A limit field for orthogonal range searches in two-dimensional random point search trees

Broutin, Nicolas; Sulzbach, Henning

DOI:
10.1016/j.spa.2018.08.014

A limit field for orthogonal range searches in two-dimensional random point search trees

Nicolas Broutin * Henning Sulzbach †

December 1, 2017

Abstract

We consider the cost of general orthogonal range queries in random quadtrees. The cost of a given query is encoded into a (random) function of four variables which characterize the coordinates of two opposite corners of the query rectangle. We prove that, when suitably shifted and rescaled, the random cost function converges uniformly in probability towards a random field that is characterized as the unique solution to a distributional fixed-point equation. Our results imply for instance that the worst case query satisfies the same asymptotic estimates as a typical query, and thereby resolve an old question of Chanzy, Devroye and Zamora-Cura [*Acta Inf.*, 37:355–383, 2000].

AMS 2010 subject classifications.
Key words. quadtree, random partition, convergence in distribution, contraction method, range query, partial match, analysis of algorithms.

1 Introduction

1.1 Quadtrees and structures for geometric data

Geometric data are central in a number of practical contexts, such as computer graphics, management of geographical data or statistical analysis. Data structures storing such information should allow for efficient dictionary operations such as updating the data base and retrieving data matching specified patterns. For general references on multidimensional data structures and more details about their various applications, see the series of monographs by Samet [26, 27, 28].

We are interested in tree-like data structures which permit efficient execution of search queries. In applications, one of the essential basic type of queries type are *(orthogonal) range queries*, which ask to report all data located inside some axis parallel rectangular region. Such queries include the case when some of the projections of the rectangular region on the axis are either reduced to a point or span the entire domain. So, in particular, range queries cover the following two cases:

- When the pattern specifies precisely all the data fields (the query rectangle is a point), we speak of an *exact match*. Such queries can typically be answered in time logarithmic in the size of the database, since only one branch needs to be explored.
- When the projections of the query rectangle on the different axes are either points or the entire domain, we speak of a *partial match*. In general, such searches explore multiple branches of the data structure to report the matching data, and the cost usually becomes polynomial.

We are interested in the comparison-based setting, where the data may be compared directly at unit cost. In this context, a few general purpose data structures generalizing binary search trees permit to answer orthogonal range queries, namely the quadtree [20], the k-d tree [1] and the relaxed k-d tree [18]. Since range queries are at the same time an essential building block of many other algorithms and still rather

*PRES Sorbonne Universités, UPMC Université Paris 06, LPMA (UMR 7599). **Email:** nicolas.broutin@upmc.fr. **Grant**: ANR-14-CE25- 0014 (ANR GRAAL)

†University of Birmingham, School of Mathematics, B15 2TT, Birmingham, UK. **Email**: henning.sulzbach@gmail.com. The author's research was partially supported by a Feodor Lynen Fellowship of the Alexander von Humboldt-Foundation.

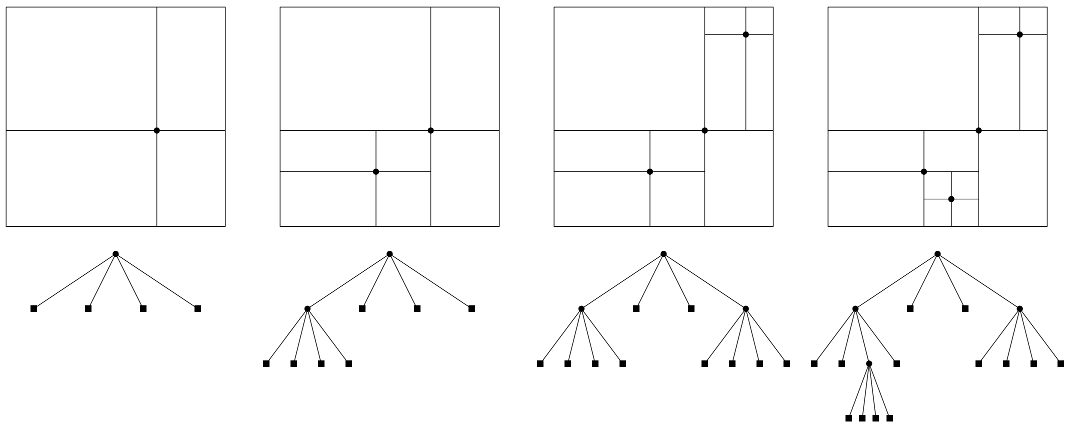

Figure 1: *The first four steps in the quadtree construction of the partition, and of the corresponding trees. The placeholders are marked with squares while the actual nodes, which 'store' one of the points, are depicted with disks.*

elementary, one would expect that their complexity should be fully understood by now. This is not the case: despite their importance, a precise quantification of the complexity of range queries in the data structures listed above is still missing. We will shortly review the literature precisely, but let us for now point out that, before the present document, even the average value in the context of uniformly random data points and a uniformly random query was only known up to a multiplicative factor.

In this paper, we provide refined analyses of the costs of orthogonal range queries in the two-dimensional data structures mentioned earlier. We mostly focus on quadtrees, but our results also apply (modulo some easy model-specific modifications) to the case of 2-d trees and relaxed 2-d trees. We only sketch the results for 2-d trees, since the phenomena at hand and the proofs are completely analogous, and the cases of partial match queries have been treated in [7] (see Section 6); the case of relaxed 2-d trees is also similar, but we leave it as an exercise to keep the present paper as concise as possible. Our results provide, for the first time, a study of the extent of the fluctuations of the complexity, precise asymptotic estimates for all moments and convergence in distribution, jointly for all axis-parallel rectangular queries.

Remarks. We emphasize the fact that we are interested in *general purpose data structures*: for instance, the range trees of Bentley [2] (and their close relatives, see [3]) are complex data structures tailored to answer range queries very fast, but this is mostly a theoretical benchmark since the space required is super-linear in the number of data points. On the other hand, the squarish k-d trees introduced by Devroye, Jabbour, and Zamora-Cura [16] would fit; unfortunately, while it is very interesting, our results do not apply to these trees. We plan to address the complexity of orthogonal range queries in squarish k-d trees in the near future.

Before going any further, let us introduce two-dimensional quadtrees. Given a sequence of points (also called *keys*) $(p_i)_{i\ge 1} \in [0,1]^2$ we define a sequence of quadtrees $(T_n)_{n\ge 1}$, together with a recursive partition of the unit square. We can see the construction algorithm as inserting the points successively and "storing" them in a node of a quaternary tree. We proceed as follows: Initially, we think of T_0 as an empty tree, which consists of a placeholder to which we assign the unit square. The first point p_1 is inserted in this placeholder and becomes the root, thereby giving rise to four placeholders. Geometrically, p_1 decomposes the unit square into four rectangular regions $Q_1, \dots, Q_4$ each of which is assigned to a child of the root (currently placeholders). Suppose that we have constructed T_n by successive insertion of $p_1, \dots, p_n$, and that T_n induces a partition of the square into $1 + 3n$ rectangles, each one assigned to one of the $1 + 3n$ placeholders of T_n. The next point p_{n+1} is then placed in the placeholder, say v, that is assigned to the rectangle containing p_{n+1}. This operation turns v into a node and creates four new placeholders just below. Geometrically, p_{n+1} divides this rectangle into four subregions that are assigned to the four newly created placeholders. See Figure 1 for an illustration of the first few steps.

1.2 The cost of queries in random quadtrees

In a standard probabilistic model, the quadtree T_n is constructed from the first n points $X_1, \dots, X_n$ of a single infinite sequence of i.i.d. random variables $(X_i)_{i\ge 1}$ with uniform distribution on $[0,1]^2$.

In this setting, an orthogonal range query asks to retrieve all points contained inside a given rectangle with (opposite) corners (a, c) and (b, d). This covers partial match queries, aiming at all elements with a given value for the first coordinate, regardless of the second one (taking the rectangle given by $(a, 0)$ and $(a, 1)$). Further, the set-up also includes a fully specified query aiming at determining whether the structure contains the element (a, c) (taking the corner (a, c) and (a, c)). The search algorithm explores recursively the nodes of the trees corresponding to regions that have a non-empty intersection with the query. As a basic measure of complexity, we consider the number of nodes inspected by the algorithm. (See Figures 2 and 3.)

Fully specified queries and distances. Here, the cost corresponds to the number of nodes on the search path, or, equivalently to the number of rectangles that contain the query point in the full refining family of nested partitions. In our probabilistic model, fully specified queries are well understood. It is known that the corresponding mean complexity is asymptotic to $\log n$ [15, 23], the variance grows like $\frac{1}{2} \log n$ [23], and that there is a central limit theorem [14, 21]. For all these results, note that this essentially corresponds to the cost of a *random* query, but adding the reference to a point in $[0, 1]^2$ would not make a difference (except on the edges, where the leading constant is different). The extreme cost is the height of the tree, and Devroye has proved that it is asymptotic to $\alpha \log n$, where $\alpha = 2.15\dots$ [13].

Partial match queries. The history case of partial match queries stretches over a much longer period and has only been finely determined very recently [6, 7]. For $t \in [0, 1]$, let $C_n(t)$ denote the number of nodes visited by a query in T_n retrieving all keys with first coordinate t. Equivalently, $C_n(t) + 1$ is given by the number of rectangles in the partition of $[0, 1]^2$ induced by T_n intersecting the vertical line at t. See Figure 2 for an illustration. Throughout the document, ξ denotes a generic random variable with uniform distribution on $[0, 1]$ which is stochastically independent of $(X_i)_{i\geq 1}$. In their seminal work on properties of random quadtrees, Flajolet, Gonnet, Puech, and Robson [23] considered the case of random queries and showed that

$$\mathbb{E}\left[C_n(\xi)\right] = \kappa n^{\beta} + O(1), \quad n \to \infty, \tag{1}$$

where

$$\beta = \frac{\sqrt{17} - 3}{2}, \qquad \text{and} \qquad \kappa = \frac{\Gamma(2\beta + 3)}{2\Gamma(\beta + 1)^3}, \tag{2}$$

and $\Gamma(\,\cdot\,)$ denotes the Gamma function. In fact, Flajolet, Gonnet, Puech, and Robson [23] provide a full asymptotic expansion for $\mathbb{E}[C_n(\xi)]$ which was generalized to the higher dimensional case by Chern and Hwang [9]. In the early 2000s, using the recursive approach underlying the results obtained by Flajolet et al. [23], there were some attempts to obtain more detailed asymptotic information such as the variance or the limit distribution of $C_n(\xi)$. However, it turns out that for a random query ξ, there is no obvious recursion for higher moments since the query couples the subproblems (the location is the same in subproblems!) To circumvent this issue, one can consider fixed queries. It is only recently that this route has been successfully explored: Curien and Joseph [12] proved that, for fixed $t \in [0, 1]$, as $n \to \infty$,

$$\mathbb{E}[C_n(t)] = K_1 h(t) n^{\beta} + o(n^{\beta}), \tag{3}$$

where

$$h(t) = (t(1-t))^{\beta/2}, \qquad \text{and} \qquad K_1 = \frac{\Gamma(2\beta + 2)\Gamma(\beta + 2)}{2\Gamma(\beta + 1)^3 \Gamma(\beta/2 + 1)^2}. \tag{4}$$

In the joint work [7] (see also [6]) with Ralph Neininger relying on the functional contraction method developed in [25], we established a distributional functional limit theorem for the rescaled process $n^{-\beta} C_n$: upon considering C_n as a right-continuous step function on $[0, 1]$, we have the following convergence in distribution

$$n^{-\beta} C_n \xrightarrow[n\to\infty]{d} \mathcal{Z}, \tag{5}$$

where the limit process $\mathcal{Z}$ is continuous and is further discussed in Section 2 below. Here, the convergence is in the space of càdlàg functions on the unit interval. See [4, Section 3] for background. Based on this result, Curien [11] was able to show that the convergence actually holds in probability (and, for any fixed $t \in [0, 1]$, even almost surely).

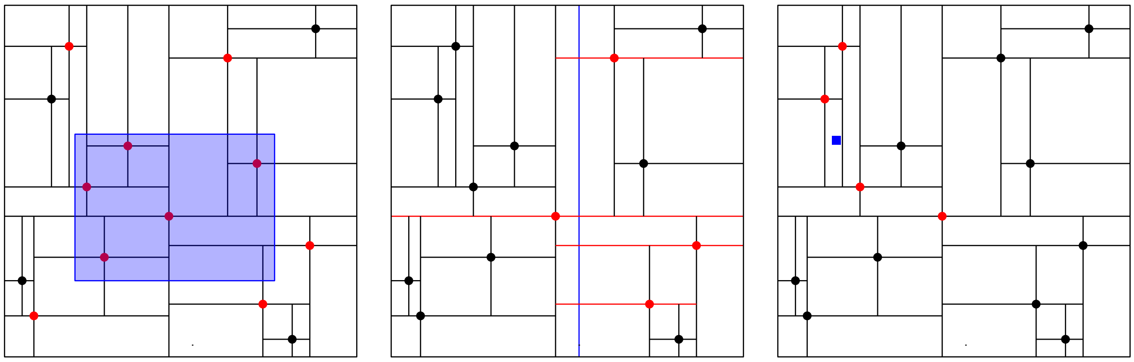

Figure 2: The partition induced by a quadtree and the different kinds of range queries. Each time, the query rectangle (which might be lower dimensional) is depicted in blue, and the nodes visited are shown in red: on the left a non-degenerate range query; in the middle, a partial match query, and on the right, a fully specified query. The corresponding tree, is depicted in Figure 3.

