma(p+1)}{n^{p-1}} \le \mathbf{E}\left\{\sum_{i\in\mathcal{F}_n} |R_i|^p\right\} \le \frac{4\Gamma(p+1)}{n^{p-1}},$$

for all n.

Proof. Let $V_1, \ldots, V_{n+1}$ be the spacings induced by n independent uniformly distributed random variables on $[0,1]$. It is known that $V_i \stackrel{\mathcal{L}}{=} \mathrm{Beta}(1,n)$. Thus, by Lemma 3, with $\mathrm{B}(s,t) = \frac{\Gamma(s)\Gamma(t)}{\Gamma(s+t)}$,

$$\begin{aligned}\mathbf{E}\left\{\sum_{i\in\mathcal{F}_n} |R_i|^p\right\} &= \mathbf{E}\left\{\sum_{i=1}^{n+1} V_i^p\right\} = \sum_{i=1}^{n+1}\int_0^1 v^p\frac{(1-v)^{n-1}}{\mathrm{B}(1,n)}dv \\ &= (n+1)\frac{\mathrm{B}(p+1,n)}{\mathrm{B}(1,n)} = \Gamma(p+1)\frac{\Gamma(n+2)}{\Gamma(p+n+1)}\ .\end{aligned}$$

Now, $\Gamma(x+1) = x\Gamma(x)$ for any $x > 0$, and for any natural number n and any $s \in [0,1]$, $n^{1-s} \le \Gamma(n+1)/\Gamma(n+s) \le (n+1)^{1-s}$ (see Mitrinović, 1970). Thus,

$$\begin{aligned}\mathbf{E}\left\{\sum_{i\in\mathcal{F}_n} |R_i|^p\right\} &= \Gamma(p+1)(n+1)\frac{\Gamma(n+1)}{(n+p)\cdots(n+p-\lfloor p\rfloor)\Gamma(n+p-\lfloor p\rfloor)} \\ &\le \frac{\Gamma(p+1)(n+1)^{2-p+\lfloor p\rfloor}}{n^{\lfloor p\rfloor+1}} \\ &= \frac{\Gamma(p+1)}{n^{p-1}}\left(\frac{n+1}{n}\right)^{2+\lfloor p\rfloor-p} \\ &\le \frac{4\Gamma(p+1)}{n^{p-1}},\end{aligned}$$

as $2 + \lfloor p \rfloor - p \le 2$. Now, for the lower bound, note that

$$
\begin{aligned}
\mathbf{E}\left\{\sum_{i\in\mathcal{F}_n} |R_i|^p\right\} &= \Gamma(p+1)(n+1)\frac{\Gamma(n+1)}{(n+p)\cdots(n+p-\lfloor p\rfloor)\Gamma(n+p-\lfloor p\rfloor)} \\
&\ge \frac{\Gamma(p+1)}{n^{p-1}}\frac{n^{\lfloor p\rfloor+1}}{(n+p)\cdots(n+p-\rfloor p\lfloor)} \\
&\ge \frac{\Gamma(p+1)}{n^{p-1}}\left(\frac{n}{n+p}\right)^{\lfloor p\rfloor+1} \\
&\ge \frac{\Gamma(p+1)}{n^{p-1}}\left(\frac{1}{1+p}\right)^{\lfloor p\rfloor+1}. \ \square
\end{aligned}
$$

Lemma 5. *In a random squarish 2-d tree constructed from the insertion of $U_1, \ldots, U_n$ independent and uniformly distributed random vectors on $[0,1]^2$ we have that for every $q \ge 1$,*

$$
\mathbf{E}\left\{\sum_{i\in\mathcal{F}_n} Y_i^q\right\} \le \begin{cases} \frac{8}{1-q/2}n^{1-q/2}, & \text{for } q\in[1,2); \\ 8e\log n, & \text{for } 2-\frac{2}{\log n}\le q\le 2; \\ \frac{5\Gamma(q/2+1)}{q/2-1}\left(\frac{q}{2}-\frac{1}{n^{q/2-1}}\right), & \text{for } q>2, \end{cases}
$$

and for $q \in [1,2)$,

$$
\mathbf{E}\left\{\sum_{i\in\mathcal{F}_n} Y_i^q\right\} \ge \left(\frac{1}{q/2+1}\right)^{\lfloor q/2\rfloor+1}\Gamma(q/2+1)n^{1-q/2}.
$$

The same result holds for $\mathbf{E}\left\{\sum_{i\in\mathcal{F}_n} X_i^q\right\}$.

Proof. Let $r > 1$, and define $S_r^{(q)} = \sum_{i\in\mathcal{F}_r} Y_i^q$. Note that, given $U_1, \ldots, U_r$, $S_{r+1}^{(q)} - S_r^{(q)}$ is distributed as Y^q when $X > Y$ and as $Y^q(U^q + (1-U)^q - 1)$ when $X \le Y$, where U is a uniform $[0,1]$ random variable, and (X,Y) are the dimensions of the rectangle split when U_{r+1} is added. Thus,

$$
\mathbf{E}\left\{S_{r+1}^{(q)} - S_r^{(q)}\right\} = \mathbf{E}\left\{\sum_{i\in\mathcal{F}_r} X_iY_i\left(\mathbb{1}_{[X_i>Y_i]}Y_i^q + \mathbb{1}_{[X_i<Y_i]}Y_i^q\left(U^q+(1-U)^q-1\right)\right)\right\}.
$$

Notice that $U^q + (1-U)^q - 1 \le 0$ for $q \ge 1$, and as $\min\{a,b\} \le \sqrt{ab}$, for $a, b \ge 0$, then by Lemmas 3 and 4,

$$
\begin{aligned}
\mathbf{E}\left\{S_{r+1}^{(q)} - S_r^{(q)}\right\} &\le \mathbf{E}\left\{\sum_{i\in\mathcal{F}_r} X_iY_i\left(\mathbb{1}_{[X_i>Y_i]}Y_i^q\right)\right\} \\
&\le \mathbf{E}\left\{\sum_{i\in\mathcal{F}_r} (X_iY_i)^{q/2+1}\right\} \le \frac{4\Gamma(q/2+2)}{r^{q/2}}.
\end{aligned}
$$

By summing the differences we get,

$$
\begin{aligned}
\mathbf{E}\left\{S_n^{(q)}\right\} &= \mathbf{E}\left\{\sum_{r=1}^{n-1}\left(S_{r+1}^{(q)} - S_r^{(q)}\right) + S_1^{(q)}\right\} \\
&\le \sum_{r=1}^{n-1} \frac{4\Gamma(q/2+2)}{r^{q/2}} + 2 \\
&\le 2 + 4\Gamma(q/2+2)\left(1 + \int_1^{n-1} \frac{1}{x^{q/2}} dx\right) \\
&\le \begin{cases} 10 + \frac{4\Gamma(q/2+2)}{1-q/2}(n^{1-q/2} - 1); & (q \in [1,2)) \\ 5\Gamma(q/2+2) + \frac{4\Gamma(q/2+2)}{q/2-1}(1 - n^{1-q/2}) & (q > 2) \end{cases} \\
&\le \begin{cases} \frac{8}{1-q/2} n^{1-q/2}; & (q \in [1,2)) \\ \frac{5\Gamma(q/2+2)}{q/2-1}\left(\frac{q}{2} - n^{1-q/2}\right) & (q > 2). \end{cases}
\end{aligned}
$$

Because $\frac{8}{1-q/2}n^{1-q/2}$, as a function of q, reaches its minimum at $q_0 = 2(1 - 1/\log n)$, and $\mathbf{E}\left\{S_n^{(q)}\right\}$ is a decreasing function of q, we have that $\mathbf{E}\left\{S_n^{(q)}\right\} \le 8e \log n$, for $q_0 \le q \le 2$. The result for $\mathbf{E}\left\{\sum_{i\in\mathcal{F}_n} X_i^q\right\}$ can be obtained similarly just by replacing the y-lengths for the x-lengths in the appropriate places.

