Random Projection Trees Revisited

Aman Dhesi[*]
Department of Computer Science
Princeton University
Princeton, New Jersey, USA.
adhesi@princeton.edu

Purushottam Kar
Department of Computer Science and Engineering
Indian Institute of Technology
Kanpur, Uttar Pradesh, INDIA.
purushot@cse.iitk.ac.in

Abstract

The Random Projection Tree (RPTREE) structures proposed in [1] are space partitioning data structures that automatically adapt to various notions of intrinsic dimensionality of data. We prove new results for both the RPTREE-MAX and the RPTREE-MEAN data structures. Our result for RPTREE-MAX gives a near-optimal bound on the number of levels required by this data structure to reduce the size of its cells by a factor $s \geq 2$. We also prove a packing lemma for this data structure. Our final result shows that low-dimensional manifolds have bounded Local Covariance Dimension. As a consequence we show that RPTREE-MEAN adapts to manifold dimension as well.

1 Introduction

The Curse of Dimensionality [2] has inspired research in several directions in Computer Science and has led to the development of several novel techniques such as dimensionality reduction, sketching etc. Almost all these techniques try to map data to lower dimensional spaces while approximately preserving useful information. However, most of these techniques do not assume anything about the data other than that they are imbedded in some high dimensional Euclidean space endowed with some distance/similarity function.

As it turns out, in many situations, the data is not simply scattered in the Euclidean space in a random fashion. Often, generative processes impose (non-linear) dependencies on the data that restrict the degrees of freedom available and result in the data having low intrinsic dimensionality. There exist several formalizations of this concept of intrinsic dimensionality. For example, [1] provides an excellent example of automated motion capture in which a large number of points on the body of an actor are sampled through markers and their coordinates transferred to an animated avatar. Now, although a large sample of points is required to ensure a faithful recovery of all the motions of the body (which causes each captured frame to lie in a very high dimensional space), these points are nevertheless constrained by the degrees of freedom offered by the human body which are very few.

Algorithms that try to exploit such non-linear structure in data have been studied extensively resulting in a large number of *Manifold Learning* algorithms for example [3, 4, 5]. These techniques typically assume knowledge about the manifold itself or the data distribution. For example, [4] and [5] require knowledge about the intrinsic dimensionality of the manifold whereas [3] requires a sampling of points that is "sufficiently" dense with respect to some manifold parameters.

Recently in [1], Dasgupta and Freund proposed space partitioning algorithms that adapt to the intrinsic dimensionality of data and do not assume explicit knowledge of this parameter. Their data structures are akin to the k-d tree structure and offer guaranteed reduction in the size of the cells after a bounded number of levels. Such a size reduction is of immense use in vector quantization [6] and regression [7]. Two such tree structures are presented in [1], each adapting to a different notion

---

[*]Work done as an undergraduate student at IIT Kanpur

of intrinsic dimensionality. Both variants have already found numerous applications in regression [7], spectral clustering [8], face recognition [9] and image super-resolution [10].

1.1 Contributions

The RPTREE structures are new entrants in a large family of space partitioning data structures such as k-d trees [11], BBD trees [12], BAR trees [13] and several others (see [14] for an overview). The typical guarantees given by these data structures are of the following types :

1. **Space Partitioning Guarantee** : There exists a bound $L(s), s \geq 2$ on the number of levels one has to go down before all descendants of a node of size Δ are of size Δ/s or less. The size of a cell is variously defined as the length of the longest side of the cell (for box-shaped cells), radius of the cell, etc.

2. **Bounded Aspect Ratio** : There exists a certain "roundedness" to the cells of the tree - this notion is variously defined as the ratio of the length of the longest to the shortest side of the cell (for box-shaped cells), the ratio of the radius of the smallest circumscribing ball of the cell to that of the largest ball that can be inscribed in the cell, etc.

3. **Packing Guarantee** : Given a fixed ball B of radius R and a size parameter r, there exists a bound on the number of disjoint cells of the tree that are of size greater than r and intersect B. Such bounds are usually arrived at by first proving a bound on the aspect ratio for cells of the tree.

These guarantees play a crucial role in algorithms for fast approximate nearest neighbor searches [12] and clustering [15]. We present new results for the RPTREE-MAX structure for all these types of guarantees. We first present a bound on the number of levels required for size reduction by any given factor in an RPTREE-MAX. Our result improves the bound obtainable from results presented in [1]. Next, we prove an "effective" aspect ratio bound for RPTREE-MAX. Given the randomized nature of the data structure it is difficult to directly bound the aspect ratios of all the cells. Instead we prove a weaker result that can nevertheless be exploited to give a packing lemma of the kind mentioned above. More specifically, given a ball B, we prove an aspect ratio bound for the smallest cell in the RPTREE-MAX that completely contains B.

Our final result concerns the RPTREE-MEAN data structure. The authors in [1] prove that this structure adapts to the *Local Covariance Dimension* of data (see Section 5 for a definition). By showing that low-dimensional manifolds have bounded local covariance dimension, we show its adaptability to the manifold dimension as well. Our result demonstrates the robustness of the notion of manifold dimension - a notion that is able to connect to a geometric notion of dimensionality such as the doubling dimension (proved in [1]) as well as a statistical notion such as Local Covariance Dimension (this paper).

1.2 Organization of the paper

In Section 2 we present a brief introduction to the RPTREE-MAX data structure and discuss its analysis. In Section 3 we present our generalized size reduction lemma for the RPTREE-MAX. In Section 4 we give an effective aspect ratio bound for the RPTREE-MAX which we then use to arrive at our packing lemma. In Section 5 we show that the RPTREE-MEAN adapts to manifold dimension.

All results cited from other papers are presented as *Facts* in this paper. We will denote by $B(x, r)$, a closed ball of radius r centered at x. We will denote by d, the intrinsic dimensionality of data and by D, the ambient dimensionality (typically $d \ll D$).

2 The RPTREE-MAX structure

The RPTREE-MAX structure adapts to the doubling dimension of data (see definition below). Since low-dimensional manifolds have low doubling dimension (see [1] Theorem 22) hence the structure adapts to manifold dimension as well.

