# Random Projection Trees Revisited
# Supplementary Material

**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

This document contains detailed proofs of theorems stated in the main paper entitled *Random Projection Trees Revisited*.

## 1 Proof of Theorem 4

**Theorem 1** (Theorem 4 restated)**.** *There is a constant $c_2$ 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 randomization in constructing the subtree rooted at $C$), for every descendant $C'$ which is more than $c_2 \cdot s \cdot d \log d$ levels below $C$, we have 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. Recall the following result.

**Fact 2** (Implicit in Theorem 3 in [1])**.** *There is a constant $c_1$ 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 $\delta > 0$, with probability at least $1 - \delta$ (over the randomization in constructing the subtree rooted at $C$), for every descendant $C'$ which is more than $c_1 d \log d + \log(1/\delta)$ levels below $C$, we have radius$(C') \leq$ radius$(C)/2$.*

Fact 2 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 $L(s) = \mathcal{O}\left(sd \log d\right)$ ☐

## 2 Proof of Lemma 6

**Lemma 3** (Lemma 6 restated)**.** *Let $B = B(x, \delta)$ be a ball contained inside an* RPTREE *cell of radius $\Delta$ 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

---

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

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

$$\begin{aligned}\frac{\sqrt{D}}{6\Delta}\int\limits_0^\infty r\mathbb{P}\left[R_B=r\right]dr &= \frac{\sqrt{D}}{6\Delta}\int\limits_0^\infty\int\limits_0^r \mathbb{P}\left[R_B=r\right]dtdr=\frac{\sqrt{D}}{6\Delta}\int\limits_0^\infty\int\limits_t^\infty \mathbb{P}\left[R_B=r\right]drdt\\ &= \frac{\sqrt{D}}{6\Delta}\int\limits_0^\infty Pr[R_B\geq t]dt\end{aligned}$$

We have the following result from [1]

**Fact 4** (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

$$\int\limits_0^\infty Pr[R_B\geq t]dt=\int\limits_0^l Pr[R_B\geq t]dt+\int\limits_l^\infty Pr[R_B\geq t]dt\leq\int\limits_0^l 1dt+\int\limits_1^0 \eta dt(\eta)$$

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

$$\frac{2\delta}{3\Delta}\left[\sqrt{2\left(d+\ln 2\right)}+\int\limits_0^1\frac{d\eta}{\sqrt{2\left(d+\ln\frac{2}{\eta}\right)}}\right]=\frac{2\delta}{3\Delta}\left[\sqrt{2\left(d+\ln 2\right)}+2\sqrt{2}e^d\int\limits_{\sqrt{\ln 2+d}}^\infty e^{-x^2}dx\right]$$

Now $\int\limits_a^\infty e^{-x^2}dx=\frac{1}{2}\left[\int\limits_{-\infty}^\infty e^{-x^2}dx-\int\limits_{-a}^a e^{-x^2}dx\right]\leq\frac{\sqrt{\pi}}{2}\left[1-\sqrt{1-e^{-a^2}}\right]\leq\frac{\sqrt{\pi}}{2}e^{-a^2}$ 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 $\frac{2\delta}{3\Delta}\left[\sqrt{2\left(d+\ln 2\right)}+\sqrt{\frac{\pi}{2}}\right]\leq\frac{3\delta\sqrt{d}}{\Delta}$. ☐

## 3 Proof of Lemma 7

**Lemma 5** (Lemma 7 restated)**.** *Let $B_1(x_1,\Delta/960s\sqrt{d})$ and $B_2(x_2,\Delta/960s\sqrt{d})$ be a pair of balls with the centers separated by atleast $\Delta/s-\Delta/960s\sqrt{d}$. Suppose these balls are contained in a ball $B(x,\Delta)$ containing data $S$ of doubling dimension $d$. Then a random split of the cell is a good split with respect to this pair with probability atleast $\frac{1}{56s}$.*

*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 6** (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 $\alpha,\beta>0$ :*

*(a)* $\mathbb{P}\left[|U\cdot x|\leq\alpha\cdot\frac{\|x\|}{\sqrt{D}}\right]\leq\sqrt{\frac{2}{\pi}}\alpha$,

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

**Fact 7** (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 $\mathbb{R}$, $\left|\text{median}\{\tilde{S}\} - \tilde{x}\right| \leq \frac{\Delta}{\sqrt{D}} \cdot \sqrt{2 \ln \frac{2}{\delta}}$.*

Projections of points, sets etc. are denoted with a tilde ( $\tilde{}$ ) sign. Applying Fact 4 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 6(a) 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 6(b) 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 7 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 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 $\frac{1}{56s}$. ☐

