


# Fixed Partial Match Queries in Quadrees

**Amalia Duch**

Universitat Politècnica de Catalunya


duch@cs.upc.edu

 <https://orcid.org/0000-0003-4371-1286>

**Gustavo Lau**

Universitat Politècnica de Catalunya


glau@cs.upc.edu

 <https://orcid.org/0000-0002-3460-9186>

**Conrado Martínez**

Universitat Politècnica de Catalunya

conrado@cs.upc.edu

 <https://orcid.org/0000-0003-1302-9067>

## Abstract

Several recent papers in the literature have addressed the analysis of the cost  $\mathcal{P}_{n,\mathbf{q}}$  of partial match search for a given fixed query  $\mathbf{q}$ —that has  $s$  out of  $K$  specified coordinates— in different multidimensional data structures. Indeed, detailed asymptotic estimates for the main term in the expected cost  $P_{n,\mathbf{q}} = \mathbb{E}\{\mathcal{P}_{n,\mathbf{q}}\}$  in standard and relaxed  $K$ -d trees are known (for any dimension  $K$  and any number  $s$  of specified coordinates), as well as stronger distributional results on  $\mathcal{P}_{n,\mathbf{q}}$  for standard 2-d trees and 2-dimensional quadrees. In this work we derive a precise asymptotic estimate for the main order term of  $P_{n,\mathbf{q}}$  in quadrees, for any values of  $K$  and  $s$ ,  $0 < s < K$ , under the assumption that the limit of  $P_{n,\mathbf{q}}/n^\alpha$  when  $n \rightarrow \infty$  exists, where  $\alpha$  is the exponent of  $n$  in the expected cost of a *random* partial match query with  $s$  specified coordinates in a random  $K$ -dimensional quadree.

**2012 ACM Subject Classification** Theory of computation  $\rightarrow$  Data structures design and analysis, Theory of computation  $\rightarrow$  Design and analysis of algorithms

**Keywords and phrases** Quadtree, Partial match queries, Associative queries, Multidimensional search, Analysis of algorithms

**Digital Object Identifier** 10.4230/LIPIcs.AofA.2018.20

**Funding** This work has been partially supported by funds from the Spanish Ministry of Economy, Industry and Competitiveness (MINECO) and the European Union (FEDER) under grant GRAMM (TIN2017-86727-C2-1-R), and by funds from the Catalan Government (AGAUR) under grant 2017SGR 786.

**Acknowledgements** We are thankful to the anonymous reviewers of the preliminary version of this paper. Their comments and advice have been very helpful to improve it in many ways, particularly in Subsection 3.3.

## 1 Introduction

One of the fundamental features of any hierarchical multidimensional data structure such as quadrees is to efficiently support partial match (PM) queries. These queries are as follows. Given a collection  $F$  of  $K$ -dimensional ( $K \geq 2$ ) tuples of the form  $\mathbf{x} = (x_0, \dots, x_{K-1})$ , with each  $x_i$  ( $0 \leq i < K$ ) belonging to a totally ordered domain  $\mathcal{D}_i$ , and a query  $\mathbf{q} = (q_0, \dots, q_{K-1})$



© Amalia Duch, Gustavo Lau, and Conrado Martínez;  
licensed under Creative Commons License CC-BY

29th International Conference on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms (AofA 2018).

Editors: James Allen Fill and Mark Daniel Ward; Article No. 20; pp. 20:1–20:18



Leibniz International Proceedings in Informatics  
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

such that  $q_i \in D_i \cup \{*\}$  ( $0 \leq i < K$ ), the goal of a PM query is to find all those tuples in  $F$  such that  $x_i$  matches  $q_i$  whenever  $q_i \neq *$ . Coordinates such that  $q_i \neq *$  are called *specified*, otherwise they are called *unspecified*; we assume that the number  $s$  of specified coordinates satisfies  $0 < s < K$ .

The average-case analysis of PM queries in random quadrees and other multidimensional data structures has a long history. In the case of quadrees, a fundamental milestone was the paper by Flajolet, Gonnet, Puech, and Robson [7] where the authors proved that the expected cost of random PM queries with  $s$  specified coordinates in random  $K$ -dimensional quadrees of  $n$  nodes is  $\beta_{s,K} n^{\alpha(s/K)} + l.o.t.$  for some constant  $\beta_{s,K}$ ; and  $\alpha = \alpha(s/K)$  the unique real solution in  $[0, 1]$  of the indicial equation

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

The exponent  $\alpha$  turns out to be exactly the same as in the expected cost of random PM queries in standard  $K$ -d trees. It was not until 2003 that Chern and Hwang [2] obtained an explicit expression for  $\beta_{s,K}$ , for general  $s$  and  $K$ , this is:

$$\beta_{s,K} = \frac{1}{(2^{K-s} - 1)\Gamma(\alpha + 1)^{K-s}\Gamma(\alpha + 2)^s} \prod_{2 \leq j \leq K} \frac{\Gamma(\alpha - \alpha_j)}{\Gamma(-\alpha_j)}, \quad (2)$$

for  $0 < s < K$  and  $K \geq 2$  and where  $\Gamma$  is the Gamma function and the  $\alpha_j$ 's are the roots of equation (1) and  $\alpha = \alpha_1 > \Re(\alpha_2) \geq \dots \geq \Re(\alpha_K)$ . Note that Chern and Hwang [2] used the indicial equation for  $\alpha + 1$  so they gave a formula for  $\beta_{s,K}$  as a function of  $\alpha'_j = \alpha_j + 1$ ,  $j = 1, \dots, K - 1$ .

In 2011 fixed PM queries were studied for the first time in 2-dimensional quadrees by Curien and Joseph [3] where the authors computed the expected cost  $\mathbb{E}\{\mathcal{P}_{n,\mathbf{q}}\}$  of a fixed PM query in 2-dimensional quadrees. In particular, they showed that if  $\mathbf{q} = (q, *)$ , then  $P_{n,\mathbf{q}} = \mathbb{E}\{\mathcal{P}_{n,\mathbf{q}}\} \sim \nu_{1,2} \cdot (q \cdot (1 - q))^{\alpha/2} \cdot n^\alpha$ , where  $\alpha = \alpha(1/2) = (\sqrt{17} - 3)/2$  is the same exponent as in the expected cost for random PM queries [7], and  $\nu_{1,2} = \frac{\Gamma(2\alpha+2)\Gamma(\alpha+2)}{2\Gamma^3(\alpha+1)\Gamma^2(\frac{\alpha}{2}+1)}$ . The asymptotic distribution was obtained for this particular case by Broutin, Neininger and Sulzbach in 2012 [1].

In this work, we extend the results of [3] to give a precise asymptotic estimate of the expected cost of a fixed PM query in random  $K$ -dimensional quadrees, for general  $K$  and  $s$ . In particular, we show that this cost is of the form

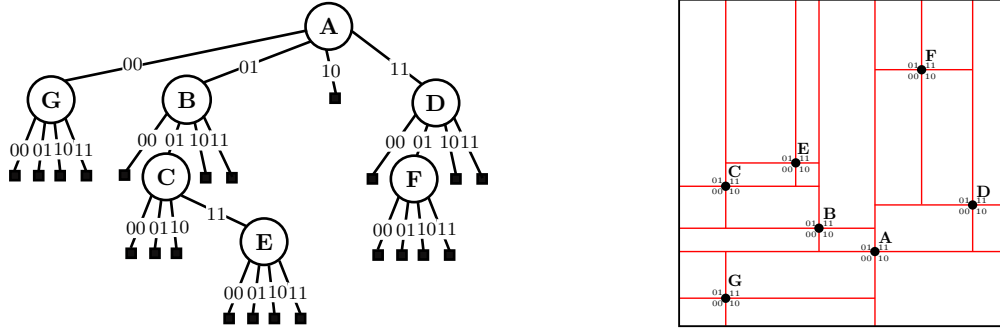
$$\nu_{s,K} \cdot \left( \prod_{i:q_i \neq *} q_i(1 - q_i) \right)^{\alpha/2} \cdot n^\alpha + l.o.t.,$$

where  $\nu_{s,K}$  is a constant that depends on  $s$ ,  $K$  and the particular query  $\mathbf{q}$  and  $\alpha = \alpha(s/K)$  is the same as for random PM queries (see above).

The paper is organised as follows. In Section 2 we give some preliminaries. We explain our methodology in Section 3 through the simplest case  $K = 2$  (Subsection 3.1). We continue with the general case of arbitrary  $s$  and  $K$  (Subsection 3.2). To complete the analysis one needs to solve an integral equation; that is the subject of Subsection 3.3. Section 4 contains some final remarks as well as some future lines of work.

## 2 Preliminaries

Let  $F$  be a collection of  $n$  multidimensional records, each one endowed with a  $K$ -dimensional key  $\mathbf{x} = (x_0, \dots, x_{K-1})$ , with coordinate  $x_j$  drawn from a totally ordered domain  $\mathcal{D}_j$ . For convenience, here we will assume that, for all  $0 \leq j < K$ ,  $\mathcal{D}_j = [0, 1]$ .



■ **Figure 1** A 2-dimensional quadtree of file  $F = \{A, B, C, D, E, F, G\}$  and the partition that it induces of the space. In this example  $F_{00} = \{G\}$ ,  $F_{01} = \{B, C, E\}$  and  $F_{0*} = \{B, C, E, G\}$ .

► **Definition 1.** A quadtree  $T$  of size  $n$  is a  $2^K$ -ary tree storing a collection  $F$  of  $n$   $K$ -dimensional records.  $T$  is either empty (when  $n = 0$ ) or each one of its  $n$  nodes holds a key from  $F$ , such that the root node of  $T$  stores a record with key  $\mathbf{x}$  and pointers to  $2^K$  subtrees, that hold the remaining  $n - 1$  records of  $F$ . Every subtree of  $T$ , let say  $T_{\mathbf{w}}$ , is associated to a bitstring  $\mathbf{w} = w_0 w_1 \dots w_{K-1} \in \{0, 1\}^K$ , in such a way that  $T_{\mathbf{w}}$  is a quadtree, and for any key  $y \in T_{\mathbf{w}}$ , it holds that  $y_j \leq x_j$  if  $w_j = 0$  and  $y_j > x_j$  if  $w_j = 1$ , for all  $0 \leq j < K$ .