Definition 1 (taken from [16]). *The doubling dimension of a set $S \subset \mathbb{R}^D$ is the smallest integer d such that for any ball $B(x, r) \subset \mathbb{R}^D$, the set $B(x, r) \cap S$ can be covered by 2^d balls of radius $r/2$.*

The RPTREE-MAX algorithm is presented data imbedded in $\mathbb{R}^D$ having doubling dimension d. The algorithm splits data lying in a cell C of radius Δ by first choosing a random direction $v \in \mathbb{R}^D$, projecting all the data inside C onto that direction, choosing a random value δ in the range $[-1, 1] \cdot 6\Delta/\sqrt{D}$ and then assigning a data point x to the left child if $x \cdot v < \text{median}(\{z \cdot v : z \in C\}) + \delta$ and the right child otherwise. Since it is difficult to get the exact value of the radius of a data set, the algorithm settles for a constant factor approximation to the value by choosing an arbitrary data point $x \in C$ and using the estimate $\tilde{\Delta} = \max(\{\|x - y\| : y \in C\})$.

The following result is proven in [1] :

Fact 2 (Theorem 3 in [1])**.** *There is a constant c_1 with the following property. Suppose an* RPTREE-MAX *is built using a data set $S \subset \mathbb{R}^D$. Pick any cell C in the* RPTREE-MAX*; suppose that $S \cap C$ has doubling dimension $\leq d$. Then with probability at least $1/2$ (over the randomization in constructing the subtree rooted at C), every descendant C' more than $c_1 d \log d$ levels below C has radius$(C') \leq$ radius$(C)/2$.*

In Sections 2, 3 and 4, we shall always assume that the data has doubling dimension d and shall not explicitly state this fact again and again. Let us consider extensions of this result to bound the number of levels it takes for the size of all descendants to go down by a factor $s > 2$. Let us analyze the case of $s = 4$. Starting off in a cell C of radius Δ, we are assured of a reduction in size by a factor of 2 after $c_1 d \log d$ levels. Hence all $2^{c_1 d \log d}$ nodes at this level have radius $\Delta/2$ or less. Now we expect that after $c_1 d \log d$ more levels, the size should go down further by a factor of 2 thereby giving us our desired result. However, given the large number of nodes at this level and the fact that the success probability in Fact 2 is just greater than a constant bounded away from 1, it is not possible to argue that after $c_1 d \log d$ more levels the descendants of all these $2^{c_1 d \log d}$ nodes will be of radius $\Delta/4$ or less. It turns out that this can be remedied by utilizing the following extension of the basic size reduction result in [1]. We omit the proof of this extension.

Fact 3 (Extension of Theorem 3 in [1])**.** *For any $\delta > 0$, with probability at least $1 - \delta$, every descendant C' which is more than $c_1 d \log d + \log(1/\delta)$ levels below C has radius$(C') \leq$ radius$(C)/2$.*

This gives us a way to boost the confidence and do the following : go down $L = c_1 d \log d + 2$ levels from C to get the the radius of all the $2^{c_1 d \log d + 2}$ descendants down to $\Delta/2$ with confidence $1 - 1/4$. Afterward, go an additional $L' = c_1 d \log d + L + 2$ levels from each of these descendants so that for any cell at level L, the probability of it having a descendant of radius $> \Delta/4$ after L' levels is less than $\frac{1}{4 \cdot 2^L}$. Hence conclude with confidence at least $1 - \frac{1}{4} - \frac{1}{4 \cdot 2^L} \cdot 2^L \geq \frac{1}{2}$ that all descendants of C after $2L + c_1 d \log d + 2$ have radius $\leq \Delta/4$. This gives a way to prove the following result :

Theorem 4. *There is a constant c_2 with the following property. For any $s \geq 2$, with probability at least $1 - 1/4$, every descendant C' which is more than $c_2 \cdot s \cdot d \log d$ levels below C has radius$(C') \leq$ radius$(C)/s$.*

Proof. Without loss of generality assume that s is a power of 2. We will prove the result by induction. Fact 3 proves the base case for $s = 2$. For the induction step, let $L(s)$ denote the number of levels it takes to reduce the size by a factor of s with high confidence. Then we have

$$L(s) \leq L(s/2) + c_1 d \log d + L(s/2) + 2 = 2L(s/2) + c_1 d \log d + 2$$

Solving the recurrence gives $L(s) = \mathcal{O}\left(sd \log d\right)$ $\qquad\square$

Notice that the dependence on the factor s is linear in the above result whereas one expects it to be logarithmic. Indeed, typical space partitioning algorithms such as k-d trees do give such guarantees. The first result we prove in the next section is a bound on the number of levels that is poly-logarithmic in the size reduction factor s.

3 A generalized size reduction lemma for RPTREE-MAX

In this section we prove the following theorem :

Theorem 5 (**Main**)**.** *There is a constant c_3 with the following property. Suppose an* RPTREE-MAX *is built using data set $S \subset \mathbb{R}^D$. Pick any cell C in the* RPTREE-MAX*; suppose that $S \cap C$ has doubling dimension $\leq d$. Then for any $s \geq 2$, with probability at least $1 - 1/4$ (over the*
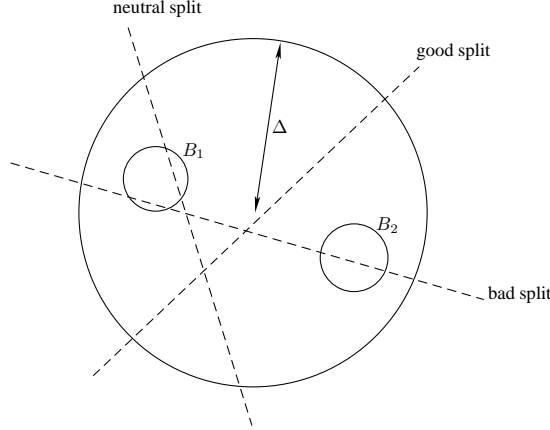
Figure 1: Balls B_1 and B_2 are of radius $\Delta/s\sqrt{d}$ and their centers are $\Delta/s - \Delta/s\sqrt{d}$ apart.

randomization in constructing the subtree rooted at C), for every descendant C' which is more than $c_3 \cdot \log s \cdot d \log sd$ levels below C, we have radius$(C') \leq$ radius$(C)/s$.

