Partial match queries in random quadtrees

Nicolas Broutin
nicolas.broutin@inria.fr
Projet Algorithms
INRIA Rocquencourt
78153 Le Chesnay
France

Ralph Neininger and Henning Sulzbach
{neiningr, sulzbach}@math.uni-frankfurt.de
Institute for Mathematics (FB 12)
J.W. Goethe University
60054 Frankfurt am Main
Germany

July 13, 2011

Abstract

We consider the problem of recovering items matching a partially specified pattern in multidimensional trees (quad trees and k-d trees). We assume the traditional model where the data consist of independent and uniform points in the unit square. For this model, in a structure on n points, it is known that the number of nodes $C_n(\xi)$ to visit in order to report the items matching an independent and uniformly on $[0,1]$ random query ξ satisfies $\mathbf{E}[C_n(\xi)] \sim \kappa n^\beta$, where κ and β are explicit constants. We develop an approach based on the analysis of the cost $C_n(x)$ of any fixed query $x \in [0,1]$, and give precise estimates for the variance and limit distribution of the cost $C_n(x)$. Our results permit to describe a limit process for the costs $C_n(x)$ as x varies in $[0,1]$; one of the consequences is that $\mathbf{E}[\max_{x\in[0,1]} C_n(x)] \sim \gamma n^\beta$; this settles a question of Devroye [Pers. Comm., 2000].

1 Introduction

Multidimensional databases arise in a number of contexts such as computer graphics, management of geographical data or statistical analysis. The question of retrieving the data matching a specified pattern is then of course of prime importance. If the pattern specifies all the data fields, the query can generally be answered in logarithmic time, and a great deal of precise analyses are available in this case [11, 13, 15, 18, 19]. We will be interested in the case when the pattern only constrains some of the data fields; we then talk of a *partial match query*.

The first investigations about partial match queries by Rivest [28] were based on digital structures. In a comparison-based setting, a few general purpose data structures generalizing binary search trees permit to answer partial match queries, namely the quadtree [10], the k-d tree [1] and the relaxed k-d tree [7]. Aside of the interest that one might have in partial match for itself, there are numerous reasons that justify the precise quantification of the cost of such general search queries in comparison-based data structures. The high dimesional trees are indeed a data structure of choice for applications that range from collision detection in motion planning to mesh generation that takes advantage of the adaptive partition of space that is produced [17, 35]. For general references on multidimensional data structures and more details about their various applications, see the series of monographs by Samet [32, 33, 34]. The cost of partial match queries also appears in (hence influences) the complexity of a number of other geometrical search questions such as range search [6] or rank selection [8].

In spite of its importance, the complexity results about partial match queries are not as precise as one could expect. In this paper, we provide novel analyses of the costs of partial match queries in some of the most important two dimensional data structures. Most of the document will focus on the special case of quadtrees ; in a final section, we discuss the case of k-d tree [1] and relaxed k-d trees [7].

QUAD TREES AND MULTIDIMENSIONAL SEARCH. The quadtree [10] allows to manage multidimensional data by extending the divide-and-conquer approach of the binary search tree. Consider the point

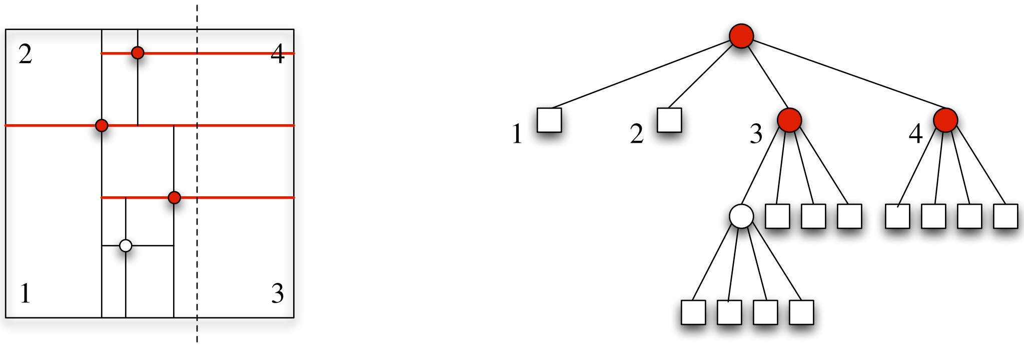

Figure 1: An example of a (point) quadtree: on the left the partition of the unit square induced by the tree data structure on the right (the children are ordered according to the numbering of the regions on the left). Answering the partial match query materialized by the dashed line on the left requires to visit the points/nodes coloured in red. Note that each one of the visited nodes correspond to a horizontal line that is crossed by the query.

sequence $p_1, p_2, \dots, p_n \in [0,1]^2$. As we build the tree, regions of the unit square are associated to the nodes where the points are stored. Initially, the root is associated with the region $[0,1]^2$ and the data structure is empty. The first point p_1 is stored at the root, and divides the unit square into four regions $Q_1, \dots, Q_4$. Each region is assigned to a child of the root. More generally, when i points have already been inserted, we have a set of $1+3^i$ (lower-level) regions that cover the unit square. The point p_{i+1} is stored in the node (say u) that corresponds to the region it falls in, divides it into four new regions that are assigned to the children of u. See Figure 1.