Any quadtree of size  $n$  induces a partition of the domain into  $(2^K - 1)n + 1$  regions, each corresponding to a leaf (or equivalently empty subtree) in the quadtree. An example of a quadtree and the partition of the space that it induces is shown in Figure 1. To build a quadtree starting from an empty tree, each insertion of a new record with key  $\mathbf{x}$  follows a path from the root to a leaf; at each step, we compare  $\mathbf{x}$  and the key at the current node to determine in which of the  $2^K$  subtrees the insertion should continue recursively, and the process ends when a leaf is reached and it is replaced by a new node containing  $\mathbf{x}$  and  $2^K$  empty subtrees. The region associated to the substituted leaf is called the *bounding box* of the subtree rooted at  $\mathbf{x}$ . Following the same convention used for the names of the subtrees, we will denote by  $B_{\mathbf{w}}$  the bounding boxes of subtrees  $T_{\mathbf{w}}$  associated to the tree rooted at  $\mathbf{x}$  and by  $F_{\mathbf{w}}$  the subset of data points of  $F$  that fall inside  $B_{\mathbf{w}}$ .

Consider a string  $\mathbf{v}$  over the alphabet  $\Sigma = \{0, 1, *\}$ . We define as  $\mathcal{L}(\mathbf{v})$  the set of binary strings matching  $\mathbf{v}$ ; that is, where each occurrence of the symbol  $*$  stands for a 0 or a 1. For instance,  $\mathcal{L}(001) = \{001\}$ ,  $\mathcal{L}(0*1) = \{001, 011\}$  and  $\mathcal{L}(1**00) = \{10000, 10100, 11000, 11100\}$ . With this notation let us define the following extension of the notion of bounding box  $B_{\mathbf{v}} = \bigcup_{\mathbf{w} \in \mathcal{L}(\mathbf{v})} B_{\mathbf{w}}$ .

Likewise  $F_{\mathbf{v}}$  is the union of the (disjoint)  $F_{\mathbf{w}}$ 's with  $\mathbf{w}$  matching  $\mathbf{v}$ . For example, in two dimensions  $B_{**} = [0, 1]^2$  is the bounding box of the root of the quadtree,  $F_{0*}$  is the subset of all those keys with first coordinate smaller than the first coordinate of the root, that is, the ones stored in  $T_{00}$  and  $T_{01}$  (see Figure 1).

To perform a PM search with query  $\mathbf{q}$ , the quadtree is recursively explored as follows. First, we check whether the root  $\mathbf{x}$  matches  $\mathbf{q}$  or not, to report it in the former case. Then, we make recursive calls in all the  $2^{K-s}$  subtrees  $T_{\mathbf{w}}$  such that the first  $s$  bits of  $\mathbf{w}$  are such that  $w_i = 0$  whenever  $q_i \neq *$  and  $q_i \leq x_i$ , and  $w_i = 1$  whenever  $q_i \neq *$  and  $q_i > x_i$ ,  $0 \leq i < s$ , and the remaining  $K - s$  bits can be either 0 or 1.

One key observation about the PM search in quadtrees (or similar data structures) is that, except for eventual matches, only the relative ranks of the coordinates matter. Let

us call the *rank vector* of a query  $\mathbf{q}$  the vector  $\mathbf{r}(\mathbf{q}) = (r_0, \dots, r_{K-1})$  such that  $r_i = *$ , if  $q_i = *$ , and  $r_i$  is the number of records  $\mathbf{x}$  in the collection  $F$  such that  $x_i \leq q_i$  ( $0 \leq r_i \leq n$ ), if  $q_i \neq *$ . Then for any two given queries  $\mathbf{q}$  and  $\mathbf{q}'$  with equal rank vectors  $\mathbf{r}(\mathbf{q}) = \mathbf{r}(\mathbf{q}')$  the PM procedure described above will visit exactly the same set of nodes of the tree. In our analysis, we shall be using rank vectors instead of the queries themselves (as done in [6]) and consider, for instance, the cost  $\mathcal{P}_{n,\mathbf{r}}$  of a PM query with given rank vector  $\mathbf{r}$  in a random quadtree of size  $n$ . The probability model for random quadrees that we will use throughout this work is that the tree is built by inserting in any order  $n$  keys drawn independently at random (coordinate by coordinate) from a continuous distribution. For the sake of simplicity, we can safely assume that the distribution is Uniform(0, 1). Because of the symmetry of the model we can also assume that the  $s$  specified coordinates of  $\mathbf{q}$  are the first  $s$  coordinates,  $0 < s < K$ , and therefore that  $\mathbf{q} = (q_0, \dots, q_{s-1}, *, \dots, *)$  and  $\mathbf{r} = (r_0, \dots, r_{s-1}, *, \dots, *)$ . We shall write hence  $\mathbf{q} = (q_0, \dots, q_{s-1})$  and  $\mathbf{r} = (r_0, r_1, \dots, r_{s-1})$  with the convention that the implicit  $K - s$  remaining components are all  $*$ 's.

### 3 Analysis

Our goal in this section is to find the expected cost  $P_{n,\mathbf{r}} = \mathbb{E}\{\mathcal{P}_{n,\mathbf{r}}\}$ , measured as the number of visited nodes, of a PM query with a fixed rank vector  $\mathbf{r}$  in a random quadtree of  $n$  nodes.

In order to show our methodology and to give some intuition on the problem we are going to start our analysis with the easiest case  $K = 2$  in Subsection 3.1. Afterwards, in Subsection 3.2, we analyze the general case.

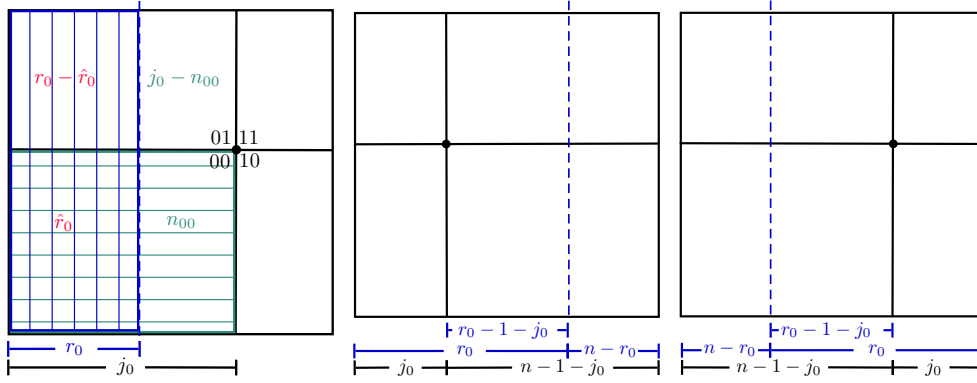
In both subsections we are going to obtain a recurrence for  $P_{n,\mathbf{r}}$ . Then, in order to solve the general recurrence, we translate it into an integral equation whose solution will give us the leading term in the asymptotic estimate for  $P_{n,\mathbf{r}}$ . The solution of the integral equation is given in Subsection 3.3.

#### 3.1 The case $K = 2$

Given a 2-dimensional quadtree  $T$ , its root splits the space into four rectangles:  $B_{00}$  (south-west of the root),  $B_{01}$  (north-west of the root),  $B_{10}$  (south-east of the root) and  $B_{11}$  (north-east of the root). These four rectangles are the corresponding *bounding boxes* of the four subtrees  $T_{00}$ ,  $T_{01}$ ,  $T_{10}$  and  $T_{11}$  from Definition 1. Recall also that  $B_{0*} = B_{00} \cup B_{01}$  and  $B_{*0} = B_{00} \cup B_{10}$  are, respectively, the rectangles west and south of the root. For any string  $\mathbf{u} \in \{0, 1, *\}^2$ , the number of data points in  $B_{\mathbf{u}}$  (equivalently, the cardinality of  $F_{\mathbf{u}}$ ) will be denoted  $\mathcal{N}_{\mathbf{u}}$ . For a random quadtree the  $\mathcal{N}_{\mathbf{u}}$ 's are random variables.

Let us now address the recurrence for  $P_{n,\mathbf{r}}$ , and to simplify let us write  $P_{n,r_0}$ , as  $\mathbf{r} = (r_0, *)$ . The basis of recursion is trivially  $P_{0,r_0} = 0$ . If  $n > 0$ , let  $\mathbf{j} = (j_0, j_1)$  be the rank vector of the root. Since  $\mathbf{q}$  contains only one specified coordinate, the relation between  $j_0$  and  $r_0$  determines whether the query intersects either  $B_{0*}$  or  $B_{1*}$ . If  $r_0 \leq j_0$ , then the query intersects  $B_{0*}$ ; otherwise it intersects  $B_{1*}$ . In our recurrence for  $P_{n,r_0}$  the value  $j_0 = \mathcal{N}_{0*} = |F_{0*}|$  run from  $r_0$  to  $n - 1$ , leading to a non-empty intersection of  $B_{0*}$  and the query, or from 0 to  $r_0 - 1$ , leading to a non-empty intersection of  $B_{1*}$  and the query. Because of the randomness assumptions, each possible value of  $\mathcal{N}_{0*}$  has probability  $1/n$  and hence this factor will weight the expected cost of the PM query conditioned to  $\mathcal{N}_{0*} = j_0$ .

The number of data points in  $B_{0*}$  is  $j_0$  by definition, and the number of data points in  $B_{1*}$  is  $n - 1 - j_0$ . If the query intersects  $B_{0*}$  then the rank of the query with respect to  $B_{0*}$  is still  $r_0$ , but if it intersects  $B_{1*}$  then its rank with respect to  $B_{1*}$  is  $r_0 - 1 - j_0$ . So the contribution to  $P_{n,r_0}$  coming from the recursive traversal of  $B_{0*}$  involves a set of  $j_0$



■ **Figure 2** A partial match in a two-dimensional quadtree. The first diagram shows the case  $r_0 \leq j_0$ , the second one the case  $j_0 < r_0$  and the third one how the east-west symmetry converts the second case into the first one.

points and the rank of the query is  $r_0$  while the contribution coming from  $B_{1*}$  involves a set  $n - 1 - j_0$  points and, because of the symmetry  $P_{n,r_0} = P_{n,n-r_0}$ , the rank of the query is  $n - r_0$ . Hence, we can reduce the case  $j_0 < r_0$  to the case  $r_0 \leq j_0$ , see Figure 2.

In the general case we would have to consider  $2^s$  regions  $B_{\mathbf{w}}$  described by bitstrings  $\mathbf{w} = w_0 \cdots w_{s-1} * \cdots *$ , where each  $w_i$  is 0 or 1 depending on whether  $r_i \leq j_i$  or not; as we consider all possible  $\mathbf{j}$ , the query will intersect these  $2^s$  different regions, and we will be able to use these “east-west” symmetry considerations to reduce their analysis to the analysis of one of them, say,  $B_{00\dots 0* \dots *}$ .