Now, for the lower bound, note that as the X_i's and the Y_i's are identically distributed:

$$
\begin{aligned}
\mathbf{E}\left\{\sum_{i\in\mathcal{F}_n} Y_i^q\right\} &= \frac{1}{2}\mathbf{E}\left\{\sum_{i\in\mathcal{F}_n}(Y_i^q + X_i^q)\right\} \\
&\ge \mathbf{E}\left\{\sum_{i\in\mathcal{F}_n}(Y_iX_i)^{q/2}\right\} \\
&\ge \left(\frac{1}{q/2+1}\right)^{\lfloor q/2\rfloor+1} \frac{\Gamma(q/2+1)}{n^{q/2-1}},
\end{aligned}
$$

by Lemma 4, for $q \in [1,2)$. □

Proof of Theorem 2. Note that the lower bound follows from Lemma 5, as $\mathbf{E}\left\{\sum_{i\in\mathcal{F}_n} Y_i\right\}$ is less than $\mathbf{E}\left\{\sum_{i=1}^{2n+1} Y_i\right\}$. For the upper bound we will use the same technique as in the proof of Lemma 5. Let $S_n = \sum_{i=1}^{2n+1} Y_i$. Note that as the sum is over all the rectangles generated by $U_1, \ldots, U_n$, we have now that for $r \ge 1$, as X_i and Y_i are identically distributed,

$$
\begin{aligned}
\mathbf{E}\{S_{r+1} - S_r\} &= \mathbf{E}\left\{\sum_{i\in\mathcal{F}_r} X_iY_i\left(\mathbb{1}_{[X_i>Y_i]}2Y_i + \mathbb{1}_{[X_i<Y_i]}(Y_iU + Y_i(1-U))\right)\right\} \\
&= 3\,\mathbf{E}\left\{\sum_{i\in\mathcal{F}_r} X_iY_i^2\right\}
\end{aligned}
$$

where $U \stackrel{\mathcal{L}}{=} \mathrm{Uniform}[0,1]$, and independent of all $U_1, \ldots, U_n$. Let $q \in (1,2)$ and $p > 1$ such that $\frac{1}{p} + \frac{1}{q} = 1$, then by Hölder's inequality used twice,

$$\mathbf{E}\left\{\sum_{i\in\mathcal{F}_r} X_i Y_i^2\right\} \le \mathbf{E}\left\{\sum_{i\in\mathcal{F}_r} (X_i Y_i)^p\right\}^{1/p} \mathbf{E}\left\{\sum_{i\in\mathcal{F}_r} Y_i^q\right\}^{1/q}$$

$$\le \left(\frac{4\Gamma(p+1)}{r^{p-1}}\right)^{1/p} \left(\frac{8}{1-q/2}\frac{1}{r^{q/2-1}}\right)^{1/q}$$

by Lemmas 4 and 5. Take $p = 3$, $q = 3/2$, and verify that the upper bound is not more than $24^{1/3}32^{2/3}/\sqrt{r} < 30/\sqrt{r}$. By summing the differences we finally obtain

$$\mathbf{E}\left\{\sum_{i=1}^{2n+1} Y_i\right\} \le \frac{5}{2} + 90\sum_{r=1}^{n-1}\frac{1}{\sqrt{r}} \le \frac{5}{2} + 90(2\sqrt{n-1} - 1) \le 180\sqrt{n}.$$

The result for $\mathbf{E}\left\{\sum_{i=1}^{2n+1} X_i\right\}$ can be obtained similarly just by replacing the y-lengths for the x-lengths in the appropriate places. □

§5. The k-dimensional case.

In this section, we obtain the k dimensional generalization of the results in the previous section by induction. Given $U_1, \ldots, U_n$, we define for each $R_i \in \mathcal{R}_n$, $X_i^* = \max_{j=1,\ldots,k} X_{ij}$ and j_i^* as the index $j \in \{1, \ldots, n\}$ for which $X_{ij} = X_i^*$. Note that j_i^* is unique w.p.1. Our main result generalizes Theorem 2 and establishes the expected time optimality of random squarish k-d trees.

Theorem 3. *Consider a random squarish k-d tree. For $\ell \in \{1, \ldots, k-1\}$, there exist $C, C' > 0$ such that*

$$C'n^{1-\frac{\ell}{k}} \le \mathbf{E}\left\{\sum_{i=1}^{2n+1}\prod_{j\in I} X_{ij}\right\} \le Cn^{1-\frac{\ell}{k}},$$

for any $I \subseteq \{1, \ldots, k\}$ of cardinality ℓ and all $n \in \mathbb{N}$. In particular, by proposition 1 the expected time of a random partial match query with s specified coordinates is $\Theta(n^{1-s/k})$.

The next lemma complements Theorem 3 when $\ell = k$.

Lemma 7. *Let $U_1, \ldots, U_n$ be independent uniformly distributed random variables over $[0,1]^k$. Let $\mathcal{R}_n = \{R_1, R_2, \ldots, R_{2n+1}\}$ be the hyper-rectangles in the partition defined by the random squarish k-d tree based on $U_1, \ldots, U_n$. Let X_{ij} be the length on the j^{th} coordinate of the i^{th} hyper-rectangle. Then,*

$$\mathbf{E}\left\{\sum_{i=1}^{2n+1} X_{i1}\cdots X_{ik}\right\} = 2h_{n+1} - 1,$$

where h_n is the n^{th} harmonic number.

We prove the following lemma that will allow us to prove the lower bound in the previous theorem.

Lemma 8. *Let $\ell \in \{1,\ldots,k\}$, then for every $x_1,\ldots,x_k > 0$,*

$$\left(\prod_{j=1}^{k} x_j\right)^{\frac{1}{k}} \le \max_{\substack{I:\ I\subseteq\{1,\ldots,k\}\\ |I|=\ell}} \left(\prod_{j\in I} x_j\right)^{\frac{1}{\ell}} .$$

Proof. Let I^* be the subset of $\{1,\ldots,k\}$ of cardinality ℓ for which the maximum above is reached. It suffices to observe that,

$$\left(\prod_{j=1}^{k} x_j\right)^{\ell} = \prod_{s=1}^{k}\left(\prod_{j=s}^{s+\ell-1} x_j\right) \le \prod_{s=0}^{k-1}\left(\prod_{j\in I^*} x_j\right) = \left(\prod_{j\in I^*} x_j\right)^{k},$$

where the subindice j must be understood as (j mod k), if $j > k$. $\square$

Proposition 2. *Let $I \subseteq \{1,\ldots,k\}$ of cardinality $\ell \in \{1,\ldots,k\}$ and $p \in [1,\frac{k}{\ell})$, then there are positive constants C and C' such that*

$$C' n^{1-p\frac{\ell}{k}} \le \mathbf{E}\left\{\sum_{i\in\mathcal{F}_n}\left(\prod_{j\in I} X_{ij}\right)^{p}\right\} \le C n^{1-p\frac{\ell}{k}},$$

for all $n \in \mathbb{N}$.

Proof. For $I \subseteq \{1,\ldots,k\}$ with $|I| = \ell$, we define

$$S_r^{I,p} = \sum_{i\in\mathcal{F}_r}\left(\prod_{j\in I} X_{ij}\right)^{p} .$$

We first look at the upper bound. We define recursively the constants $C_k(\ell,p)$ for any integer $k > 0$, $\ell \in \{1,\ldots,k\}$ and real number $p \in [1,\frac{k}{\ell})$ as follows,

$$C_k(\ell,p) = \begin{cases} 4\Gamma(p+1), & \text{if } \ell = k; \\ (k-\ell)\left(\frac{1}{1-\frac{p\ell}{k}}\right) C_k(k,\tilde{q})^{1/\tilde{q}} C_k(\ell+1, p\tilde{p}\ell/(\ell+1))^{1/\tilde{p}} + 2 & \text{if } \ell < k, \end{cases}$$

where $\tilde{p},\tilde{q} > 1$ depend on p,k and ℓ, they are such that $\frac{1}{\tilde{p}} + \frac{1}{\tilde{q}} = 1$, and $1 \le p\tilde{p}\frac{\ell}{\ell+1} < \frac{k}{\ell+1}$. For the sake of clarity we will choose $\tilde{p}$ later.