Compared to this, data structures such as [12] give deterministic guarantees for such a reduction in $D \log s$ levels which can be shown to be optimal (see [1] for an example). Thus our result is optimal but for a logarithmic factor. Moving on with the proof, let us consider a cell C of radius Δ in the RPTREE-MAX that contains a dataset S having doubling dimension $\leq d$. Then for any $\epsilon > 0$, a repeated application of Definition 1 shows that the S can be covered using at most $2^{d \log(1/\epsilon)}$ balls of radius $\epsilon\Delta$. We will cover $S \cap C$ using balls of radius $\frac{\Delta}{960s\sqrt{d}}$ so that $\mathcal{O}\left((sd)^d\right)$ balls would suffice. Now consider all pairs of these balls, the distance between whose centers is $\geq \frac{\Delta}{s} - \frac{\Delta}{960s\sqrt{d}}$. If random splits separate data from all such pairs of balls i.e. for no pair does any cell contain data from both balls of the pair, then each resulting cell would only contain data from pairs whose centers are closer than $\frac{\Delta}{s} - \frac{\Delta}{960s\sqrt{d}}$. Thus the radius of each such cell would be at most Δ/s.

We fix such a pair of balls calling them B_1 and B_2. A split in the RPTREE-MAX is said to be *good* with respect to this pair if it sends points inside B_1 to one child of the cell in the RPTREE-MAX and points inside B_2 to the other, *bad* if it sends points from both balls to both children and *neutral* otherwise (See Figure 1). We have the following properties of a random split :

Lemma 6. *Let $B = B(x, \delta)$ be a ball contained inside an* RPTREE-MAX *cell of radius Δ that contains a dataset S of doubling dimension d. Lets us say that a random split splits this ball if the split separates the data set S into two parts. Then a random split of the cell splits B with probability atmost $\frac{3\delta\sqrt{d}}{\Delta}$.*

Proof. The RPTREE-MAX splits proceed by randomly projecting the data in a cell onto the real line and then choosing a split point in an interval of length $12\Delta/\sqrt{D}$. It is important to note that the random direction and the split point are chosen independently. Hence, suppose data inside the ball B gets projected onto an interval $\tilde{B}$ of radius r, then the probability of it getting split is atmost $r\sqrt{D}/6\Delta$ since the split point is chosen randomly in an interval of length $12\Delta/\sqrt{D}$ independently of the projection. Let R_B be the random variable that gives the radius of the interval $\tilde{B}$. Hence the probability of B getting split is the following

$$\frac{\sqrt{D}}{6\Delta}\int_0^\infty r\mathbb{P}\left[R_B = r\right]dr = \frac{\sqrt{D}}{6\Delta}\int_0^\infty\int_0^r \mathbb{P}\left[R_B = r\right]dtdr = \frac{\sqrt{D}}{6\Delta}\int_0^\infty\int_t^\infty \mathbb{P}\left[R_B = r\right]drdt$$

$$= \frac{\sqrt{D}}{6\Delta}\int_0^\infty Pr[R_B \geq t]dt$$

We have the following result from [1]

4

Fact 7 (Lemma 6 of [1]). $\mathbb{P}\left[R_B \geq \frac{4\delta}{\sqrt{D}}\sqrt{2\left(d + \ln\frac{2}{\eta}\right)}\right] \leq \eta$

Fix the value $l = \frac{4\delta}{\sqrt{D}}\sqrt{2\left(d + \ln 2\right)}$. Using the fact that for any t, $Pr[R_B \geq t] \leq 1$ and making the change of variables $t = \frac{4\delta}{\sqrt{D}}\sqrt{2\left(d + \ln\frac{2}{\eta}\right)}$ we get

$$\int_0^\infty Pr[R_B \geq t]dt = \int_0^l Pr[R_B \geq t]dt + \int_l^\infty Pr[R_B \geq t]dt \leq \int_0^l 1 dt + \int_1^0 \eta dt(\eta)$$

Simplifying the above expression, we get the split probability to be atmost

$$\frac{2\delta}{3\Delta}\left[\sqrt{2\left(d + \ln 2\right)} + \int_0^1 \frac{d\eta}{\sqrt{2\left(d + \ln\frac{2}{\eta}\right)}}\right] = \frac{2\delta}{3\Delta}\left[\sqrt{2\left(d + \ln 2\right)} + 2\sqrt{2}e^d\int_{\sqrt{\ln 2 + d}}^\infty e^{-x^2}dx\right]$$

Now $\int_a^\infty e^{-x^2}dx = \frac{1}{2}\left[\int_{-\infty}^\infty e^{-x^2}dx - \int_{-a}^a e^{-x^2}dx\right] \leq \frac{\sqrt{\pi}}{2}\left[1 - \sqrt{1 - e^{-a^2}}\right] \leq \frac{\sqrt{\pi}}{2}e^{-a^2}$ since $1 - \sqrt{1 - x} < x$ for $0 < x < 1$. Using $d \geq 1$, we get the probability of the ball B getting split to be atmost $\frac{2\delta}{3\Delta}\left[\sqrt{2\left(d + \ln 2\right)} + \sqrt{\frac{\pi}{2}}\right] \leq \frac{3\delta\sqrt{d}}{\Delta}$. □

Lemma 8. *Let B_1 and B_2 be a pair of balls as described above contained in the cell C that contains data of doubling dimension d. Then a random split of the cell is a good split with respect to this pair with probability at least $\frac{1}{56s}$.*

Proof. The techniques used in the proof of this lemma are the same as those used to prove a similar result in [1]. We are giving a proof sketch here for completeness. We use the following two results from [1]

Fact 9 (Lemma 5 of [1]). *Fix any $x \in \mathbb{R}^D$. Pick a random vector $U \sim \mathcal{N}\left(0, (1/D)I_D\right)$. Then for any $\alpha, \beta > 0$:*

1. $\mathbb{P}\left[|U \cdot x| \leq \alpha \cdot \frac{\|x\|}{\sqrt{D}}\right] \leq \sqrt{\frac{2}{\pi}}\alpha,$

2. $\mathbb{P}\left[|U \cdot x| \geq \beta \cdot \frac{\|x\|}{\sqrt{D}}\right] \leq \frac{2}{\beta}e^{-\beta^2/2}.$

