Random projection trees and low dimensional manifolds

Sanjoy Dasgupta and Yoav Freund [1]

Abstract

We present a simple variant of the k-d tree which automatically adapts to intrinsic low dimensional structure in data without having to explicitly learn this structure.

1 Introduction

The k-d tree (Bentley, 1975) is a spatial data structure that partitions $\mathbb{R}^D$ into hyperrectangular cells. It is built in a recursive manner, splitting along one coordinate direction at a time (Figure 1, left). The succession of splits corresponds to a binary tree whose leaves contain the individual cells in $\mathbb{R}^D$.

These trees are among the most widely-used spatial partitionings in machine learning and statistics. To understand their application, consider Figure 1(left), for instance, and suppose that the dots are points in a database, while the cross is a query point q. The cell containing q, which we will henceforth denote cell(q), can quickly be identified by moving q down the tree. If the diameter of cell(q) is small (where the diameter is taken to mean the distance between the furthest pair of data points in the cell), then the points in it can be expected to have similar properties, for instance similar labels. Point q can then be classified according to the majority vote of the labels in its cell, or the label of its nearest neighbor in the cell. The statistical theory around k-d trees is thus centered on *the rate at which the diameter of individual cells drops as you move down the tree*; for details, see for instance Chapter 20 of the textbook by Devroye, Gyorfi, and Lugosi (1996).

It is an empirical observation that the usefulness of k-d trees diminishes as the dimension D increases. This is easy to explain in terms of cell diameter; specifically, we will show that:

> There is a data set in $\mathbb{R}^D$ for which a k-d tree requires D levels in order to halve the cell diameter.

In other words, if the data lie in $\mathbb{R}^{1000}$, then it could take 1000 levels of the tree to bring the diameter of cells down to half that of the entire data set. This would require 2^{1000} data points!

Thus k-d trees are susceptible to the same curse of dimensionality that has been the bane of other nonparametric statistical methods. However, a recent positive development in statistics and machine learning has been the realization that a lot of data which superficially lies in a very high-dimensional space $\mathbb{R}^D$, actually has low *intrinsic* dimension, in the sense of lying close to a manifold of dimension $d \ll D$. There has thus been a huge interest in algorithms which learn this manifold from data, with the intention that future data can then be transformed into this low-dimensional space, in which the usual nonparametric (and other) methods will work well. This field is quite recent and yet the literature on it is already voluminous; some

[1] Email: dasgupta,yfreund@cs.ucsd.edu

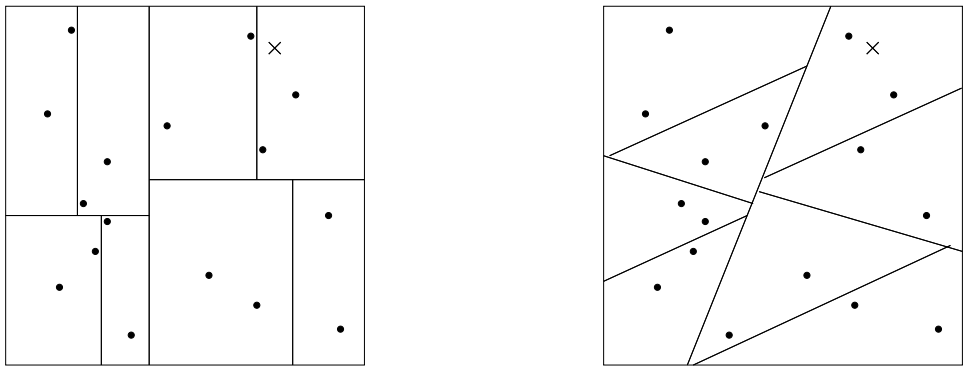


Figure 1: Left: A spatial partitioning of $\mathbb{R}^2$ induced by a k-d tree with three levels. The first split is along the x-axis, the second along the y-axis, and the third along the x-axis. The dots are data points; the cross marks a query point q. Right: Partitioning induced by an RP tree.

early foundational work includes that of Tenenbaum *et al* (2000), Roweis and Saul (2000) and Belkin and Niyogi (2003).