For $\ell \in \{2,\ldots,k\}$, we define the hypothesis $\mathcal{H}_\ell$ stating that the upper bound holds for all $n \in \mathbb{N}$, all $I \subseteq \{1,\ldots,k\}$ such that $|I| = \ell$, and all $p \in [1,\frac{k}{\ell})$, with constant $C_k(\ell,p)$. We will prove $\mathcal{H}_\ell$ with an inductive argument. First, note that $\mathcal{H}_k$ holds by Lemma 4. Assuming that $\mathcal{H}_\ell$ is true, we will prove $\mathcal{H}_{\ell-1}$. Let $I \subseteq \{1,\ldots,k\}$ such that $\ell - 1 = |I| \ge 1$, and $p \in [1,\frac{k}{\ell-1})$. Then for any integer $r \ge 1$ we have,

$$\begin{aligned}\mathbf{E}\left\{S_{r+1}^{I,p} - S_r^{I,p} | U_1,\ldots,U_r\right\} = \sum_{i\in\mathcal{F}_r}\left(\prod_{j=1}^{k} X_{ij}\right)&\left\{\mathbb{1}_{[j_i^*\notin I]}\left(\prod_{j\in I} X_{ij}\right)^{p}\right. \\ &\left. +\mathbb{1}_{[j_i^*\in I]}\left(\prod_{j\in I} X_{ij}\right)^{p}\int_0^1 (x^p + (1-x)^p - 1)dx\right\},\end{aligned}$$

as we are using the longest edge cut method. Since $\int_0^1 (x^p + (1-x)^p - 1)dx \le 0$ for any $p \ge 1$, we can drop the second term above and take expected values so that,

$$\mathbf{E}\left\{S_{r+1}^{I,p} - S_r^{I,p}\right\} \le \sum_{t \notin I} \mathbf{E}\left\{\sum_{i \in \mathcal{F}_r} \left(\prod_{j=1}^k X_{ij}\right) \mathbb{1}_{[j_i^*=t]} \left(\prod_{j \in I} X_{ij}\right)^p\right\}.$$

Let us denote by $E(t)$ the expected value of the t^{th} term above. Observe that $\mathbb{1}_{[j_i^*=t]} X_{ij} \le X_{ij}^{\frac{\ell-1}{\ell}} X_{it}^{\frac{1}{\ell}}$. Thus we can bound each $E(t)$ as follows,

$$E(t) \le \mathbf{E}\left\{\sum_{i \in \mathcal{F}_r} \left(\prod_{j=1}^k X_{ij}\right) \left(\prod_{j \in I \cup \{t\}} X_{ij}\right)^{\frac{\ell-1}{\ell}p}\right\}.$$

Now, for any $\tilde{p}, \tilde{q} > 1$ such that $\frac{1}{\tilde{p}} + \frac{1}{\tilde{q}} = 1$, we have by applying Hölder's inequality twice that,

$$E(t) \le \mathbf{E}\left\{\sum_{i \in \mathcal{F}_r} \left(\prod_{j=1}^k X_{ij}\right)^{\tilde{q}}\right\}^{\frac{1}{\tilde{q}}} \mathbf{E}\left\{\sum_{i \in \mathcal{F}_r} \left(\prod_{j \in I \cup \{t\}} X_{ij}\right)^{\frac{\ell-1}{\ell}p\tilde{p}}\right\}^{\frac{1}{\tilde{p}}}.$$

We can apply hypothesis $\mathcal{H}_\ell$ to bound the second term above, if we can choose $\tilde{p} > 1$ such that $p\tilde{p}\frac{\ell-1}{\ell} \in [1, k/\ell)$. Note that $\frac{k}{p(\ell-1)} > 1$, as $p \in [1, \frac{k}{\ell-1})$. Let us define $\tilde{p} = \max\left\{\sqrt{k/p(\ell-1)}, \frac{\ell}{(\ell-1)p}\right\}$, so that $\tilde{p} > 1$, yet $1 \le p\tilde{p}\frac{\ell-1}{\ell} < \frac{k}{\ell}$. This completely defines the constant $C_k(\ell, p)$. We can therefore use hypothesis $\mathcal{H}_\ell$ and obtain,

$$\begin{aligned} E(t) &\le \left(\frac{C_k(k,\tilde{q})}{r^{\tilde{q}-1}}\right)^{1/\tilde{q}} \left(\frac{C_k(\ell, p\tilde{p}(\ell-1)/\ell)}{r^{\frac{\ell-1}{k}p\tilde{p}-1}}\right)^{1/\tilde{p}} \\ &= \frac{C_k(k,\tilde{q})^{1/\tilde{q}} C_k(\ell, p\tilde{p}(\ell-1)/\ell)^{\frac{1}{\tilde{p}}}}{r^{\frac{\ell-1}{k}p}}. \end{aligned}$$

We can thus bound the differences as follows,

$$\mathbf{E}\left\{S_{r+1}^{I,p} - S_r^{I,p}\right\} \le \sum_{t \notin I} E(t) \le \frac{(k-\ell+1) C_k(k,\tilde{q})^{1/\tilde{q}} C_k(\ell, p\tilde{p}(\ell-1)/\ell)^{1/\tilde{p}}}{r^{\frac{\ell-1}{k}p}}.$$

Since $p < \frac{k}{\ell-1}$, we have that $\sum_{r=1}^n \frac{1}{r^{p\frac{\ell-1}{k}}} \le \frac{1}{1-p\frac{\ell-1}{k}} \frac{n}{n^{p\frac{\ell-1}{k}}}$. So, by summing the differences we get,

$$\mathbf{E}\left\{S_n^{I,p}\right\} \le \left[C_k(\ell-1,p) - 2\right] n^{1-p\frac{\ell-1}{k}} + 2 \le C_k(\ell-1,p) n^{1-\frac{p(\ell-1)}{k}}$$

as $\mathbf{E}\left\{S_1^{I,p}\right\} \le 2$, for every $p \ge 1$, and any nonempty $I \subseteq \{1, \ldots, k\}$. Thus, hypothesis $\mathcal{H}_{\ell-1}$ is proved.

We now prove the lower bound. As we flip a perfect coin at the beginning of the process to choose the side of R_0 that we cut, all the coordinates $X_{i1}, \ldots, X_{ik}$ of a hyper-rectangle

R_i are exchangeable. So, denoting by $\mathcal{S}$ the set of all $I' \subseteq \{1, \ldots, k\}$ of cardinality ℓ, all the random variables $\sum_{i \in \mathcal{F}_n} \prod_{j \in I'} X_{ij}^p$ are equally distributed so that:

$$\mathbf{E}\left\{\sum_{i \in \mathcal{F}_n} \prod_{j \in I} X_{ij}^p\right\} = \frac{1}{|\mathcal{S}|} \mathbf{E}\left\{\sum_{I' \in \mathcal{S}} \sum_{i \in \mathcal{F}_n} \prod_{j \in I'} X_{ij}^p\right\}.$$

Then, by Lemma 4 and Lemma 8,

$$\mathbf{E}\left\{\sum_{i \in \mathcal{F}_n} \left(\prod_{j \in I} X_{ij}\right)^p\right\} \geq \frac{1}{|\mathcal{S}|} \mathbf{E}\left\{\sum_{i \in \mathcal{F}_n} \left(\prod_{j=1}^{k} X_{ij}\right)^{\frac{p\ell}{k}}\right\} \geq C' \frac{n}{n^{\frac{p\ell}{k}}}. \square$$

We must note that by Lemma 4, if $\ell = k$, then for any $p \geq 0$, there are positive constants C and C', depending on p such that the previous result holds. We are now ready to prove Theorem 3.