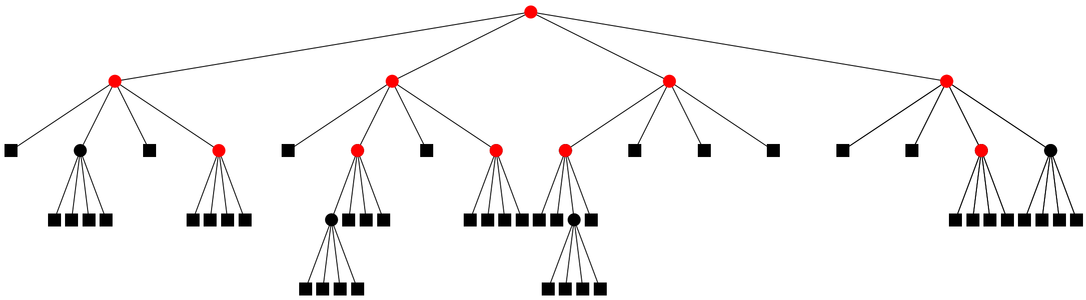

Figure 3: The quadtree corresponding to the partition shown in Figure 2. The nodes in red are the nodes visited by the range query illustrated in the left picture there, and correspond to the red points in that picture.

Orthogonal range queries. Despite their importance, the cost of range searches has been much less studied. For the closely related structure of k-d trees, the first important contribution is due to Chanzy, Devroye, and Zamora-Cura [8] who consider queries anchored at a uniformly random point in the unit square with specific predetermined dimensions Δ_1 and Δ_2 along the first and second coordinates. (These queries are allowed to exit $[0,1]^2$.) For the average cost m_n of such a query, they obtain the following explicit bounds

$$\gamma \le \frac{m_n}{\Delta_1 \Delta_2 n + (\Delta_1 + \Delta_2) n^\beta + \log n} \le \gamma', \quad 0 < \gamma < \gamma' < \infty. \tag{6}$$

This exhibits a contribution of the "volume" of nodes to report, and a "perimeter" effect of the order of magnitude of a partial match. Chanzy, Devroye, and Zamora-Cura [8, Section 8] also mention that an analogous result should hold for random quadtrees.

The aim of the present paper is to prove a limit theorem for the joint complexity of all range searches simultaneously, just as (5) for the partial match queries. Such an approach will also give access to the cost of extreme queries, which depend on the data points in an intricate way. We first define the set of all queries for which one should have joint convergence. Let $I = \{(a,b,c,d) \in [0,1]^4 : a \le b, c \le d\}$. For $(a,b,c,d) \in I$, let $Q(a,b,c,d) = (a,b] \times (c,d]$ and denote by $O_n(a,b,c,d)$ the number of nodes visited by the algorithm to answer the query with rectangle $Q(a,b,c,d)$ in T_n. The random variable $O_n(a,b,c,d)$ is well-defined except on lower-dimensional subsets of I, and we agree to extend it to these sets by imposing right continuity in all coordinates. Note that, by this convention, the range query is a genuine generalization of partial match queries since, for $t \in [0,1]$, we have $O_n(t,t,0,1) = C_n(t)$. Our main results are presented in the following section.

1.3 Main results: joint convergence of all orthogonal range queries

We are interested in the joint asymptotic cost of all range queries with rectangle $Q(a,b,c,d)$, $(a,b,c,d) \in I$, in T_n, that is, in the asymptotic behavior of the family of random variables $O_n(a,b,c,d), (a,b,c,d) \in$

I. Let $\mathcal{C}_4^+$ be space of continuous functions on I equipped with supremum norm $\|\cdot\|$ such that $\|f\| = \sup_{(a,b,c,d)\in I} |f(a,b,c,d)|$. For $(a,b,c,d) \in I$, let $\mathrm{Vol}(a,b,c,d) = (b-a)(d-c)$ denote the area of the query rectangle $Q(a,b,c,d)$. Recall that the sequence of random fields $O_n = O_n(a,b,c,d), n \geq 1$ relies on the sequence of quadtrees $T_n, \geq 1$, constructed from the same sequence $(X_i)_{i\geq 1}$. Our main result is the following theorem.

Theorem 1.1. *There exists a random continuous $\mathcal{C}_4^+$-valued random variable $\mathcal{O}$ (a random field) such that, in probability and with convergence of all moments,*

$$\left\| \frac{O_n - n\mathrm{Vol}}{n^\beta} - \mathcal{O} \right\| \xrightarrow[n\to\infty]{} 0\,.$$

Observe that the statement in Theorem 1.1 also covers query rectangles with zero Lebesgue measure ($\mathrm{Vol}(a,b,c,d) = 0$). The following straightforward consequence settles the question of Chanzy, Devroye, and Zamora-Cura [8] about the average cost of the worst-case range query:

Corollary 1.2. *We have the following convergence, in probability with convergence of all moments:*

$$n^{-\beta} \cdot \sup_{(a,b,c,d)\in I} \Big\{ O_n(a,b,c,d) - n(b-a)(d-c) \Big\} \xrightarrow[n\to\infty]{} \sup_{a,b,c,d} \mathcal{O}(a,b,c,d)\,.$$

The limit field $\mathcal{O}$ is uniquely characterized as the solution to a stochastic fixed-point equation.

Proposition 1.3. *Up to a multiplicative constant, the process $\mathcal{O}$ is the unique $\mathcal{C}_4^+$-valued random field (in distribution) with $\mathbb{E}[\|\mathcal{O}\|^2] < \infty$ satisfying the stochastic fixed-point equation*

$$\mathcal{O} \stackrel{d}{=} \sum_{r=1}^{4} D_r(\mathcal{O}^{(r)}), \tag{7}$$

where $\mathcal{O}^{(1)}, \dots, \mathcal{O}^{(4)}$ are copies of $\mathcal{O}$, $D_1, \dots, D_r$ are random linear operators defined in (35) and the random variables $\mathcal{O}^{(1)}, \dots, \mathcal{O}^{(4)}$, and $(D_1, \dots, D_4)$ are independent.

Proposition 7 only characterizes the distribution of $\mathcal{O}$ up to a multiplicative constant. The following proposition identifies the limit mean, and hence the missing multiplicative constant.

Proposition 1.4. *Let $(a,b,c,d) \in I$. Then, we have*

$$\mathbb{E}[\mathcal{O}(a,b,c,d)] = \frac{1}{2}\Big(\mu(a,d) - \mu(a,c) + \mu(b,d) - \mu(b,c) + \mu(c,b) - \mu(c,a) + \mu(d,b) - \mu(d,a)\Big),$$

where $\mu(t,s) = K_1 h(t) g(s)$, the constant K_1 defined in (4), and g is a continuous and monotonically increasing bijection on $[0,1]$ satisfying $g(s) = 1 - g(1-s)$ for every $s \in [0,1]$. Furthermore g is $\mathcal{C}^\infty$ on $(0,1)$ and the unique bounded measurable function on $[0,1]$ satisfying, for every $s \in [0,1]$,

$$g(s) = \frac{\beta+1}{2}\left(\int_s^1 v^\beta g\Big(\frac{s}{v}\Big)\, dv + \int_0^s (1-v)^\beta g\Big(\frac{s-v}{1-v}\Big)\, dv \right) + \frac{1}{2} s^{\beta+1}\,. \tag{8}$$

About the higher-dimensional case. All results in the present paper crucially rely on the limit theorems for partial match complexities formulated in [7], whose proofs are based on the functional contraction method developed in [25]. While high-dimensional analogues of the results summarized on fully specified queries, the mean complexity of partial match queries (1) and inequalities of type (6) are known [8, 9, 14, 15, 21, 23, 24], generalizations of our results would require to extend the contraction method to functions of multiple variables which is technically more demanding. For instance, even the high-dimensional analogues of the results in [7] are unknown. We intend to carry out an analysis of partial match and range queries in the d-dimensional case, for $d \geq 3$, elsewhere. Here, it is important to observe that, while we are dealing with random functions of multiple variables, we never use directly the contraction method for these fields thanks to a number of couplings.

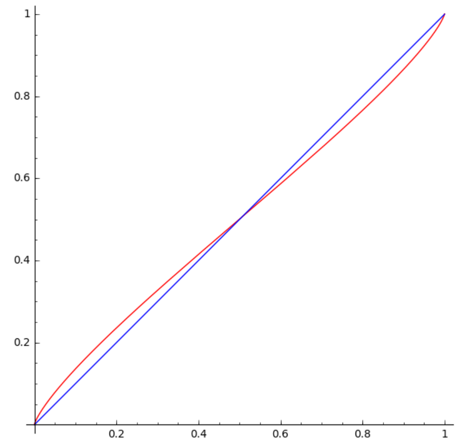

Figure 4: *The function g obtained by iteration in red, and the identity in blue for comparison.*

1.4 Strategy of the proof and combinatorial identities

The proof of Theorem 1.1 relies on a representation of $O_n(a,b,c,d)$ as a sum of different terms which aims at accounting precisely for the contributions of the "volume" and "perimeter" effects that are visible in the upper and lower bounds in (6).

Every point $X = (x_1, x_2) \in [0,1]^2$ partitions the unit square into four regions

$$
\begin{aligned}
\mathrm{SW}(X) &= [0, x_1) \times [0, x_2] \\
\mathrm{NW}(X) &= [0, x_1] \times (x_2, 1] \\
\mathrm{SE}(X) &= (x_1, 1] \times [0, x_2] \\
\mathrm{NE}(X) &= (x_1, 1] \times (x_2, 1] \,.
\end{aligned}
$$

For $(a,b,c,d) \in I$, the lines $\{x = a\}$, $\{x = b\}$, $\{y = c\}$, and $\{y = c\}$ partition the unit square $[0,1]^2$. Let $R_1 = \mathrm{SW}(a,c) = [0,a] \times [0,c]$, $R_2 = \mathrm{NW}(a,d) = [0,a) \times (d,1]$, $R_3 = \mathrm{SE}(b,c) = (b,1] \times [0,c)$, $R_4 = \mathrm{NE}(b,d) = (b,1] \times (d,1]$, be the regions south-west, north-west, south-east and north-east of $Q(a,b,c,d)$, respectively. (See Figure 5 for an illustration.) Let also $S_1 = [a,b] \times [0,c)$, $S_2 = [0,a] \times [c,d]$, $S_3 = (b,1] \times [c,d]$ and $S_4 = [a,b] \times (d,1]$ denote the regions south, east, north and west of $Q(a,b,c,d)$. Then $(0,1]^2$ is the disjoint union of $Q(a,b,c,d)$, R_1, R_2, R_3, R_4 and S_1, S_2, S_3, S_4.

Fix $n \geq 1$. Let $N_n(a,b,c,d)$ denote the number of points among $X_1, X_2, \dots, X_n$ that lie within the query rectangle $Q(a,b,c,d)$. For $s,t \in [0,1]$, let $Y_n^<(t,s)$ denote the number of nodes visited to answer the partial match query $\{x = t\}$ such that the corresponding point in $[0,1]^2$ lies in $\mathrm{SW}(t,s)$. So, for instance, the number of points lying in S_4 that are visited when answering the partial match query $\{x = a\}$ is $Y_n^<(a,d) - Y_n^<(a,c)$. Similarly, define $Y_n^\geq(t,s)$ as the number of points lying in $\mathrm{SE}(t,s)$ that are visited by a partial match query at t. The functions $\bar{Y}_n^<(t,s)$ and $\bar{Y}_n^\geq(t,s)$ are defined in a symmetric way when exchanging the first and second coordinates of every point $X_1, X_2, \dots, X_n$.

Let $D_n^{(1)}(a,b,c,d)$ denote the number of points among $X_1, X_2, \dots, X_n$ that are lying in R_1 and are visited by a fully specified search query in T_n retrieving (a,c). Analogously, let $D_n^{(2)}(a,b,c,d)$, $D_n^{(3)}(a,b,c,d)$ and $D_n^{(4)}(a,b,c,d)$ denote the number of points lying respectively in R_2, R_3 and R_4 that are visited to answer a fully-specified query for (a,d), (b,c) and (b,d), respectively.

We agree that, on horizontal and vertical lines containing points in the set $\{X_1, X_2, \dots\}$, all functions introduced in the previous paragraph are continuous from the right in all coordinates.

The following representation is reminiscent of the decomposition in sliced queries by [17].

Lemma 1.1. *Fix $n \geq 1$ and a quadtree on n points. For $(a, b, c, d) \in I$, we have*

$$\begin{aligned} O_n(a,b,c,d) = N_n(a,b,c,d) \\ &+ Y_n^{\geq}(b,d) - Y_n^{\geq}(b,c) + Y_n^{<}(a,d) - Y_n^{<}(a,c) \\ &+ \bar{Y}_n^{\geq}(d,b) - \bar{Y}_n^{\geq}(d,a) + \bar{Y}_n^{<}(c,b) - \bar{Y}_n^{<}(c,a) \\ &+ D_n^{(1)}(a,b,c,d) + D_n^{(2)}(a,b,c,d) + D_n^{(3)}(a,b,c,d) + D_n^{(4)}(a,b,c,d). \end{aligned} \tag{9}$$

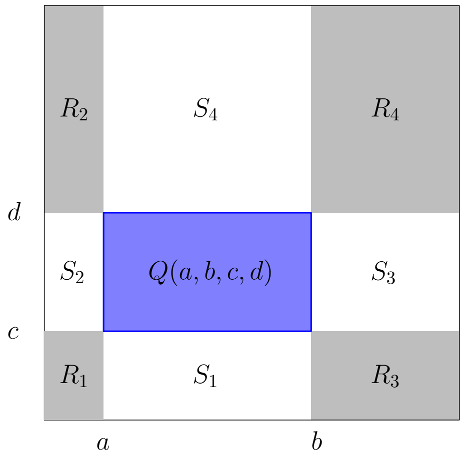

Figure 5: *The decomposition in Lemma 1.1 of the points visited to answer the query with rectangle $Q(a, b, c, d)$, here depicted in blue, and which are counted by $O_n(a, b, c, d)$.*

Proof. Consider $X_1, X_2, \ldots, X_n$ as fixed, and fix $(a, b, c, d) \in I$. The points that are visited to answer some range query may be grouped into different subsets depending on to which part of the unit square among $Q(a, b, c, d), R_1, R_2, R_3, R_4$ or S_1, S_2, S_3, S_4 they belong.