ANALYSIS OF PARTIAL MATCH RETRIEVAL. For the analysis, we will focus on the model of *random quadtrees*, where the data points are uniformly distributed in the unit square. In the present case, the data are just points, and the problem of partial match retrieval consists in reporting all the data with one of the coordinates (say the first) being $s \in [0,1]$. It is a simple observation that the number of nodes of the tree visited when performing the search is precisely $C_n(s)$, the number of regions in the quadtree that insersect a vertical line at s. The first analysis of partial match in quadtrees is due to Flajolet et al. [14] (after the pioneering work of Flajolet and Puech [12] in the case of k-d trees). They studied the singularities of a differential system for the generating functions of partial match cost to prove that, for a random query ξ, being independent of the tree and uniformly distributed on $[0,1]$,

$$\mathbf{E}[C_n(\xi)] \sim \kappa\, n^\beta \qquad \text{where} \qquad \kappa = \frac{\Gamma(2\beta+2)}{2\Gamma(\beta+1)^3}, \quad \beta = \frac{\sqrt{17}-3}{2}, \tag{1}$$

and $\Gamma(x)$ denotes the Gamma function $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$. This has since been strengthened by Chern and Hwang [3], who provided the order of the error term (together with the values of the leading constant in all dimensions). The most precise result is (6.2) there, saying that

$$\mathbf{E}[C_n(\xi)] = \kappa\, n^\beta - 1 + O(n^{\beta-1}). \tag{2}$$

To gain a refined understanding of the cost beyond the level of expectations we pursue two directions. First, to justify that the expected value is a reasonable estimate of the cost, one would like a guarantee that the cost of partial match retrieval are actually close to their mean. However, deriving higher moments turns out to be more subtle than it seems. In particular, when the query line is random (like in the uniform case) although the four subtrees at the root really are independent given their sizes, the contributions of the two subtrees that do hit the query line are *dependent*! The relative location of the query line inside these two subtrees, is again uniform, but unfortunately it is same in both regions. This issue has not yet been addressed appropriately, and there is currently no result on the variance of or higher moments for $C_n(\xi)$.

The second issue lies in the very definition of the cost measure: even if the data follow some distribution (here uniform), should one really assume that the query also satisfies this distribution? In other words, should we focus on $C_n(\xi)$? Maybe not. But then, what distribution should one use for the query line?

One possible approach to overcome both problems is to consider the query line to be fixed and to study $C_n(s)$ for $s \in [0,1]$. This raises another problem: even if s is fixed at the top level, as the search is performed, the *relative* location of the the queries in the recursive calls varies from a node to another! Thus, in following this approach, one is led to consider the entire process $C_n(s), s \in [0,1]$; this is the method we use here.

Recently Curien and Joseph [4] obtained some results in this direction. They proved that for every fixed $s \in (0,1)$,

$$\mathbf{E}[C_n(s)] \sim K_1(s(1-s))^{\beta/2}n^\beta, \qquad K_1 = \frac{\Gamma(2\beta+2)\Gamma(\beta+2)}{2\Gamma(\beta+1)^3\Gamma\left(\beta/2+1\right)^2}. \tag{3}$$

On the other hand, Flajolet et al. [14, 15] prove that, along the edge one has $\mathbf{E}[C_n(0)] = \Theta(n^{\sqrt{2}-1}) = o(n^\beta)$ (see also [4]). The behaviour about the x-coordinate U of the first data point certainly resembles that along the edge, so that one has $\mathbf{E}[C_n(U)] = o(n^\beta)$. It suggests that $C_n(s)$ should not be concentrated around its mean, and that $n^{-\beta}C_n(s)$ should converge to a non-trivial random variable as $n \to \infty$. This random variable would of course carry much information about the asymptotic properties of the cost of partial match queries in quadtrees. Below, we identify these limit random variables and obtain refined asymptotic information on the complexity of partial match queries in quadtrees from them.

2 Main results and implications

Our main contribution is to prove the following convergence result:

Theorem 1. *Let $C_n(s)$ be the cost of a partial match query at a fixed line s in a random quadtree. Then, there exists a random continuous function Z such that, as $n \to \infty$,*

$$\left(\frac{C_n(s)}{K_1 n^\beta}, s \in [0,1]\right) \xrightarrow{d} (Z(s), s \in [0,1]). \tag{4}$$

This convergence in distribution holds in the Banach space $(\mathcal{D}[0,1], \|\cdot\|)$ of right-continuous functions with left limits (càdlàg) equipped with the supremum norm defined by $\|f\| = \sup_{s\in[0,1]}|f(s)|$.

Note that the convergence in (4) above is stronger than the convergence in distribution of the finite dimensional marginals

$$\left(\frac{C_n(s_1)}{K_1 n^\beta}, \frac{C_n(s_2)}{K_1 n^\beta}, \dots, \frac{C_n(s_k)}{K_1 n^\beta}\right) \xrightarrow{d} (Z(s_1), Z(s_2), \dots, Z(s_k))$$

as $n \to \infty$, for any natural number k and points $s_1, s_2, \dots, s_k \in [0,1]$ [see, e.g., 2]. Theorem 1 has a myriad of consequences in terms of estimates of the costs of partial match queries in random quadtrees. Of course, Theorem 1 would be of less practical interest if we could not characterize the distribution of the random function Z (see Figure 2 for a simulation):

Proposition 2. *The distribution of the random function Z in (4) is a fixed point of the following recursive functional equation*

$$\begin{aligned} Z(s) \overset{d}{=} &\mathbf{1}_{\{s<U\}}\left[(UV)^\beta Z^{(1)}\left(\frac{s}{U}\right) + (U(1-V))^\beta Z^{(2)}\left(\frac{s}{U}\right)\right] \\ &+ \mathbf{1}_{\{s\geq U\}}\left[((1-U)V)^\beta Z^{(3)}\left(\frac{s-U}{1-U}\right) + ((1-U)(1-V))^\beta Z^{(4)}\left(\frac{s-U}{1-U}\right)\right], \end{aligned} \tag{5}$$

where U and V are independent $[0,1]$-uniform random variables and $Z^{(i)}$, $i = 1, \dots, 4$ are independent copies of the process Z, which are also independent of U and V. Furthermore, Z in (4) is the only solution of (5) such that $\mathbf{E}[Z(s)] = (s(1-s))^{\beta/2}$ for all $s \in [0,1]$ and $\mathbf{E}[\|Z\|^2] < \infty$.