Proof of Theorem 3. The lower bound follows immediately from the previous proposition. For any subset $I \subseteq \{1, \ldots, k\}$ of cardinality $\ell \in \{1, \ldots, k-1\}$, we define:

$$S_n^I = \sum_{i=1}^{2n} \prod_{j \in I} X_{ij}.$$

As we are using the longest edge cut method we have that,

$$\begin{aligned}\mathbf{E}\left\{S_{r+1}^I - S_r^I | U_1, \ldots, U_n\right\} &= \sum_{i \in \mathcal{F}_r} \prod_{j=1}^{k} X_{ij} \left\{\mathbb{1}_{[j_i^* \notin I]} 2 \prod_{j \in I} X_{ij} + \mathbb{1}_{[j_i^* \in I]} \prod_{j \in I} X_{ij}\right\} \\ &\leq 3 \sum_{i \in \mathcal{F}_r} \prod_{j=1}^{k} X_{ij} \prod_{j \in I} X_{ij}.\end{aligned}$$

We choose now $p = \sqrt{k/\ell}$, $q = 1/(1 - \sqrt{\ell/k})$, so that $\frac{1}{p} + \frac{1}{q} = 1$, and apply Hölder inequality with these values to get,

$$\mathbf{E}\left\{S_{r+1}^I - S_r^I\right\} \leq 3\, \mathbf{E}\left\{\sum_{i \in \mathcal{F}_r} \left(\prod_{j=1}^{k} X_{ij}\right)^p\right\}^{1/p} \mathbf{E}\left\{\sum_{i \in \mathcal{F}_r} \left(\prod_{j \in I} X_{ij}\right)^q\right\}^{1/q}.$$

Then by Lemma 4 and proposition 2, there exists a positive constant C depending upon ℓ and k such that

$$\mathbf{E}\left\{S_{r+1}^I - S_r^I\right\} \leq \frac{C}{r^{\frac{\ell}{k}}}.$$

We add the differences to get

$$\mathbf{E}\left\{S_n^I\right\} \leq C\left(\sum_{r=1}^{n} \frac{1}{r^{\frac{\ell}{k}}}\right) + 2 \leq \frac{C}{1 - \frac{\ell}{k}} \left(\frac{n}{n^{\frac{\ell}{k}}}\right) + 2. \square$$

PROOF OF LEMMA 7. First, note that for any $1 \le i \le n$, $X_{i1} \cdots X_{ik}$ is the volume $|R_i|$ of the hyper-rectangle R_i. Note that if $U_1, \ldots, U_i$ have already been inserted in $[0,1]^k$, and U_{i+1} is a new point, then the size of the two hyper-rectangles generated by U_{i+1} is equal to the size of the hyper-rectangle in the final partition of $[0,1]^k$ in which U_{i+1} falls. Let us denote by $R(U_{i+1})$ this hyper-rectangle. Thus,

$$\mathbf{E}\left\{\sum_{i=1}^{2n+1} X_{i1} \cdots X_{ik}\right\} = 1 + \sum_{i=0}^{n-1} \mathbf{E}\left\{\mathbf{E}\left\{|R(U_{i+1})| \mid U_1, \ldots U_i\right\}\right\},$$

where the 1 accounts for the root hyper-rectangle. We claim that $\mathbf{E}\{|R(U_{i+1})|\} = \frac{2}{i+2}$. Note that the claim is obviously true for $i = 0$. Now, suppose that $U_1, \ldots, U_i$ have already been inserted in the squarish k-d tree, so that there are $i+1$ external nodes. These external nodes represent the $i+1$ hyper-rectangles partitioning $[0,1]^k$. Let these hyper-rectangles be $S_1, \ldots, S_{i+1}$, and let the numbering be so that the leaves are taken from left to right, in order of appearance as leaves in the squarish k-d tree of $U_1, \ldots, U_i$. Then,

$$\begin{aligned}\mathbf{E}\{|R(U_{i+1})|\} &= \mathbf{E}\left\{\mathbf{E}\left\{\sum_{\ell=1}^{i+1} \mathbb{1}_{[U_{i+1} \in S_\ell]} |S_\ell| \;\middle|\; U_1, \ldots U_i\right\}\right\} \\ &= \mathbf{E}\left\{\sum_{\ell=1}^{i+1} |S_\ell| \, \mathbf{P}\left\{U_{i+1} \in S_\ell \;\middle|\; U_1, \ldots, U_i\right\}\right\} \\ &= \mathbf{E}\left\{\sum_{\ell=1}^{i+1} |S_\ell|^2\right\}.\end{aligned}$$

By Lemma 3, $(|S_1|, \ldots, |S_{i+1}|)$ are jointly distributed as uniform spacings. All these spacings are identically distributed following a $\mathrm{Beta}(1, i)$ distribution. If B is a $\mathrm{Beta}(1, i)$ random variable, then we have $\mathbf{E}\{B\} = 1/(i+1)$ and $\mathbf{E}\{B^2\} = 2/((i+1)(i+2))$. Therefore,

$$\mathbf{E}\{|R(U_{i+1})|\} = (i+1)\,\mathbf{E}\{B^2\} = \frac{2}{i+2}.$$

and thus,

$$1 + \sum_{i=0}^{n-1} \mathbf{E}\{|R(U_{i+1})|\} = 1 + 2(H_{n+1} - 1). \;\square$$

§6. Orthogonal range search

In this section, we obtain tight upper bounds for the expected complexity for Bentley's range search algorithm. For random orthogonal range search, the following theorem establishes the standard for comparisons. Theorem 5 below then states that random squarish k-d trees are superior to random k-d trees for any kind of random orthogonal range search.

Theorem 4 (Chanzy, Devroye and Zamora-Cura, 1999). *Given is a random k-d tree of size n. Let Q be a random query hyper-rectangle of dimensions $\Delta_1 \times \cdots \times \Delta_k$*

(which are deterministic functions of n taking values in $[0,1]$), with center at Z which is uniformly distributed on $[0,1]^k$, and independent of the k-d tree. Let N_n be the number of comparisons that Bentley's orthogonal range search algorithm performs. Then, there exist constants $\gamma > \gamma' > 0$ depending upon k only such that

$$\gamma' \leq \frac{\mathbf{E}\{N_n\}}{\left(\log n + \sum_{\substack{I \subseteq \{1,\ldots,k\} \\ 0 \leq |I| < k}} \left(\prod_{j \notin I} \Delta_j\right) n^{\alpha(|I|/k)}\right)} \leq \gamma \ ,$$

where $\alpha(\cdot)$ is the function defined in Theorem 1.

Theorem 5. *Given is a random squarish k-d tree of size n. Let Q be a random query hyper-rectangle of dimensions $\Delta_1 \times \cdots \times \Delta_k$ (which are deterministic functions of n taking values in $[0,1]$), with center at Z which is uniformly distributed on $[0,1]^k$, and independent of the k-d tree. Let N_n be the number of comparisons that Bentley's orthogonal range search algorithm performs. Then, there exist constants $\gamma > \gamma' > 0$ depending upon k only such that*

$$\gamma' \leq \frac{\mathbf{E}\{N_n\}}{\left(\log n + \sum_{\substack{I \subseteq \{1,\ldots,k\} \\ 0 \leq |I| < k}} \prod_{j \notin I} \Delta_j n^{1-\frac{|I|}{k}}\right)} \leq \gamma \ .$$

We can rewrite the previous result as follows,

$$\mathbf{E}\{N_n\} \leq \gamma \left(n \prod_{j=1}^{k} \Delta_j + \sum_{\ell=1}^{k-1} n^{1-\frac{\ell}{k}} \sum_{\substack{I \subseteq \{1,\ldots,k\} \\ |I| = \ell}} \prod_{j \notin I} \Delta_j + \log n \right),$$

and therefore by allowing any r of the Δ_j's to be zero, the term that will dominate the previous bound is,

$$n^{1-\frac{r}{k}} \sum_{I; |I| = r} \prod_{j \notin I} \Delta_j.$$

For example, when $k = 2$, $\Delta = \Theta(1/n^\alpha)$, and $\Delta' = \Theta(1/n^\beta)$, then

$$\mathbf{E}\{N_n\} \leq \gamma \left(n^{1-\alpha-\beta} + n^{\frac{1}{2}-\alpha} + n^{\frac{1}{2}-\beta} + \log n\right).$$

By looking at the different regions in the α-β plane we obtain,

$$\mathbf{E}\{N_n\} \leq \begin{cases} \Theta(\log n), & \text{for } \alpha \geq 1/2 \text{ and } \beta \geq 1/2; \\ \Theta(\max\{n^{1/2-\alpha} n^{1/2-\beta}\}), & \text{for } \alpha > 1/2, \beta < 1/2, \text{ or } \alpha < 1/2, \beta > 1/2; \\ \Theta(n^{1-\alpha-\beta}), & \text{for } \alpha \leq 1/2, \beta \leq 1/2. \end{cases}$$

Note that if $\alpha = 0$ and $\beta \geq 1/2$, or $\beta = 0$ and $\alpha \geq 1/2$, we recover the expected complexity time of the random partial match query problem (see figure 5).