Now, a node of T_n storing the point X_i is visited when answering the query with rectangle $Q(a, b, c, d)$ if and only if, at the time of its insertion in T_n, the point X_i falls in a rectangle of the partition induced by the quadtree T_{i-1} that intersects $Q(a, b, c, d)$ (See [8]). It follows that the nodes corresponding to the points X_i lying within $Q(a, b, c, d)$ are all visited. Furthermore, the nodes corresponding to the points lying in R_1, R_2, R_3 and R_4 are visited when answering the query $Q(a, b, c, d)$ if and only if they are visited when answering the fully specified queries for the points (a, c), (a, d), (b, c) and (b, d), respectively. Finally, we need to consider the points among $X_1, X_2, \ldots, X_n$ lying in the rectangles S_1, S_2, S_3 and S_4. For these, the rectangle corresponding to a point X_i intersects $Q(a, b, c, d)$ if and only if it intersects the line segment separating S_i, $1 \leq i \leq 4$, from $Q(a, b, c, d)$ (open at one of the end points). It follows easily that, aside from the points that may lie on one of the boundaries of the regions, the contributions are precisely the number of points lying in S_1, S_2, S_3 or S_4 reported by one of the partial match queries $\{x = a\}$, $\{x = b\}$, $\{y = c\}$, or $\{y = d\}$. If follows that the claim holds for all $(a, b, c, d) \in I$ provided that no data point lies on any of these lines; the right-continuity of the functions completes the proof. □

With Lemma 1.1 under our belt, it is now relatively easy to get an intuition for why Theorem 1.1 should hold, and what is needed to turn this intuition into a proof. One first observes that $N_n(a, b, c, d)$ is distributed as a binomial random variable with parameters n and $\mathrm{Vol}(a, b, c, d)$, so that $N_n = n\mathrm{Vol} + O(\sqrt{n})$ uniformly in I. Since $\beta > 1/2$, the error term for the rescaled process is $O(n^{1/2-\beta}) = o(1)$, again uniformly in I. On the other hand, since $D_n^{(i)}$, $1 \leq i \leq 4$, are visited by a fully specified query retrieving some point, $\max_{1 \leq i \leq 4} D_n^{(i)}$ is bounded above by the height of the tree T_n, and thus $O(\log n)$ (see [13], but much weaker bounds would also suffice). As a consequence, for any fixed $(a, b, c, d) \in I$, the limit of

$$\frac{O_n(a,b,c,d) - n\mathrm{Vol}(a,b,c,d)}{n^{\beta}}$$

should be the limit of sum of the terms of the form $n^{-\beta} Y_n^{<}$, $n^{-\beta} Y_n^{\geq}$, $n^{-\beta} \bar{Y}_n^{<}$ and $n^{-\beta} Y_n^{\geq}$. Note that, of course, these terms are not independent; furthermore, the values for different $(a, b, c, d) \in I$ are dependent as well. We will prove that each of these terms converges uniformly on I in probability. The four types of terms we now have to deal with are all "partial-match"-like, and the technology we have developed in [6, 7] will come in handy.

1.5 Plan of the paper

The remainder of the paper is organized as follows: In Section 2, we introduce the relevant background and constructions about partial match queries and their limit process that needs to be generalized to consider the terms involved in Lemma 1.1. Section 3 deals with the case of one-sided partial match queries, where only the point lying on one or the other side of the query line are reported (terms of the form $Y_n^{<}(t,1)$ and $Y_n^{\geq}(t,1)$). Section 4 deals with the further restriction of the count of points depending on their second coordinate (along the direction parallel to the query line), namely terms of the form $Y_n^{<}(t,s)$ and $Y_n^{\geq}(t,s)$. We put everything together and complete the proofs in Section 5. Finally, the extensions to 2-d are presented in Section 6.

2 The cost of partial match queries in random quadtrees

In this section, we introduce the notation and review the relevant constructions and results about partial match queries that are central to our approach. We refer the reader to [6, 7] for more details and the proofs.

We call a set $Q \subseteq [0,1]^2$ a *half-open* rectangle if, for some real numbers a, b, c, d such that $0 < a < b \leq 1$ and $0 < c < d \leq 1$, we have

$$Q = \begin{cases} (a,b] \times (c,d], & \text{or} \\ [0,b] \times (c,d], & \text{or} \\ (a,b] \times [0,d], & \text{or} \\ [0,b] \times [0,d]. \end{cases}$$

Let $\mathcal{Q}$ be the set of all half-open rectangles and $\mathbb{T} = \bigcup_{k\geq 0}\{1,2,3,4\}^k$ be the infinite quaternary tree of words on the alphabet $\{1,2,3,4\}$: the finite words on $\{1,2,3,4\}$ are the nodes, and the ancestors of some word are its prefixes (including the empty word $\varnothing$). For $u \in \mathbb{T} \setminus \{\varnothing\}$, the word $\bar{u}$ obtained by removing the last letter is called the parent of u. A subset $A \in \mathbb{T}$ is called a tree if it is closed by taking prefixes. For a finite tree $A \subset \mathbb{T}$, let ∂A be the set of nodes $u \in \mathbb{T}$ such that $u \notin A$, but $\bar{u} \in A$.

From the sequence $(X_i)_{i\geq 1}$ of random points in $[0,1]^2$, we recursively construct

(i) a bijection $\pi : \mathbb{N} \to \mathbb{T}$ inducing node labels $X_{\pi^{-1}(v)}$ associated with $v \in \mathbb{T}$,

(ii) a family $\{Q_v \in \mathcal{Q} : v \in \mathbb{T}\}$ where $Q_\varnothing = [0,1]^2$, and for all $v \in \mathbb{T}$, Q_v is the disjoint union of the four half-open rectangles $Q_{v1}, \dots, Q_{v4}$, and

(iii) an increasing (for inclusion) sequence of trees $(T_n)_{n\geq 1}$ such that, for every $n \geq 1$, $\{Q_v : v \in \partial T_n\}$ is a partition of $[0,1]^2$.

We proceed as follows. First, let $\pi(1) = \varnothing$, T_1 consist of the root node $\varnothing$ and $Q_1 := \mathrm{SW}(X_1)$, $Q_2 := \mathrm{NW}(X_1)$, $Q_3 := \mathrm{SE}(X_1)$ and $Q_4 := \mathrm{NE}(X_1)$ be the four half-open rectangles generated by insertion of X_1 in $[0,1]^2$. Next, for $n \geq 1$, having defined $\pi(j)$ for all $j \leq n$, the tree T_n and rectangles Q_v for all $v \in \partial T_n$ such that $\{Q_v : v \in \partial T_n\}$ is a partition of $[0,1]^2$, we let $\pi(n+1)$ be the unique node $v \in \partial T_n$ with $X_{n+1} \in Q_v$. Further, $T_{n+1} := T_n \cup \{v\}$ and $Q_{v1}, \dots, Q_{v4}$ are defined respectively as $Q_v \cap \mathrm{SW}(X_{n+1})$, $Q_v \cap \mathrm{NW}(X_{n+1})$, $Q_v \cap \mathrm{SE}(X_{n+1})$ and $Q_v \cap \mathrm{NE}(X_{n+1})$. The partition of the unit square induced by T_{n+1} is then given by $\{Q_v : v \in \partial T_{n+1}\}$.

Let $Q_v^{(i)}, i = 1,2$ be the projection of Q_v on the ith component. For $v \in \mathbb{T}$, we define the time-transformations φ_v, φ'_v quantifying the position of a point (t,s) relative to the boundary of Q_v: we set

$$\varphi_v(t) = \mathbf{1}_{Q_v^{(1)}}(t) \frac{t - \inf Q_v^{(1)}}{\sup Q_v^{(1)} - \inf Q_v^{(1)}}, \quad t \in [0,1], \tag{10}$$

and

$$\varphi'_v(s) = \mathbf{1}_{Q_v^{(2)}}(s) \frac{s - \inf Q_v^{(2)}}{\sup Q_v^{(2)} - \inf Q_v^{(2)}}, \quad s \in [0,1]. \tag{11}$$

With $X_{\pi^{-1}(v)} = (x_1^v, x_2^v)$, we set $U^v := \varphi_v(x_1^v)$ and $V^v := \varphi'_v(x_2^v)$. By construction, $\{(U^v, V^v) : v \in \mathbb{T}\}$ is a family of independent random variables with the uniform distribution on $[0,1]^2$. To keep the notation simple, we write $U := U^\varnothing$ and $V := V^\varnothing$; more generally, we usually drop the reference to $\varnothing$ when the meaning is clear from the context.

The limit process $\mathcal{Z}$. In the remainder of the manuscript, we write $\mathcal{C}_k, k = 1, 2$ for the space of continuous functions on $[0,1]^k$ endowed with the supremum norm $\|f\| = \sup_{t\in[0,1]^k} |f(t)|$. Recall the constants K_1, β and the function h from equations (2) and (4). For $v \in \mathbb{T}$, let $|v|$ denote the length of the word (the distance between v and the root $\varnothing$). Further, for $Q \in \mathcal{Q}$, let $|Q|$ denote its area. Set $\mathcal{Z}_0^v = K_1 h$ for all $v \in \mathbb{T}$, and, recursively,

$$\mathcal{Z}_{n+1}^v(t) = \sum_{r=1}^{4} \mathbf{1}_{Q_{vr}^{(1)}}(t) A_{vr}^{\beta} \mathcal{Z}_n^{vr}(\varphi_{vr}(t)), \quad v \in \mathbb{T}, \tag{12}$$

where $A_v = |Q_v|/|Q_{\bar{v}}|$ for $v \in \mathbb{T}, v \neq \varnothing$. In other words, for all $v \in \mathbb{T}$, we have

$$A_{v1} = U^v V^v, \quad A_{v2} = U^v(1 - V^v), \quad A_{v3} = (1 - U^v)V^v, \quad \text{and } A_{v4} = (1 - U^v)(1 - V^v).$$

By the results stated in Proposition 2, Theorem 5, Proposition 9, Lemma 10 and Proposition 11 in [7], there exist random continuous functions $\mathcal{Z}^v, v \in \mathbb{T}$, such that

(i) the random variables $\mathcal{Z}^v, v \in \mathbb{T}$, are identically distributed,

(ii) $\|\mathcal{Z}_n^v - \mathcal{Z}^v\| \to 0$ almost surely and with convergence of all moments,

(iii) $\mathbb{E}[\mathcal{Z}^v(t)^m] = c_m h(t)^m$ for appropriate constants $c_m > 0$ where $c_1 = K_1$,

(iv) $\mathbb{E}[\|\mathcal{Z}^v\|^p] < \infty$ for all $p > 0$,

(v) $\mathcal{Z}^{v1}, \ldots, \mathcal{Z}^{v4}, U^v, V^v$ are stochastically independent and, almost surely, for all $t \in [0,1]$,

$$\mathcal{Z}^v(t) = \sum_{r=1}^{4} \mathbf{1}_{Q_{vr}^{(1)}}(t) A_{vr}^{\beta} \mathcal{Z}^{vr}(\varphi_{vr}(t)),$$

(vi) up to a multiplicative constant, $\mathcal{Z}^v$ is the unique continuous process (in distribution) with $\mathbb{E}[\|\mathcal{Z}^v\|^2] < \infty$ satisfying the stochastic fixed-point equation

$$\mathcal{Z}^v \stackrel{d}{=} \left(\sum_{r=1}^{4} \mathbf{1}_{Q_r^{(1)}}(t) A_r^{\beta} \mathcal{Z}^{(r)}(\varphi_r(t)) \right)_{t\in[0,1]}. \tag{13}$$

Here, $\mathcal{Z}^{(1)}, \ldots, \mathcal{Z}^{(4)}$ are copies of $\mathcal{Z}^v$, and $\mathcal{Z}^{(1)}, \ldots, \mathcal{Z}^{(4)}, U, V$ are independent.