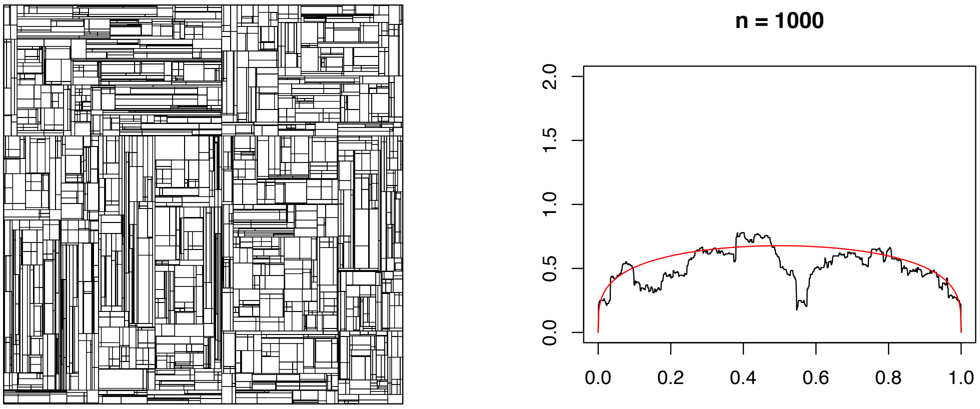

Figure 2: A random quadtree on 1000 points and the corresponding partial match process on the right ; in red we have shown the expected value.

This is indeed relevant since the convergence that implies Theorem 1 is strong enough to guarantee convergence of the variance of the costs of partial match queries. The following theorem for uniform queries ξ is the direct extension of the pioneering work of Flajolet and Puech [12], Flajolet et al. [14] for the cost of partial match queries at a uniform line in random multidimensional trees.

Theorem 3. *If ξ is uniformly distributed on $[0,1]$, independent of (C_n) and Z, then*

$$\frac{C_n(\xi)}{K_1 n^\beta} \xrightarrow{d} Z(\xi),$$

in distribution with convergence of the first two moments. In particular

$$\mathbf{Var}\,(C_n(\xi)) \sim K_4 n^{2\beta} \qquad \textit{where} \qquad K_4 := K_1^2 \cdot \mathbf{Var}(Z(\xi)) \approx 0.447363034.$$

In particular, Theorem 3 identifies the asymptotic order of $\mathbf{Var}(C_n(\xi))$ which is to be compared with studies that neglected the dependence between the contributions of the subtrees mentioned above [20, 21, 23]. We also have an asymptotic for the variance of the cost at a fixed query:

Theorem 4. *We have for all $s \in (0,1)$, as $n \to \infty$,*

$$\mathbf{Var}\,(C_n(s)) \sim \left(2\mathrm{B}(\beta+1,\beta+1)\frac{2\beta+1}{3(1-\beta)} - 1\right)(s(1-s))^\beta n^{2\beta}. \tag{6}$$

Here, $\mathrm{B}(a,b) := \int_0^1 x^{a-1}(1-x)^{b-1}\,dx$ denotes the Eulerian beta integral ($a, b > 0$).

Some of the most striking consequence concerns the cost of the *worst query* in a random plane quadtree. Note in particular that the supremum does not induce any extra logarithmic terms in the asymptotic cost.

Theorem 5. *Let $S_n = \sup_{s\in[0,1]} C_n(s)$. Then, as $n \to \infty$,*

$$n^{-\beta} S_n \xrightarrow{d} S \overset{d}{=} \sup_{s\in[0,1]} Z(s) \qquad \textit{and} \qquad \mathbf{E}[S_n] \sim n^\beta \mathbf{E}[S], \qquad \mathbf{Var}(S_n) \sim n^{2\beta}\mathbf{Var}(S).$$

Finally we note that the one-dimension marginals of the limit process $(Z(s), s \in [0,1])$ are all the same up to a multiplicative constant.

Theorem 6. *There is a random variable $Z \geq 0$ such that for all $s \in [0,1]$,*

$$Z(s) \stackrel{d}{=} (s(1-s))^{\beta/2} Z. \tag{7}$$

The distribution of Z is characterized by its moments $c_m := \mathbf{E}[Z^m]$, $m \in \mathbb{N}$. They are given by $c_1 = 1$ and the recurrence

$$c_m = \frac{2(\beta m + 1)}{(m-1)\left(m+1-\frac{3}{2}\beta m\right)} \sum_{\ell=1}^{m-1} \binom{m}{\ell} \mathrm{B}(\beta\ell+1, \beta(m-\ell)+1) c_\ell c_{m-\ell}, \quad m \geq 2.$$

Plan of the paper. Our approach requires to work with random functions; as one might expect, proving convergence in a space of functions involves a fair amount of unavoidable technicalities. Here, we try to keep the discussion at a rather high level, to avoid diluting the main ideas in an ocean of intricate details. In Section 3, we give an overview of our main tool, the contraction method. In Section 4, we identify the variance and the supremum of the limit process Z, and deduce the large n asymptotics for $C_n(s)$ in Theorems 3 and 5.

3 Contraction method: from the real line to functional spaces

3.1 Overview

The aim of this section is give an overview of the method we employ to prove Theorem 1. The idea is very natural and relies on a contraction argument in a certain space of probability distributions. In the context of the analysis of performance of algorithms, the method was first employed by Rösler [29] who proved convergence in distribution for the rescaled total cost of the randomized version of quicksort. The method was then further developed by Rachev and Rüschendorf [27], Rösler [30], and later on in [5, 9, 22, 24, 25, 31] and has permitted numerous analyses in distribution for random discrete structures.

So far, the method has mostly been used to analyze random variables taking real values, though a few applications on functions spaces have been made, see [5, 9, 16]. Here we are interested in the function space $\mathcal{D}[0,1]$ with the uniform topology, but the main idea persists: (1) devise a recursive equation for the quantity of interest (here the process$(C_n(s), s \in [0,1])$), and (2) prove that a properly rescaled version of the quantity converges to a fixed point of a certain map related to the recursive equation ; (3) if the map is a contraction in a certain metric space, then a fixed point is unique and may be obtained by iteration. We now move on to the first step of this program.