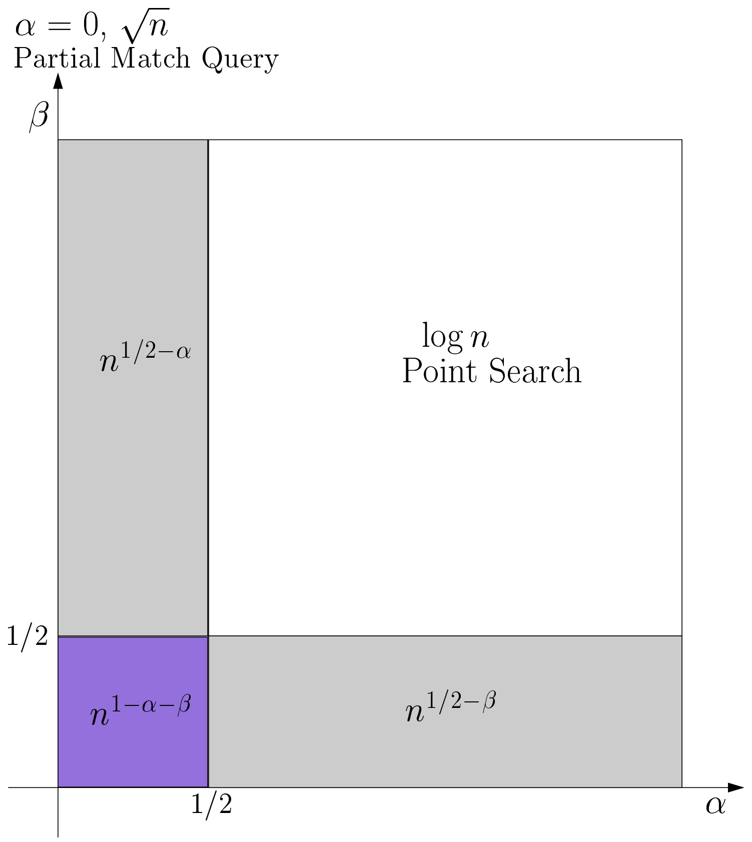

FIGURE 5. The complexity regions for $\Delta = \Theta(1/n^{\alpha})$ and $\Delta' = \Theta(1/n^{\beta})$.

Lemma 9. *Let $U_1, \ldots, U_n$ be independent and uniformly distributed over $[0,1]^k$ random variables, let X_i^* be the largest side of the i^{th} hyper-rectangle generated by $U_1, \ldots, U_n$. Then, for all $n \geq 0$,*

$$\mathbf{E}\left\{\sum_{i\in\mathcal{F}_n} \mathbb{1}_{[X_i^* > \frac{1}{2}]}\right\} \leq 2^{4k-3}.$$

PROOF. Note that $\mathbf{E}\left\{\sum_{i\in\mathcal{F}_n} \mathbb{1}_{[X_i^*>\frac{1}{2}]}\right\} \leq 2^k\, \mathbf{E}\left\{\sum_{i\in\mathcal{F}_n} \prod_{j\in I_i} X_{ij}\right\}$, where $I_i = \{j : X_{ij} > \frac{1}{2}\}$. Define $S_n = \sum_{i\in\mathcal{F}_n}\left(\prod_{j:X_{ij}>\frac{1}{2}} 8X_{ij}\right)$. We are going to prove that $\mathbf{E}\{S_n\}$ is decreasing so that for $n \geq 1$,

$$\mathbf{E}\left\{\sum_{i\in\mathcal{F}_n} \mathbb{1}_{[X_i^*>\frac{1}{2}]}\right\} \leq 2^{k-3}\,\mathbf{E}\{S_n\} \leq 2^{k-3}\,\mathbf{E}\{S_0\} = 2^{4k-3}.$$

To show $\mathbf{E}\{S_n\} \leq \mathbf{E}\{S_0\}$, we look at the differences once again. Set $P_i = \prod_{j\in I_i} 8X_{ij}$. Then,

$$S_{r+1} - S_r = \sum_{i\in\mathcal{F}_r} |R_i| \mathbb{1}_{[X_i^*>\frac{1}{2}]}\left\{-P_i + \mathbb{1}_{[XX_i^*>\frac{1}{2}]}\left(P_iX + \mathbb{1}_{[|I_i|>1]}\frac{P_i}{8X_i^*}\right)\right.$$
$$\left. + \mathbb{1}_{[(1-X)X_i^*>\frac{1}{2}]}\left(P_i(1-X) + \mathbb{1}_{[|I_i|>1]}\frac{P_i}{8X_i^*}\right) + \mathbb{1}_{[XX_i^*\leq\frac{1}{2};\ (1-X)X_i^*\leq\frac{1}{2}]}\left(2\mathbb{1}_{[|I_i|>1]}\frac{P_i}{8X_i^*}\right)\right\},$$

where $X \stackrel{\mathcal{L}}{=} \text{Uniform}[0,1]$, and it is independent of $U_1, \ldots, U_r$. Therefore,

$$\begin{aligned}
\mathbf{E}\left\{S_{r+1} - S_r | U_1, \ldots, U_r\right\} &\leq \sum_{i \in \mathcal{F}_r} |R_i| \mathbb{1}_{[X_i^* > \frac{1}{2}]} P_i \left(-1 + \int_{\frac{1}{2X_i^*}}^{1} (x + \frac{1}{4}) dx \right. \\
&\quad \left. + \int_0^{1-\frac{1}{2X_i^*}} ((1-x) + 1/4) dx + \int_{1-\frac{1}{2X_i^*}}^{\frac{1}{2X_i^*}} 1/2 dx \right) \\
&= \sum_{i \in \mathcal{F}_r} |R_i| \mathbb{1}_{[X_i^* > \frac{1}{2}]} P_i \left(\frac{1}{4X_i^*} - \frac{1}{(2X_i^*)^2} \right) \\
&\leq 0. \square
\end{aligned}$$

Proof of Theorem 5. Let T be the squarish k-d tree constructed from $U_1, \ldots, U_n$. Note that a node U_i in T is visited if and only if the query hyper-rectangle Q intersects R_i, where R_i is the hyper-rectangle in the final partition of $[0,1]^k$ generated by $U_1, \ldots, U_{i-1}$, in which U_i falls. Thus, the running time of the range search algorithm is exactly the number of hyper-rectangles in $\mathcal{R}_n$ that Q intersects,

$$N_n = \sum_{i=0}^{2n} \mathbb{1}_{[R_i \cap Q \neq \emptyset]}.$$

Also, given $U_1, \ldots, U_n$, the probability that Q intersects R_i is the probability that Z has some coordinate that is within distance $\Delta_j/2$ of R_i, and this probability is clearly bounded by the volume of R_i expanded by Δ_j in the j^{th} direction, for all j. Therefore,

$$\begin{aligned}
\mathbf{E}\left\{N_n\right\} &\leq \mathbf{E}\left\{ \sum_{i=1}^{2n} \prod_{j=1}^{k} (X_{ij} + \Delta_j) \right\} \\
&= \sum_{I \subseteq \{1,\ldots,k\}} \prod_{j \notin I} \Delta_j \, \mathbf{E}\left\{ \sum_{i=1}^{2n} \prod_{j \in I} X_{ij} \right\} + 1 \\
&\leq \gamma \left(\sum_{\substack{I \subseteq \{1,\ldots,k\} \\ 0 \leq |I| < k}} \prod_{j \notin I} \Delta_j n^{1-\frac{|I|}{k}} + \log n \right),
\end{aligned}$$

for some $\gamma > 0$ by Theorem 3 and the Lemma 7. For the lower bound we may assume that