In this paper, we are interested in techniques that automatically adapt to intrinsic low dimensional structure without having to explicitly learn this structure. The most obvious first question is, do k-d trees adapt to intrinsic low dimension? The answer is no: in fact the bad example mentioned above has an intrinsic dimension of just $O(\log D)$. But we show that a simple variant of k-d trees does, in fact, possess this property. Instead of splitting along coordinate directions at the median, we do the following: we split along a random direction in S^{D-1} (the unit sphere in $\mathbb{R}^D$), and instead of splitting exactly at the median, we introduce a small amount of "jitter". We call these *random projection trees* (Figure 1, right), or RP trees for short, and we can show the following.

> Pick any cell C in the RP tree. If the data in C have intrinsic dimension d, then all descendant cells $\geq d \log d$ levels below will have at most half the diameter of C.

There is no dependence at all on the extrinsic dimensionality (D) of the data.

2 Detailed overview

In what follows, we will always assume the data lie in $\mathbb{R}^D$.

2.1 k-d trees and RP trees

Both k-d and random projection (RP) trees are built by recursive binary splits. They differ only in the nature of the split, which we define in a subroutine called CHOOSERULE. The core tree-building algorithm is called MAKETREE, and takes as input a data set $S \subset \mathbb{R}^D$.

procedure MAKETREE(S)
 if $|S| < MinSize$
  then return $(Leaf)$
  else
$Rule \leftarrow$ CHOOSERULE(S)
$LeftTree \leftarrow$ MAKETREE($\{x \in S : Rule(x) = \text{true}\}$)
$RightTree \leftarrow$ MAKETREE($\{x \in S : Rule(x) = \text{false}\}$)
return $([Rule, LeftTree, RightTree])$

The k-d tree CHOOSERULE picks a coordinate direction (at random, or by cycling through the coordinates in a fixed order) and then splits the data on its median value for that coordinate.

procedure CHOOSERULE(S)
 comment: k-d tree version

 choose a random coordinate direction i
 $Rule(x) := x_i \leq \text{median}(\{z_i : z \in S\})$
 return $(Rule)$

On the other hand, an RPTree chooses a direction uniformly at random from the unit sphere S^{D-1} and splits the data into two roughly equal-sized sets using a hyperplane orthogonal to this direction. We describe two variants, which we call RPTree-Max and RPTree-Mean. The first is more theoretically motivated (in the sense that we can prove a slightly stronger type of bound for it) while the second is more practically motivated. They are both adaptive to intrinsic dimension, although the proofs are in different models and use very different techniques.

We start with the CHOOSERULE for RPTree-Max.

procedure CHOOSERULE(S)
 comment: RPTree-Max version

 choose a random unit direction $v \in \mathbb{R}^D$
 pick any point $x \in S$, and let y be the farthest point from it in S
 choose δ uniformly at random in $[-1, 1] \cdot 6\|x - y\|/\sqrt{D}$
 $Rule(x) := x \cdot v \leq (\text{median}(\{z \cdot v : z \in S\}) + \delta)$
 return $(Rule)$

(In this paper, $\|\cdot\|$ always denotes Euclidean distance.) A tree of this kind, with boundaries that are arbitrary hyperplanes, is generically called a binary space partition (BSP) tree (Fuchs, Kedem, and Naylor, 1980). Our particular variant is built using two kinds of randomness, in the split directions as well in the perturbations. Both are crucial for the bounds we give.

The RPTree-Mean is similar to RPTree-Max, but differs in two critical respects. First, it occasionally performs a different kind of split: a cell is split into two pieces based on distance from the mean. Second, when splitting according to a random direction, the actual split location is optimized more carefully.

procedure CHOOSERULE(S)
comment: RPTree-Mean version

if $\Delta^2(S) \leq c \cdot \Delta_A^2(S)$

then
- choose a random unit direction v
- sort projection values: $a(x) = v \cdot x \ \ \forall x \in S$, generating the list $a_1 \leq a_2 \leq \cdots \leq a_n$
- **for** $i = 1, \ldots, n-1$ compute
  - $\mu_1 = \frac{1}{i}\sum_{j=1}^{i} a_j,\ \mu_2 = \frac{1}{n-i}\sum_{j=i+1}^{n} a_j$
  - $c_i = \sum_{j=1}^{i}(a_j - \mu_1)^2 + \sum_{j=i+1}^{n}(a_j - \mu_2)^2$
- find i that minimizes c_i and set $\theta = (a_i + a_{i+1})/2$
- $Rule(x) := v \cdot x \leq \theta$