The complexity of partial match queries. The main result in [7] is the limit law (5) with limit process $\mathcal{Z} := \mathcal{Z}^\varnothing$. The identity (13) for the distribution of $\mathcal{Z}$ is reminiscent of the following distributional recurrence for the discrete process C_n: letting $N_1, \ldots, N_4$ denote the subtree sizes of the quadtree T_n (such that $N_1 + \cdots + N_4 = n - 1$), we have

$$C_n \stackrel{d}{=} \left(\sum_{r=1}^{4} \mathbf{1}_{Q_r^{(1)}}(t) C_{N_r}^{(r)}(\varphi_r(t)) + 1 \right)_{t\in[0,1]}. \tag{14}$$

Here, the random sequences $(C_n^{(1)})_{n\ge0}, \ldots, (C_n^{(4)})_{n\ge0}$ stemming from the complexities in the four subtrees are independent and identically distributed and also independent from $N_1, \ldots, N_4, U, V$. Further, given (U, V), the vector $(N_1, \ldots, N_4)$ has the multinomial distribution with parameters $(n - 1; A_1, \ldots, A_4)$. This recurrence is at the heart of the proof of (5), and similar recurrences also play key roles in the present work. As already mentioned in the introduction, by [11, Corollary 1.2], in probability (and therefore with convergence of all moments by [7, Theorem 4]),

$$\|n^{-\beta} C_n - \mathcal{Z}\| \xrightarrow[n\to\infty]{} 0. \tag{15}$$

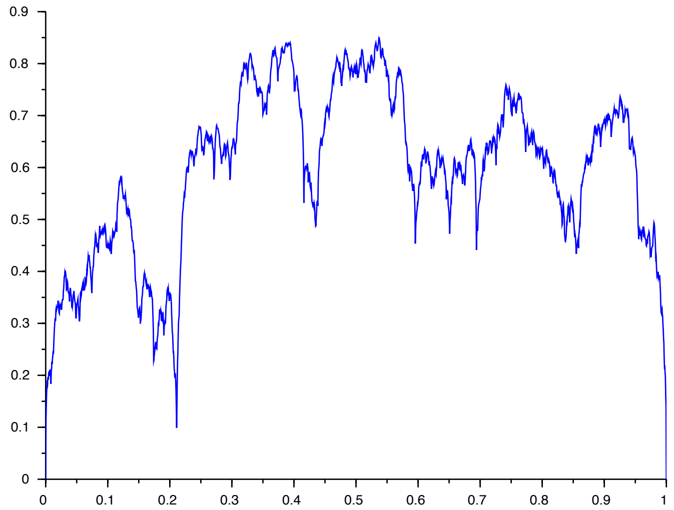

Figure 6: *A simulation of the limit process $\mathcal{Z}$ for the cost of partial match queries.*

3 One-sided partial match queries

For $t \in [0,1]$, let $C_n^{\geq}(t)$ denote the number of nodes with first coordinate at least t visited by a partial match query retrieving all keys with first component t in T_n. In the partition of the unit square induced by T_n, $C_n^{\geq}(t)$ can be identified as the number of horizontal lines intersecting the vertical line at t stemming from the points lying to the right of t. Again, we agree $C_n^{\geq}(\cdot)$ to be a right-continuous step function. Set $C_n^{<}(t) = C_n(t) - C_n^{\geq}(t)$. By construction, $C_n^{\geq}$ satisfies a recurrence very similar to (14), namely

$$(C_n^{\geq}(t))_{t\in[0,1]} \stackrel{d}{=} \left(\sum_{r=1}^{4} \mathbf{1}_{Q_r^{(1)}}(t) C_{N_r}^{\geq,(r)}(\varphi_r(t)) + \mathbf{1}_{[0,U)}(t) \right)_{t\in[0,1]}, \tag{16}$$

with conditions on independence and distributions as in (14). Since the only difference between (14) and (16) concerns the additive term which asymptotically vanishes after rescaling by n^β, one might guess that $C_n^{\geq}$ admits the same (distributional) scaling limit as C_n modulo a multiplicative constant. In fact, Proposition 3.1 below shows much more. The following lemma is an important ingredient of the proof.

Lemma 3.1. *There exists $\varepsilon > 0$ such that, uniformly in $t \in [0,1]$ and as $n \to \infty$,*

$$\mathbb{E}[C_n^{\geq}(t)] = \frac{K_1}{2} h(t) n^\beta + O(n^{\beta-\varepsilon}), \qquad \textit{and} \qquad \mathbb{E}[C_n^{<}(t)] = \frac{K_1}{2} h(t) n^\beta + O(n^{\beta-\varepsilon}). \tag{17}$$

Here, K_1 and β are the constants given in (2) and (4).

The proof of Lemma 3.1 is postponed until the end of the section.

Proposition 3.1. *In probability and with convergence of all moments, as $n \to \infty$,*

$$\|n^{-\beta} C_n^{\geq} - \mathcal{Z}/2\| \to 0, \qquad \textit{and} \qquad \|n^{-\beta} C_n^{<} - \mathcal{Z}/2\| \to 0.$$

Proof. Since $\|n^{-\beta} C_n - \mathcal{Z}\| \to 0$ by (15), it suffices to show that $n^{-\beta}\|C_n^{\geq} - C_n^{<}\| \to 0$ in probability and with respect to all moments. To this end, write $\mu_n^{\geq}(t) = \mathbb{E}[C_n^{\geq}(t)]$, $\mu_n^{<}(t) = \mathbb{E}[C_n^{<}(t)]$ and let

$$X_n(t) := \frac{C_n^{\geq}(t) - \mu_n^{\geq}(t)}{n^\beta} - \frac{C_n^{<}(t) - \mu_n^{<}(t))}{n^\beta}.$$

Note that $\mathbb{E}[X_n(t)] = 0$ for all $t \in [0,1]$. By construction, the process X_n satisfies the following functional distributional recurrence:

$$X_n \stackrel{d}{=} \left(\sum_{r=1}^{4} \mathbf{1}_{Q_r^{(1)}}(t) \left(\frac{N_r}{n}\right)^{\beta} X_{N_r}^{(r)}(\varphi_r(t)) \right. \\ \left. + \frac{\sum_{r=1}^4 \mathbf{1}_{Q_r^{(1)}}(t)\left[\mu_{N_r}^{\geq}(\varphi_r(t)) - \mu_{N_r}^{<}(\varphi_r(t))\right] + \mathbf{1}_{[U,1]}(t) - \mathbf{1}_{[0,U)}(t)}{n^{\beta}} \right)_{t\in[0,1]}, \tag{18}$$

again with assumptions on independence and distributions as in (14).

By Lemma 3.1, uniformly in $t \in [0,1]$, the additive term (18) converges to zero almost surely and with respect to all moments. Therefore, and as for every $r \in \{1,2,3,4\}$, $N_r/n \to A_r$ almost surely by the concentration of the binomial distribution, one would expect that, if X_n admits a limit process X, then X should satisfy the distributional fixed-point equation obtained by taking limits in the recurrence above:

$$X \stackrel{d}{=} \left(\sum_{r=1}^{4} \mathbf{1}_{Q_r^{(1)}}(t) A_r^{\beta} X^{(r)}(\varphi_r(t)) \right)_{t\in[0,1]}. \tag{19}$$

Here, $X^{(1)}, \dots, X^{(4)}$ are copies of X, and $X^{(1)}, \dots, X^{(4)}, U, V$ are stochastically independent (U and V appear in the definition of A_r, $r = 1, \dots, 4$). As this equation is homogeneous, it is solved by the process which is identical to zero. Furthermore, an application of [25, Lemma 18] shows that the zero process is the unique solution (in distribution) of (19) among all random processes with zero mean and finite absolute second moment. As the proof of (5) in [7], the rigorous verification of that the convergence $\|X_n\| \to 0$ holds in probability makes use of the functional contraction method. The application of [25, Theorem 22] requires to verify a set of conditions (C1)–(C5) formulated in that paper. By the similarity of the processes C_n and $C_n^{\geq}$ and their distributional recurrences, conditions (C1), (C2), (C4) and (C5) can be verified exactly in the same way as it was done in [7] for the process C_n, and we omit the details. The fact that the zero process is a solution of (19) guarantees (C3). This shows distributional (or, equivalently, stochastic) convergence. The convergence of moments follows by monotonicity since $\max\{C_n^{<}, C_n^{\geq}\} \leq C_n$ and $\sup_{n\geq 1} \mathbb{E}[\|C_n\|^p] < \infty$ by [7, Theorem 4]. ☐

Corollary 3.2. *In probability and with respect to all moments, we have*

$$\sup_{0\leq s,t\leq 1} \left| \frac{O_n(s,t,0,1) - n(t-s)}{n^{\beta}} - \frac{\mathcal{Z}(t) + \mathcal{Z}(s)}{2} \right| \xrightarrow[n\to\infty]{} 0.$$

Proof. Let $N_n(t)$ denote the total number of points among $X_1, X_2, \dots, X_n$ that have a first coordinate at least t. By Donsker's classical theorem for empirical distribution functions, we have

$$\left(\frac{N_n(t) - n(1-t)}{\sqrt{n}} \right)_{t\in[0,1]} \xrightarrow[n\to\infty]{d} B, \tag{20}$$

where B is a Brownian bridge, that is, $B(t) = W(t) - tW(1), t \in [0,1]$ for a standard Brownian motion W. It is well-known that this convergence is also with respect to all moments (this follows, e.g., from the Dvoretzky–Kiefer–Wolfowitz inequality [19]). Thus, since $\beta > 1/2$, uniformly for $s, t \in [0,1]$, we have $n^{-\beta}(N_n(t) - N_n(s)) \to 0$ in probability and with respect to all moments. As $O_n(s,t,0,1) = N_n(t) - N_n(s) + C_n^{\geq}(t) + C_n^{<}(s)$, the assertion follows from Proposition 3.1. ☐

Finally, we prove Lemma 3.1 that was instrumental in the proof of Proposition 3.1.

Proof of Lemma 3.1. The most technical ingredient in the proof of (5) that appears in [7] is the following strengthening of (3): there exists $\varepsilon > 0$ such that, uniformly in $t \in [0,1]$, and as $n \to \infty$

$$\mathbb{E}[C_n(t)] = K_1 h(t) n^{\beta} + O(n^{\beta-\varepsilon}). \tag{21}$$

This result heavily relies on the methods developed by Curien and Joseph in [12]. Note that the fact that the first point X_1 falls on one side of the line $\{x = t\}$ induces an asymmetry between $C_n^{<}(t)$ and $C_n^{\geq}(t)$ for n fixed, and the means $\mathbb{E}[C_n^{<}(t)]$ and $\mathbb{E}[C_n^{\geq}(t)]$ are different (unless $t = 1/2$). Somewhat as a consequence of this inherent asymmetry, we have no simple/soft argument to deduce (17) directly from (21), and it is necessary to repeat all steps in the verification of the latter. To work out all details would go beyond the scope of this note, and we confine ourselves to the discussion of the main steps: first of all, as in [7, Section 5], one considers a Poissonized model in which the points $X_1, X_2, \dots$ are inserted following the arrival times of a homogeneous Poisson process on $[0, \infty)$. Consequently, we deal with a family of increasing quadtrees $(T_s)_{s\geq 0}$ built on the points arrived before time s (there are a Poisson(s) number of such points), and the corresponding partial match query complexities $(C_s(t))_{s\geq 0}$ and $(C_s^{\geq}(t))_{s\geq 0}$. Standard de-poissonization arguments based on the concentration of the Poisson distribution imply that it is sufficient to show the corresponding statements for the continuous-time process, namely there exists $\varepsilon > 0$ such that

$$\sup_{0\leq t\leq 1} \left| n^{-\beta}\mathbb{E}[C_s^{\geq}(t)] - \frac{K_1}{2}h(t) \right| = O(s^{-\varepsilon}), \quad s \to \infty.$$

Following [7], one distinguishes between the behavior at the boundary ($t \in [0, \delta] \cup [1-\delta, 1]$ for small $\delta > 0$) and away from the boundary ($t \in [\delta, 1-\delta]$). In [7], it was enough to consider $t \leq 1/2$ by symmetry, but here, this is not the case. Lemmas 14 and 15 in [7] contain the corresponding bounds for the process C_s, and we now argue that these bounds apply with the *same* involved constants to the one-sided quantity $C_s^{\geq}$.

Concerning the behavior at the boundary, we have the trivial bound

$$\begin{aligned}\sup_{t\in[0,\delta]\cup[1-\delta,1]} \left| s^{-\beta}\mathbb{E}[C_s^{\geq}(t)] - \frac{K_1}{2}h(t) \right| &\leq \sup_{t\in[0,\delta]\cup[1-\delta,1]} s^{-\beta}\mathbb{E}[C_s^{\geq}(t)] + \sup_{t\in[0,\delta]\cup[1-\delta,1]} K_1 h(t) \\ &\leq \sup_{t\in[0,\delta]} s^{-\beta}\mathbb{E}[C_s(t)] + \sup_{t\in[0,\delta]} \frac{K_1}{2}h(t)\,,\end{aligned}$$

which brings us back to the situation of the two-sided problem and explains why the bound in [7, Lemma 14] also applies in our case. The main part of the proof of (21) is contained in [7, Lemma 14] and relies on a coupling argument between $\varphi_v(t)$ for the unique node $v = v_1 v_2 \dots v_k, k \geq 1$, with $v_i \in \{1, 3\}$ for all $i = 1, \dots, k$ and $t \in Q_v$ and a uniformly distributed random variable ξ. We do not explain this step in detail but mention that one distinguishes two cases: first, in case 1 (coupling has not yet happened), similarly to the last display, one uses the crucial uniform upper bound in [12, Lemma 2]:

$$\sup_{s\geq 0} \sup_{t\in[0,1]} s^{-\beta}\mathbb{E}[C_s(t)] < \infty.$$

By monotonicity, this bound can also be applied to $C_s^{\geq}$. In case 2 (coupling has happened), the main ingredient in the proof is the expansion in (2). (Here, the second order term is important.) Since for a uniform random variable ξ independent of $(X_i)_{i\geq 1}$, we have $\mathbb{E}[C_n^{\geq}(\xi)] = \mathbb{E}[C_n^{<}(\xi)] = \mathbb{E}[C_n(\xi)]/2$, the same arguments apply in our case. □

4 Constrained partial match queries

4.1 Preliminary considerations

For $t, s \in [0, 1]$, let $Y_n(t, s) = O_n(t, t, 0, s)$ be the number of nodes visited by the partial match query retrieving the points with first coordinate equal to t and second coordinate at most s. On lower-dimensional subsets where $Y_n(t, s)$ is not well-defined, we assume the function to be right continuous in both coordinates. Note that $Y_n(t, 1) = C_n(t)$ for $t \in [0, 1]$. To prove a functional limit theorem for Y_n, we use a variant of the functional contraction method which makes explicit use of our encoding. A very similar approach was taken in [5] when studying the dual tree of a partitioning of the disc by sequential insertions of non-crossing random chords. As a result, we provide a proof of convergence of $n^{-\beta}Y_n$ that avoids another application of the complex machinery developed in [25].