Write $I_1^{(n)}, \dots, I_4^{(n)}$ for the number of points falling in the four regions created by the point stored at the root. Then, given the coordinates of the first data point (U, V), we have, cf. Figure 1,

$$(I_1^{(n)}, \dots, I_4^{(n)}) \stackrel{d}{=} \mathrm{Mult}(n-1; UV, U(1-V), (1-U)(1-V), (1-U)V).$$

Observe that, for the cost inside a subregion, what matters is the location of the query line *relative* to the region. Thus a decomposition at the root yields the following recursive relation, for any $n \geq 1$,

$$C_n(s) \stackrel{d}{=} 1 + \mathbf{1}_{\{s<U\}} \left[C^{(1)}_{I_1^{(n)}}\left(\frac{s}{U}\right) + C^{(2)}_{I_2^{(n)}}\left(\frac{s}{U}\right) \right] + \mathbf{1}_{\{s \geq U\}} \left[C^{(3)}_{I_3^{(n)}}\left(\frac{1-s}{1-U}\right) + C^{(4)}_{I_4^{(n)}}\left(\frac{1-s}{1-U}\right) \right], \tag{8}$$

where $U, I_1^{(n)}, \dots, I_4^{(n)}$ are the quantities already introduced and $(C_k^{(1)}), \dots, (C_k^{(4)})$ are independent copies of the sequence $(C_k, k \geq 0)$, independent of $(U, V, I_1^{(n)}, \dots, I_4^{(n)})$. We stress that this equation does not only hold true pointwise for fixed s but also as cádlág functions on the unit interval. The relation in (8) is the fundamental equation for us.

Letting $n \to \infty$ (formally) in (8) suggests that, if $n^{-\beta} C_n(s)$ does converge to a random variable $Z(s)$ in a sense to be precised, then the distribution of the process $(Z(s), 0 \leq s \leq 1)$ should satisfy the following fixed point equation

$$\begin{aligned} Z(s) \stackrel{d}{=} & \mathbf{1}_{\{s<U\}} \left[(UV)^\beta Z^{(1)}\left(\frac{s}{U}\right) + (U(1-V))^\beta Z^{(2)}\left(\frac{s}{U}\right)\right] \\ & + \mathbf{1}_{\{s\geq U\}} \left[((1-U)V)^\beta Z^{(3)}\left(\frac{s-U}{1-U}\right) + ((1-U)(1-V))^\beta Z^{(4)}\left(\frac{s-U}{1-U}\right)\right], \end{aligned} \tag{9}$$

where U and V are independent $[0,1]$-uniform random variables and $Z^{(i)}$, $i = 1,\dots,4$ are independent copies of the process Z, which are also independent of U and V.

The last step leading to the fixed point equation (9) needs now to be made rigorous. This is at this point that the contraction method enters the game. The distribution of a solution to our fixed-point equation (9) lies in the set of probability measures on the Banach space $(\mathcal{D}[0,1], \|\cdot\|)$, which is the set we have to endow with a metric. The recursive equation (8) is an example for the following, more general setting of random additive recurrences: Let (X_n) be $\mathcal{D}[0,1]$-valued random variables with

$$X_n \stackrel{d}{=} \sum_{r=1}^{K} A_r^{(n)}\Big(X^{(r)}_{I_r^{(n)}}\Big) + b^{(n)}, \quad n \geq 1, \tag{10}$$

where $(A_1^{(n)},\dots,A_K^{(n)})$ are random linear and continuous operators on $\mathcal{D}[0,1]$, $b^{(n)}$ is a $\mathcal{D}[0,1]$-valued random variable, $I_1^{(n)},\dots,I_K^{(n)}$ are random integers between 0 and $n-1$ and $(X_n^{(1)}),\dots,(X_n^{(K)})$ are distributed like (X_n). Moreover $(A_1^{(n)},\dots,A_K^{(n)},b^{(n)},I_1^{(n)},\dots,I_K^{(n)}),(X_n^{(1)}),\dots,(X_n^{(K)})$ are independent.

To establish Theorem 1 as a special case of this setting we use Proposition 7 below. Proposition 7 is covered by the forthcoming paper [26]. We first state conditions needed to deal with the general recurrence (10); we will then justify that it can indeed be used in the case of cost of partial match queries. Consider the following assumptions, where, for a random linear operator A we write $\|A\|_2 := \mathbf{E}[\|A\|_{\mathrm{op}}^2]^{1/2}$ with $\|A\|_{\mathrm{op}} := \sup_{\|x\|=1}\|A(x)\|$. Suppose (X_n) obeys (10) and

(A1) Convergence and contraction. We have $\|A_r^{(n)}\|_2, \|b_n\|_2 < \infty$ for all $r = 1,\dots,K$ and $n \geq 0$ and there exist random operators $A_1,\dots,A_K$ on $\mathcal{D}[0,1]$ and a $\mathcal{D}[0,1]$-valued random variable b with, for some positive sequence $R(n) \downarrow 0$, as $n \to \infty$,

$$\|b^{(n)} - b\|_2 + \sum_{r=1}^{K}\left(\|A_r^{(n)} - A_r\|_2 + \left\|\mathbf{1}_{\{I_r^{(n)} \leq n_0\}} A_r^{(n)}\right\|_2\right) = O(R(n)) \tag{11}$$

and for all $\ell \in \mathbb{N}$,

$$\mathbf{E}\left[\mathbf{1}_{\{I_r^{(n)} \in \{0,\dots,\ell\}\cup\{n\}\}} \|A_r^{(n)}\|_{\mathrm{op}}^2\right] \to 0$$

and

$$L^* = \limsup_{n\to\infty} \mathbf{E}\left[\sum_{r=1}^{K} \|A_r^{(n)}\|_{\mathrm{op}}^2 \frac{R(I_r^{(n)})}{R(n)}\right] < 1. \tag{12}$$

(A2) Existence and equality of moments. $\mathbf{E}[\|X_n\|^2] < \infty$ for all n and $\mathbf{E}[X_{n_1}(t)] = \mathbf{E}[X_{n_2}(t)]$ for all $n_1, n_2 \in \mathbb{N}_0, t \in [0,1]$.