## 4 Proof of Lemma 9

**Lemma 8** (Lemma 9 restated). *Consider a cell $C$ of radius $\Delta$ in the* RPTREE-MAX *containing data of doubling dimension $d$ and fix a pair of balls $B_1(x_1, \Delta/960s\sqrt{d})$ and $B_2(x_2, \Delta/960s\sqrt{d})$ with the centers separated by atleast $\Delta/s - \Delta/960s\sqrt{d}$. 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 $p_k^0 \leq \left(1 - \frac{1}{68s}\right)^l p_{k-l}^l$.*

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

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

☐

## 5 Proof of Lemma 11

**Lemma 9** (Lemma 11 restated)**.** *There exists a constant $c_5$ such that the probability of a ball of radius $R$ in a cell of radius $\Delta$ getting split before it lands up in a cell of radius $\Delta/2$ is at most $\frac{c_5 R d\sqrt{d}\log d}{\Delta}$.*

*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 time in the $i^{\text{th}}$ split and the preceding $i-1$ splits are incapable of separating $B$ from all the $B_j$'s. Thus $\mathbb{P}\left[E\right] \leq \sum\limits_{i>0} \mathbb{P}\left[E[i]\right]$. 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 $B_j$s 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 $\mathbb{P}\left[E[i]\right] \leq \min\left\{1, \left(1-\frac{1}{192}\right)^{i-1}\cdot N\right\}\cdot\frac{3R\sqrt{d}}{\Delta}$. Let $k$ be such that $\left(1-\frac{1}{192}\right)^{k-1} \leq \frac{1}{4N}$. Clearly $k = \mathcal{O}\left(d\log d\right)$ suffices. Thus we have

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

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

## 6 Proof of Theorem 12

**Theorem 10** (Theorem 12 restated)**.** *There exists a constant $c_6$ such that with probability $> 1-1/4$, a given 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 $\mathbb{P}\left[F[m]|F[m-1]\right]$ is exactly $\mathbb{P}\left[E\right]$ where $E$ is the event described in Lemma 11 for appropriately set value of radius $\Delta$. Also $\mathbb{P}\left[F[m]|\neg F[m-1]\right] = 0$. Thus we have

$$\begin{aligned}\mathbb{P}\left[F[m]\right] &= \prod_{i=0}^{m-1}\mathbb{P}\left[F[i+1]|\,F[i]\right] = \prod_{i=0}^{m-1}\left(1-\frac{c_5 Rd\sqrt{d}\log d}{\Delta_{\max}/2^i}\right) \geq 1-\sum_{i=0}^{m-1}\frac{c_5 Rd\sqrt{d}\log d}{\Delta_{\max}/2^i}\\ &= 1-\sum_{i=0}^{m-1}\frac{c_5 Rd\sqrt{d}\log d}{2^{m-i}\Delta^*} = 1-\frac{1}{4}\sum_{i=0}^{m-1}\frac{1}{2^{m-i}} \geq 1-\frac{1}{4}\end{aligned}$$

Setting $c_6 = 4c_5$ gives us the desired result. □

## 7 Proof of Theorem 14

Let us first recall a result about smooth manifolds being used to prove this result.

**Fact 11** (Implicit in Lemma 5.3 of [2])**.** *Suppose $\mathcal{M}$ is a Riemannian manifold with condition number $\tau$. 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 $x, y\in\mathcal{M}'$, $\|x_{||}(T) - y_{||}(T)\|^2 \geq (1-\epsilon)\|x-y\|^2$.*

We will now prove the following result

**Theorem 12** (Theorem 14 restated)**.** *Given a data set $S\subset\mathcal{M}$ where $\mathcal{M}$ is a $d$-dimensional Riemannian manifold with condition number $\tau$, then for any $\epsilon \leq \frac{1}{4}$, $S$ has local covariance dimension $\left(d, \epsilon, \frac{\sqrt{\epsilon}\tau}{3}\right)$.*

*Proof.* 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 $q = \underset{x \in \mathcal{M}}{\arg\min} \|\mu - x\|$ 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 11 between the vectors $x_i$ and $q$.

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

$$\begin{aligned}
\sum_{x \in S} \|(x - \mu)_{\|}(T)\|^2 &= \sum_{x \in S} \|f(x - \mu)\|^2 = \sum_{x \in S} \|f(x - q)\|^2 = \sum_{x \in S} \|f(x) - f(q)\|^2 \\
&\geq \sum_{x \in S} (1 - \epsilon)\|x - q\|^2 \geq (1 - \epsilon) \sum_{x \in S} \|x - \mu\|^2
\end{aligned}$$

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

## 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] 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.