Recalling the time-transformations (10) and (11), we have the following distributional recursive equation on the level of random fields with parameter space $[0,1]^2$:

$$\begin{aligned}\big(Y_n(t,s)\big)_{t,s\in[0,1]} \stackrel{d}{=} \Big(&\mathbf{1}_{Q_1}(t,s)Y^{(1)}_{N_1}(\varphi_1(t),\varphi_1'(s))\\&+\mathbf{1}_{Q_2}(t,s)\left[Y^{(1)}_{N_1}(\varphi_1(t),1)+Y^{(2)}_{N_2}(\varphi_2(t),\varphi_2'(s))\right]+\mathbf{1}_{[V,1]}(s)\\&+\mathbf{1}_{Q_3}(t,s)Y^{(3)}_{N_3}(\varphi_3(t),\varphi_3'(s))\\&+\mathbf{1}_{Q_4}(t,s)\left[Y^{(3)}_{N_3}(\varphi_3(t),1)+Y^{(4)}_{N_4}(\varphi_4(t),\varphi_4'(s))\right]\Big)_{t,s\in[0,1]}.\end{aligned} \tag{22}$$

Here, we have the same conditions on independence and distributions as in (14). Thus, if $n^{-\beta}Y_n$ converges, we expect the distribution of the limit Y to satisfy the following fixed-point equation

$$\begin{aligned}\big(Y(t,s)\big)_{t,s\in[0,1]} \stackrel{d}{=} \Big(&\mathbf{1}_{Q_1}(t,s)A_1^\beta Y^{(1)}(\varphi_1(t),\varphi_1'(s))\\&+\mathbf{1}_{Q_2}(t,s)\left[A_1^\beta Y^{(1)}(\varphi_1(t),1)+A_2^\beta Y^{(2)}(\varphi_2(t),\varphi_2'(s))\right]\\&+\mathbf{1}_{Q_3}(t,s)A_3^\beta Y^{(3)}(\varphi_3(t),\varphi_3'(s))\\&+\mathbf{1}_{Q_4}(t,s)\left[A_3^\beta Y^{(3)}(\varphi_3(t),1)+A_4^\beta Y^{(4)}(\varphi_4(t),\varphi_4'(s))\right]\Big)_{t,s\in[0,1]},\end{aligned}$$

where $Y^{(1)},\dots,Y^{(4)}$ are copies of Y, and the random variables $Y^{(1)},\dots,Y^{(4)},U,V$ are independent. Since $Y_n(t,1)=C_n(t)$, and this process is well understood, the crucial observation is that, for any fixed $(t,s)\in[0,1]^2$, only *one* of the processes $(Y_n^{(1)}),\dots,(Y_n^{(4)})$ contributes to the recursive decomposition (22) at a point whose second coordinate differs from one. The same can be said about the associated stochastic fixed-point equation. It is this fact why we do not need to engage the methodology of [25] to show convergence. (We do however need some ideas of this work to characterize the distribution of Y. See Proposition 4.4 below.)

4.2 Construction of the limit process and convergence

We proceed as in the construction of the process $\mathcal{Z}$ described in Section 2. To simplify the notation, let us introduce the following operators: for $v\in\mathbb{T}$, and for a function $f:[0,1]^2\to\mathbb{R}$, define

$$\begin{aligned}B_1^v(f)(t,s)&=A_{v1}^\beta\left[\mathbf{1}_{Q_{v1}}(t,s)f(\varphi_{v1}(t),\varphi_{v1}'(s))+\mathbf{1}_{Q_{v2}}(t,s)f(\varphi_{v1}(t),1)\right],\\B_2^v(f)(t,s)&=A_{v2}^\beta\mathbf{1}_{Q_{v2}}(t,s)f(\varphi_{v2}(t),\varphi_{v2}'(s)),\\B_3^v(f)(t,s)&=A_{v3}^\beta\left[\mathbf{1}_{Q_{v3}}(t,s)f(\varphi_{v3}(t),\varphi_{v3}'(s))+\mathbf{1}_{Q_{v4}}(t,s)f(\varphi_{v4}(t),1)\right],\\B_4^v(f)(t,s)&=A_{v4}^\beta\mathbf{1}_{Q_{v4}}(t,s)f(\varphi_{v4}(t),\varphi_{v4}'(s)).\end{aligned} \tag{23}$$

For all $v\in\mathbb{T}$, let $\mathcal{Y}_0^v(t,s)=K_1h(t)$ for all $t,s\in[0,1]$. Then, recursively, set

$$\mathcal{Y}_{n+1}^v(t,s)=\sum_{r=1}^4 B_r^v(\mathcal{Y}_n^{vr})(t,s),\qquad v\in\mathbb{T}. \tag{24}$$

This definition extends the construction of $\mathcal{Z}_n^v$ in (12) since we have $\mathcal{Y}_n^v(t,1)=\mathcal{Z}_n^v(t)$ for $t\in[0,1]$. We first verify that this indeed allows to construct a family of processes $(\mathcal{Y}^v)_{v\in\mathbb{T}}$ that have the required properties:

Proposition 4.1. *There exist random continuous $\mathcal{C}_2$-valued fields $\mathcal{Y}^v, v\in\mathbb{T}$, such that*

(i) the random variables $\mathcal{Y}^v, v\in\mathbb{T}$, are identically distributed,

(ii) $\|\mathcal{Y}_n^v-\mathcal{Y}^v\|\to 0$ almost surely and with convergence of all moments,

(iii) $\mathcal{Y}^v(t,1)=\mathcal{Z}^v(t)$ for all $t\in[0,1]$,

(iv) $\mathbb{E}[\|\mathcal{Y}^v\|^p]<\infty$ for all $p>0$, and

(v) $\mathcal{Y}^{v1},\dots,\mathcal{Y}^{v4},U^v,V^v$ are stochastically independent and, almost surely, for all $t,s\in[0,1]$,

$$\mathcal{Y}^v(t,s)=\sum_{r=1}^4 B_r^v(\mathcal{Y}^{vr})(t,s), \tag{25}$$

where the operators B_r^v, $r\in\{1,2,3,4\}$, $v\in\mathbb{T}$, are defined in (23).

Proof. By the definition in (24) of the family of process $(\mathcal{Y}_n^v)_{n\geq 0}$, $v \in \mathbb{T}$, we have

$$
\begin{aligned}
&[\mathcal{Y}_{n+1}^v(t,s) - \mathcal{Y}_n^v(t,s)]^2 \\
&= \sum_{r=1}^{4} \mathbf{1}_{Q_{vr}}(t,s) A_{vr}^{2\beta} [\mathcal{Y}_n^{vr}(\varphi_{vr}(t), \varphi'_{vr}(s)) - \mathcal{Y}_{n-1}^{vr}(\varphi_{vr}(t), \varphi'_{vr}(s))]^2 \\
&+ 2\mathbf{1}_{Q_{v2}}(t,s) A_{v1}^{\beta} A_{v2}^{\beta} [\mathcal{Y}_n^{v1}(\varphi_1(t),1) - \mathcal{Y}_{n-1}^{vr}(\varphi_1(t),1)] \cdot [\mathcal{Y}_n^{v2}(\varphi_2(t), \varphi'_2(s)) - \mathcal{Y}_{n-1}^{v2}(\varphi_2(t), \varphi'_2(s))] \\
&+ 2\mathbf{1}_{Q_{v4}}(t,s) A_{v3}^{\beta} A_{v4}^{\beta} [\mathcal{Y}_n^{v3}(\varphi_3(t),1) - \mathcal{Y}_{n-1}^{v3}(\varphi_3(t),1)] \cdot [\mathcal{Y}_n^{v4}(\varphi_4(t), \varphi'_4(s)) - \mathcal{Y}_{n-1}^{v4}(\varphi_4(t), \varphi'_4(s))].
\end{aligned}
$$

Let $\Delta_n = \mathbb{E}[\|\mathcal{Y}_{n+1}^v - \mathcal{Y}_n^v\|^2]$ for $n \geq 0$. By the Cauchy–Schwarz inequality, it follows that, for $n \geq 1$,

$$
\Delta_n \leq \Delta_{n-1} \sum_{r=1}^{r} \mathbb{E}[A_{vr}^{2\beta}] + 4\sqrt{\Delta_{n-1} \cdot \mathbb{E}\left[\sup_{t\in[0,1]} |\mathcal{Y}_{n+1}^v(t,1) - \mathcal{Y}_n^v(t,1)|^2\right]}.
$$

Now, by [7, Proposition 9], there exist constants $C > 0$ and $q \in (0,1)$ such that, for every $n \geq 0$,

$$
\mathbb{E}\left[\sup_{t\in[0,1]} |\mathcal{Y}_{n+1}^v(t,1) - \mathcal{Y}_n^v(t,1)|^2\right] \leq C^2 q^n \,. \tag{26}
$$

As a consequence, setting

$$
\gamma := \sum_{r=1}^{4} \mathbb{E}[A_r^{2\beta}] = \frac{4}{(2\beta+1)^2} < 1, \tag{27}
$$

we obtain

$$
\Delta_n \leq \gamma \Delta_n + C q^{n/2} \sqrt{\Delta_n}.
$$

By induction on $n \geq 0$ it follows that $\Delta_n = O(r^n)$ for all $\sqrt{q} < r < 1$. (This argument is worked out in the proof of [5, Lemma 2.3].) Uniform almost sure convergence follows straightforwardly from the completeness of $\mathcal{C}_2$. (These details are explained in the proof of [5, Theorem 1.2].) Convergence of the pth moment follows by induction on p along the same lines. (Here, one uses that an exponential bound of the form (26) is valid for any higher moment.) ☐

Recall the definitions of the fields $Y_n^{\geq}$ and $Y_n^{<}$ given in Section 1.4. Note that $Y_n^{\geq} + Y_n^{<} = Y_n$.

Proposition 4.2. *Let $\mathcal{Y} := \mathcal{Y}^{\varnothing}$. In probability and with convergence of all moments,*

$$
\|n^{-\beta} Y_n - \mathcal{Y}\| \xrightarrow[n\to\infty]{} 0.
$$

In the same sense,

$$
\|n^{-\beta} Y_n^{\geq} - \mathcal{Y}/2\| \xrightarrow[n\to\infty]{} 0, \qquad \textit{and} \qquad \|n^{-\beta} Y_n^{<} - \mathcal{Y}/2\| \xrightarrow[n\to\infty]{} 0.
$$

Proof. For $v \in \mathbb{T}$, let $(a_n^v)_{n\geq 1}$ be the (maximal) increasing sequence of indices with such that $X_{a_n^v} \in Q_v$. Then if $X_n = (x_n^1, x_n^2)$, $n \geq 1$, we set, for each $n \geq 1$, $X_n^v = (\varphi_v(x_{a_n^v}^1), \varphi'_v(x_{a_n^v}^2))$. This sequence plays the same role in the construction of the subtree rooted at v as the sequence $(X_i)_{i\geq 1}$ in the entire tree (the subtree rooted at $\varnothing$). In other words, if we define $C_n^v(t)$ and $Y_n^v(t)$ analogously to (15), but relying on this sequence, we obtain

$$
\mathbb{E}[\|n^{-\beta} C_n^v(t) - \mathcal{Z}^v(t)\|^p] \to 0 \tag{28}
$$

for all $p > 0$. By construction, almost surely, for all $t, s \in [0,1]$,

$$
Y_n^v(t,s) = \sum_{r=1}^{4} B_r(Y_{N_r}^{vr})(t,s) + \mathbf{1}_{[V,1]}(s).
$$

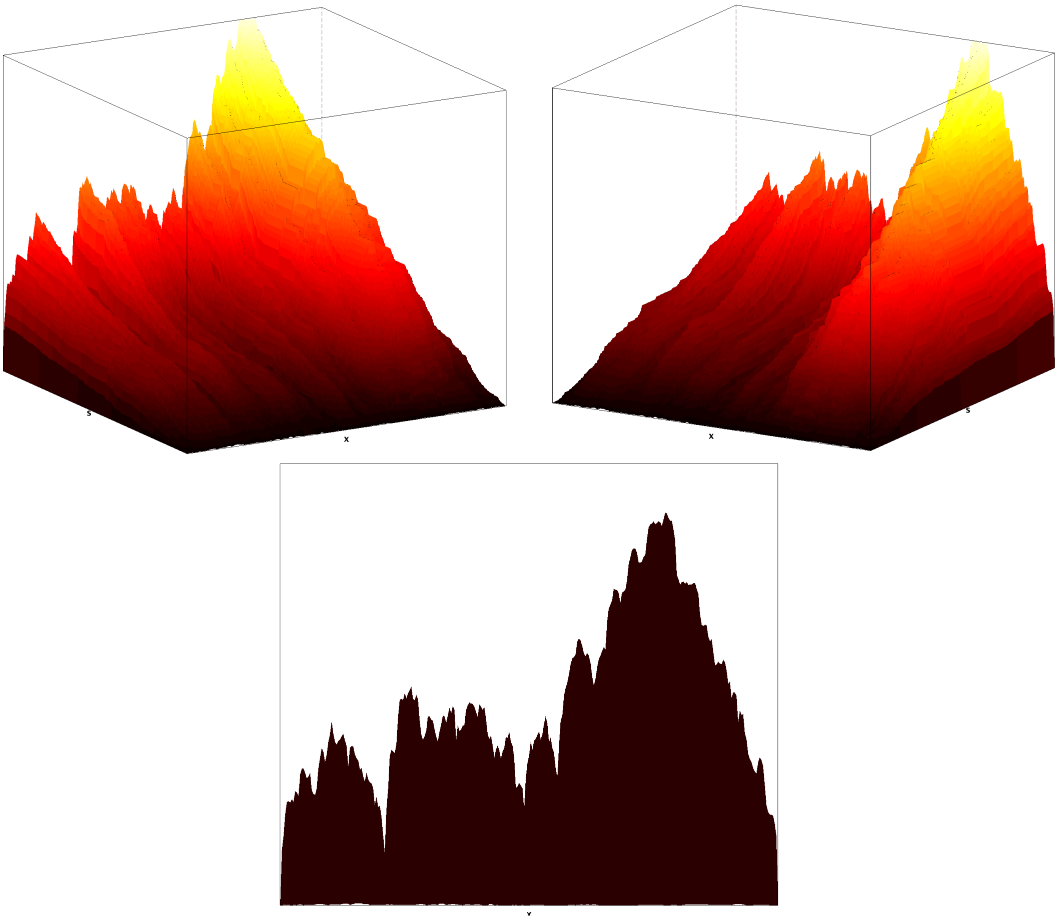

Figure 7: *A simulation of the process $\mathcal{Y}$: on the left, the process $\mathcal{Y}$; on the right, $\mathcal{Y}(\cdot, 1)$ which is distributed like $\mathcal{Z}$.*