(A3) Existence of a continuous solution. There exists a solution X of the fixed-point equation

$$X \stackrel{d}{=} \sum_{r=1}^{K} A_r(X^{(r)}) + b \tag{13}$$

with continuous paths, $\mathbf{E}[\|X\|^2] < \infty$ and $\mathbf{E}[X(t)] = \mathbf{E}[X_1(t)]$ for all $t \in [0,1]$. Again $(A_1,\dots,A_K,b),X^{(1)},\dots,X^{(K)}$ are independent and $X^{(1)},\dots,X^{(K)}$ are distributed like X.

(A4) PERTURBATION CONDITION. $X_n = W_n + h_n$ where $\|h_n - h\| \to 0$ with $h \in \mathcal{D}[0,1]$ and random variables W_n in $\mathcal{D}[0,1]$ such that there exists a sequence (r_n) with, as $n \to \infty$,

$$\mathbf{P}\left(W_n \notin \mathcal{D}_{r_n}[0,1]\right) \to 0.$$

Here, $\mathcal{D}_{r_n}[0,1] \subset \mathcal{D}[0,1]$ denotes the set of functions on the unit interval, for which there is a decomposition of $[0,1]$ into intervals of length as least r_n on which they are constant.

(A5) RATE OF CONVERGENCE. $R(n) = o\left(\log^{-m}(1/r_n)\right)$.

The crucial part that makes everything work consists in choosing a probability metric in such a way that the limiting map is indeed a contraction. The contraction method presented here for the Banach space $(\mathcal{D}[0,1], \|\cdot\|)$ is based on the Zolotarev metric ζ_s and, for our fixed-point equation, we indeed obtain contraction with $s = 2$. This follows by our modified assumption A1 since

$$\mathbf{E}\left[\sum_{r=1}^{K} \|A_r\|^2\right] = \lim_n \mathbf{E}\left[\sum_{r=1}^{K} \|A_r^{(n)}\|^2\right] \leq \limsup_n \mathbf{E}\left[\sum_{r=1}^{K} \|A_r^{(n)}\|^2 \frac{R(I_r^{(n)})}{R(n)}\right] < 1.$$

The amounts of details to be verified prevents us to provide a complete proof of all the assumptions in the present case. In the remainder of the section, we will not come back on the method and Proposition 7 itself but show how it can be applied; we will however, discuss and outline the proof of the main assumptions (A1), (A2), (A3) and (A5).

Proposition 7. *Let X_n fulfill (10). Provided that Assumptions (A1)–(A3) are satisfied, the solution X of the fixed-point equation (13) is unique.*

i. For all $t \in [0,1]$, $X_n(t) \to X(t)$ in distribution, with convergence of the first two moments;

ii. If U is independent of $(X_n), X$ and distributed on $[0,1]$ then $X_n(U) \to X(U)$ in distribution again with convergence of the first two moments.

iii. If also (A4) and (A5) hold, then $X_n \to X$ in distribution in $(\mathcal{D}[0,1], \|\cdot\|)$.

3.2 Existence of a continuous solution

In this section, we outline the proof of existence of a continuous process Z that satisfies the distributional fixed point equation (9) as it is needed for assumption (A3). We construct the process Z as the pointwise limit of martingales. We then show that the convergence is actually almost surely uniform, which allows us to conclude that Z is actually continuous with probability one. Write $\mathcal{C}[0,1]$ for the space of continuous functions on $[0,1]$.

Consider the infinite 4-ary tree $\mathcal{T} = \bigcup_{n\geq 0}\{1,2,3,4\}^n$. For a node $u \in \mathcal{T}$, we write $|u|$ for its depth, i.e. the distance between u and the root $\varnothing$. The descendants of $u \in \mathcal{T}$ correspond to all the words in $\mathcal{T}$ with prefix u. Let $\{U_v, v \in \mathcal{T}\}$ and $\{V_v, v \in \mathcal{T}\}$ be two independent families of i.i.d. $[0,1]$-uniform random variables.

CONSTRUCTION BY ITERATION. Define the operator $G : (0,1)^2 \times \mathcal{C}[0,1]^4 \to \mathcal{C}[0,1]$ by

$$\begin{aligned} G(x,y,f_1,f_2,f_3,f_4)(s) =& \mathbf{1}_{\{s<x\}}\left[(xy)^\beta f_1\left(\frac{s}{x}\right) + (x(1-y))^\beta f_2\left(\frac{s}{x}\right)\right] \\ &+ \mathbf{1}_{\{s\geq x\}}\left[((1-x)y)^\beta f_3\left(\frac{s-x}{1-x}\right) + ((1-x)(1-y))^\beta f_4\left(\frac{s-x}{1-x}\right)\right]. \end{aligned} \tag{14}$$

Let h be the map defined by $h(s) = (s(1-s))^{\beta/2}$, where $2\beta = \sqrt{17} - 3$. For every node $u \in \mathcal{T}$, let $Z_0^u = h$. Then define recursively

$$Z_{n+1}^u = G(U_u, V_u, Z_n^{u1}, Z_n^{u2}, Z_n^{u3}, Z_n^{u4}). \tag{15}$$

Finally, define $Z_n = Z_n^\varnothing$ to be the value observed at the root of $\mathcal{T}$ when the iteration has been started with h in all the nodes at level n.