Let us come back to  $K = 2$ . The region  $B_{0*}$  is the union of the two bounding boxes  $B_{00}$  and  $B_{01}$  (in general we will consider regions  $B_{\mathbf{w}}$  that contain  $2^{K-s}$  bounding boxes) and our goal is to use further symmetries to reduce the analysis of the cost of traversing both bounding boxes to the analysis of just traversing one of them, say,  $B_{00}$ .

Let  $Q_{j_0, r_0}$  be the contribution to the expected cost of a PM query due to the recursive call in  $T_{00}$ , when the query has rank  $r_0$  in the first coordinate and given that there are  $j_0 \geq r_0$  nodes to the west of the root.

Suppose that  $\mathcal{N}_{00} = n_{00}$ . The rank vector of the query in the recursive call to  $T_{00}$  will be  $(\hat{r}_0, *)$ , and the contribution to the expected cost will then be  $P_{n_{00}, \hat{r}_0}$ . So it only remains to determine: a) the probability that  $\mathcal{N}_{00} = n_{00}$ , given the rank vector of the root  $\mathbf{j}$  and, b) the probability that the rank vector of the query with respect to  $B_{00}$  is  $(\hat{r}_0, *)$ . Let us define the subsets of data points  $F'_{\mathbf{v}}$  and the corresponding bounding boxes  $B'_{\mathbf{v}}$  like  $F_{\mathbf{v}}$  and  $B_{\mathbf{v}}$ , but with respect to the given query, instead of the root. The value  $\hat{r}_0$  is the number of data points in the intersection between  $B_{00}$  and  $B'_{0*}$ , see Figure 2. We will use  $\mathcal{R}_{\langle \mathbf{0} \rangle} := |F_{00} \cap F'_{0*}|$ .

In general,  $\langle \mathbf{i} \rangle := *^i 0 *^{K-1-i}$ , so using this convention, we can also write  $\mathcal{N}_{\langle \mathbf{0} \rangle} = j_0$  and  $|F'_{\langle \mathbf{0} \rangle}| = r_0$ . Conditioned on the sizes of  $F_{00}$ ,  $F_{\langle \mathbf{0} \rangle}$  and  $F'_{\langle \mathbf{0} \rangle}$ , the random variable  $\mathcal{R}_{\langle \mathbf{0} \rangle}$  obeys a hypergeometric distribution:

$$\Pr \left\{ \mathcal{R}_{\langle \mathbf{0} \rangle} = \hat{r}_0 \mid \mathcal{N}_{00} = n_{00}, \mathcal{N}_{\langle \mathbf{0} \rangle} = j_0, |F'_{\langle \mathbf{0} \rangle}| = r_0 \right\} = \frac{\binom{n_{00}}{\hat{r}_0} \binom{j_0 - n_{00}}{r_0 - \hat{r}_0}}{\binom{j_0}{r_0}}.$$

Now if we look at the contribution to the expected cost due to the traversal of  $T_{01}$ , we have that  $\mathcal{N}_{01} = j_0 - n_{00}$  and the rank of the query with respect to  $B_{01}$  is  $(r_0 - \hat{r}_0, *)$ . The fact that the second coordinate is unspecified allow us to do the analysis above with  $n_{01}$  instead of  $n_{00}$  and we would have obtained symmetric formulas. We can exploit this

north-south symmetry that will give us a factor of 2. Taking into account the visit to the root and our discussion so far we can write

$$P_{n,r_0} = 1 + \frac{2}{n} \left( \sum_{j_0=0}^{r_0-1} Q_{n-1-j_0, n-r_0} + \sum_{j_0=r_0}^{n-1} Q_{j_0, r_0} \right), \quad (3)$$

where, for  $n_{0*} \geq r$ , we have

$$Q_{j_0, r_0} = \sum_{n_{00}=0}^{j_0} \Pr \{ \mathcal{N}_{00} = n_{00} \mid \mathcal{N}_{\langle 0 \rangle} = j_0 \} \sum_{\hat{r}_0=0}^{r_0} \left( \frac{\binom{n_{00}}{\hat{r}_0} \binom{j_0-n_{00}}{r_0-\hat{r}_0}}{\binom{j_0}{r_0}} P_{n_{00}, \hat{r}_0} \right). \quad (4)$$

To complete the recurrence for  $P_{n,r_0}$  we need only to obtain the probability that  $\mathcal{N}_{00} = n_{00}$ , conditioned on  $\mathcal{N}_{\langle 0 \rangle} = j_0$ . Since  $\mathcal{N}_{\langle 1 \rangle}$  can take any value in  $[0..n-1]$  with identical probability, the number of points in  $B_{00}$  will take any value between 0 and  $j_0$  with identical probability  $1/(j_0+1)$ . Plugging this probability and (4) into (3) yields to the desired recurrence for  $P_{n,r_0}$ .

An asymptotic estimate of the main term of  $P_{n,r_0}$  follows by deriving an integral equation for  $f(z_0) := \lim_{n \rightarrow \infty} P_{n,z_0 n}/n^\alpha$  and solving that integral equation. We give the details of the derivation of the integral equation in the case of  $K=2$  in Lemma 4.

### 3.2 The general case

Let  $\mathbf{r} = (r_0, r_1, \dots, r_{s-1})$  be the query rank vector and let  $\mathbf{j} = (j_0, \dots, j_{s-1})$  be the first  $s$  coordinates of the rank vector for the root of the random quadtree. Thus we have that  $j_i$  is the value of  $|F_{\langle \mathbf{i} \rangle}| = \mathcal{N}_{\langle \mathbf{i} \rangle}$ . These  $K$  strings of the form  $\langle \mathbf{i} \rangle$  constitute a “basis” in the sense that we can obtain any region  $B_{\mathbf{w}}$  by complementation ( $B_{*^i 1 *^{K-1-i}} = B_{*...*} \setminus B_{\langle \mathbf{i} \rangle}$ ) and intersection of the appropriate  $B_{\langle \mathbf{i} \rangle}$ ’s.

Like we did for  $K=2$  our goal is to use the symmetries of the problem to reduce the whole analysis to the analysis of the contribution to the total cost of one particular subtree, namely,  $T_{0^s}$ . Again, call  $Q_{\mathbf{j}, \mathbf{r}}$  the contribution of the recursive call in  $T_{0^s}$ , conditioned to  $r_i \leq j_i$  for all  $i$ ,  $0 \leq i < s$ . This condition guarantees that the PM search will recursively continue in that subtree.

Then, because of the  $K-s$  symmetries on unspecified coordinates (like the north-south symmetry of the case  $K=2$ ) and because of the  $s$  symmetries for specified coordinates (like the east-west symmetry when  $K=2$ ), we can express  $P_{n, \mathbf{r}}$  in terms of  $Q_{\mathbf{j}, \mathbf{r}}$ ’s. In particular, considering all the possibilities for  $\mathbf{j}$  gives a factor  $1/n^s$ , and a summation over all bitstrings  $\mathbf{w}$  of length  $s$  to cover the cases where the query intersects  $B_{\mathbf{w}}$ . Finally the factor  $2^{K-s}$  stems from the  $2^{K-s}$  bounding boxes that each  $B_{\mathbf{w}}$  contains. Hence,

$$P_{n, \mathbf{r}} = 1 + \frac{2^{K-s}}{n^s} \sum_{\mathbf{w} \in \{0,1\}^s} \sum_{j_0} \cdots \sum_{j_{s-1}} Q_{\mathbf{j}'_{\mathbf{w}}(\mathbf{j}), \mathbf{r}'_{\mathbf{w}}(\mathbf{r})}, \quad (5)$$

where the summation ranges are  $r_i \leq j_i \leq n-1$  if  $w_i = 0$ , and  $0 \leq j_i \leq r_i - 1$  if  $w_i = 1$ , and the rank vectors  $\mathbf{j}'_{\mathbf{w}} = (j'_0, \dots, j'_{s-1})$  and  $\mathbf{r}'_{\mathbf{w}} = (r'_0, \dots, r'_{s-1})$  are defined as follows: if  $w_i = 0$  then  $j'_i = j_i$  and  $r'_i = r_i$ , otherwise if  $w_i = 1$  then  $j'_i = n-1-j_i$  and  $r'_i = n-r_i$ .

For any  $i$ ,  $0 \leq i < K$ , we will denote  $\mathbf{0}^i$  the string  $0^i *^{K-i}$ , that is, a string of length  $K$  consisting of  $i$  zeros, followed by  $K-i$  \*’s.

The method to obtain a formula for  $Q_{\mathbf{j}, \mathbf{r}}$  consists of the following steps: 1) First we use Lemma 5 to obtain the probability distribution of the number of data points  $\mathcal{N}_{\mathbf{0}^s}$  in the

“corner” hyperrectangle, by intersecting the sets  $F_{\langle 0 \rangle}, F_{\langle 1 \rangle}, \dots, F_{\langle s-1 \rangle}$ , with sizes  $j_0, \dots, j_{s-1}$ , respectively. This will be expressed by  $s-1$  “hypergeometric” sums that will give us the probability that  $\mathcal{N}_{\mathbf{0}^s} = \ell_s$ ; 2) Given that the last  $K-s$  coordinates are unspecified, and conditioned on  $j_i = \mathcal{N}_{\langle i \rangle}$ ,  $0 \leq i < s$ , all the potential sizes of  $\mathcal{N}_{\langle i \rangle} = |F_{\langle i \rangle}|$ ,  $s \leq i < K$ , are equiprobable. This will be expressed by  $K-s$  “uniform” sums that will allow us to derive the probability distribution for  $\mathcal{N}_{\mathbf{0}^K}$ , and 3) Now conditioning on  $\mathcal{N}_{\mathbf{0}^K} = |F_{\mathbf{0}^K}|$ , and given  $\mathbf{r}$  we intersect  $F_{\mathbf{0}^K}$  with each of  $F'_{\langle 0 \rangle}, F'_{\langle 1 \rangle}, \dots, F'_{\langle s-1 \rangle}$  to obtain the components of  $\mathbf{r}_{\mathbf{0}^K} = (\hat{r}_0, \dots, \hat{r}_{s-1})$ . We will denote  $\mathcal{R}_{\langle i \rangle} = |F_{\mathbf{0}^K} \cap F'_{\langle i \rangle}|$  the random variable that gives the  $i$ -th component of  $\mathbf{r}_{\mathbf{0}^K}$ . As in the case  $K=2$ , the probability distribution of the  $\mathcal{R}_{\langle i \rangle}$ ’s is hypergeometric and it will lead to  $s$  additional “hypergeometric” sums.