Hence, with $\gamma_n := \mathbb{E}[\|n^{-\beta} Y_n - \mathcal{Y}\|^2]$, by the same steps relying on the Cauchy–Schwarz inequality that we used in the proof of Proposition 4.1 (but this time, for conditional expectations), we obtain

$$\gamma_n \leq \mathbb{E}\left[\sum_{r=1}^{4} \gamma_{N_r} \cdot \left(\frac{N_r}{n}\right)^{2\beta}\right] + 4\mathbb{E}\left[\sqrt{\gamma_{N_2} \cdot \mathbb{E}\left[\sup_{t\in[0,1]} |C^1_{N_1}(t) - \mathcal{Z}^1(t)|^2 \middle| N_1\right]}\right] + 8n^{-\beta}(1 + \mathbb{E}\left[\gamma_{N_1}\right]).$$

From here, since $N_r/n \to A_r$ for $r \in \{1, 2, 3, 4\}$ and $\gamma < 1$ with γ given in (27), an easy induction on n shows that $(\gamma_n)_{n\geq 1}$ is a bounded sequence. Therefore, the last display and the convergence of $\|C^1_n - \mathcal{Z}^1\|$ to zero in (15) imply that

$$\gamma_n \leq 4\mathbb{E}\left[\gamma_{N_1} \cdot \left(\frac{N_1}{n}\right)^{2\beta}\right] + o(1).$$

The verification that $\gamma_n \to 0$ is now standard in the framework of the contraction method. As in the previous proposition, since the convergence (28) also holds for higher moments, by induction over p, one verifies the convergence of the pth moment. Analogously, relying on (3.1), one then shows that $n^{-\beta p}\mathbb{E}[\|Y^{\geq}_n - Y^{<}_n\|^p] \to 0$ for all $p > 0$, thereby concluding the proof. □

4.3 Properties of the limit process $\mathcal{Y}$

Proposition 4.3. *For every $t, s \in [0, 1]$, we have $\mathbb{E}[\mathcal{Y}(t, s)] = K_1 h(t) g(s)$ where g is the unique bounded and measurable function on $[0, 1]$ satisfying (8). The function g is continuous and monotonically increasing.*

Proof. Write $\nu(t, s) = \mathbb{E}[\mathcal{Y}(t, s)]$. Taking the expectation in (25) and using the fact that $\nu(t, 1) = K_1 h(t)$

yields $\nu = G(\nu)$ where G is the functional operator given by

$$\begin{aligned} G(f)(t,s) = &\int_t^1\int_s^1 (uv)^\beta f\left(\frac{t}{u},\frac{s}{v}\right) dvdu + \int_0^t\int_0^s ((1-u)(1-v))^\beta f\left(\frac{t-u}{1-u},\frac{s-v}{1-v}\right) dvdu \\ &+ \int_t^1\int_0^s (u(1-v))^\beta f\left(\frac{t}{u},\frac{s-v}{1-v}\right) dvdu + \int_0^t\int_s^1 ((1-u)v)^\beta f\left(\frac{t-u}{1-u},\frac{s}{v}\right) dvdu \\ &+ K_1\int_t^1\int_0^s (u(1-v))^\beta h\left(\frac{t}{u}\right) dvdu + K_1\int_0^t\int_0^s ((1-u)v)^\beta h\left(\frac{t-u}{1-u}\right) dvdu. \end{aligned} \tag{29}$$

It should be clear from the structure of the terms on the right-hand side that the summands factorize if f is proportional to $h(t)y(s)$ for a bounded, measurable function y. More precisely, if $h\otimes y$ denotes the function on $[0,1]^2$ such that $h\otimes y(t,s) = h(t)y(s)$, we have

$$\begin{aligned} G(K_1 h\otimes y)(t,s) = K_1&\left(\int_t^1 u^\beta h\left(\frac{t}{u}\right) du + \int_0^t (1-u)^\beta h\left(\frac{t-u}{1-u}\right) du\right) \\ &\times\left(\int_s^1 v^\beta y\left(\frac{s}{v}\right) dv + \int_0^s (1-v)^\beta y\left(\frac{s-v}{1-v}\right) dv + \int_0^s v^\beta dv\right). \end{aligned}$$

As $\mathbb{E}[\mathcal{Z}(t)] = K_1 h(t)$, the fixed-point equation (13) implies that the first factor in the last display equals $K_1(\beta+1)h(t)/2$. (The details of the calculation are worked out in [7, Lemma 8].) Hence, if we choose $y = g$ with g as in the statement of the proposition, then the last display equals $K_1 h(t)g(s)$. To conclude the proof if suffices to show that:

(i) there exists at most one fixed-point of G in the set of bounded measurable functions on $[0,1]^2$,

(ii) there exists a unique bounded measurable function g satisfying (8), and

(iii) the function g in (ii) is continuous and increasing.

The first two claims follow from standard contraction arguments. We start with (ii). Let G' be the operator

$$G'(y)(s) = \frac{\beta+1}{2}\int_s^1 v^\beta y\left(\frac{s}{v}\right) dv + \int_0^s (1-v)^\beta y\left(\frac{s-v}{1-v}\right) dv + \frac{s^{\beta+1}}{2}.$$

Let g_1, g_2 be bounded measurable functions. Observing that the map $s \mapsto \int_s^1 v^\beta dv + \int_0^s (1-v)^\beta dv$ considered on $[0,1]$ attains its maximum which has value $(2-2^{-\beta})/(\beta+1)$ at $s = 1/2$, we obtain

$$\begin{aligned} \|G'(g_1) - G'(g_2)\| &= \frac{\beta+1}{2}\cdot\sup_{0\le s\le 1}\left|\int_s^1 v^\beta (g_1-g_2)\left(\frac{s}{v}\right) dv + \int_0^s (1-v)^\beta (g_1-g_2)\left(\frac{s-v}{1-v}\right) dv\right| \\ &\le \frac{\beta+1}{2}\|g_1-g_2\|\cdot\sup_{0\le s\le 1}\left|\int_s^1 v^\beta dv + \int_0^s (1-v)^\beta dv\right| \\ &\le (1-2^{-\beta-1})\|g_1-g_2\|. \end{aligned}$$

As $1-2^{-\beta-1} < 1$ and the space of bounded measurable functions on $[0,1]$ is complete with respect to the supremum norm, it follows from Banach's fixed-point theorem that there exists a unique bounded measurable solution of (8). Continuity of this solution follows easily from the theorem of dominated convergence. Monotonicity follows once we have verified (i) since, for any $n\ge 1$, the process $Y_n(t,s)$ (and hence its mean) is increasing in s. We move on to (i). Let f_1, f_2 be bounded and measurable functions on $[0,1]^2$. Then, we have

$$\begin{aligned} \left|G(f_1)(t,s) - G(f_2)(t,s)\right|^2 &= \left|\mathbb{E}\left[\sum_{r=1}^4 \mathbf{1}_{Q_r}(t,s)A_r^\beta\cdot(f_1-f_2)(\varphi_r(t),\varphi_r'(s))\right]\right|^2 \\ &\le \left|\mathbb{E}\left[\sum_{r=1}^4 \mathbf{1}_{Q_r}(t,s)A_r^\beta\cdot\left|(f_1-f_2)(\varphi_r(t),\varphi_r'(s))\right|\right]\right|^2 \\ &\le \mathbb{E}\left[\sum_{r=1}^4 \mathbf{1}_{Q_r}(t,s)A_r^{2\beta}\cdot\left|(f_1-f_2)(\varphi_r(t),\varphi_r'(s))\right|^2\right] \\ &\le \gamma\|f_1-f_2\|^2, \end{aligned}$$

where we used Jensen's inequality in the third step, and γ is the constant defined in (27). As $\gamma < 1$, taking the supremum over t, s on the left-hand side shows that G has at most one fixed-point. This proves (i). $\square$

To conclude the section, we show that the distribution of $\mathcal{Y}$ is characterized by the identity (25).

Proposition 4.4. *Up to a multiplicative constant, the process $\mathcal{Y}$ is the unique continuous field (in distribution) with $\mathbb{E}[\|\mathcal{Y}\|^2] < \infty$ satisfying the stochastic fixed-point equation*

$$\mathcal{Y} \stackrel{d}{=} \sum_{r=1}^{4} B_r(\mathcal{Y}^{(r)}), \tag{30}$$

with operators $B_1, \dots, B_4$ defined in (23), where $\mathcal{Y}^{(1)}, \dots, \mathcal{Y}^{(4)}, U, V$ are independent and $\mathcal{Y}^{(1)}, \dots, \mathcal{Y}^{(4)}$ are distributed like $\mathcal{Y}$.

Proof. While most parts of the functional contraction method in [25] are developed in the space $\mathcal{C}_1$ (and the space of càdlàg functions on $[0, 1]$) many results are formulated for general separable Banach spaces. Let $\mathcal{M}$ be the set of probability distributions on $\mathcal{C}_2$. Solutions of (30) in the space $\mathcal{M}$ are then fixed-points of the map $T : \mathcal{M} \to \mathcal{M}$ such that for $\mu \in \mathcal{M}$, we have

$$T(\mu) = \mathcal{L}\left(\sum_{r=1}^{4} B_r(X^{(r)}) \right),$$

where we write $\mathcal{L}(\,\cdot\,)$ for the distribution of a random variable, $X^{(1)}, \dots, X^{(4)}, U, V$ are independent, and $X^{(1)}, \dots, X^{(4)}$ are distributed like ν.

Let $\mathcal{M}_2$ denote the set of $\mu \in \mathcal{M}$ with $\int \|f\|^2 d\mu(f) < \infty$. For a probability distribution $\mu \in \mathcal{M}_2$, let $\mathcal{M}_2(\mu)$ be the set of probability distributions $\nu \in \mathcal{M}_2$ satisfying

$$\mathbb{E}[\psi(X)] = \mathbb{E}[\psi(Y)] \quad \text{for all } \psi \in \mathcal{C}_2^*, \tag{31}$$

where $\mathcal{L}(Y) = \nu$, $\mathcal{L}(X) = \mu$, and $\mathcal{C}_2^*$ denotes the topological dual space of $\mathcal{C}_2$, that is, the space of linear maps $\psi : \mathcal{C}_2 \to \mathbb{R}$ with

$$\|\psi\| := \sup_{f \in \mathcal{C}_2, \|f\|=1} |\psi(f)| < \infty.$$

By [25, Lemma 18] (applied with $s = 2$ in the notation there), a sufficient condition to ensure that there exists at most one fixed-point of T in $\mathcal{M}_2(\mathcal{L}(\mathcal{Y}))$ is that

(i) $T(\mathcal{M}_2(\mathcal{L}(\mathcal{Y}))) \subseteq \mathcal{M}_2(\mathcal{L}(\mathcal{Y}))$, and

(ii) $\sum_{r=1}^{4} \mathbb{E}[\|B_r\|^2] < 1$.

The second claim is easy to verify as $\|B_r\| = A_r^\beta$, and the sum therefore equals γ defined in (27). To prove (i), since $\|B_r(X^{(r)})\| \leq \|B_r\| \cdot \|X^{(r)}\|$ and these factors are independent, a simple application of Minkowski's inequality shows $T(\mathcal{M}_2) \subseteq \mathcal{M}_2$. Next, we recall the Riesz representation theorem: for $\psi \in \mathcal{C}_2^*$ there exists a (unique) finite signed measure η on $[0, 1]^2$ such that $\psi(f) = \int f(t) d\eta(t)$. Thus, by Fubini's theorem, condition (31) is satisfied if X and Y have the same mean functions. As we already know by Proposition 4.1 that $\mathcal{L}(\mathcal{Y})$ is a fixed-point of T in the set $\mathcal{M}_2(\mathcal{L}(\mathcal{Y}))$, the map T preserves the mean function on this set. This finishes the proof of (i).

To conclude the proof of the proposition, it remains to show that the mean function $m_\eta(t, s) := \int f(t)g(s) d\eta(f, g)$ of a fixed point η of T with $\eta \in \mathcal{M}_2$ is equal to the mean function of $\mathcal{Y}$ up to a multiplicative constant. By homogeneity, it suffices to consider the case when $\mathbb{E}[m_\eta(\xi, 1)] = \mathbb{E}[\mathcal{Y}(\xi, 1)] = \kappa$ with κ defined in (2). The claim follows from the previous proof as the map G defined in (8) has at most one fixed-point in $\mathcal{C}_2$ upon verifying that $m_\eta(t, 1) = K_1 h(t)$. The map $f(t) = m_\eta(t, 1)$ satisfies $\mathbb{E}[f(\xi)] = \kappa$ and

$$f(t) = \frac{2}{\beta + 1} \left(\int_t^1 u^\beta h\left(\frac{t}{u}\right) du + \int_0^t (1 - u)^\beta h\left(\frac{t - u}{1 - u}\right) du \right), \quad t \in [0, 1].$$

By the argument in [12, Section 5], the unique continuous (or only bounded and measurable) function satisfying these two properties is $K_1 h(t)$. This concludes the proof. $\square$

5 Proofs of the main results

The representation in Lemma 1.1 requires us to also consider partial match queries when the first coordinate is arbitrary, and the second should equal a specified value. Of course, all the results we have devised in Sections 2, 3 and 4 apply by symmetry, and we only need to set the notations to avoid confusions.