$\Delta_j \le 1/2$ and do the following:

$$\begin{aligned}
\mathbf{E}\{N_n\} &\ge \mathbf{E}\left\{\sum_{i\in\mathcal{F}_n} \mathbb{1}_{[Q\cap R_i]}\mathbb{1}_{[\forall j\in\{1,\dots,k\}:X_{ij}\le 1/2]}\right\} \\
&\ge \mathbf{E}\left\{\sum_{i\in\mathcal{F}_n}\prod_{j=1}^{k}\left(X_{ij}+\frac{\Delta_j}{2}\right)\mathbb{1}_{[\forall j\in\{1,\dots,k\}:X_{ij}\le 1/2]}\right\} \\
&= \mathbf{E}\left\{\sum_{i\in\mathcal{F}_n}\prod_{j=1}^{k}\left(X_{ij}+\frac{\Delta_j}{2}\right)\right\} - \mathbf{E}\left\{\sum_{i\in\mathcal{F}_n}\prod_{j=1}^{k}\left(X_{ij}+\frac{\Delta_j}{2}\right)\mathbb{1}_{[\exists j\in\{1,\dots,k\}:X_{ij}> 1/2]}\right\} \\
&= \sum_{I\subseteq\{1,\dots,k\}}\prod_{j\notin I}\frac{\Delta_j}{2}\,\mathbf{E}\left\{\sum_{i\in\mathcal{F}_n}\prod_{j\in I}X_{ij}\right\} \\
&\qquad - \sum_{I\subseteq\{1,\dots,k\}}\prod_{j\notin I}\frac{\Delta_j}{2}\,\mathbf{E}\left\{\sum_{i\in\mathcal{F}_n}\prod_{j\in I}X_{ij}\mathbb{1}_{[\exists j\in\{1,\dots,k\}:X_{ij}> 1/2]}\right\}.
\end{aligned}$$

We can bound the second term above for any given $I\subseteq\{1,\dots,k\}$ as follows,

$$\mathbf{E}\left\{\sum_{i\in\mathcal{F}_n}\prod_{j\in I}X_{ij}\mathbb{1}_{[\exists j\in\{1,\dots,k\}:X_{ij}> 1/2]}\right\} \le \mathbf{E}\left\{\sum_{i\in\mathcal{F}_n}\mathbb{1}_{[X_i^*>1/2]}\right\} \le 2^{4k-3},$$

by the previous lemma. Thus, for all n large enough, we can choose $\gamma'>0$ such that,

$$\mathbf{E}\{N_n\} \ge \gamma'\left(\sum_{\substack{I\subseteq\{1,\dots,k\}\\ 0\le |I|<k}}\prod_{j\notin I}\Delta_j n^{1-\frac{|I|}{k}} + \log n\right). \ \square$$

§7. Nearest neighbor search

We consider two natural nearest neighbor search algorithms. In algorithm $\mathbf{A}$, start with an orthogonal range search with a square box of size $1/n^{1/k}$ centered at the query point X. Repeat with boxes of sizes $k^{i/2}/n^{1/k}$ for $i=0,1,2,3,\dots$ until $i+1$, where i is the index of the first nonempty box. Report the nearest point in the $i+1$-st box. Each orthogonal range search taken individually (for fixed i) takes expected time $O(\log n)$ by Theorem 5. We show in fact that the total expected time is $O(\log n \log\log n)$:

Theorem 6. *Let X be a point uniformly distributed on $[0,1]^2$. Consider a squarish 2-d tree based on n i.i.d. points on $[0,1]^2$. Then the expected time of algorithm $\mathbf{A}$ is $O(\log n\log\log n)$.*

PROOF. Let $\mathcal{T}$ be the total time it takes algorithm $\mathbf{A}$ to finish. Let $\mathcal{T}_i$ be the running time of Bentley's range search algorithm on n i.i.d. points on $[0,1]^k$ and a cube Q_i centered at

X of length $k^{i/2}/n^{1/k}$, and let M_i be the number of points in Q_i. Note that,

$$\mathbf{E}\{\mathcal{T}\} \le O(\log n) + \mathbf{E}\left\{\mathcal{T}_1 + \mathcal{T}_2 + \sum_{i=3}^{m} \mathcal{T}_i \mathbb{1}_{[M_{i-2}=0]}\right\},$$

where $m = \lfloor \frac{2}{k}\log_k(2^k n)\rfloor$ bounds the maximum number of iterations the algorithms can perform. Thus, it is enough to prove that $\mathbf{E}\left\{\sum_{i=3}^{m} \mathcal{T}_i \mathbb{1}_{[M_{i-2}=0]}\right\} = O(\log n \log\log n)$. Let $t = \lceil \frac{2}{k}\log_k(2^k \log n)\rceil$, then

$$\mathbf{E}\left\{\sum_{i=3}^{m} \mathcal{T}_i \mathbb{1}_{[M_{i-2}=0]}\right\} \le (t+1)\,\mathbf{E}\{\mathcal{T}_{t+1}\} + 2n \sum_{i=t+2}^{m} \mathbf{P}\{M_{i-2}=0\}.$$

Now, by Theorem 5,

$$\begin{aligned}
&(t+1)\,\mathbf{E}\{\mathcal{T}_{t+1}\}\\
&\le \gamma\left(\frac{2}{k}\log_k(2^k\log n)+2\right)\left(k^{(t+1)k/2} + \sum_{\ell=1}^{k-1} n^{1-\ell/k} \sum_{\substack{I\subseteq\{1,\ldots,k\}\\ |I|=\ell}} \prod_{j\notin I} \frac{k^{(t+1)/2}}{n^{1/k}} + \log n\right)\\
&= \gamma\left(\frac{2}{k}\log_k(2^k\log n)+2\right)\left(k^{(t+1)k/2} + \sum_{\ell=1}^{k-1} n^{1-\ell/k}\binom{k}{\ell}\frac{k^{(t+1)(k-\ell)/2}}{n^{1-\ell/k}} + \log n\right)\\
&= \gamma\left(\frac{2}{k}\log_k(2^k\log n)+2\right)\left(k^{(t+1)k/2} + k^{(t+1)k/2}\sum_{\ell=1}^{k-1}\binom{k}{\ell}k^{-(t+1)\ell/2} + \log n\right)\\
&\le \gamma\left(\frac{2}{k}\log_k(2^k\log n)+2\right)\left(k^{(t+1)k/2}\left(k^{-(t+1)/2}+1\right)^k + \log n\right)\\
&\le \gamma\left(\frac{2}{k}\log_k(2^k\log n)+2\right)\left(k^k 2^k \log n\left(\frac{1}{\sqrt{k}(2^k\log n)^{1/k}}+1\right)^k + \log n\right)\\
&= O(\log n\log\log n),
\end{aligned}$$

for all $n \ge e$. Finally, for $i \le m$,

$$\mathbf{P}\{M_{i-2}=0\} \le \left(1-\frac{k^{k(i-2)/2}}{2^k n}\right)^n \le e^{-k^{k(i-2)/2}/2^k},$$

and therefore $\mathbf{P}\{M_{t+2}=0\} \le 1/n$. Thus,

$$2n\sum_{i=t+2}^{m}\mathbf{P}\{M_{i-2}=0\} \le 2m = O(\log n). \;\square$$

Theorem 6 is in contrast with the situation for standard random k-d trees, where algorithm **A** is shown to take expected time $\Theta(n^\rho)$, where $\rho \in (0.061, 0.064)$ depends upon k only (Chanzy, Devroye and Zamora-Cura, 1999). In algorithm **B**, insert X in the squarish k-d tree, and let Q be the rectangle associated with X. Let X' be the parent of X in the tree (note: $X' \in Q$). Perform an orthogonal range search centered at X with dimensions $2\|X'-X\|$ in all directions. Report the nearest neighbor among all points returned by this orthogonal range search. We will analyze this algorithm for $k=2$ only.

Theorem 7. *Let X be a point uniformly distributed on $[0,1]^2$. Consider a squarish 2-d tree based on n i.i.d. points on $[0,1]^2$. Then the expected time of algorithm* **B** *is $O(\log^2 n)$.*

The bound on algorithm **B** is a bit worse than that for algorithm **A**, because while most rectangles are squarish, a sufficient number of them are elongated. In fact, for given $M > 1$, about $1/M$ of the final (leaf) rectangles or more should have an edge ratio exceeding M. For edge ratio M, and considering that all rectangle areas are about $1/n$, we see that the orthogonal range search should take about M points (the longest edge is about $\sqrt{M/n}$). The expected number of returned elements is at least $\Theta(\log n)$. And the expected number of leaf rectangles visited is of the same order. But each visited leaf rectangle also induces a visit to all of its ancestors, and there are about $\log n$ of those, hence the claim. The remainder of this section contains the proof of Theorem 7.