Therefore the general formula for  $Q_{\mathbf{j}, \mathbf{r}}$  is:

$$Q_{\mathbf{j}, \mathbf{r}} = \sum_{\ell_s=0}^{j_{s-1}} \Pr \left\{ \mathcal{N}_{\mathbf{0}^s} = \ell_s \left| \bigwedge_{i=0}^{s-1} \mathcal{N}_{\langle i \rangle} = j_i \right. \right\} \times \sum_{\ell_K=0}^{\ell_s} \Pr \left\{ \mathcal{N}_{\mathbf{0}^K} = \ell_K \left| \mathcal{N}_{\mathbf{0}^s} = \ell_s \right. \right\} \\ \times \sum_{\mathbf{r}_{\mathbf{0}^K}=(\hat{r}_0, \dots, \hat{r}_{s-1})} \Pr \left\{ \bigwedge_{i=0}^{s-1} \mathcal{R}_{\langle i \rangle} = \hat{r}_i \left| \mathcal{N}_{\mathbf{0}^K} = \ell_K, \bigwedge_{i=0}^{s-1} |F'_{\langle i \rangle}| = r_i \right. \right\} \times P_{\ell_K, \mathbf{r}_{\mathbf{0}^K}}. \quad (6)$$

We can expand this last expression as:

$$Q_{\mathbf{j}, \mathbf{r}} = \sum_{\ell_s=0}^{j_{s-1}} \dots \sum_{\ell_2=0}^{j_1} \left( \frac{\binom{j_0}{\ell_2} \binom{n-1-j_0}{j_1-\ell_2}}{\binom{n-1}{j_1}} \dots \frac{\binom{\ell_{s-1}}{\ell_s} \binom{n-1-\ell_{s-1}}{j_{s-1}-\ell_s}}{\binom{n-1}{j_{s-1}}} \right) \\ \times \frac{1}{\ell_s+1} \sum_{\ell_{s+1}=0}^{\ell_s} \dots \frac{1}{\ell_{K-1}+1} \sum_{\ell_K=0}^{\ell_{K-1}} \\ \sum_{\hat{r}_0=0}^{\ell_K \wedge r_0} \frac{\binom{\ell_K}{\hat{r}_0} \binom{j_0-\ell_K}{r_0-\hat{r}_0}}{\binom{j_0}{r_0}} \dots \sum_{\hat{r}_{s-1}=0}^{\ell_K \wedge r_{s-1}} \frac{\binom{\ell_K}{\hat{r}_{s-1}} \binom{j_{s-1}-\ell_K}{r_{s-1}-\hat{r}_{s-1}}}{\binom{j_{s-1}}{r_{s-1}}} P_{\ell_K, (\hat{r}_0, \dots, \hat{r}_{s-1})}, \quad (7)$$

where we have used  $x \wedge y = \min(x, y)$  to stress the intersections that are involved in each case, e.g.  $\hat{r}_i$  ranges from 0 to  $\ell_K \wedge r_i$  since the number of data points is given by  $|F_{\mathbf{0}^K} \cap F'_{\langle i \rangle}|$ ; with  $|F_{\mathbf{0}^K}| = \mathcal{N}_{\mathbf{0}^K} = \ell_K$  and  $|F'_{\langle i \rangle}| = r_i$ .

To derive the integral equation corresponding to the recurrence above we can use arguments similar to those in the case  $K=2$ . We give all the details of this derivation, as well as other necessary technical lemmas in Appendix A.

► **Lemma 2.** *If  $f(z_0, \dots, z_{s-1}) = \lim_{n \rightarrow \infty} \frac{P_{n, \mathbf{r}}}{n^\alpha}$  exists, with  $\alpha = \alpha(s/K)$  the solution of the indicial equation (1) and  $z_i = \lim_{n \rightarrow \infty} r_i/n$ ,  $0 < z_i < 1$ , for all  $i$ ,  $0 \leq i < s$ , then  $f(z_0, \dots, z_{s-1})$  is the unique solution of*

$$f(z_0, \dots, z_{s-1}) = \left( \frac{2}{\alpha+1} \right)^{K-s} \times \sum_{\mathbf{w} \in (0+1)^s} \left\{ \int_{I_{w_0}(z_0)} \dots \int_{I_{w_{s-1}}(z_{s-1})} f\left(\varphi_{w_0}(z_0, u_0), \dots, \varphi_{w_{s-1}}(z_{s-1}, u_{s-1})\right) \right. \\ \left. \cdot \left( \psi_{w_0}(u_0) \dots \psi_{w_{s-1}}(u_{s-1}) \right)^\alpha du_{s-1} \dots du_0 \right\}, \quad (8)$$

where  $I_0(z) = [0, z]$ ,  $I_1(z) = [z, 1]$ ,  $\psi_0(u) = 1-u$ ,  $\psi_1(u) = u$ ,  $\varphi_0(z, u) = (1-z)/(1-u)$  and  $\varphi_1(z, u) = z/u$ , which satisfies the following boundary conditions:

1.  $f(z_0, \dots, z_{s-1})$  is symmetric on all variables, that is, for any  $i$  and  $j$ ,

$$f(z_0, \dots, z_i, \dots, z_j, \dots, z_{s-1}) = f(z_0, \dots, z_j, \dots, z_i, \dots, z_{s-1}).$$

2. For any  $z_i \in (0, 1)$ ,  $0 \leq i < s$ ,  $f$  is symmetric with respect to the axis  $z_i = 1/2$ , that is,

$$f(z_0, \dots, z_i, \dots, z_{s-1}) = f(z_0, \dots, 1 - z_i, \dots, z_{s-1}).$$

3. For any  $i$ ,  $0 \leq i < s$ ,

$$\lim_{z_i \rightarrow 0^+} f(z_0, \dots, z_i, \dots, z_{s-1}) = \lim_{z_i \rightarrow 1^-} f(z_0, \dots, z_i, \dots, z_{s-1}) = 0.$$

- 4.

$$\int_0^1 \int_0^1 \cdots \int_0^1 f(z_0, \dots, z_{s-1}) dz_0 \cdots dz_{s-1} = \beta_{s,K}.$$

**Proof.** We will follow a procedure similar to the one in the proof of Lemma 4, which covers the case  $K = 2$ .

The steps that we will give to obtain the integral equation for general  $K$  are:

1. Apply Lemma 6 to (7)  $s$  times in the  $s$  hypergeometric sums (the last sums over the  $\hat{r}_i$ 's)
2. Convert the  $K - s$  uniform sums (the middle sums over the  $\ell_i$ 's,  $s < i \leq K$ ) into the corresponding integral by passing to the limit. That gives  $K - s$  factors  $1/(\alpha + 1)$ .
3. Apply Lemma 7 once to the first  $s - 1$  hypergeometric sums (over the  $\ell_i$ 's,  $2 \leq i \leq s$ ).
4. Convert all the sums in (5) into integrals by passing to the limit.

Here, we use  $\ell_i$  to denote the values that the random variables  $\mathcal{N}_{0^i}$  can take, like we did in subsection 3.2, and in particular in (6) and successive.

Defining  $f\left(\frac{r_0}{n}, \dots, \frac{r_{s-1}}{n}\right) := P_{n,r}/n^\alpha$ , where  $\alpha$  is the solution of the indicial equation for quadrees, we get:

$$\begin{aligned} \frac{Q_{j,r}}{n^\alpha} &= \sum_{\ell_s=0}^{j_{s-1}} \cdots \sum_{\ell_2=0}^{j_1} \left( \frac{\binom{j_0}{\ell_2} \binom{n-1-j_0}{j_1-\ell_2}}{\binom{n-1}{j_1}} \cdots \frac{\binom{\ell_{s-1}}{\ell_s} \binom{n-1-\ell_{s-1}}{j_{s-1}-\ell_s}}{\binom{n-1}{j_{s-1}}} \right) \\ &\quad \times \frac{1}{\ell_s + 1} \sum_{\ell_{s+1}=0}^{\ell_s} \cdots \frac{1}{\ell_{K-1} + 1} \sum_{\ell_K=0}^{\ell_{K-1}} \\ &\quad \sum_{\hat{r}_0=0}^{\ell_K \wedge r_0} \frac{\binom{\ell_K}{\hat{r}_0} \binom{j_0-\ell_K}{r_0-\hat{r}_0}}{\binom{j_0}{r_0}} \cdots \sum_{\hat{r}_{s-1}=0}^{\ell_K \wedge r_{s-1}} \frac{\binom{\ell_K}{\hat{r}_{s-1}} \binom{j_1-\ell_K}{r_{s-1}-\hat{r}_{s-1}}}{\binom{j_{s-1}}{r_{s-1}}} \times f\left(\frac{\hat{r}_0}{\ell_K}, \dots, \frac{\hat{r}_{s-1}}{\ell_K}\right) \left(\frac{\ell_K}{n}\right)^\alpha. \end{aligned}$$

Hence, defining  $u_{0^i} = \lim_{n \rightarrow \infty} (\ell_i/n)$  for  $s \leq i \leq K$ ,  $z_i = \lim_{n \rightarrow \infty} (r_i/n)$  and  $u_i =$



$\lim_{n \rightarrow \infty} (j_i/n)$  for  $0 \leq i < K$  and applying Lemma 6  $s$  times:

$$\begin{aligned}
\lim_{n \rightarrow \infty} \frac{Q_{\mathbf{j}, \mathbf{r}}}{n^\alpha} &= \lim_{n \rightarrow \infty} \sum_{\ell_s=0}^{j_{s-1}} \cdots \sum_{\ell_2=0}^{j_1} \left( \frac{\binom{j_0}{\ell_2} \binom{n-1-j_0}{j_1-\ell_2}}{\binom{n-1}{j_1}} \cdots \frac{\binom{\ell_{s-1}}{\ell_s} \binom{n-1-\ell_{s-1}}{j_{s-1}-\ell_s}}{\binom{n-1}{j_{s-1}}} \right) \\
&\quad \times \frac{1}{\ell_s+1} \sum_{\ell_{s+1}=0}^{\ell_s} \cdots \frac{1}{\ell_{K-1}+1} \sum_{\ell_K=0}^{\ell_{K-1}} f\left(\frac{r_0}{j_0}, \dots, \frac{r_{s-1}}{j_{s-1}}\right) \left(\frac{\ell_K}{n}\right)^\alpha \\
&= \lim_{n \rightarrow \infty} \sum_{\ell_s=0}^{j_{s-1}} \cdots \sum_{\ell_2=0}^{j_1} \left( \frac{\binom{j_0}{\ell_2} \binom{n-1-j_0}{j_1-\ell_2}}{\binom{n-1}{j_1}} \cdots \frac{\binom{\ell_{s-1}}{\ell_s} \binom{n-1-\ell_{s-1}}{j_{s-1}-\ell_s}}{\binom{n-1}{j_{s-1}}} \right) \\
&\quad \times \frac{1}{u_0^s} \int_0^{u_0^s} \cdots \frac{1}{u_{0^{K-1}}} \int_0^{u_{0^{K-1}}} f\left(\frac{z_0}{u_0}, \dots, \frac{z_{s-1}}{u_{s-1}}\right) u_{0^K}^\alpha du_{0^K} \cdots du_{0^{s+1}} \\
&= \lim_{n \rightarrow \infty} \sum_{\ell_s=0}^{j_{s-1}} \cdots \sum_{\ell_2=0}^{j_1} \left( \frac{\binom{j_0}{\ell_2} \binom{n-1-j_0}{j_1-\ell_2}}{\binom{n-1}{j_1}} \cdots \frac{\binom{\ell_{s-1}}{\ell_s} \binom{n-1-\ell_{s-1}}{j_{s-1}-\ell_s}}{\binom{n-1}{j_{s-1}}} \right) \\
&\quad \times f\left(\frac{z_0}{u_0}, \dots, \frac{z_{s-1}}{u_{s-1}}\right) \frac{u_{0^s}^\alpha}{(\alpha+1)^{K-s}}.
\end{aligned}$$