To this end, for $i \geq 1$, let the point $\overline{X}_i \in [0,1]^2$ be obtained from X_i be swapping the two coordinates. Switching from $(X_i)_{i\geq 1}$ to $(\overline{X}_i)_{i\geq 1}$ does not alter the shape of the trees: there is a consistent family of relabellings that transforms $(\overline{T}_n)_{n\geq 1}$ into $(T_n)_{n\geq 1}$. However, the corresponding partitions of the unit square are modified (and obtained from one another by a simple symmetry with respect to the principle diagonal of $[0,1]^2$). More specifically, for each $v \in \mathbb{T}$, the region $\overline{Q}_v$ is obtained from Q_v by swapping the coordinates of the four corners. Further, for the time-transformations defined in (10) and (11), we have $\overline{\varphi}_v = \varphi'_v$ and $\overline{\varphi}'_v = \varphi_v$. In particular, $\overline{A}_v = A_v$ for all $v \in \mathbb{T}$. We define the operators $\overline{B}^v_r, r = 1, \ldots, r$ analogously to B^v_r in (23) upon replacing Q_v, φ_v and φ'_v there by their analogues $\overline{Q}_v, \overline{\varphi}_v$ and $\overline{\varphi}'_v$.

Finally, we shall define the processes $\overline{C}_n, \overline{Y}_n$ and their one-sided versions in the process with swapped coordinates analogously to C_n, Y_n. Of course, all results proved above hold analogously for these quantities, and we denote the corresponding limits by $\overline{\mathcal{Z}}, \overline{\mathcal{Y}}$. The joint distributions of these quantities are intricate, but we can characterize them by distributional fixed-point equations.

Proposition 5.1. *(a) Up to a multiplicative constant, the pair $(\mathcal{Z}, \overline{\mathcal{Z}})$ is the unique $\mathcal{C}_1^2$-valued process (in distribution) satisfying $\mathbb{E}[\|\mathcal{Z}\|^2]$, $\mathbb{E}[\|\overline{\mathcal{Z}}\|^2] < \infty$ and*

$$(\mathcal{Z}, \overline{\mathcal{Z}}) \stackrel{d}{=} \sum_{r=1}^{4} A_r^\beta \left(\mathbf{1}_{Q_r^{(1)}}(\,\cdot\,) \mathcal{Z}^{(r)}(\varphi_r(\,\cdot\,)), \mathbf{1}_{Q_r^{(2)}}(\,\cdot\,) \overline{\mathcal{Z}}^{(r)}(\varphi'_r(\,\cdot\,)) \right). \tag{32}$$

Here, $(\mathcal{Z}^{(1)}, \overline{\mathcal{Z}}^{(1)}), \ldots, (\mathcal{Z}^{(4)}, \overline{\mathcal{Z}}^{(4)})$ are independent copies of $(\mathcal{Z}, \overline{\mathcal{Z}})$, independent of U, V.
(b) Similarly, up to a multiplicative constant, the pair $(\mathcal{Y}, \overline{\mathcal{Y}})$ is the unique $\mathcal{C}_2^2$-valued process satisfying $\mathbb{E}[\|\mathcal{Y}\|^2]$, $\mathbb{E}[\|\overline{\mathcal{Y}}\|^2] < \infty$ and

$$(\mathcal{Y}, \overline{\mathcal{Y}}) \stackrel{d}{=} \sum_{r=1}^{4} \left(B_r(\mathcal{Y}^{(r)}), \overline{B}_r(\overline{\mathcal{Y}}^{(r)}) \right). \tag{33}$$

Here, $(\mathcal{Y}^{(1)}, \overline{\mathcal{Y}}^{(1)}), \ldots, (\mathcal{Y}^{(4)}, \overline{\mathcal{Y}}^{(4)})$ are independent copies of $(\mathcal{Y}, \overline{\mathcal{Y}})$, independent of U, V.

Proof. Both statements are proved following the lines of the proof of Proposition 4.4. To show the first, define $\mathcal{M}, \mathcal{M}_2$ and $\mathcal{M}_2(\mu)$ as in that proof, but in the space $\mathcal{C}_1^2$. The only step of that proof one needs to take a closer look at is the verification that, for $\mu, \nu \in \mathcal{M}_2$, we have $\nu \in \mathcal{M}_2(\mu)$ (that is, condition (31) holds), if $\mathbb{E}[X_i(t,s)] = \mathbb{E}[Y_i(t,s)]$ for $i = 1, 2$ and $t, s \in [0,1]$, where $\mathcal{L}(X_1, X_2) = \mu$ and $\mathcal{L}(Y_1, Y_2) = \nu$. As in the proof of Proposition 4.4, this follows from a simple application of Fubini's theorem since every bounded linear form $\psi \in (\mathcal{C}_2^2)^*$ can be written as $\psi(f,g) = \int f d\eta_1 + \int g d\eta_2$ for finite signed measures η_1, η_2 on $[0,1]^2$. The second assertion for the process $(\mathcal{Y}, \overline{\mathcal{Y}})$ follows analogously. □

Proofs of Theorem 1.1 and Proposition 1.4. Recall the representation in Lemma 1.1. We are now ready to make formal the arguments at the end of Section 1.4: Theorem 1.1 follows with

$$\mathcal{O}(a,b,c,d) = \frac{\mathcal{Y}(b,d) - \mathcal{Y}(b,c) + \mathcal{Y}(a,d) - \mathcal{Y}(a,c) + \overline{\mathcal{Y}}(d,b) - \overline{\mathcal{Y}}(d,a) + \overline{\mathcal{Y}}(c,b) - \overline{\mathcal{Y}}(c,a)}{2}, \tag{34}$$

since

(i) the summands involving $Y_n^{\geq}, Y_n^{<}, \overline{Y}_n^{\geq}$ and $\overline{Y}_n^{<}$ converge uniformly after rescaling by Proposition 4.2,

(ii) $\|n^{-\beta}(N_n - n\mathrm{Vol})\| \to 0$ in probability and with convergence of moments by the argument given in the proof of Corollary 3.2, and

(iii) $\|n^{-\beta} D_n^{(i)}\| \to 0$ for $i = 1, \ldots, 4$, in probability and with convergence of moments since $D_n^{(i)}$ is bounded from above by the height of T_n which is well-known to be $O(\log n)$ [see, e.g,. 13].

Proposition 1.4 follows from (34) and Proposition 4.3. □

To characterize the limit process $\mathcal{O}$, we consider the distributional decomposition of O_n. As for the process Y_n, this requires the definition of four linear operators, this time on the space $\mathcal{C}_4^+$. We set

$$\begin{aligned}
D_1(f)(a,b,c,d) = A_1^\beta\big[&\mathbf{1}_{Q_1}(a,c)\mathbf{1}_{Q_1}(b,d)f(\varphi_1(a),\varphi_1(b),\varphi_1'(c),\varphi_1'(d))\\
&+\mathbf{1}_{Q_1}(a,c)\mathbf{1}_{Q_2}(b,d)f(\varphi_1(a),\varphi_1(b),\varphi_1'(c),1)\\
&+\mathbf{1}_{Q_1}(a,c)\mathbf{1}_{Q_3}(b,d)f(\varphi_1(a),1,\varphi_1'(c),\varphi_1'(d))\\
&+\mathbf{1}_{Q_1}(a,c)\mathbf{1}_{Q_4}(b,d)f(\varphi_1(a),1,\varphi_1'(c),1)\big]\\
D_2(f)(a,b,c,d) = A_2^\beta\big[&\mathbf{1}_{Q_2}(a,c)\mathbf{1}_{Q_2}(b,d)f(\varphi_2(a),\varphi_2(b),\varphi_2'(c),\varphi_2'(d))\\
&+\mathbf{1}_{Q_1}(a,c)\mathbf{1}_{Q_2}(b,d)f(\varphi_2(a),\varphi_2(b),0,\varphi_2'(d))\\
&+\mathbf{1}_{Q_2}(a,c)\mathbf{1}_{Q_3}(b,d)f(\varphi_2(a),1,\varphi_2'(c),\varphi_2'(d))\\
&+\mathbf{1}_{Q_1}(a,c)\mathbf{1}_{Q_4}(b,d)f(\varphi_2(a),1,0,\varphi_2'(d))\big]
\end{aligned} \tag{35}$$

$$\begin{aligned}
D_3(f)(a,b,c,d) = A_3^\beta\big[&\mathbf{1}_{Q_3}(a,c)\mathbf{1}_{Q_3}(b,d)f(\varphi_3(a),\varphi_3(b),\varphi_3'(c),\varphi_3'(d))\\
&+\mathbf{1}_{Q_1}(a,c)\mathbf{1}_{Q_3}(b,d)f(0,\varphi_3(b),\varphi_3'(c),\varphi_3'(d))\\
&+\mathbf{1}_{Q_3}(a,c)\mathbf{1}_{Q_4}(b,d)f(\varphi_3(a),\varphi_3(b),\varphi_3'(c),1)\\
&+\mathbf{1}_{Q_1}(a,c)\mathbf{1}_{Q_4}(b,d)f(0,\varphi_3(b),\varphi_3'(c),1)\big]\\
D_3(f)(a,b,c,d) = A_4^\beta\big[&\mathbf{1}_{Q_4}(a,c)\mathbf{1}_{Q_4}(b,d)f(\varphi_4(a),\varphi_4(b),\varphi_4'(c),\varphi_4'(d))\\
&+\mathbf{1}_{Q_2}(a,c)\mathbf{1}_{Q_4}(b,d)f(0,\varphi_4(b),\varphi_4'(c),\varphi_3'(d))\\
&+\mathbf{1}_{Q_3}(a,c)\mathbf{1}_{Q_4}(b,d)f(\varphi_4(a),\varphi_4(b),0,\varphi_3'(d))\\
&+\mathbf{1}_{Q_1}(a,c)\mathbf{1}_{Q_4}(b,d)f(0,\varphi_4(b),0,\varphi_4'(d))\big].
\end{aligned} \tag{36}$$

Proof of Proposition 1.3. By construction, we have, for every $n \geq 1$,

$$O_n \stackrel{d}{=} \left(\sum_{r=1}^{4} D_r(O_{N_r}^{(r)})(a,b,c,d)+1\right)_{(a,b,c,d)\in I}$$

with conditions on independence and distributions as in (14). By Theorem 1.1 it follows that the limit field $\mathcal{O}$ is solution to the fixed-point equation (7). By the same argument used in the proofs of Propositions 4.4 and 5.1, one shows that $\mathcal{L}(\mathcal{O})$ is the unique solution of (7) in the set $\mathcal{M}_2(\mathcal{L}(\mathcal{O}))$, where this set is defined as in the proof of Proposition 4.4 but with $\mathcal{C}_4^+$ instead of $\mathcal{C}_2$. It remains to check that, up to a multiplicative constant, the operator G'' given by $G''(f)(a,b,c,d) = \sum_{r=1}^4 \mathbb{E}[D_r(f)(a,b,c,d)]$, has a unique fixed-point in the space $\mathcal{C}_4^+$. This follows from several applications of the contraction arguments used in the proof of Proposition 4.4 whose structure we now describe. Let $\varrho_1, \varrho_2 \in \mathcal{C}_4^+$ be two fixed-points of G'' satisfying $\mathbb{E}[\varrho_1(\xi,\xi,0,1)] = \mathbb{E}[\varrho_1(\xi,\xi,0,1)] = \kappa$. (a) By [12, Section 5], this implies $\varrho_1(x,x,0,1) = \varrho_2(x,x,0,1) = K_1 h(x)$. (b) Then, setting $(a,c) = (0,0)$, the contraction argument in the proof of Proposition 4.4 shows that $\varrho_1(0,x,0,y) = \varrho_2(0,x,0,y)$; proceeding analogously for the choices $(a,d) = (0,1)$, $(b,c) = (1,0)$ and $(b,d) = (1,1)$ yields $\varrho_1(0,x,y,1) = \varrho_2(0,x,y,1)$, $\varrho_1(x,1,0,y) = \varrho_2(x,1,0,y)$ and $\varrho_1(x,1,y,1) = \varrho_2(x,1,y,1)$. (c) Then, in the next step, one sets $a = 0$ and proves along the same lines that $\varrho_1(0,x,y,z) = \varrho_2(0,x,y,z)$; analogously for $b = 1, c = 0$ and $d = 1$ yields $\varrho_1(x,1,y,z) = \varrho_2(x,1,y,z)$, $\varrho_1(x,y,0,z) = \varrho_2(x,y,0,z)$ and $\varrho_1(x,y,z,1) = \varrho_2(x,y,z,1)$. Finally, with these identities in hand, one can verify that $\|\varrho_1 - \varrho_2\| = \|G''(\varrho_1) - G''(\varrho_2)\| \leq \sqrt{\gamma}\|\varrho_1 - \varrho_2\|$, where γ is defined in (27), just as for the operator G in Proposition 4.4. Since $\gamma < 1$, this gives $\varrho_1 = \varrho_2$ and concludes the proof. □

6 Orthogonal range queries in random 2-d trees

The 2-d trees have been introduced by Bentley [1]. As for quadtrees, the data are partitioned recursively, but the splits in 2-d trees are only binary; since the data is two-dimensional, one alternates between vertical and horizontal splits, depending on the parity of the level in the tree. Given a sequence of points $p_1, p_2, \ldots \in [0,1]^2$, the tree and the regions associated to each node are constructed as follows. Initially, T_0 is an empty tree, which consists of a placeholder, to which we assign the entire square $[0,1]^2$. The first point