Fact 10 (Corollary 8 of [1]). *Suppose $S \subset \mathbb{R}^D$ lies within ball $B(x, \Delta)$. Pick any $0 < \delta < 2/e^2$. Let this set be projected randomly onto the real line. Let us denote by $\tilde{x}$, the projection of x by $\tilde{S}$, the projection of the set S. Then with probability atleast $1 - \delta$ over the choice of random projection onto $\mathbb{R}$, $\left|median\{\tilde{S}\} - \tilde{x}\right| \leq \frac{\Delta}{\sqrt{D}} \cdot \sqrt{2\ln\frac{2}{\delta}}$.*

Projections of points, sets etc. are denoted with a tilde ($\tilde{}$) sign. Applying Fact 7 with $\eta = \frac{2}{e^{31}}$, we get that with probability $> 1 - \frac{2}{e^{31}}$, the ball B_1 gets projected to an interval of length atmost $\frac{\Delta}{30s\sqrt{D}}$ centered at $\tilde{x}_1$. The same holds for B_2. Applying Fact 91 with $\alpha = \frac{384}{959}$ gives us $|\tilde{x}_1 - \tilde{x}_2| \geq \frac{\Delta}{2s\sqrt{D}}$ with probability $1 - \frac{1536}{4795}$. Furthermore, an application of Fact 92 with $\beta = \sqrt{2\ln 40}$ shows that with probability atleast $1 - \frac{1}{54}$, $|\tilde{x}_1 - \tilde{x}| \leq \frac{3\Delta}{\sqrt{D}}$. The same holds true for $\tilde{x}_2$ as well. Finally an application of Fact 10 with $\delta = \frac{1}{20}$ shows that the median of the projected set $\tilde{S}$ will lie within a distance $\frac{3\Delta}{\sqrt{D}}$ of $\tilde{x}$ (i.e. the projection of the center of the cell) with probability atleast $1 - \frac{1}{20}$.

Simple calculations show that the preceding guarantees imply that with probability atleast $\frac{1}{2}$ over the choice of random projections, the projections of both the balls will lie within the interval from which a split point would be chosen. Further more there would be a gap of atleast $\frac{\Delta}{2s\sqrt{D}} - 2\frac{\Delta}{30s\sqrt{D}}$ between

5

the projections of the two balls. Hence, given that these good events take place, with probability atleast $\frac{\sqrt{D}}{12\Delta}\left(\frac{\Delta}{2s\sqrt{D}} - 2\frac{\Delta}{30s\sqrt{D}}\right)$ over the choice of the split point, the balls will get cleanly separated. Note that this uses independence of the choice of projection and the choice of the split point. Thus the probability of a good split is atleast $\frac{1}{56s}$. $\qquad\square$

Lemma 11. *Let B_1 and B_2 be a pair of balls as described above contained in the cell C that contains data of doubling dimension d. Then a random split of the cell is a bad split with respect to this pair with probability at most $\frac{1}{320s}$.*

Proof. The proof of a similar result in [1] uses a conditional probability argument. However the technique does not work here since we require a bound that is inversely proportional to s. We instead make a simple observation that the probability of a bad split is upper bounded by the probability that one of the balls is split since for any two events A and B, $\mathbb{P}[A \cap B] \leq \min\{\mathbb{P}[A], \mathbb{P}[B]\}$. The result then follows from an application of Lemma 6. $\qquad\square$

We are now in a position to prove Theorem 5. What we will prove is that starting with a pair of balls in a cell C, the probability that some cell k levels below has data from both the balls is exponentially small in k. Thus, after going enough number of levels we can take a union bound over all pairs of balls whose centers are well separated (which are $\mathcal{O}\left((sd)^{2d}\right)$ in number) and conclude the proof.

Proof. (of Theorem 5) Consider a cell C of radius Δ in the RPTREE-MAX and fix a pair of balls contained inside C with radii $\Delta/960s\sqrt{d}$ and centers separated by at least $\Delta/s - \Delta/960s\sqrt{d}$. Let p_j^i denote the probability that a cell i levels below C has a descendant j levels below itself that contains data points from both the balls. Then the following holds :

Lemma 12. $p_k^0 \leq \left(1 - \frac{1}{68s}\right)^l p_{k-l}^l$.

Proof. We have the following expression for p_k^0 :

$$
\begin{aligned}
p_k^0 \quad \leq \quad & \mathbb{P}\left[\text{split at level 0 is a good split}\right] \cdot 0 + \\
& \mathbb{P}\left[\text{split at level 0 is a bad split}\right] \cdot 2p_{k-1}^1 + \\
& \mathbb{P}\left[\text{split at level 0 is a neutral split}\right] \cdot p_{k-1}^1 \\
\leq \quad & \frac{1}{320s} \cdot 2p_{k-1}^1 + \left(1 - \frac{1}{320s} - \frac{1}{56s}\right) \cdot p_{k-1}^1 \\
= \quad & \left(1 + \frac{1}{320s} - \frac{1}{56s}\right) \cdot p_{k-1}^1 \\
= \quad & \left(1 - \frac{1}{68s}\right) p_{k-1}^1 \\
\leq \quad & \left(1 - \frac{1}{68s}\right)^2 p_{k-2}^2 \qquad \left(\text{Similarly } p_{k-1}^1 \leq \left(1 - \frac{1}{68s}\right) p_{k-2}^2\right) \\
& \vdots \\
\leq \quad & \left(1 - \frac{1}{68s}\right)^l p_{k-l}^l
\end{aligned}
$$

$\qquad\square$

Note that this gives us $p_k^0 \leq \left(1 - \frac{1}{68s}\right)^k$ as a corollary. However using this result would require us to go down $k = \Omega(sd\log(sd))$ levels before $p_k^0 = \frac{1}{\Omega((sd)^{2d})}$ which results in a bound that is worse (by a factor logarithmic in s) than the one given by Theorem 4. This can be attributed to the small probability of a good split for a tiny pair of balls in large cells. However, here we are completely neglecting the fact that as we go down the levels, the radii of cells go down as well and good splits become more frequent.