Replacing  $u_{0^s}$  by  $\ell_s/n$  and applying Lemma 7 once to the first  $s-1$  hypergeometric sums we obtain:

$$\lim_{n \rightarrow \infty} \frac{Q_{\mathbf{j}, \mathbf{r}}}{n^\alpha} = \frac{1}{(\alpha+1)^{K-s}} f\left(\frac{z_0}{u_0}, \dots, \frac{z_{s-1}}{u_{s-1}}\right) \prod_{i=0}^{s-1} u_i^\alpha. \quad (9)$$

Finally, introduce the following notation:  $I_0(z) = [0, z]$ ,  $I_1(z) = [z, 1]$ ,  $\varphi_0(z, u) = (1-z)/(1-u)$  and  $\varphi_1(z, u) = z/u$ . Plugging (9) into (5)) and passing to the limit (the fourth step in the procedure that we have described) yields the stated integral equation.  $\blacktriangleleft$

Conditions 1 and 2 in the lemma follow from the combinatorics of the problem. By symmetry,  $P_{n, \mathbf{r}} = P_{n, \mathbf{r}'}$  for any permutation  $\mathbf{r}'$  of the rank vector  $\mathbf{r}$ . Likewise, if  $\mathbf{r} = (r_0, \dots, r_i, \dots, r_{s-1})$  and  $\mathbf{r}' = (r_0, \dots, r_{i-1}, n-r_i, r_{i+1}, \dots, r_{s-1})$  then  $P_{n, \mathbf{r}} = P_{n, \mathbf{r}'}$ . Condition 3 needs an inductive argument in the number of non-extreme ( $z_i \neq 0$  and  $z_i \neq 1$ ) coordinates. When all specified coordinates are extreme, say,  $z_0 = z_1 = \dots = z_{s-1} = 0$  we must have  $f = 0$ ; indeed, it is very easy to prove that  $P_{n, (0, \dots, 0)} = o(n^\alpha)$ . We do not give here a complete and detailed analysis when  $s_0 \leq s$  specified coordinates are extreme; the computations and the reasoning is analogous to that carried out in [6] for  $K$ -d trees. Last but not least, Condition 4 follows by summing the expected cost  $P_{n, \mathbf{r}}$  over all possible rank vectors  $\mathbf{r}$  and dividing by  $(n+1)^s$ : it must yield the known expected cost of a random partial match query  $\beta_{s, K} n^\alpha + o(n^\alpha)$ . In terms of  $f$ , we must integrate  $f$  in the domain  $[0, 1]^s$  to obtain  $\beta_{s, K}$ . For a detailed justification the reader can refer to [6]: it is straightforward to adapt the discussion there to the case of quadrees.

### 3.3 Solving the integral equation

From the integral equation (8) in Lemma 2 we can obtain an equivalent partial differential equation (PDE) by application of the differential operators

$$\Phi_j(f) = z_j(1-z_j) \frac{\partial^2 f}{\partial z_j^2} + \alpha(2z_j-1) \frac{\partial f}{\partial z_j} - \alpha(\alpha+1)f.$$

Indeed, if we define the operator

$$I_i(f) = z_i^{\alpha+1} \int_{z_i}^1 f(z_0, \dots, z_{i-1}, u_i, z_{i+1}, \dots, z_{s-1}) \frac{du_i}{u_i^{\alpha+2}} + \\ (1 - z_i)^{\alpha+1} \int_0^{z_i} f(z_0, \dots, z_{i-1}, v_i, z_{i+1}, \dots, z_{s-1}) \frac{dv_i}{(1 - v_i)^{\alpha+2}}$$

then the integral equation (8) in Lemma 2 can be written as

$$f = \left( \frac{2}{\alpha + 1} \right)^{K-s} I_0(I_1(\dots(I_{s-1}(f))\dots)),$$

using the changes of variables  $u_i := z_i/u_i$  and  $v_i := (1 - z_i)/(1 - u_i)$ .

Then, as

$$\Phi_i(I_j(g)) = \Psi_i(g) = (2z_i - 1) \frac{\partial g}{\partial z_i} - 2\alpha g$$

it follows that

$$\Phi_0(\Phi_1(\dots(\Phi_{s-1}(f))\dots)) = \left( \frac{2}{\alpha + 1} \right)^{K-s} \Phi_0(\Phi_1(\dots(\Phi_{s-1}(I_0(I_1(\dots(I_{s-1}(f))\dots))\dots))).$$

Now, since  $\Phi_i$ 's and  $\Psi_i$ 's commute –  $\Phi_i(\Phi_j(g)) = \Phi_j(\Phi_i(g))$ ,  $\Psi_i(\Psi_j(g)) = \Psi_j(\Psi_i(g))$  – and  $\Phi_i(\Psi_j(g)) = \Psi_j(\Phi_i(g))$  for any  $i \neq j$ , we can manipulate the equation above to get

$$\Phi_0(\Phi_1(\dots(\Phi_{s-1}(f))\dots)) = \left( \frac{2}{\alpha + 1} \right)^{K-s} \Psi_0(\Psi_1(\dots(\Psi_{s-1}(f))\dots))$$

or

$$\left( \Phi_0 \circ \Phi_1 \circ \dots \circ \Phi_{s-1} - \left( \frac{2}{\alpha + 1} \right)^{K-s} \Psi_0 \circ \Psi_1 \circ \dots \circ \Psi_{s-1} \right)(f) = 0, \quad (10)$$

which is the sought PDE, succinctly expressed in terms of the linear differential operators  $\Phi_i$  and  $\Psi_i$ ,  $i = 0, \dots, s-1$ .

The resulting PDE is homogeneous and linear, hence it is natural to try to solve it by separation of variables. The shape of equation (10) also cries out to try a solution in separated variables. Therefore, we will assume that the solution to the integral equation (8) is a function:  $f(z_0, z_1, \dots, z_{s-1}) = \phi_0(z_0) \cdot \phi_1(z_1) \cdot \dots \cdot \phi_{s-1}(z_{s-1})$ .

Given that the function  $f$  is symmetric with respect to any permutation of its arguments, we can also safely assume that all the functions  $\phi_0, \phi_1, \dots, \phi_{s-1}$  are the same function  $\phi$ . Rather than working with the PDE itself, we may use our assumption to rewrite equation (8) as:

$$\phi(z_0) \cdot \phi(z_1) \cdot \dots \cdot \phi(z_{s-1}) = \left( \frac{2}{\alpha + 1} \right)^{K-s} \prod_{i=0}^{s-1} \left( \int_0^{z_i} \phi\left(\frac{1-z_i}{1-u_i}\right) (1-u_i)^\alpha du_i + \int_{z_i}^1 \phi\left(\frac{z_i}{u_i}\right) u_i^\alpha du_i \right). \quad (11)$$

If  $\phi$  is a solution of the following equation

$$\phi(z) = \left( \frac{2}{\alpha + 1} \right)^{\frac{K-s}{s}} \left( \int_0^z \phi\left(\frac{1-z}{1-u}\right) (1-u)^\alpha du + \int_z^1 \phi\left(\frac{z}{u}\right) u^\alpha du \right), \quad (12)$$

then it would be a solution of equation (11). As shown in [4],

$$\phi(z) = \mu(z(1-z))^{\delta-1}, \quad \delta = \left(\frac{2}{\alpha+1}\right)^{\frac{K-s}{s}},$$

is such a solution, where  $\mu$  is an arbitrary constant and we have discarded additional terms in the general solution based on symmetry considerations.

Because the exponent  $\alpha$  is a solution to the indicial equation (1) it follows that  $\delta = \frac{\alpha}{2} + 1$  and hence the solution to (8) is:

$$f(z_0, \dots, z_{s-1}) = \nu_{s,K} \cdot \prod_{i=0}^{s-1} (z_i(1-z_i))^{\alpha/2},$$

where  $\nu_{s,K}$  is a constant that depends on  $s$  and  $K$  only. To finish our derivation and to obtain the value of  $\nu_{s,K}$  we replace  $f$  by the expression above in Condition 4 of Lemma 2 and we get:

$$\nu_{s,K} \left( \int_0^1 (z(1-z))^{\alpha/2} dz \right)^s = \nu_{s,K} \left( \frac{\Gamma(\alpha/2 + 1)^2}{\Gamma(\alpha + 2)} \right)^s = \beta_{s,K},$$

so we can use the expression for  $\beta_{s,K}$  in Equation (2) to find an explicit formula for  $\nu_{s,K}$ .

To argue unicity of the solution, we should begin noticing that the linear homogeneous PDE satisfied by the function  $f$  has all real-analytic coefficients in the domain  $(0, 1)^s$ , because the coefficients of the operators  $\Psi_i$  and  $\Phi_i$  are analytic too in that domain and the PDE results from the composition of such operators.

Moreover, the highest derivative in the PDE is  $\partial^{2s} f / \partial z_0^2 \cdots \partial z_{s-1}^2$  and its coefficient  $\prod_{0 \leq i < s} z_i(1-z_i)$  is clearly always positive in  $(0, 1)^s$ , hence, the PDE is elliptic. Then, by Holmgren's theorem, any solution is real-analytic; and from Cauchy-Kovalevskaya theorem it follows that it must be unique, since this last theorem guarantees that there is a unique real-analytic solution (see for instance [8, 11]). Altogether, these results tell us that the solution that we have found, starting from the *ansatz* that it admitted a representation in separable variables, is unique.

It remains to verify by direct substitution that  $P_{n,\mathbf{r}} = f(\mathbf{r}/n)n^\alpha$  is a solution of recurrence (5) replacing the independent term by  $o(1)$ , which is the error resulting from approximating the summations by integrals. With this our main result follows.