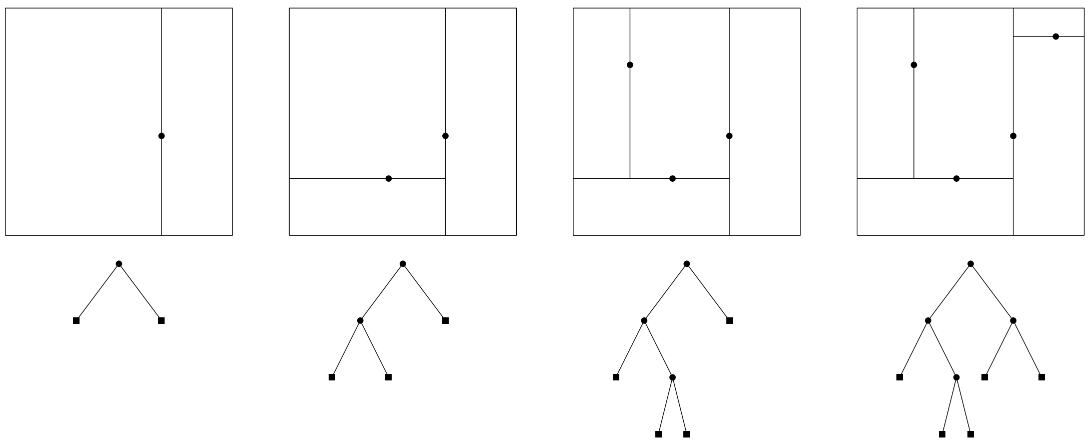

Figure 8: *The first four steps of the construction of a* 2*-d tree, and the corresponding partitions of the unit square.*

p_1 is inserted in this placeholder, and becomes the root, thereby giving rise to two new placeholders. Geometrically, p_1 splits *vertically* the unit square in two rectangles, which are associated with the two children of the root. More generally, when i points have already been inserted, the tree T_i has i internal nodes, and induces a partition of the unit square into $i+1$ regions, each one associated to one of the $i+1$ placeholders of T_i. The point p_{i+1} is then stored in the placeholder, say v, that is assigned to the rectangle of the partition containing p_{i+1}. This operation turns v into an internal node, and creates two new placeholders just below. Geometrically, p_{i+1} divides this rectangle into two subregions that are assigned to the two newly created placeholders; that last partition step depends on the parity of the depth of v in the tree: if it is odd we partition horizontally, if it is even we partition vertically. See Figure 8 for an illustration. (Of course, one could start at the root with a *horizontal* split, and then splits would occur horizontally at even levels and vertically at odd levels.)

We now consider a sequence $(X_i)_{i\ge 1}$ of i.i.d. uniform random points in $[0,1]$, and the 2-d trees obtained by sequential insertion of $X_1, X_2, \dots$ into an initially empty 2-d tree. It is convenient to consider the trees as subtrees of the infinite binary tree $\mathcal{T} = \cup_{m\ge 0}\{1,2\}^m$. We can actually construct at the same time two sequences of trees $(T_n^{=})_{n\ge 0}$ and $(T_n^{\perp})_{n\ge 0}$ as well as the corresponding refining partitions of $[0,1]^2$ encoded in the collections $\{Q_v^{=} : v \in \mathcal{T}\}$ and $\{Q_v^{\perp} : v \in \mathcal{T}\}$, which correspond to the two cases where the split at the root is horizontal or vertical respectively. Put aside the binary splitting, the construction is similar to the one in Section 2 and we omit the details, and only mention that if $X_1 = (U, V)$, then

$$\begin{cases} Q_1^{=} = [0,1]\times[0,V] \\ Q_2^{=} = [0,1]\times(V,1] \end{cases} \qquad \text{and} \qquad \begin{cases} Q_1^{\perp} = [0,U]\times[0,1] \\ Q_2^{\perp} = (U,1]\times[0,1]\,. \end{cases}$$

For $(a,b,c,d)\in I$, let $O_n^{=}(a,b,c,d)$ and $O_n^{\perp}(a,b,c,d)$ denote the number of nodes of the 2-d tree visited to perform the query with rectangle $Q(a,b,c,d)$ when the partition at the root is horizontal or vertical respectively. Define $\overline{O}_n^{=}$ and $\overline{O}_n^{\perp}$ analogously in the 2-d tree constructed from the sequence $\overline{X}_i, i\ge 1$, obtained by swapping the two coordinates of each point. By construction, we have $O_n^{=}(a,b,c,d) = \overline{O}_n^{\perp}(c,d,a,b)$ and $O_n^{\perp}(a,b,c,d) = \overline{Q}_n^{=}(c,d,a,b)$. In particular, with

$$\iota : I \to I, \quad \iota(a,b,c,d) = (c,d,a,b), \tag{37}$$

the sequences $(O_i^{=})_{i\ge 1}$ and $(O_i^{\perp}\circ\iota)_{i\ge 1}$ are identically distributed. Hence, it suffices to focus on the sequence $O_i^{=}, i \ge 1$.

Theorem 6.1. *There exist random continuous $\mathcal{C}_4^{+}$-valued random variables $\mathcal{O}^{=}$ and $\mathcal{O}^{\perp}$ (random fields) such that, in probability and with convergence of all moments,*

$$\left\| \frac{O_n^{=} - n\mathrm{Vol}}{n^{\beta}} - \mathcal{O}^{=} \right\| \xrightarrow[n\to\infty]{} 0 \qquad \textit{and} \qquad \left\| \frac{O_n^{\perp} - n\mathrm{Vol}}{n^{\beta}} - \mathcal{O}^{\perp} \right\| \xrightarrow[n\to\infty]{} 0.$$

The processes $\mathcal{O}^{\perp}$ and $\mathcal{O}^{=}\circ\iota$ have the same distribution.

To characterize the limit field $\mathcal{O}^{=}$ as the solution to a stochastic fixed-point equation let us keep track of the relative positions with respect to cells of the partition as follows: for $s \in [0,1]$, we set

$$\psi_1^{=}(s) = \mathbf{1}_{[0,V]}(s)\frac{s}{V} \qquad \text{and} \qquad \psi_2^{=}(s) = \mathbf{1}_{[V,1]}(s)\frac{s-V}{1-V}\,.$$

Note that the volumes of the rectangular regions at the first level of the partition are

$$A_{=,1} = V, \qquad A_{=,2} = 1-V\,.$$

We now define the operators $D_1^{=}$ and $D_2^{=}$

$$\begin{aligned} D_1^{=}(f)(a,b,c,d) &= A_{=,1}^{\beta}\left[\mathbf{1}_{[0,V]}(d) f(a,b,\psi_1^{=}(c),\psi_1^{=}(d)) + \mathbf{1}_{[0,V]\times(V,1]}(c,d) f(a,b,\psi_1^{=}(c),1)\right] \\ D_2^{=}(f)(a,b,c,d) &= A_{=,2}^{\beta}\left[\mathbf{1}_{(V,1]}(c) f(a,b,\psi_2^{=}(c),\psi_2^{=}(d)) + \mathbf{1}_{[0,V]\times(V,1]}(c,d) f(a,b,0,\psi_2^{=}(d))\right]. \end{aligned} \tag{38}$$

Proposition 6.2. *Up to a multiplicative constant, $\mathcal{O}^{=}$ is the unique $\mathcal{C}_4^{+}$-valued random field (in distribution) such that $\mathbb{E}[\|\mathcal{O}^{=}\|^2] < \infty$ satisfying the stochastic fixed-point equation*

$$\mathcal{O}^{=} \stackrel{d}{=} D_1^{=}(\mathcal{O}^{=,(1)} \circ \iota) + D_2^{=}(\mathcal{O}^{=,(2)} \circ \iota) \tag{39}$$

where $\mathcal{O}^{=,(1)}$ and $\mathcal{O}^{=,(2)}$ are copies of $\mathcal{O}^{=}$, $D_1^{=}, D_2^{=}$ are random linear operators defined in (38) and $\iota : I \to I$ is defined in (37). Furthermore, the random variables $\mathcal{O}^{=,(1)}, \mathcal{O}^{=,(2)}$, and $(D_1^{=}, D_2^{=})$ are independent.

Proposition 6.2 only characterizes the distribution of $\mathcal{O}^{=}$ up to a multiplicative constant. The next proposition identifies the limit mean, and hence the missing multiplicative constant. Here, the values of the constants appearing are reminiscent of the fact that, for uniform partial match queries in the trees $T_n^{=}$ and $T_n^{\perp}$, Flajolet and Puech [22] proved the analogue of expansion (1) and Chern and Hwang [10] identified the leading constants as

$$\kappa^{=} = \frac{13(3-5\beta)}{2}\kappa \qquad \text{and} \qquad \kappa^{\perp} = 13(2\beta-1)\kappa\,.$$

Note that $2\kappa^{=} = (\beta+1)\kappa^{\perp}$.

Proposition 6.3. *Let $(a,b,c,d) \in I$. Then, we have*

$$\begin{aligned} \mathbb{E}[\mathcal{O}^{=}(a,b,c,d)] = &\frac{13(3-5\beta)}{4}\Big(\mu(a,d)-\mu(a,c)+\mu(b,d)-\mu(b,c)\Big) \\ &+ \frac{13}{2(2\beta-1)}\Big(\mu(c,b)-\mu(c,a)+\mu(d,b)-\mu(d,a)\Big), \end{aligned}$$

where $\mu(t,s)$ is the function from Proposition 1.4.

References

[1] J. L. Bentley. Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 18:509–517, 1975.

[2] J. L. Bentley. Decomposable searching problems. *Inform. Process. Lett.*, 8(5):244–251, 1979.

[3] J. L. Bentley and J. H. Friedman. Data structures for range searching. *ACM Computing Surveys*, 11:397–409, 1979.

[4] P. Billingsley. *Convergence of Probability Measures*. Wiley Series in Probability and Mathematical Statistics. Wiley, second edition, 1999.

[5] N. Broutin and H. Sulzbach. The dual tree of a recursive triangulation of the disk. *Ann. Probab.*, 43(2):738–781, 2015.

[6] N. Broutin, R. Neininger, and H. Sulzbach. Partial match queries in random quadtrees. In *Proceedings of the Twenty-Third Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1056–1065. ACM, New York, 2012.

[7] N. Broutin, R. Neininger, and H. Sulzbach. A limit process for partial match queries in random quadtrees and 2-d trees. *Ann. Appl. Probab.*, 23(6):2560–2603, 2013.

[8] P. Chanzy, L. Devroye, and C. Zamora-Cura. Analysis of range search for random k-d trees. *Acta Informatica*, 37:355–383, 2000.

[9] H.-H. Chern and H.-K. Hwang. Partial match queries in random quadtrees. *SIAM J. Comp.*, 32:904–915, 2003.

[10] H.-H. Chern and H.-K. Hwang. Partial match queries in random k-d trees. *SIAM J. Comp.*, 35:1440–1466, 2006.

[11] N. Curien. Strong convergence of partial match queries in random quadtrees. *Combin. Probab. Comput.*, 21(5): 683–694, 2012.

[12] N. Curien and A. Joseph. Partial match queries in two-dimensional quadtrees: A probabilistic approach. *Adv. in Appl. Probab.*, 43:178–194, 2011.

[13] L. Devroye. Branching processes in the analysis of the heights of trees. *Acta Informatica*, 24:277–298, 1987.

[14] L. Devroye. Universal limit laws for depth in random trees. *SIAM J. Comp.*, 28(2):409–432, 1998.

[15] L. Devroye and L. Laforest. An analysis of d-dimensional quad trees. *SIAM J. Comp.*, 19:821–832, 1990.

[16] L. Devroye, J. Jabbour, and C. Zamora-Cura. Squarish k-d trees. *SIAM J. Comp.*, 30:1678–1700, 2001.

[17] A. Duch and C. Martínez. On the average performance of orthogonal range search in multidimensional data structures. *J. Algorithms*, 44(1):226–245, 2002.

[18] A. Duch, V. Estivill-Castro, and C. Martínez. Randomized k-dimensional binary search trees. In K.-Y. Chwa and O.H. Ibarra, editors, *Proc. of the 9th International Symposium on Algorithms and Computation (ISAAC'98)*, volume 1533 of *Lecture Notes in Computer Science*, pages 199–208. Springer Verlag, 1998.

[19] A. Dvoretzky, J. Kiefer, and J. Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Ann. Math. Statist.*, 27:642–669, 1956.

[20] R. A. Finkel and J. L. Bentley. Quad trees, a data structure for retrieval on composite keys. *Acta Informatica*, 4: 1–19, 1974.

[21] P. Flajolet and T. Lafforgue. Search costs in quadtrees and singularity perturbation asymptotics. *Discrete Comp. Geom.*, 12:151–175, 1994.

[22] P. Flajolet and C. Puech. Partial match retrieval of multidimensional data. *J. Assoc. Comput. Mach.*, 33(2): 371–407, 1986.

[23] P. Flajolet, G. H. Gonnet, C. Puech, and J. M. Robson. Analytic variations on quadtrees. *Algorithmica*, 10: 473–500, 1993.

[24] P. Flajolet, G. Labelle, L. Laforest, and B. Salvy. Hypergeometrics and the cost structure of quadtrees. *Random Structures Algorithms*, 7:117–144, 1995.

[25] Ralph Neininger and Henning Sulzbach. On a functional contraction method. *Ann. Probab.*, 43(4), 2015.

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

[27] H. Samet. *Applications of Spatial Data Structures: Computer Graphics, Image Processing, and GIS*. Addison-Wesley, Reading, MA, 1990.

[28] H. Samet. *Foundations of multidimensional and metric data structures*. Morgan Kaufmann, 2006.