Indeed setting $s = 2$ in Theorems 8 and 11 tells us that if the pair of balls were to be contained in a cell of radius $\frac{\Delta}{s/2}$ then the good and bad split probabilities are $\frac{1}{112}$ and $\frac{1}{640}$ respectively. This paves

way for an inductive argument : assume that with probability $> 1 - 1/4$, in $L(s)$ levels, the size of all descendants go down by a factor s. Denote by p_g^l the probability of a good split in a cell at depth l and by p_b^l the corresponding probability of a bad split. Set $l^* = L(s/2)$ and let E be the event that the radius of every cell at level l^* is less than $\frac{\Delta}{s/2}$. Let C' represent a cell at depth l^*. Then,

$$
\begin{aligned}
p_g^{l^*} &\geq \mathbb{P}\left[\text{good split in } C'|E\right] \cdot \mathbb{P}\left[E\right] \geq \frac{1}{112} \cdot \left(1 - \frac{1}{4}\right) \geq \frac{1}{150} \\
p_b^{l^*} &= \mathbb{P}\left[\text{bad split in } C'|E\right] \cdot \mathbb{P}\left[E\right] + \mathbb{P}\left[\text{bad split in } C'|\neg E\right] \cdot \mathbb{P}\left[\neg E\right] \\
&\leq \frac{1}{640} \cdot 1 + \frac{1}{640} \cdot \frac{1}{4} \leq \frac{1}{512}
\end{aligned}
$$

Notice that now, for any $m > 0$, we have $p_m^{l^*} \leq \left(1 - \frac{1}{213}\right)^m$. Thus, for some constant c_4, setting $k = l^* + c_4 d \log(sd)$ and applying Lemma 12 gives us $p_k^0 \leq \left(1 - \frac{1}{68s}\right)^{l^*} \left(1 - \frac{1}{213}\right)^{c_4 d \log(sd)} \leq \frac{1}{4(sd)^{2d}}$. Thus we have

$$
L(s) \leq L(s/2) + c_4 d \log(sd)
$$

which gives us the desired result on solving the recurrence i.e. $L(s) = \mathcal{O}\left(d \log s \log sd\right)$. $\qquad\square$

4 A packing lemma for RPTREE-MAX

In this section we prove a probabilistic packing lemma for RPTREE-MAX. A formal statement of the result follows :

Theorem 13 (Main). *Given any fixed ball $B(x, R) \subset \mathbb{R}^D$, with probability greater than $1/2$ (where the randomization is over the construction of the RPTREE-MAX), the number of disjoint RPTREE-MAX cells of radius greater than r that intersect B is at most $\left(\frac{R}{r}\right)^{\mathcal{O}(d \log d \log(dR/r))}$.*

Data structures such as BBD-trees give a bound of the form $\mathcal{O}\left(\frac{R}{r}\right)^D$ which behaves like $\left(\frac{R}{r}\right)^{\mathcal{O}(1)}$ for fixed D. In comparison, our result behaves like $\left(\frac{R}{r}\right)^{\mathcal{O}\left(\log \frac{R}{r}\right)}$ for fixed d. We will prove the result in two steps : first of all we will show that with high probability, the ball B will be completely inscribed in an RPTREE-MAX cell C of radius no more than $\mathcal{O}\left(Rd\sqrt{d} \log d\right)$. Thus the number of disjoint cells of radius at least r that intersect this ball is bounded by the number of descendants of C with this radius. To bound this number we then invoke Theorem 5 and conclude the proof.

4.1 An effective aspect ratio bound for RPTREE-MAX cells

In this section we prove an upper bound on the radius of the smallest RPTREE-MAX cell that completely contains a given ball B of radius R. Note that this effectively bounds the aspect ratio of this cell. Consider any cell C of radius Δ that contains B. We proceed with the proof by first showing that the probability that B will be split before it lands up in a cell of radius $\Delta/2$ is at most a quantity inversely proportional to Δ. Note that we are not interested in all descendants of C - only the ones ones that contain B. That is why we argue differently here. We consider balls of radius $\Delta/512\sqrt{d}$ surrounding B at a distance of $\Delta/2$ (see Figure 2). These balls are made to cover the annulus centered at B of mean radius $\Delta/2$ and thickness $\Delta/512\sqrt{d}$ – clearly $d^{\mathcal{O}(d)}$ balls suffice. Without loss of generality assume that the centers of all these balls lie in C.

Notice that if B gets separated from all these balls without getting split in the process then it will lie in a cell of radius $< \Delta/2$. Fix a B_i and call a random split of the RPTREE-MAX *useful* if it separates B from B_i and *useless* if it splits B. Using a proof technique similar to that used in Lemma 8 we can show that the probability of a useful split is at least $\frac{1}{192}$ whereas Lemma 6 tells us that the probability of a useless split is at most $\frac{3R\sqrt{d}}{\Delta}$.

Lemma 14. *There exists a constant c_5 such that the probability of a ball of radius R in a cell of radius Δ getting split before it lands up in a cell of radius $\Delta/2$ is at most $\frac{c_5 Rd\sqrt{d} \log d}{\Delta}$.*

Proof. The only bad event for us is the one in which B gets split before it gets separated from all the B_j's. Call this event E. Also, denote by $E[i]$ the bad event that B gets split for the first
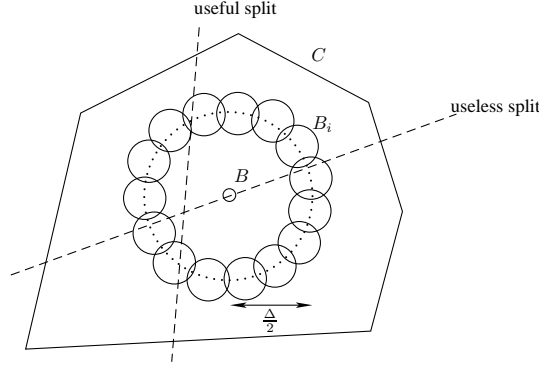
Figure 2: Balls B_i are of radius $\Delta/512\sqrt{d}$ and their centers are $\Delta/2$ far from the center of B.