► **Theorem 3.** *If  $\lim_{n \rightarrow \infty} \frac{P_{n,\mathbf{r}}}{n^\alpha}$  exists then the expected cost  $P_{n,\mathbf{r}}$  of a PM query with given rank vector  $\mathbf{r}$  such that  $r_i = z_i n + o(n)$  for some  $z_i \in (0, 1)$ ,  $0 \leq i < s$ , in a random  $K$ -dimensional quadtree of size  $n$  is*

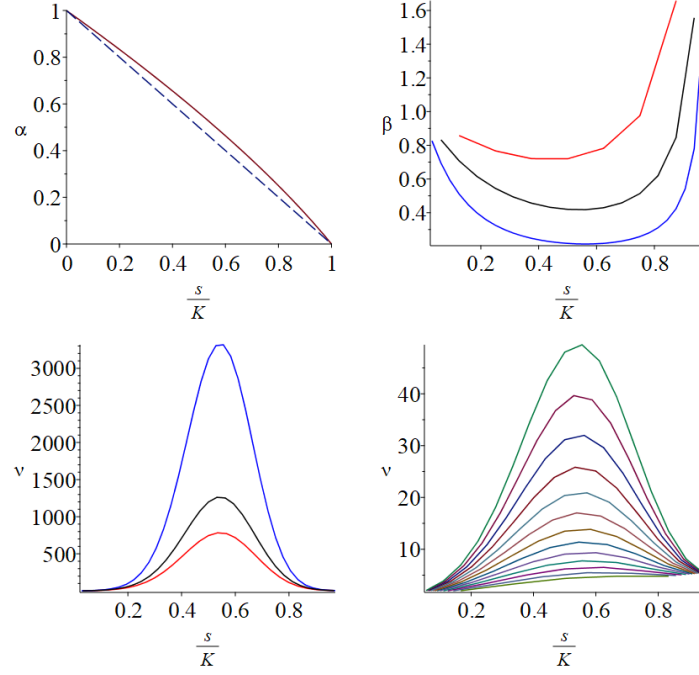
$$P_{n,\mathbf{r}} = \nu_{s,K} \left( \prod_{i=0}^{s-1} z_i(1-z_i) \right)^{\alpha/2} n^\alpha + o(n^\alpha),$$

where  $\alpha$  is the unique solution in  $(0, 1)$  of

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

$$\nu_{s,K} = \frac{1}{(2^{K-s} - 1) \Gamma(\alpha + 1)^{K-s} \Gamma(\alpha/2 + 1)^{2s}} \prod_{2 \leq j \leq K} \frac{\Gamma(\alpha - \alpha_j)}{\Gamma(-\alpha_j)},$$

and the  $\alpha_j$ 's, with  $\alpha = \alpha_1 > \Re(\alpha_2) \geq \cdots \geq \Re(\alpha_K)$ , are the roots of the indicial equation above.



■ **Figure 3** Variation of the exponent  $\alpha(s/K)$  (top-left),  $\beta(s, K)$  for  $K \in \{8, 16, 32\}$  (top-right) and  $\nu(s, K)$  for  $K \in \{30, 32, 36\}$  (bottom-left), as well as  $\nu(s, K)$  for all  $6 \leq K \leq 18$  (bottom-right).

Figure 3 depicts how the exponent  $\alpha = \alpha(s/K)$ , and the constants  $\beta(s, K)$  and  $\nu(s, K)$  vary with respect to  $s$  and  $K$ . In all cases, the  $x$ -axis is  $s/K$  to ease the comparison –  $\alpha$  is a function of  $s/K$  alone, but  $\beta$  and  $\nu$  depend on both  $s$  and  $K$ . In the graphs for  $\beta(s, K)$  and  $\nu(s, K)$  we have drawn three curves in each case, corresponding to  $K = 8$  (red),  $K = 16$  (black) and  $K = 32$  (blue) in the graph for  $\beta(s, K)$ , and  $K = 30$  (red),  $K = 32$  (black) and  $K = 36$  (blue) in the graph for  $\nu(s, K)$ . Moreover in the graph of  $\alpha(s/K)$  we have also plotted  $1 - s/K$  (dashed line) for reference. For fixed  $K$ ,  $\beta(s, K)$  is a convex function with a minimum close to  $s = K/2$  but slowly shifted to the right. Likewise, for fixed  $K$ ,  $\nu(s, K)$  is a bell-shaped function with a single global maximum near  $s = K/2$  but also slightly shifted to the right ( $\nu(s, K)$  is not defined for  $s = K$ ). If we denote  $\nu^*(K) = \max_{0 < s < K} \nu(s, K)$  the graph shows that  $\nu^*(K)$  grows with  $K$ . On the other hand, the graph and further numerical computations suggest that there is a limiting curve  $\beta_\infty(x) = \lim_{K \rightarrow \infty} \beta(\lfloor xK \rfloor, K)$  that is a lower bound for any  $\beta(s, K)$  as  $K \rightarrow \infty$ .

When  $s = 0$  (no coordinate is specified), we have  $\alpha(0) = \beta(0, K) = \nu(0, K) = 1$ , despite all these constant are not well defined when  $s = 0$ . Notice that for  $s = 0$  the partial match degenerates to a full traversal of the quadtree and visits its  $n$  nodes.

In the opposite situation, when all coordinates are specified,  $s = K$ ,  $\beta$  and  $\nu$  are undefined, and  $\alpha(1) = 0$ . The expected cost of a partial match is not  $\Theta(1) = \Theta(n^0)$  but  $\Theta(\log n)$  as it is actually an exact search.

## 4 Conclusions and Future Work

Our main result, Theorem 3, gives the main order term of the expected cost  $P_{n,r}$  of a PM search with a fixed query of rank vector  $\mathbf{q}$ , for quadrees of any dimension  $K$  and any number

of specified coordinates. It can be easily translated to an equivalent result in terms of the coordinates  $q_i$  of the query, namely,

$$P_{n,\mathbf{q}} = \nu_{s,K} \cdot \left( \prod_{i:q_i \neq *} q_i(1 - q_i) \right)^{\alpha/2} \cdot n^\alpha + \text{l.o.t.}$$

under the assumption of uniformity of the coordinates of the data points (see, for instance, [6]).

We show that quadrees behave qualitatively as standard and relaxed  $K$ -d trees [6]. There we conjectured that the form of the expected cost of a PM search with fixed query would have the same “shape” for a wide variety of multidimensional data structures, excluding those producing very balanced partitions of the space (e.g., quadtries, squarish  $K$ -d trees). Duch and Lau [5] have disproved the conjecture, in its broadest terms, as it does not apply to locally balanced  $K$ -d trees. However, it seems that the conjecture might hold for hierarchical multidimensional data structures where: 1) no balancing of subtrees occurs; 2) the partition at each node follows a fixed rule independent of the current data point.

From the methodological viewpoint, we systematically exploit the many symmetries that appear in the problem to simplify its formulation and to make its mathematical manipulation feasible.

Several open problems remain. To begin with, the existence of  $\lim_{n \rightarrow \infty} \frac{P_{n,\mathbf{r}}}{n^\alpha}$ , which has been rigorously proved for  $K = 2$  in [3] (also in [1]); our result in that case coincides with the previous ones. We are currently working in the proof of the existence of the required limit for general  $K$ ; meanwhile, our results follow from the – yet unproven – assumption that such limit exists. We shall mention that there is compelling evidence that this is the case. On the other hand, the existence of a limiting distribution for  $\mathcal{P}_{n,\mathbf{r}}/n^\alpha$  has been shown only for the case of standard 2-d trees and 2-dimensional quadrees, but not for other data structures or larger dimensions, and this is a question worth of further study.

Another goal for future research, more technical in nature but also more ambitious, is to develop tools that would allow a straightforward, (semi-)automatic derivation of the recurrences or distributional equations, the proof of the existence of the limiting distribution, the corresponding integral equations for the expectation and other higher order moments, etc. This kind of techniques would ease the obtainment of results, such as the ones in previous literature and the ones in this paper, for many other multidimensional data structures and it might also open the door for “universality” results such as the ones conjectured in [6].

---

## References

- 1 Nicolas Broutin, Ralph Neininger, and Henning Sulzbach. Partial match queries in random quadrees. In Yuval Rabani, editor, *Proc. of the 23<sup>rd</sup> Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1056–1065, 2012.
- 2 H.-H. Chern and H.-K. Hwang. Partial match queries in random quadrees. *SIAM J. Comput.*, 32:904–915, 2003.
- 3 N. Curien and A. Joseph. Partial match queries in two-dimensional quadrees: A probabilistic approach. *Advances in Applied Probability*, 43:178–194, 2011.
- 4 A. Duch, R. M. Jiménez, and C. Martínez. Selection by rank in  $k$ -dimensional binary search trees. *Random Structures and Algorithms*, 2012. doi:10.1002/rsa.20476.
- 5 Amalia Duch and Gustavo Lau. Partial match queries in relaxed  $K$ -dt trees. In *Proc. of the Fourteenth ACM-SIAM Workshop on Analytic Algorithmics and Combinatorics (ANALCO)*, pages 131–138, 2017. doi:10.1137/1.9781611974775.13.

- 6 Amalia Duch, Gustavo Lau, and Conrado Martínez. On the cost of fixed partial match queries in k-d trees. *Algorithmica*, 75(4):684–723, 2016. doi:10.1007/s00453-015-0097-4.
- 7 Philippe Flajolet, Gaston Gonnet, Claude Puech, and John Michael Robson. Analytic variations on quad trees. *Algorithmica*, 10:473–500, 1993.
- 8 Gerald B. Folland. *Introduction to Partial Differential Equations*. Princeton University Press, 2nd edition, 1995.
- 9 Steven G. Krantz and Harold R. Parks. *A Primer of Real Analytic Functions*. Birkhäuser, 2nd edition, 2002.
- 10 S. Ross. *A First Course in Probability*. Prentice Hall, Upper Saddle River, New Jersey, 8th edition, 2010.
- 11 Daniel Zwillinger. *Handbook of Differential Equations*. Academic Press, 3rd edition, 1997.