Lemma 10. *Let $Z, U_1, \ldots, U_n$ be independent and uniformly distributed random variables on $[0,1]^2$. Let $X_n(Z)$ and $Y_n(Z)$ be the x-length and y-length of the rectangle in the final partition (of the squarish 2-d tree) induced by $U_1, \ldots, U_n$ in which Z falls. Then, both $n\,\mathbf{E}\left\{X_n^2(Z)\right\}$ and $n\,\mathbf{E}\left\{Y_n^2(Z)\right\}$ are $O(\log^2 n)$.*

Proof. By Lemmas 4 and 5, for any $p, q > 1$ such that $\frac{1}{p} + \frac{1}{q} = 1$, we have that

$$\begin{aligned}
\mathbf{E}\left\{X_n^2(Z)\right\} &= \mathbf{E}\left\{\sum_{i \in \mathcal{F}_n} X_i^3 Y_i\right\} \\
&\le \mathbf{E}\left\{\sum_{i \in \mathcal{F}_n} (X_i Y_i)^p\right\}^{1/p} \mathbf{E}\left\{\sum_{i \in \mathcal{F}_n} X_i^{2q}\right\}^{1/q} \\
&\le \left(\frac{4\Gamma(p+1)}{n^{p-1}}\right)^{1/p} \left(\frac{5\Gamma(q+1)}{q-1}\left(q - \frac{1}{n^{q-1}}\right)\right)^{1/q} \\
&= \frac{4^{1/p} 5^{1/q} \left(\Gamma(p+1)\right)^{1/p} \left(\Gamma(q+1)\right)^{1/q}}{(q-1)^{1/q}} \frac{(qn^{q-1} - 1)^{1/q}}{n}.
\end{aligned}$$

Let us choose $q = 1 + \frac{1}{\log n}$, $p = \log n + 1$, and assume $n > e$. As $\Gamma(p+1) \le \sqrt{2\pi}\left(\frac{p}{e}\right)^p e^{1/12p}$ (see for example, Abramowitz and Stegun, 1970), there is $c > 0$, such that $(\Gamma(p+1))^{1/p} \le cp = c(\log n + 1)$, and there is $c' > 0$, such that $(\Gamma(q+1))^{1/q} \le c'q \le 4c'$. Furthermore, $(q-1)^{-1/q} = (\log n)^{\frac{\log n}{\log n + 1}} \le \log n$, and $(qn^{q-1} - 1)^{1/q} \le 2e - 1$. Therefore $n\,\mathbf{E}\left\{X_n^2(Z)\right\} = O(\log^2 n)$. The result for $n\,\mathbf{E}\left\{Y_n^2(Z)\right\}$ follows in the same manner. □

Lemma 11 (Devroye, 1986). *Let H_n be the height of a random binary search tree of size n, then for any integer $k \ge \max\{1, \log n\}$ we have*

$$\mathbf{P}\left\{H_n \ge k\right\} \le \frac{1}{n}\left(\frac{2e\log n}{k}\right)^k.$$

Lemma 12. *Let $Z, U_1, \ldots, U_n$ be independent and uniformly distributed random variables over $[0,1]^2$. Let $X_n(Z)$ and $Y_n(Z)$ be the x-length and y-length of the rectangle* *in the final partition induced by* $U_1, \ldots, U_n$ *in which* Z *falls. Then* $\mathbf{E}\left\{X_n(Z)\sum_{i=1}^{2n} X_i\right\}$, $\mathbf{E}\left\{Y_n(Z)\sum_{i=1}^{2n} Y_i\right\}$, $\mathbf{E}\left\{X_n(Z)\sum_{i=1}^{2n} Y_i\right\}$, *and* $\mathbf{E}\left\{Y_n(Z)\sum_{i=1}^{2n} X_i\right\}$ *are* $O(\log^2 n)$.

PROOF. Let $\mathcal{F}_n$ denote the collection of final rectangles in the squarish 2-d tree T constructed from $U_1, \ldots, U_n$. For a final rectangle R_i, denote by $D(R_i)$ its depth. Then $\sum_{i=1}^{2n} X_i \le \sum_{i\in\mathcal{F}_n} D(R_i)X_i + 1$. Thus if H_n is the height of T,

$$
\begin{aligned}
\mathbf{E}\left\{\sum_{i=1}^{2n} X_i X_n(Z)\right\} &\le \mathbf{E}\left\{\sum_{i\in\mathcal{F}_n} D(R_i)X_i \sum_{j\in\mathcal{F}_n} X_j^2 Y_j\right\} + 1 \\
&\le \mathbf{E}\left\{H_n \sum_{i\in\mathcal{F}_n} X_i \sum_{j\in\mathcal{F}_n} X_j^2 Y_j\right\} + 1 \\
&\le t\log n\,\mathbf{E}\left\{\sum_{i\in\mathcal{F}_n} X_i \sum_{j\in\mathcal{F}_n} X_j^2 Y_j\right\} + 1 \\
&\quad + \mathbf{E}\left\{\mathbb{1}_{[H_n\ge t\log n]} H_n \sum_{i\in\mathcal{F}_n} X_i \sum_{j\in\mathcal{F}_n} X_j^2 Y_j\right\} + 1,
\end{aligned}
$$

for any $t > 1$. Using Lemma 11, we see that,

$$
\mathbf{E}\left\{\mathbb{1}_{[H_n\ge t\log n]} H_n \sum_{i\in\mathcal{F}_n} X_i \sum_{j\in\mathcal{F}_n} X_j^2 Y_j\right\} \le n^3\,\mathbf{P}\left\{H_n \ge t\log n\right\} \le n^2 n^{t\log\left(\frac{2e}{t}\right)}.
$$

We choose t such that $t\log\left(\frac{2e}{t}\right) < -2$ so that

$$
\mathbf{E}\left\{\mathbb{1}_{[H_n\ge t\log n]} H_n \sum_{i\in\mathcal{F}_n} X_i \sum_{j\in\mathcal{F}_n} X_j^2 Y_j\right\} = O(1).
$$

We complete the proof by showing that $\mathbf{E}\left\{\sum_{i\in\mathcal{F}_n} X_i \sum_{j\in\mathcal{F}_n} X_j^2 Y_j\right\} = O(\log n)$. For this, let $S_r = \sum_{i\in\mathcal{F}_r} X_i \sum_{j\in\mathcal{F}_r} X_j^2 Y_j$, for $r = 1, \ldots, n-1$. Note that

$$
\begin{aligned}
S_{r+1} - S_r = \sum_{m\in\mathcal{F}_r} X_m Y_m \Bigg[&\mathbb{1}_{[X_m<Y_m]} X_m \sum_{j\in\mathcal{F}_r} X_j^2 Y_j \\
&+ \mathbb{1}_{[X_m>Y_m]}((XX_m)^2 Y_m + ((1-X)X_m)^2 Y_m - X_m^2 Y_m) \sum_{i\in\mathcal{F}_r} X_i \Bigg],
\end{aligned}
$$

where $X \stackrel{\mathcal{L}}{=} \text{Uniform}[0,1]$, and is independent of all $U_1, \ldots, U_n$. Now, as $(XX_m)^2 Y_m + ((1-X)X_m)^2 Y_m - X_m^2 Y_m \le 0$, we have that

$$
S_{r+1} - S_r \le \sum_{i\in\mathcal{F}_r} (X_i Y_i)^{3/2} \sum_{j\in\mathcal{F}_r} X_j^2 Y_j.
$$

Note that for any $p, q > 1$, such that $\frac{1}{p} + \frac{1}{q} = 1$,

$$\mathbf{E}\left\{S_{r+1} - S_r\right\} \le \mathbf{E}\left\{\left(\sum_{i\in\mathcal{F}_r} (X_i Y_i)^{3/2}\right)^p\right\}^{1/p} \mathbf{E}\left\{\left(\sum_{j\in\mathcal{F}_r} X_j^2 Y_j\right)^q\right\}^{1/q},$$

and again by Hölder's inequality, and Lemma 4, by choosing $q = \sqrt{1.4}$, and $p = \frac{\sqrt{1.4}}{\sqrt{1.4}-1}$,

$$\mathbf{E}\left\{\left(\sum_{i\in\mathcal{F}_r} (X_i Y_i)^{3/2}\right)^p\right\}^{1/p} \le \mathbf{E}\left\{r^{p/q} \sum_{i\in\mathcal{F}_r} (X_i Y_i)^{3p/2}\right\}^{1/p} \le \frac{12}{\sqrt{r}} .$$