time in the ith split and the preceding $i-1$ splits are incapable of separating B from all the B_j's. Thus $\mathbb{P}[E] \leq \sum_{i>0} \mathbb{P}[E[i]]$. Since any given split is a useful split (i.e. separates B from a fixed B_j) with probability $> \frac{1}{192}$, the probability that $i-1$ splits will fail to separate all B_js from the B (while not splitting B) is at most $\min\left\{1, \left(1-\frac{1}{192}\right)^{i-1} \cdot N\right\}$ where $N = d^{\mathcal{O}(d)}$ is the number of balls B_j. Since all splits in an RPTREE-MAX are independent of each other, we have $\mathbb{P}[E[i]] \leq \min\left\{1, \left(1-\frac{1}{192}\right)^{i-1} \cdot N\right\} \cdot \frac{3R\sqrt{d}}{\Delta}$. Let k be such that $\left(1-\frac{1}{192}\right)^{k-1} \leq \frac{1}{4N}$. Clearly $k = \mathcal{O}(d\log d)$ suffices. Thus we have

$$\mathbb{P}[E] \leq \frac{3R\sqrt{d}}{\Delta} \sum_{i>0} \min\left\{1, \left(1-\frac{1}{192}\right)^{i-1} \cdot N\right\} \leq \frac{3R\sqrt{d}}{\Delta}\left(\sum_{i=1}^{k} 1 + \sum_{i=1}^{\infty} \frac{1}{4}\left(1-\frac{1}{192}\right)^i\right)$$

which gives us $\mathbb{P}[E] = \mathcal{O}\left(\frac{Rd\sqrt{d}\log d}{\Delta}\right)$ since the second summation is just a constant. $\qquad\square$

We now state our result on the "effective" bound on aspect ratios of RPTREE-MAX cells.

Theorem 15. *There exists a constant c_6 such that with probability $> 1 - 1/4$, a given (fixed) ball B of radius R will be completely inscribed in an* RPTREE-MAX *cell C of radius no more than $c_6 \cdot Rd\sqrt{d}\log d$.*

Proof. Let $\Delta^* = 4c_5 Rd\sqrt{d}\log d$ and $\Delta_{\max}$ be the radius of the entire dataset. Denote by $F[i]$ the event that B ends up unsplit in a cell of radius $\frac{\Delta_{\max}}{2^i}$. The event we are interested in is $F[m]$ for $m = \log\frac{\Delta_{\max}}{\Delta^*}$. Note that $\mathbb{P}[F[m]|F[m-1]]$ is exactly $\mathbb{P}[E]$ where E is the event described in Lemma 14 for appropriately set value of radius Δ. Also $\mathbb{P}[F[m]|\neg F[m-1]] = 0$. Thus we have

$$\mathbb{P}[F[m]] = \prod_{i=0}^{m-1} \mathbb{P}[F[i+1]|F[i]] = \prod_{i=0}^{m-1}\left(1 - \frac{c_5 Rd\sqrt{d}\log d}{\Delta_{\max}/2^i}\right) \geq 1 - \sum_{i=0}^{m-1} \frac{c_5 Rd\sqrt{d}\log d}{\Delta_{\max}/2^i}$$

$$= 1 - \sum_{i=0}^{m-1} \frac{c_5 Rd\sqrt{d}\log d}{2^{m-i}\Delta^*} = 1 - \frac{1}{4}\sum_{i=0}^{m-1}\frac{1}{2^{m-i}} \geq 1 - \frac{1}{4}$$

Setting $c_6 = 4c_5$ gives us the desired result. $\qquad\square$

Proof. (of Theorem 13) Given a ball B of radius R, Theorem 15 shows that with probability at least $3/4$, B will lie in a cell C of radius at most $R' = \mathcal{O}\left(Rd\sqrt{d}\log d\right)$. Hence all cells of radius atleast r that intersect this ball must be either descendants or ancestors of C. Since we want
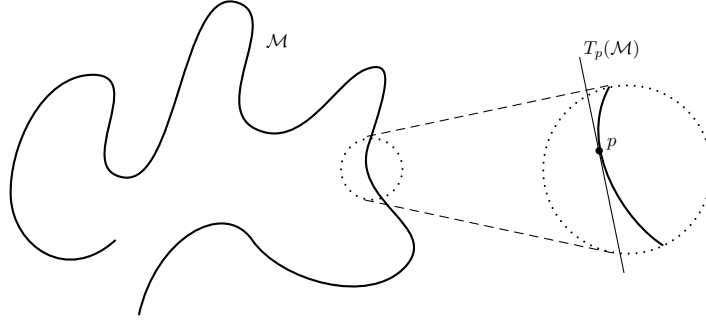
8

Figure 3: Locally, almost all the energy of the data is concentrated in the tangent plane.

an upper bound on the largest number of such disjoint cells, it suffices to count the number of descendants of C of radius no less than r. We know from Theorem 5 that with probability at least $3/4$ in $\log(R'/r)d\log(dR'/r)$ levels the radius of all cells must go below r. The result follows by observing that the RPTREE-MAX is a binary tree and hence the number of children can be at most $2^{\log(R'/r)d\log(dR'/r)}$. The success probability is at least $(3/4)^2 > 1/2$. $\qquad\square$

5 Local covariance dimension of a smooth manifold

The second variant of RPTREE, namely RPTREE-MEAN, adapts to the local covariance dimension (see definition below) of data. We do not go into the details of the guarantees presented in [1] due to lack of space. Informally, the guarantee is of the following kind : given data that has small local covariance dimension, on expectation, a data point in a cell of radius r in the RPTREE-MEAN will be contained in a cell of radius $c_7 \cdot r$ in the next level for some constant $c_7 < 1$. The randomization is over the construction of RPTREE-MEAN as well as choice of the data point. This gives per-level improvement albeit in expectation whereas RPTREE-MAX gives improvement in the worst case but after a certain number of levels.

We will prove that a d-dimensional Riemannian submanifold $\mathcal{M}$ of $\mathbb{R}^D$ has bounded local covariance dimension thus proving that RPTREE-MEAN adapts to manifold dimension as well.