## A Technical Lemmas

► **Lemma 4.** If  $f(z) = \lim_{n \rightarrow \infty} \frac{P_{n,r}}{n^\alpha}$  exists, with  $\alpha = \alpha(1/2)$  the solution of the indicial equation (1) when  $s = 1$  and  $K = 2$ , and  $z = \lim_{n \rightarrow \infty} r/n$ ,  $0 < z < 1$ , then

$$f(z) = \frac{2}{\alpha + 1} \left( \int_0^z f\left(\frac{1-z}{1-u}\right) (1-u)^\alpha du + \int_z^1 f\left(\frac{z}{u}\right) u^\alpha du \right). \quad (13)$$

The symmetry  $P_{n,r_0} = P_{n,n-r_0}$  implies that in general  $f(z) = f(1-z)$  and in particular  $f\left(\frac{1-z}{1-u}\right) = f\left(1 - \frac{1-z}{1-u}\right)$  from where it follows that equation (13) is the same as the one for standard 2d-trees (see [4]).

**Proof.** Let  $f(r_0/n) := P_{n,r_0}/n^\alpha$ . Then we have that

$$\frac{P_{a,(b,*)}}{n^\alpha} = f\left(\frac{b}{a}\right) \left(\frac{a}{n}\right)^\alpha$$

and therefore, substituting into (4)

$$\begin{aligned} \frac{Q_{j_0,r_0}}{n^\alpha} &= \frac{1}{j_0 + 1} \sum_{n_{00}=0}^{j_0} \sum_{\hat{r}_0=0}^{r_0} \left( \frac{\binom{n_{00}}{\hat{r}_0} \binom{j_0-n_{00}}{r_0-\hat{r}_0}}{\binom{j_0}{r_0}} f\left(\frac{\hat{r}_0}{n_{00}}\right) \left(\frac{n_{00}}{n}\right)^\alpha \right) \\ &= \frac{1}{j_0 + 1} \sum_{n_{00}=0}^{j_0} \sum_{\hat{r}_0=0}^{r_0} \left( \frac{\binom{n_{00}}{\hat{r}_0} \binom{j_0-n_{00}}{r_0-\hat{r}_0}}{\binom{j_0}{r_0}} f\left(\frac{\hat{r}_0}{j_0} \frac{j_0}{n} \frac{n}{n_{00}}\right) \left(\frac{n_{00}}{n}\right)^\alpha \right) \end{aligned}$$

The last sum is the expected value of a function of a hypergeometric random variable. Passing to the limit when  $n \rightarrow \infty$ , Lemma 6 allows us to exchange the expected value and the function. Therefore passing to the limit when  $n \rightarrow \infty$ , with  $z = \lim_{n \rightarrow \infty} (r/n)$ ,  $u_{0*} = \lim_{n \rightarrow \infty} (j_0/n)$ ,  $u_{00} = \lim_{n \rightarrow \infty} (n_{00}/n)$ , and assuming that  $f$  is real analytic in Lemma 6 we can apply it to get:

$$\begin{aligned} \lim_{n \rightarrow \infty} \frac{Q_{j_0,r_0}}{n^\alpha} &= \frac{1}{u_{0*}} \int_0^{u_{0*}} f\left(\frac{u_{00}}{u_{0*}} \frac{z}{u_{0*}} \frac{u_{0*}}{u_{00}}\right) u_{00}^\alpha du_{00} = \frac{1}{u_{0*}} \int_0^{u_{0*}} f\left(\frac{z}{u_{0*}}\right) u_{00}^\alpha du_{00} \\ &= \frac{1}{\alpha + 1} f\left(\frac{z}{u_{0*}}\right) u_{0*}^\alpha. \end{aligned}$$

and similarly

$$\begin{aligned} \lim_{n \rightarrow \infty} \frac{Q_{n-1-j_0,n-r_0}}{n^\alpha} &= \frac{1}{1-u_{0*}} \int_0^{1-u_{0*}} f\left(\frac{1-z}{1-u_{0*}}\right) u_{00}^\alpha du_{00} \\ &= \frac{1}{\alpha + 1} f\left(\frac{1-z}{1-u_{0*}}\right) (1-u_{0*})^\alpha \end{aligned}$$

Since  $j_0 = 0 \implies u_{0*} = 0$ ,  $j_0 = r_0 \implies u_{0*} = z_0$  and in the limit  $j_0 = r_0 - 1 \implies u_{0*} = z_0$ ,  $j_0 = n - 1 \implies u_{0*} = 1$  and  $\frac{\Delta j_0}{n} \rightarrow du_{0*}$  replacing in (3) and passing to the limit we obtain this integral equation:

$$\begin{aligned} f(z_0) &= 2 \int_0^{z_0} \frac{1}{1-u_{0*}} f\left(\frac{1-z_0}{1-u_{0*}}\right) \int_0^{1-u_{0*}} u_{00}^\alpha du_{00} du_{0*} \\ &\quad + 2 \int_{z_0}^1 \frac{1}{u_{0*}} f\left(\frac{z_0}{u_{0*}}\right) \int_0^{u_{0*}} u_{00}^\alpha du_{00} du_{0*} \\ &= 2 \int_0^{z_0} \frac{1}{1-u_{0*}} f\left(\frac{1-z_0}{1-u_{0*}}\right) \frac{(1-u_{0*})^{\alpha+1}}{\alpha+1} du_{0*} + 2 \int_{z_0}^1 \frac{1}{u_{0*}} f\left(\frac{z_0}{u_{0*}}\right) \frac{u_{0*}^{\alpha+1}}{\alpha+1} du_{0*}. \end{aligned}$$

Replacing now in (3)), passing to the limit  $n \rightarrow \infty$  and, to simplify, replacing  $u_{0*}$  by  $u$  we get the integral equation (13) in the statement of the Lemma.  $\blacktriangleleft$

► **Lemma 5.** *Given a random  $K$  dimensional quadtree with  $n$  data points the conditional probability that  $\mathcal{N}_{0^K} = \ell_K$  given that  $\mathcal{N}_{\langle i \rangle} = n_{\langle i \rangle}$  for  $0 \leq i \leq K-1$  is:*

$$\Pr \left\{ \mathcal{N}_{0^K} = \ell_K \mid \bigwedge_{i=0}^{K-1} \mathcal{N}_{\langle i \rangle} = n_{\langle i \rangle} \right\} = \sum_{\ell_{K-1}=0}^{n_{\langle i \rangle K-2}} \cdots \sum_{\ell_3=0}^{n_{\langle 2 \rangle}} \sum_{\ell_2=0}^{n_{\langle 1 \rangle}} \left( \frac{\binom{\ell_1}{\ell_2} \binom{n-1-\ell_1}{n_{\langle 1 \rangle}-\ell_2}}{\binom{n-1}{n_{\langle 1 \rangle}}} \right. \\ \left. \frac{\binom{\ell_2}{\ell_3} \binom{n-1-\ell_2}{n_{\langle 2 \rangle}-\ell_3}}{\binom{n-1}{n_{\langle 2 \rangle}}} \cdots \frac{\binom{\ell_{K-2}}{\ell_{K-1}} \binom{n-1-\ell_{K-2}}{n_{\langle K-2 \rangle}-\ell_{K-1}}}{\binom{n-1}{n_{\langle K-2 \rangle}}} \frac{\binom{\ell_{K-1}}{\ell_K} \binom{n-1-\ell_{K-1}}{n_{\langle K-1 \rangle}-\ell_K}}{\binom{n-1}{n_{\langle K-1 \rangle}}} \right). \quad (14)$$

**Proof.** In the base case  $K = 2$  given  $n$ ,  $\mathcal{N}_{\langle 0 \rangle} \equiv \mathcal{N}_{0*} = n_{0*}$  and  $\mathcal{N}_{\langle 1 \rangle} \equiv \mathcal{N}_{*0} = n_{*0}$ , the probability that the intersection of the rectangles  $B_{\langle 0 \rangle} = B_{0*}$  and  $B_{\langle 1 \rangle} = B_{*0}$  contains  $\ell_2 = n_{00}$  nodes is the probability of having  $\ell_2 = n_{00}$  successes in  $n_{*0}$  draws without replacement from a population of size  $n-1$  that contains  $n_{0*}$  successes. It is  $n-1$  instead of  $n$  because the root cannot be in the intersections. Therefore the distribution is hypergeometric:

$$\Pr \{ \mathcal{N}_{00} = n_{00} \mid \mathcal{N}_{0*} = n_{0*}, \mathcal{N}_{*0} = n_{*0} \} = \frac{\binom{n_{0*}}{n_{00}} \binom{n-1-n_{0*}}{n_{*0}-n_{00}}}{\binom{n-1}{n_{*0}}}.$$

Assume that the lemma is true for  $K$  dimensions. We can do the inductive step based on writing the intersection of  $K+1$  sets as an intersection of  $K$  sets followed by the intersection of two sets:

$$\bigcap_{i=0}^K F_{*^i 0^{*K-i}} = \left( \bigcap_{i=0}^{K-1} F_{*^i 0^{*K-i}} \right) \cap F_{*^K 0} = F_{0^K *} \cap F_{*^K 0} = F_{0^K+1}.$$

Taking into account all the possible values of  $\mathcal{N}_{0^K *}$ , we have:

$$\begin{aligned} &\Pr \left\{ \mathcal{N}_{0^{K+1}} = n_{0^{K+1}} \mid \bigwedge_{i=0}^K \mathcal{N}_{*^i 0^{*K-i}} = n_{*^i 0^{*K-i}} \right\} \\ &= \sum_{n_{0^K *}=0}^{n_{*^K-1} 0^{*}} \left( \Pr \left\{ \mathcal{N}_{0^K *} = n_{0^K *} \mid \bigwedge_{i=0}^{K-1} \mathcal{N}_{*^i 0^{*K-i}} = n_{*^i 0^{*K-i}} \right\} \right. \\ &\quad \times \Pr \left\{ \mathcal{N}_{0^{K+1}} = n_{0^{K+1}} \mid \mathcal{N}_{0^K *} = n_{0^K *}, \mathcal{N}_{*^K 0} = n_{*^K 0} \right\} \Bigg) \\ &= \sum_{n_{0^K *}=0}^{n-1} \left( \Pr \left\{ \mathcal{N}_{0^K *} = n_{0^K *} \mid \bigwedge_{i=0}^{K-1} \mathcal{N}_{*^i 0^{*K-i}} = n_{*^i 0^{*K-i}} \right\} \times \frac{\binom{n_{0^K *}}{n_{0^K+1}} \binom{n-1-n_{0^K *}}{n_{*^K 0}-n_{0^K+1}}}{\binom{n-1}{n_{*^K 0}}} \right), \end{aligned}$$

applying the inductive hypothesis (14) (adding a  $*$  to the end of each string) completes the proof. Notice that we have used  $\ell_i$  instead of  $n_{0^i * K-i}$  and  $n_{\langle 1 \rangle} = n_{*^i 0 * K-1-i}$  in the statement of the theorem.  $\blacktriangleleft$