A SERIES REPRESENTATION FOR Z_n. For $s \in [0,1]$, $Z_n(s)$ is the sum of exactly 2^n terms, each one being the contribution of one of the boxes at level n that is cut by the line at s. Let $\{Q_i^n(s), 1 \le i \le 2^n\}$ be the set of rectangles at level n whose first coordinate intersect s. Suppose that the projection of $Q_i^n(s)$ on the first coordinate yields the interval $[\ell_i^n, r_i^n]$. Then

$$Z_n(s) = \sum_{i=1}^{2^n} \mathrm{Leb}(Q_i^n(s))^\beta \cdot h\left(\frac{s-\ell_i^n}{r_i^n - \ell_i^n}\right), \tag{16}$$

where $\mathrm{Leb}(Q_i^n(s))$ denotes the volume of the rectangle $Q_i^n(s)$. The difference between Z_n and Z_{n+1} only relies in the functions appearing the boxes $Q_i^n(s)$: We have

$$Z_{n+1}(s) - Z_n(s) = \sum_{i=1}^{2^n} \mathrm{Leb}(Q_i^n(s))^\beta \cdot \left[G(U_i', V_i', h, h, h, h)\left(\frac{s-\ell_i^n}{r_i^n - \ell_i^n}\right) - h\left(\frac{s-\ell_i^n}{r_i^n - \ell_i^n}\right)\right], \tag{17}$$

where U_i', V_i', $1 \le i \le 2^n$ are i.i.d. $[0,1]$-uniform random variables. In fact, U_i' and V_i' are some of the variables U_u, V_u for nodes u at level n. Observe that, although $Q_i^n(s)$ is *not* a product of n independent terms of the form UV because of size-biasing, U_i', V_i' are in fact *unbiased*, i.e. uniform. Let $\mathscr{F}_n$ denote the σ-algebra generated by $\{U_u, V_u : |u| < n\}$. Then the family $\{U_i', V_i' : 1 \le i \le 2^n\}$ is independent of $\mathscr{F}_n$.

A MARTINGALE. Let $s \in [0,1]$ be fixed. We show that the sequence $(Z_n(s), n \ge 0)$ is a non-negative discrete time martingale ; so it converges with probability one to a finite limit $Z(s)$. To prove that $Z_n(s)$ is a indeed a martingale, it suffices to prove that, for $1 \le i \le 2^n$,

$$\mathbf{E}\left[\, G(U_i', V_i', h, h, h, h)\left(\frac{s-\ell_i^n}{r_i^n - \ell_i^n}\right) \,\middle|\, \mathscr{F}_n \,\right] = h\left(\frac{s-\ell_i^n}{r_i^n - \ell_i^n}\right).$$

Since $U_i', V_i', 1 \le i \le 2^n$ are independent of $\mathscr{F}_n$, this clearly reduces to the following lemma.

Lemma 8. *For the operator G defined in (14) and U, V two independent $[0,1]$-uniform random variables, and any $s \in [0,1]$, we have* $\mathbf{E}\left[G(U,V,h,h,h,h)(s)\right] = h(s)$.

ALMOST SURE CONTINUITY. Assume for the moment that there exist constants $a, b \in (0,1)$ and C such that

$$\mathbf{P}\left(\sup_{s\in[0,1]} |Z_{n+1}(s) - Z_n(s)| \ge a^n\right) \le C \cdot b^n. \tag{18}$$

Then, by the Borel–Cantelli lemma, the sequence (Z_n) is almost surely cauchy with respect to the supremum norm. Completeness of $(\mathcal{C}[0,1], \|\cdot\|)$ yields the existence of a random process Z with continuous paths such that $Z_n \to Z$ uniformly on $[0,1]$. We now move on to showing that there exist constants a and b such that (18) is satisfied. We start by a bound for a fixed value $s \in [0,1]$.

Lemma 9. *For every $s \in [0,1]$, any $a \in (0,1)$, and any integer n large enough, we have the bound*

$$\mathbf{P}\left(|Z_{n+1}(s) - Z_n(s)| \ge a^n\right) \le 4(16e\log(1/a))^n.$$

Then, in order to handle the supremum over $s \in [0,1]$, in (18) note that the number of values taken by Z_n is at most the number of boxes at level n, i.e. 4^n. To avoid unnecessary technicalities, we use fixed points (much more than 4^n) to control the extent of $\sup_{s\in[0,1]} |Z_{n+1}(s) - Z_n(s)|$. Consider the set V_n of x-coordinates of the vertical boundaries of all the rectangles at level n. Let $L_n = \inf\{|x-y| : x, y \in V_n\}$. Then, on the event that $L_n \ge \gamma^n$, we have

$$\sup_{s\in[0,1]} |Z_{n+1}(s) - Z_n(s)| \le \sup_{1\le i\le\lfloor\gamma^{-n}\rfloor} |Z_{n+1}(i\gamma^n) - Z_n(i\gamma^n)|.$$

In particular, it follows by the union bound that, for any $\gamma \in (0,1)$,

$$\mathbf{P}\left(\sup_{s\in[0,1]} |Z_{n+1}(s) - Z_n(s)| \geq a^n\right) \leq \gamma^{-n} \sup_{s\in[0,1]} \mathbf{P}\left(|Z_{n+1}(s) - Z_n(s)| \geq a^n\right) + \mathbf{P}\left(L_n < \gamma^n\right).$$

The following lemma then yields (18) which completes the proof.