Definition 16. *A set $S \subset \mathbb{R}^D$ has local covariance dimension (d, ϵ, r) if there exists an isometry M of $\mathbb{R}^D$ under which the set S when restricted to any ball of radius r has a covariance matrix for which some d diagonal elements contribute a $(1 - \epsilon)$ fraction of its trace.*

This is a more general definition than the one presented in [1] which expects the top d eigenvalues of the covariance matrix to account for a $(1 - \epsilon)$ fraction of its trace. However, all that [1] requires for the guarantees of RPTREE-MEAN to hold is that there exist d orthonormal directions such that a $(1 - \epsilon)$ fraction of the energy of the dataset i.e. $\sum_{x \in S} \|x - mean(S)\|^2$ is contained in those d dimensions. This is trivially true when $\mathcal{M}$ is a d-dimensional affine set. However we also expect that for small neighborhoods on smooth manifolds, most of the energy would be concentrated in the tangent plane at a point in that neighborhood (see Figure 3). Indeed, we can show the following :

Theorem 17 (Main). *Given a data set $S \subset \mathcal{M}$ where $\mathcal{M}$ is a d-dimensional Riemannian manifold with condition number τ, then for any $\epsilon \leq \frac{1}{4}$, S has local covariance dimension $\left(d, \epsilon, \frac{\sqrt{\epsilon}\tau}{3}\right)$.*

For manifolds, the local curvature decides how small a neighborhood should one take in order to expect a sense of "flatness" in the non-linear surface. This is quantified using the *Condition Number* τ of $\mathcal{M}$ (introduced in [17]) which restricts the amount by which the manifold can curve locally. The condition number is related to more prevalent notions of local curvature such as the second fundamental form [18] in that the inverse of the condition number upper bounds the norm of the second fundamental form [17]. Informally, if we restrict ourselves to regions of the manifold of radius τ or less, then we get the requisite flatness properties. This is formalized in [17] as follows. For any hyperplane $T \subset \mathbb{R}^D$ and a vector $v \in \mathbb{R}^d$, let $v_{\parallel}(T)$ denote the projection of v onto T.

9

Fact 18 (Implicit in Lemma 5.3 of [17]). *Suppose $\mathcal{M}$ is a Riemannian manifold with condition number τ. For any $p \in \mathcal{M}$ and $r \leq \sqrt{\epsilon}\tau, \epsilon \leq \frac{1}{4}$, let $\mathcal{M}' = B(p,r) \cap \mathcal{M}$. Let $T = T_p(\mathcal{M})$ be the tangent space at p. Then for any $x, y \in \mathcal{M}'$, $\|x_\|(T) - y_\|(T)\|^2 \geq (1 - \epsilon)\|x - y\|^2$.*

This already seems to give us what we want - a large fraction of the length between any two points on the manifold lies in the tangent plane - i.e. in d dimensions. However in our case we have to show that for some d-dimensional plane P, $\sum_{x \in S} \|(x - \mu)_\|(P)\|^2 > (1 - \epsilon) \sum_{x \in S} \|x - \mu\|^2$ where $\mu = mean(S)$. The problem is that we cannot apply Fact 18 since there is no surety that the mean will lie on the manifold itself. However it turns out that certain points on the manifold can act as "proxies" for the mean and provide a workaround to the problem.

Proof. (of Theorem 17) Suppose $\mathcal{M}' = B(x_0, r) \cap \mathcal{M}$ for $r = \frac{\sqrt{\epsilon}\tau}{3}$ and we are given data points $S = \{x_1, \ldots x_n\} \subset \mathcal{M}'$. Let $q = \arg\min_{x \in \mathcal{M}} \|\mu - x\|$ be the closest point on the manifold to the mean. The smoothness properties of $\mathcal{M}$ tell us that the vector $(\mu - q)$ is perpendicular to $T_q(\mathcal{M})$, the d-dimensional tangent space at q (in fact any point q at which the function $g : x \in \mathcal{M} \longmapsto \|x - \mu\|$ attains a local extrema would also have the same property). This has interesting consequences - let f be the projection map onto $T_q(\mathcal{M})$ i.e. $f(v) = v_\|(T_q(\mathcal{M}))$.

Then $f(\mu - q) = 0$ since $(\mu - q) \perp T_q(\mathcal{M})$. This implies that for any vector $v \in \mathbb{R}^D$, $f(v - \mu) = f(v - q) + f(q - \mu) = f(v - q) = f(v) - f(q)$ since f is a linear map. We now note that $\min_i \|\mu - x_i\| \leq r$. If this were not true then we would have $\sum_i \|\mu - x_i\| > nr^2$ whereas we know that $\sum_i \|\mu - x_i\| \leq \sum_i \|x_0 - x_i\| \leq nr^2$ since for any random variable $X \in \mathbb{R}^D$ and fixed $v \in \mathbb{R}^D$, we have $\mathbb{E}\left[\|X - v\|^2\right] \geq \mathbb{E}\left[\|X - \mathbb{E}[X]\|^2\right]$. Since $\|\mu - x_i\| \leq r$ for some $x_i \in \mathcal{M}$, we know, by definition of q, that $\|\mu - q\| \leq r$ as well.

We also have $\|\mu - x_0\| \leq r$ (since the convex hull of the points is contained in the ball B and the mean, being a convex combination of the points, is contained in the hull) and $\|x_i - x_0\| \leq r$ for all points x_i. Hence we have for any point x_i, $\|x_i - q\| \leq \|x_i - x_0\| + \|x_0 - \mu\| + \|\mu - q\| \leq 3r$ and conclude that $S \subset B(q, 3r) \cap \mathcal{M} = B(q, \sqrt{\epsilon}\tau) \cap \mathcal{M}$ which means we can apply Fact 18 between the vectors x_i and q.

Let $T = T_q(\mathcal{M})$ and q as chosen above. We have

$$\sum_{x \in S} \|(x - \mu)_\|(T)\|^2 = \sum_{x \in S} \|f(x - \mu)\|^2 = \sum_{x \in S} \|f(x - q)\|^2 = \sum_{x \in S} \|f(x) - f(q)\|^2$$

$$\geq \sum_{x \in S} (1 - \epsilon)\|x - q\|^2 \geq (1 - \epsilon) \sum_{x \in S} \|x - \mu\|^2$$

where the last inequality again uses the fact that for a random variable $X \in \mathbb{R}^D$ and fixed $v \in \mathbb{R}^D$, $\mathbb{E}\left[\|X - v\|^2\right] \geq \mathbb{E}\left[\|X - \mathbb{E}[X]\|^2\right]$. □

6 Conclusion

In this paper we considered the two random projection trees proposed in [1]. For the RPTREE-MAX data structure, we provided an improved bound (Theorem 5) on the number of levels required to decrease the size of the tree cells by any factor $s \geq 2$. However the bound we proved is poly-logarithmic in s. It would be nice if this can be brought down to logarithmic since it would directly improve the packing lemma (Theorem 13) as well. More specifically the packing bound would become $\left(\frac{R}{r}\right)^{\mathcal{O}(1)}$ instead of $\left(\frac{R}{r}\right)^{\mathcal{O}\left(\log \frac{R}{r}\right)}$ for fixed d.