By applying Hölder's inequality inside the expected value,

$$\begin{aligned}
\mathbf{E}\left\{\left(\sum_{j\in\mathcal{F}_r} X_j^2 Y_j\right)^q\right\}^{1/q} &\le \mathbf{E}\left\{r^{q/p} \sum_{j\in\mathcal{F}_r} (X_j^2 Y_j)^q\right\}^{1/q} \\
&\le r^{1/p}\left(\mathbf{E}\left\{\sum_{j\in\mathcal{F}_r} (X_j Y_j)^{qp}\right\}^{1/p} \mathbf{E}\left\{\sum_{j\in\mathcal{F}_r} X_j^{q^2}\right\}^{1/q}\right)^{1/q} \\
&\le 46\, r^{1/p}\left(\left(\frac{1}{r^{qp-1}}\right)^{1/p}\left(\frac{1}{r^{q^2/2-1}}\right)^{1/q}\right)^{1/q} \\
&= \frac{46}{\sqrt{r}} .
\end{aligned}$$

Thus, $\mathbf{E}\left\{S_{r+1} - S_r\right\} \le 552/r$, and by summing the differences we finally can conclude that $\mathbf{E}\left\{\sum_{i\in\mathcal{F}_n} X_i \sum_{j\in\mathcal{F}_n} X_j^2 Y_j\right\}$ is indeed $O(\log n)$. The other expected values can be bounded in the same way. □

Proof of Theorem 7. Given $U_1, \ldots, U_n$, we define $L_n(Z) = 2(X_n(Z) + Y_n(Z))$. Note that as the expected height of T is $O(\log n)$, the expected time complexity of the nearest neighbor algorithm is bounded by $O(\log n)$ plus the expected time of random orthogonal range search with query rectangle Q having all sides of length $L_n(Z)$, and centered at Z. Let N_n be the time complexity of a range search. By the same arguments followed in Theorem 3 we have,

$$\mathbf{E}\left\{N_n\right\} \le \mathbf{E}\left\{\sum_{i=1}^{2n+1} X_i Y_i\right\} + 2\,\mathbf{E}\left\{\sum_{i=1}^{2n+1} L_n(Z)(X_i + Y_i)\right\} + 8n\,\mathbf{E}\left\{L_n^2(Z)\right\} + 1.$$

By Lemma 5, $\mathbf{E}\left\{\sum_{i=1}^{2n+1} X_i Y_i\right\} = O(\log n)$. For $\mathbf{E}\left\{\sum_{i=1}^{2n+1} L_n(Z)(X_i + Y_i)\right\}$, Lemma 12 above shows that it is $O(\log^2 n)$. As $n\,\mathbf{E}\left\{X_n(Z) Y_n(Z)\right\} = n\,\mathbf{E}\left\{\sum_{i\in\mathcal{F}_n} (X_i Y_i)^2\right\}$, Lemma 4 shows that it is $O(1)$. Finally, by Lemma 10 we have that $n\,\mathbf{E}\left\{L_n^2(Z)\right\} = O(\log^2 n)$. Thus the expected running time of algorithm $\mathbf{B}$ is $O(\log^2 n)$. □

§8. Further work and open problems

QUADTREES. For quadtree splitting in k dimensions (Finkel and Bentley, 1974; Bentley and Stanat, 1975), it is easy to see that the analysis and thus Theorem 1 are not valid. In fact, for random quadtrees, the expected performance for partial match queries was shown to be of the order of that for standard random k-d trees (Flajolet, Gonnet, Puech and Robson, 1991, 1992). For orthogonal range search with query rectangles depending upon n, see Chanzy, Devroye and Zamora-Cura (1999).

EXPECTED WORST-CASE COMPLEXITY. We conjecture that the expected worst case complexity over all range search rectangles of dimensions Δ_i (but with worst-case location of the center) is also bounded from above by the bound given in Theorem 2. And the expected worst-case time for an s-dimensional partial match query is conjectured to be $O(n^{1-s/k})$ for $s < k$ (for $s = k$, the complexity is clearly bounded by the expected height of the tree, $O(\log n)$).

NON-UNIFORM DISTRIBUTIONS. Finally, we also intend to study the behavior of squarish k-d trees for nonuniform distributions, although it appears once again that the upper bound of Theorem 2 remains valid for all distributions with a joint density on $[0, 1]^k$.

§9. References

M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, Dover Publications, New York, 1970.

P. K. Agarwal, "Range searching," in: *Handbook of Discrete and Computational Geometry*, ed. J. E. Goodman and J. O'Rourke, pp. 575–598, CRC Press, Boca Raton, FL, 1997.

J. L. Bentley, "Multidimensional binary search trees used for associative searching," *Communications of the ACM*, vol. 18, pp. 509–517, 1975.

J. L. Bentley, "Multidimensional binary search trees in database applications," *IEEE Transactions on Software Engineering*, vol. SE-5, pp. 333–340, 1979.

J. L. Bentley and J. H. Friedman, "Data structures for range searching," *ACM Computing Surveys*, vol. 11, pp. 397–409, 1979.

J. L. Bentley and D. F. Stanat, "Analysis of range searches in quad trees," *Information Processing Letters*, vol. 3, pp. 170–173, 1975.

P. Chanzy, L. Devroye, and C. Zamora-Cura, "Analysis of range search for random k-d trees," Technical Report, School of Computer Science, McGill University, Montreal, 1999.

L. Devroye, "A note on the height of binary search trees," *Journal of the ACM*, vol. 33, pp. 489–498, 1986.

L. Devroye, "Branching processes in the analysis of the heights of trees," *Acta Informatica*, vol. 24, pp. 277–298, 1987.

R. A. Finkel and J. L. Bentley, "Quad trees: a data structure for retrieval on composite keys," *Acta Informatica*, vol. 4, pp. 1–9, 1974.

P. Flajolet, G. Gonnet, C. Puech, and J. M. Robson, "The analysis of multidimensional searching in quad-trees," in: *Proceedings of the Second Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 100–109, SIAM, Philadelphia, 1991.

P. Flajolet, G. Gonnet, C. Puech, and J. M. Robson, "Analytic variations on quadtrees," *Algorithmica*, vol. 10, pp. 473–500, 1992.

P. Flajolet and C. Puech, "Partial match retrieval of multidimensional data," *Journal of the ACM*, vol. 33, pp. 371–407, 1986.

D. Gardy, P. Flajolet and C. Puech, "Average cost of orthogonal range queries in multiattribute trees," *Information Systems*, vol. 14, pp. 341–350, 1989.

G. H. Gonnet and R. Baeza-Yates, *Handbook of Algorithms and Data Structures*, Addison-Wesley, Workingham, 1991.

D. E. Knuth, *The Art of Computer Programming, Vol. 3 : Sorting and Searching (2nd ed)*, Addison-Wesley, Reading, MA, 1997.

D. T. Lee and C. K. Wong, "Worst-case analysis for region and partial region searches in multidimensional binary search trees and quad trees," *Acta Informatica*, vol. 9, pp. 23–29, 1977.

H. M. Mahmoud, *Evolution of Random Search Trees*, John Wiley, New York, 1992.

D. S. Mitrinović, *Analytic Inequalities*, Springer-Verlag, New York, N.Y., 1970.

B. Pittel, "On growing random binary trees," *Journal of Mathematical Analysis and Applications*, vol. 103, pp. 461–480, 1984.

H. Samet, *Applications of Spatial Data Structures*, Addison-Wesley, Reading, MA, 1990a.

H. Samet, *The Design and Analysis of Spatial Data Structures*, Addison-Wesley, Reading, MA, 1990b.

J. S. Vitter and P. Flajolet, "Average-case analysis of algorithms and data structures," in: *Handbook of Theoretical Computer Science, Volume A: Algorithms and Complexity*, ed. J. van Leeuwen, pp. 431–524, MIT Press, Amsterdam, 1990.

F. F. Yao, "Computational geometry," in: *Handbook of Theoretical Computer Science, Volume A: Algorithms and Complexity*, ed. J. van Leeuwen, pp. 343–389, MIT Press, Amsterdam, 1990.