Lemma 10. *For any positive real number γ small enough, it exists an integer $n_1(\gamma)$ with*

$$\mathbf{P}\left(L_n < \gamma^n\right) \leq 6 \cdot 4^n \gamma^{n/201}, \qquad n \geq n_1(\gamma).$$

3.3 Uniform convergence of the mean

The proof of Theorem 1 requires to show uniform convergence of the first moment $n^{-\beta}\mathbf{E}\left[C_n(s)\right]$ towards $\mu_1(s) = K_1(s(1-s))^{\beta/2}$ uniformly on $[0,1]$ in order to verify assumption (A1), in particular the rate $R(n)$ in (11). Note that, since $C_n(s)$ is continuous in any fixed s almost surely, the function $s \to \mathbf{E}\left[C_n(s)\right]$ is continuous for any n. Curien and Joseph [4] only show pointwise convergence, and proving uniform convergence requires a good deal of additional arguments.

The first step is to prove a Poissonized version, the fixed-n version is then obtained by a routine Tauberian argument. Consider a Poisson point process with unit intensity on $[0,1]^2 \times [0,\infty)$. The first two coordinates represent the location inside the unit square; the third one represents the time of arrival of the point. Let $P_t(s)$ denote the partial match cost for a query at $x = s$ in the quad tree built from the points arrived by time t.

Proposition 11. *There exists $\varepsilon > 0$ such that*

$$\sup_{s\in[0,1]} |t^{-\beta}\mathbf{E}[P_t(s)] - \mu_1(s)| = O(t^{-\varepsilon}).$$

The proof of Proposition 11 relies crucially on two main ingredients: first, a strengthening of the arguments developed by Curien and Joseph [4], and the speed of convergence $\mathbf{E}[C_n(\xi)]$ to $\mathbf{E}[\mu_1(\xi)]$ for a uniform query line ξ, see (2), by Chern and Hwang [3]. By symmetry, we write for any $\delta \in (0,1/2)$

$$\sup_{s\in[0,1]} |t^{-\beta}\mathbf{E}[P_t(s)] - \mu_1(s)| \leq \sup_{s\leq\delta} \left|t^{-\beta}\mathbf{E}[P_t(s)] - \mu_1(s)\right| + \sup_{s\in(\delta,1/2]} \left|t^{-\beta}\mathbf{E}[P_t(s)] - \mu_1(s)\right|. \tag{19}$$

The two terms in the right-hand side above are controlled by the following two lemmas.

Lemma 12 (Behavior on the edge)**.** *There exists a constant C_1 such that*

$$\limsup_{t\to\infty} \sup_{s\leq\delta} \left|t^{-\beta}\mathbf{E}[P_t(s)] - \mu_1(s)\right| \leq C_1\delta^{\beta/2}. \tag{20}$$

Lemma 13 (Behavior away from the edge)**.** *There exist constants C_2, C_3, η with $0 < \eta < \beta$ and $\gamma \in (0,1)$ such that, for any integer k, and real number $\delta \in (0,1/2)$ we have, for any $t > 0$,*

$$\sup_{s\geq\delta} |t^{-\beta}\mathbf{E}[P_t(s)] - \mu_1(s)| \leq C_2\delta^{-1}(1-\gamma)^k + C_3 k 2^k (\beta-\eta)^{-2k} t^{-\eta}.$$

BEHAVIOUR ALONG THE EDGE. The behaviour away from the edge is rather involved and we do not describe how the bound in Lemma 13 is obtained. To deal with the term for involving the values of $s \in [0,\delta]$, we relate the value $\mathbf{E}[P_t(s)]$ to $\mathbf{E}[P_t(\delta)]$. Note that the limit first moment $\mu_1(s) = \lim_{n\to\infty} \mathbf{E}[P_t(s)]$ is monotonic for $s \in [0,1/2]$. It seems, at least intuitively, that for any fixed real number $t > 0$, $\mathbf{E}[P_t(s)]$ should also be monotonic for $s \in [0,1/2]$, but we were unable to prove it. The following weaker version is sufficient for our purpose.

Proposition 14 (Almost monotonicity)**.** *For any $s < 1/2$ and $\varepsilon \in [0, 1-2s)$, we have*

$$\mathbf{E}[P_t(s)] \leq \mathbf{E}\left[P_{t(1+\varepsilon)}\left(\frac{s+\varepsilon}{1+\varepsilon}\right)\right].$$

4 Second moment and supremum

In this section, we obtain explicit expressions about the limit, proving that our general approach also turns out to yield effective and computable results.

VARIANCE OF THE COST. We first focus on the result in Theorem 3. Our main result implies the convergence $n^{-2\beta}\mathbf{E}[C_n(s)^2] \to \mathbf{E}[Z(s)^2]$. Write $h(s) = \mathbf{E}[Z(s)] = (s(1-s))^{\beta/2}$. Taking second moments in (9) and writing it as an integral in terms of $\mu_2(s) = \mathbf{E}[Z(s)^2]$ yields that we have the following integral equation, for every $s \in [0,1]$,

$$\mu_2(s) = \frac{2}{2\beta+1}\left\{\int_s^1 x^{2\beta}\mu_2\left(\frac{s}{x}\right)dx + \int_0^s (1-x)^{2\beta}\mu_2\left(\frac{1-s}{1-x}\right)dx\right\} + 2\mathrm{B}(\beta+1,\beta+1)\cdot\frac{h(s)^2}{\beta+1}.$$

One easily verifies that the function f given by $f(s) = c_2 h(s)^2$ solves the above equation provided that the constant c_2 satisfies

$$c_2 = \frac{2}{(2\beta+1)(\beta+1)}c_2 + 2\frac{\mathrm{B}(\beta+1,\beta+1)}{\beta+1} \qquad \text{that is} \qquad c_2 = 2\mathrm{B}(\beta+1,\beta+1)\frac{2\beta+1}{3(1-\beta)},$$

since $\beta^2 = 2 - 3\beta$. So if we were sure that $\mu_2(s)$ is indeed $c_2 h(s)^2$, we would have by integration

$$\mathbf{Var}(Z(\xi)) = c_2\mathrm{B}(\beta+1,\beta+1) - \mathrm{B}(\beta/2+1,\beta/2+1)^2.$$