As far as dependence on d is concerned, there is room for improvement in the packing lemma. We have shown that the smallest cell in the RPTREE-MAX that completely contains a fixed ball B of radius R has an aspect ratio no more than $\mathcal{O}\left(d\sqrt{d}\log d\right)$ since it has a ball of radius R inscribed in it and can be circumscribed by a ball of radius no more than $\mathcal{O}\left(Rd\sqrt{d}\log d\right)$. Any improvement in the aspect ratio of the smallest cell that contains a given ball will also directly improve the packing lemma.

Moving on to our results for the RPTREE-MEAN, we demonstrated that it adapts to manifold dimension as well. However the constants involved in our guarantee are pessimistic. For instance, the radius parameter in the local covariance dimension is given as $\frac{\sqrt{\epsilon\tau}}{3}$ - this can be improved to $\frac{\sqrt{\epsilon\tau}}{2}$ if one can show that there will always exists a point $q \in B(x_0, r) \cap \mathcal{M}$ at which the function $g : x \in \mathcal{M} \longmapsto \|x - \mu\|$ attains a local extrema.

We conclude with a word on the applications of our results. As we already mentioned, packing lemmas and size reduction guarantees for arbitrary factors are typically used in applications for nearest neighbor searching and clustering. However, these applications (viz [12], [15]) also require that the tree have bounded depth. The RPTREE-MAX is a pure space partitioning data structure that can be coerced by an adversarial placement of points into being a primarily left-deep or right-deep tree having depth $\Omega(n)$ where n is the number of data points.

Existing data structures such as BBD Trees remedy this by alternating space partitioning splits with data partitioning splits. Thus every alternate split is forced to send at most a constant fraction of the points into any of the children thus ensuring a depth that is logarithmic in the number of data points. A similar technique is used by [7] to bound the depth of the version of RPTREE-MAX used in that paper. However it remains to be seen if the same trick can be used to bound the depth of RPTREE-MAX while maintaining the packing guarantees because although such "space partitioning" splits do not seem to hinder Theorem 5, they do hinder Theorem 13 (more specifically they hinder Theorem 14).

We leave open the question of a possible augmentation of the RPTREE-MAX structure, or a better analysis, that can simultaneously give the following guarantees :

1. **Bounded Depth** : depth of the tree should be $o(n)$, preferably $(\log n)^{\mathcal{O}(1)}$

2. **Packing Guarantee** : of the form $\left(\frac{R}{r}\right)^{\left(d \log \frac{R}{r}\right)^{\mathcal{O}(1)}}$

3. **Space Partitioning Guarantee** : assured size reduction by factor s in $(d \log s)^{\mathcal{O}(1)}$ levels

Acknowledgments

References

[1] Sanjoy Dasgupta and Yoav Freund. Random Projection Trees and Low dimensional Manifolds. In *40th Annual ACM Symposium on Theory of Computing*, pages 537–546, 2008.

[2] Piotr Indyk and Rajeev Motwani. Approximate Nearest Neighbors : Towards Removing the Curse of Dimensionality. In *30th Annual ACM Symposium on Theory of Computing*, pages 604–613, 1998.

[3] Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(22):2319–2323, 2000.

[4] Piotr Indyk and Assaf Naor. Nearest-Neighbor-Preserving Embeddings. *ACM Transactions on Algorithms*, 3, 2007.

[5] Richard G. Baraniuk and Michael B. Wakin. Random Projections of Smooth Manifolds. *Foundations of Computational Mathematics*, 9(1):51–77, 2009.

[6] Yoav Freund, Sanjoy Dasgupta, Mayank Kabra, and Nakul Verma. Learning the structure of manifolds using random projections. In *Twenty-First Annual Conference on Neural Information Processing Systems*, 2007.

[7] Samory Kpotufe. Escaping the curse of dimensionality with a tree-based regressor. In *22nd Annual Conference on Learning Theory*, 2009.

[8] Donghui Yan, Ling Huang, and Michael I. Jordan. Fast Approximate Spectral Clustering. In *15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 907–916, 2009.

[9] John Wright and Gang Hua. Implicit Elastic Matching with Random Projections for Pose-Variant Face Recognition. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1502–1509, 2009.

[10] Jian Pu, Junping Zhang, Peihong Guo, and Xiaoru Yuan. Interactive Super-Resolution through Neighbor Embedding. In *9th Asian Conference on Computer Vision*, pages 496–505, 2009.

[11] Jon Louis Bentley. Multidimensional Binary Search Trees Used for Associative Searching. *Communications of the ACM*, 18(9):509–517, 1975.

[12] Sunil Arya, David M. Mount, Nathan S. Netanyahu, Ruth Silverman, and Angela Y. Wu. An Optimal Algorithm for Approximate Nearest Neighbor Searching Fixed Dimensions. *Journal of the ACM*, 45(6):891–923, 1998.

[13] Christian A. Duncan, Michael T. Goodrich, and Stephen G. Kobourov. Balanced Aspect Ratio Trees: Combining the Advantages of k-d Trees and Octrees. *Journal of Algorithms*, 38(1):303–333, 2001.

[14] Hanan Samet. *Foundations of Multidimensional and Metric Data Structures*. Morgan Kaufmann Publishers, 2005.

[15] Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman, and Angela Y. Wu. A local search approximation algorithm for k-means clustering. *Computational Geometry*, 28(2-3):89–112, 2004.

[16] Patrice Assouad. Plongements Lipschitziens dans $\mathbb{R}^n$. *Bulletin Société Mathématique de France*, 111:429–448, 1983.

[17] Partha Niyogi, Stephen Smale, and Shmuel Weinberger. Finding the Homology of Submanifolds with High Confidence from Random Samples. *Discrete & Computational Geometry*, 39(1-3):419–441, 2008.

[18] Sebastián Montiel and Antonio Ros. *Curves and Surfaces*, volume 69 of *Graduate Studies in Mathematics*. American Mathematical Society and Real Sociedad Matemática Epañola, 2005.