► **Lemma 6.** *Given a random two dimensional quadtree let  $\mathcal{N}_{0*}$ ,  $\mathcal{N}_{*0}$  and  $\mathcal{N}_{00}$  be respectively the random variables of the number of nodes west, south and south-west of the root. If  $f$  is a real analytic function [9] in  $(0, 1)$ ,  $\lim_{n \rightarrow \infty} n_{0*}/n = u_{0*}$  and  $\lim_{n \rightarrow \infty} n_{*0}/n = u_{*0}$ , where  $u_{0*}, u_{*0} \in (0, 1)$ , then*

$$\begin{aligned} \lim_{n \rightarrow \infty} \mathbb{E} \left\{ f \left( \frac{\mathcal{N}_{00}}{n} \right) \middle| \mathcal{N}_{0*} = n_{0*}, \mathcal{N}_{*0} = n_{*0} \right\} &= \lim_{n \rightarrow \infty} \sum_{n_{00}=0}^{n_{*0}} \left( \frac{\binom{n_{0*}}{n_{00}} \binom{n-n_{0*}}{n-n_{00}}}{\binom{n}{n_{*0}}} f \left( \frac{n_{00}}{n} \right) \right) \\ &= \lim_{n \rightarrow \infty} \sum_{n_{00}=0}^{n_{*0}} \left( \frac{\binom{n_{0*}}{n_{00}} \binom{n-n_{0*}}{n-n_{00}}}{\binom{n}{n_{*0}}} f \left( \frac{n_{00}}{n} \right) \right) \\ &= f(u_{0*} u_{*0}). \end{aligned} \quad (15)$$

**Proof.** For simplicity, in the hypergeometric probability formulas we have replaced  $n-1$  by  $n$  as in the limit they are the same.

Since  $f$  is real analytic all derivatives of  $f$  exist in  $(0, 1)$  and we can write, for some  $x_0 \in (0, 1)$ ,

$$f(x) = \sum_{i=0}^{\infty} a_i (x - x_0)^i = \sum_{i=0}^{\infty} a_i \sum_{k=0}^i \binom{i}{k} (-x_0)^{i-k} x^k.$$

Since the series on the right side converges we can use the linearity of expectations:

$$\mathbb{E} \{ f(x) \} = \sum_{i=0}^{\infty} a_i \sum_{k=0}^i \binom{i}{k} (-x_0)^{i-k} \mathbb{E} \{ x^k \}.$$

Therefore we only need to prove the lemma for  $f(x) = x^k$ . If  $X_{n,m,N}$  is a hypergeometric random variable with parameters  $n$ ,  $m$ , and  $N$  then [10]:

$$\mathbb{E} \{ X_{n,m,N}^k \} = \frac{nm}{N} \mathbb{E} \{ (X_{n-1,m-1,N-1} + 1)^{k-1} \}.$$

Based on that it is easy to prove by induction that for every  $k \in \mathbb{N}$  there are integers  $c_{k,i}$ , with  $c_{k,k} = 1$ , such that:

$$\mathbb{E} \{ X_{n,m,N}^k \} = \sum_{i=0}^k c_{k,i} \frac{n^i m^i}{N^i}.$$

Therefore if  $f(x) = x^k$ :

$$\begin{aligned} \mathbb{E} \left\{ f \left( \frac{\mathcal{N}_{00}}{n} \right) \middle| \mathcal{N}_{0*} = n_{0*}, \mathcal{N}_{*0} = n_{*0} \right\} &= \mathbb{E} \left\{ \frac{\mathcal{N}_{00}^k}{n^k} \middle| \mathcal{N}_{0*} = n_{0*}, \mathcal{N}_{*0} = n_{*0} \right\} \\ &= \frac{\sum_{i=0}^k c_{k,i} \frac{n_{0*}^i n_{*0}^i}{n^i}}{n^k} = \sum_{i=0}^k c_{k,i} \frac{n_{0*}^i n_{*0}^i}{n^i n^k}. \end{aligned}$$

In the last sum the only term that does not go to zero as  $n \rightarrow \infty$  is the last one, where  $i = k$ . Given that  $c_{k,k} = 1$ , we have:

$$\lim_{n \rightarrow \infty} \mathbb{E} \left\{ \frac{\mathcal{N}_{00}^k}{n^k} \middle| \mathcal{N}_{0*} = n_{0*}, \mathcal{N}_{*0} = n_{*0} \right\} = \lim_{n \rightarrow \infty} \frac{\frac{n_{0*}^k n_{*0}^k}{n^k}}{n^k} = \lim_{n \rightarrow \infty} \left( \frac{n_{0*}}{n} \right)^k \left( \frac{n_{*0}}{n} \right)^k = u_{0*}^k u_{*0}^k.$$

That proves the lemma for  $f(x) = x^k$ .  $\blacktriangleleft$



The lemma can be generalised to any dimension  $K$  using mathematical induction on the number of dimensions, again assuming that the function  $f$  is real analytic (in several variables).

► **Lemma 7.** *Given a random quadtree let  $\mathcal{N}_{(i)}$  be the random variable of the number of data points that have their  $i$ -th coordinate less than the  $i$ -th coordinate of the root and the rest of the coordinates undetermined and let  $\mathcal{N}_{0^K}$  be the random variable of the size of the cuboid where all the coordinates have values lower than the respective coordinates of the root. If  $f$  is real analytic in  $(0, 1)^K$ ,  $\lim_{n \rightarrow \infty} n_{(i)}/n = u_i$  for  $0 \leq i < K$ , where  $u_i \in (0, 1)$ , then*

$$\lim_{n \rightarrow \infty} \mathbb{E} \left\{ f\left(\frac{\mathcal{N}_{0^K}}{n}\right) \middle| \bigwedge_{i=0}^{K-1} \mathcal{N}_{(i)} = n_{(i)} \right\} = f\left(\prod_{i=0}^{K-1} u_i\right).$$

**Proof.** The base case  $K = 2$  has been proved. Assume that the lemma is true for  $K$  dimensions. Then:

$$\begin{aligned} & \lim_{n \rightarrow \infty} \mathbb{E} \left\{ f\left(\frac{\mathcal{N}_{0^{K+1}}}{n}\right) \middle| \bigwedge_{i=0}^K \mathcal{N}_{*^i 0^{*K-i}} = n_{*^i 0^{*K-i}} \right\} \\ &= \lim_{n \rightarrow \infty} \sum_{n_{0^{K+1}}=0}^{n-1} \Pr \left\{ \mathcal{N}_{0^{K+1}} = n_{0^{K+1}} \middle| \bigwedge_{i=0}^K \mathcal{N}_{*^i 0^{*K-i}} = n_{*^i 0^{*K-i}} \right\} f\left(\frac{n_{0^{K+1}}}{n}\right) \\ &= \lim_{n \rightarrow \infty} \sum_{n_{0^{K+1}}=0}^{n-1} \sum_{n_{0^{K*}}=0}^{n-1} \left( \Pr \left\{ \mathcal{N}_{0^{K*}} = n_{0^{K*}} \middle| \bigwedge_{i=0}^{K-1} \mathcal{N}_{*^i 0^{*K-i}} = n_{*^i 0^{*K-i}} \right\} \right. \\ & \quad \times \Pr \left\{ \mathcal{N}_{0^{K+1}} = n_{0^{K+1}} \middle| \mathcal{N}_{0^{K*}} = n_{0^{K*}}, \mathcal{N}_{*^K 0} = n_{*^K 0} \right\} f\left(\frac{n_{0^{K+1}}}{n}\right) \\ &= \lim_{n \rightarrow \infty} \sum_{n_{0^{K*}}=0}^{n-1} \Pr \left\{ \mathcal{N}_{0^{K*}} = n_{0^{K*}} \middle| \bigwedge_{i=0}^{K-1} \mathcal{N}_{*^i 0^{*K-i}} = n_{*^i 0^{*K-i}} \right\} \\ & \quad \times \mathbb{E} \left\{ f\left(\frac{\mathcal{N}_{0^{K+1}}}{n}\right) \middle| \mathcal{N}_{0^{K*}} = n_{0^{K*}}, \mathcal{N}_{*^K 0} = n_{*^K 0} \right\} \\ &= \lim_{n \rightarrow \infty} \sum_{n_{0^{K*}}=0}^{n-1} \Pr \left\{ \mathcal{N}_{0^{K*}} = n_{0^{K*}} \middle| \bigwedge_{i=0}^{K-1} \mathcal{N}_{*^i 0^{*K-i}} = n_{*^i 0^{*K-i}} \right\} \times f\left(\lim_{n \rightarrow \infty} \frac{n_{0^{K*}} n_{*^K 0}}{(n-1)n}\right) \\ &= \lim_{n \rightarrow \infty} \mathbb{E} \left\{ f\left(\lim_{n \rightarrow \infty} \frac{\mathcal{N}_{0^{K*}} n_{*^K 0}}{(n-1)n}\right) \middle| \bigwedge_{i=0}^{K-1} \mathcal{N}_{*^i 0^{*K-i-1} 0} = n_{*^i 0^{*K-i-1} 0} \right\}. \end{aligned}$$

Replacing  $n-1$  by  $n$ , because in the limit they are equivalent, and using the induction hypothesis (adding 0 at the end of each string) we have:

$$\begin{aligned} & \lim_{n \rightarrow \infty} \mathbb{E} \left\{ f\left(\frac{\mathcal{N}_{0^{K+1}}}{n}\right) \middle| \bigwedge_{i=0}^K \mathcal{N}_{*^i 0^{*K-i}} = n_{*^i 0^{*K-i}} \right\} = f\left(\lim_{n \rightarrow \infty} \left( \prod_{i=0}^{K-1} \frac{n_{*^i 0^{*K-i-1} 0}}{n_{*^K 0}} \right) \frac{n_{*^K 0}}{n}\right) \\ &= f\left(\lim_{n \rightarrow \infty} \prod_{i=0}^K \frac{n_{*^i 0^{*K-i}}}{n}\right) = f\left(\prod_{i=0}^K u_{*^i 0^{*K-i}}\right). \end{aligned}$$

◀

► **Lemma 8.** *The real function  $f(x) = x^a(1-x)^a$  is real analytic, i. e. it is infinitely differentiable and agrees with its Taylor series, in the interval  $(0, 1)$  for any real number  $a$ .*

## 20:18 Fixed Partial Match Queries in Quadrees

**Proof.** By the binomial series, or Newton's generalized binomial theorem,  $f_1(x) = (1 - x)^a$  is real analytic in  $(-1, 1)$  and  $f_2(x) = x^a = (1 + (x - 1))^a$  is real analytic in  $(0, 2)$ . Therefore their product,  $f(x)$ , is real analytic in  $(0, 1)$ . ◀