To complete the proof, it suffices to show that the integral equation satisfied by μ_2 actually admits a unique solution. To this aim, we show that the map K defined below is a contraction for the supremum norm (the details are omitted)

$$Kf(s) = \frac{2}{2\beta+1}\left\{\int_s^1 x^{2\beta} f\left(\frac{s}{x}\right)dx + \int_0^s (1-x)^{2\beta} f\left(\frac{1-s}{1-x}\right)dx\right\} + 2\mathrm{B}(\beta+1,\beta+1)\frac{[s(1-s)]^\beta}{\beta+1}.$$

COST OF THE WORST QUERY. The uniform convergence of $n^{-\beta}C_n(\cdot)$ to the process $Z(\cdot)$ directly implies (continuous mapping theorem) the first claim of Theorem 5,

$$\frac{S_n}{K_1 n^\beta} \xrightarrow{d} S := \sup_{s\in[0,1]} Z(s). \tag{21}$$

The convergence in the Zolotarev metric ζ_2 on which the contraction method is based here, is strong enough to imply convergence of the first two moments of S_n to the corresponding moments of S.

5 Concluding remarks

The method we exposed here to obtain refined results about the costs of partial match queries in quadtrees also applies to other geometric data structures based on the divide-and-conquer approach. In particular, similar results can be obtained for the k-d trees of Bentley [1] or the relaxed k-d trees of Duch et al. [7].

We conclude by mentioning some open questions. The supremum of the process is of great interest since it upperbounds the cost of any query. Can one identify the moments of the supremum $\sup_{s\in[0,1]} Z(s)$ (first and second)? In the course of our proof, we had to construct a continuous solution of the fixed point equation. We prove convergence in distribution, but conjecture that the convergence actually holds almost surely.

References

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

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

[3] H. Chern and H. Hwang. Partial match queries in random quadtrees. *SIAM Journal on Computing*, 32: 904–915, 2003.

[4] N. Curien and A. Joseph. Partial match queries in two-dimensional quadtrees: A probabilistic approach. *Advances in Applied Probability*, 43:178–194, 2011.

[5] M. Drmota, S. Janson, and R. Neininger. A functional limit theorem for the profile of search trees. *Ann. Appl. Probab.*, 18(1):288–333, 2008. ISSN 1050-5164.

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

[7] A. Duch, V. Estivill-Castro, and C. Martínez. Randomized k-dimensional binary search trees. In K.-Y. Chwa and O. 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.

[8] A. Duch, R. Jiménez, and C. Martínez. Rank selection in multidimensional data. In A. López-Ortiz, editor, *Proceedings of LATIN*, volume 6034 of *Lecture Notes in Computer Science*, pages 674–685, Berlin, 2010. Springer.

[9] K. Eickmeyer and L. Rüschendorf. A limit theorem for recursively defined processes in L^p. *Statist. Decisions*, 25(3):217–235, 2007. ISSN 0721-2631.

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

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

[12] P. Flajolet and C. Puech. Partial match retrieval of multidimensional data. *Jounal of the ACM*, 33(2):371–407, 1986.

[13] P. Flajolet and R. Sedgewick. *Analytic Combinatorics*. Cambridge University Press, Cambridge, UK, 2009.

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

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

[16] R. Grübel. On the silhouette of binary search trees. *Ann. Appl. Probab.*, 19(5):1781–1802, 2009. ISSN 1050-5164.

[17] K. Ho-Le. Finite element mesh generation methods: a review and classification. *Computer-Aided Design*, 20:27–38, 1988.

[18] D. E. Knuth. *The Art of Computer Programming: Sorting and Searching*, volume 3. Addison-Wesley, 2d edition, 1998.

[19] H. Mahmoud. *Evolution of Random Search Trees*. Wiley, New York, 1992.

[20] C. Martínez, A. Panholzer, and H. Prodinger. Partial match in relaxed multidimensional search trees. *Algorithmica*, 29(1–2):181–204, 2001.

[21] R. Neininger. Asymptotic distributions for partial match queries in K-d trees. *Random Structures and Algorithms*, 17:403–427, 2000.

[22] R. Neininger. On a multivariate contraction method for random recursive structures with applications to Quicksort. *Random Structures Algorithms*, 19(3-4):498–524, 2001. ISSN 1042-9832. Analysis of algorithms (Krynica Morska, 2000).

[23] R. Neininger and L. Rüschendorf. Limit laws for partial match queries in quadtrees. *The Annals of Applied Probability*, 11:452–469, 2001.

[24] R. Neininger and L. Rüschendorf. A general limit theorem for recursive algorithms and combinatorial structures. *Ann. Appl. Probab.*, 14(1):378–418, 2004. ISSN 1050-5164.

[25] R. Neininger and L. Rüschendorf. On the contraction method with degenerate limit equation. *Ann. Probab.*, 32(3B):2838–2856, 2004. ISSN 0091-1798.

[26] R. Neininger and H. Sulzbach. On a functional contraction method. 2011. Manuscript in preparation.

[27] S. Rachev and L. Rüschendorf. Probability metrics and recursive algorithms. *Advances in Applied Probability*, 27:770–799, 1995.

[28] R. Rivest. Partial-match retrieval algorithms. *SIAM Journal on Computing*, 5(19–50), 1976.

[29] U. Rösler. A limit theorem for "quicksort". *RAIRO Informatique théorique et Applications*, 25:85–100, 1991.

[30] U. Rösler. A fixed point theorem for distributions. *Stochastic Processes and their Applications*, 37:195–214, 1992.

[31] U. Rösler. On the analysis of stochastic divide and conquer algorithms. *Algorithmica*, 29(1-2):238–261, 2001. ISSN 0178-4617. Average-case analysis of algorithms (Princeton, NJ, 1998).

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

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

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

[35] M. Yerry and M. Shephard. A modified quadtree approach to finite element mesh generation. *IEEE Computer Graphics and Applications*, 3:39–46, 1983.