else $\{Rule(x) := \|x - \text{mean}(S)\| \leq \text{median}\{\|z - \text{mean}(S)\| : z \in S\}$

return $(Rule)$

In the code, c is a constant, $\Delta(S)$ is the diameter of S (the distance between the two furthest points in the set), and $\Delta_A(S)$ is the *average* diameter, that is, the average distance between points of S:

$$\Delta_A^2(S) = \frac{1}{|S|^2}\sum_{x,y \in S} \|x - y\|^2.$$

2.2 Low-dimensional manifolds

The increasing ubiquity of massive, high-dimensional data sets has focused the attention of the statistics and machine learning communities on the curse of dimensionality. A large part of this effort is based on exploiting the observation that many high-dimensional data sets have low *intrinsic dimension*. This is a loosely defined notion, which is typically used to mean that the data lie near a smooth low-dimensional manifold.

For instance, suppose that you wish to create realistic animations by collecting human motion data and then fitting models to it. You could put a large number of sensors on a human being, and then measure the three-dimensional position of each sensor (relative to a chosen reference sensor on a relatively stable location such as the pelvis) while the person is walking or running. The number of sensors, say N, might be large in order to get dense coverage of the body. Each data point is then a $(3N)$-dimensional vector, measured at a certain point of time. However, despite this seeming high dimensionality, the number of degrees of freedom is small, corresponding to the dozen-or-so joint angles in the body. The sensor positions are more or less deterministic functions of these joint angles. The number of degrees of freedom becomes even smaller if we *double* the dimension of the embedding space by including for each sensor its relative velocity vector. In this space of dimension $6N$ the measured points will lie very close to the *one* dimensional manifold describing the combinations of locations and speeds that the limbs go through during walking or running.

In machine learning and statistics, almost all the work on exploiting intrinsic low dimensionality consists of algorithms for learning the structure of these manifolds; or more precisely, for learning embeddings of these manifolds into low-dimensional Euclidean space. In this work we devise a simple and compact data structure that automatically exploits the low intrinsic dimensionality of data on a local level without explicitly learning the global manifold structure.

2.3 Intrinsic dimensionality

How should intrinsic dimensionality be formalized? In this paper we explore three definitions: doubling dimension, manifold dimension, and local covariance dimension.

The idea for *doubling dimension* appeared in Assouad (1983).

Definition 1 *For any point $x \in \mathbb{R}^D$ and any $r > 0$, let $B(x, r) = \{z : \|x - z\| < r\}$ denote the open ball of radius r centered at x. The doubling dimension of $S \subset \mathbb{R}^D$ 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 $r/2$.*

(It is well known that the doubling dimension is at most $O(D)$.) This definition is convenient to work with, and has proved fruitful in recent work on embeddings of metric spaces (Assouad, 1983; Heinonen, 2001; Gupta, Krauthgamer, and Lee, 2003). It implies, for instance, the existence of ϵ-nets of size $O((1/\epsilon)^d)$, for all ϵ. In our analysis, we will repeatedly use such approximations, and take union bounds over these rather than over the data set.

How does this relate to manifolds? In other words, what is the doubling dimension of a d-dimensional Riemannian submanifold of $\mathbb{R}^D$? We show (Theorem 13 in Section 4) that it is $O(d)$, subject to a bound on the second fundamental form of the manifold.

We also consider a statistical notion of dimension: we say that a set S has *local covariance dimension* (d, ϵ, r) if neighborhoods of radius $\leq r$ have $(1-\epsilon)$ fraction of their variance concentrated in a d-dimensional subspace. To make this precise, start by letting $\sigma_1^2, \sigma_2^2, \ldots, \sigma_D^2$ denote the eigenvalues of the covariance matrix; these are the variances in each of the eigenvector directions.

Definition 2 *Set $S \subset \mathbb{R}^D$ has local covariance dimension (d, ϵ, r) if its restriction to any ball of radius $\leq r$ has covariance matrix whose largest d eigenvalues satisfy*

$$\sigma_i^2 \;\geq\; \frac{1-\epsilon}{d} \cdot (\sigma_1^2 + \cdots + \sigma_D^2).$$

2.4 Main results

Suppose an RP tree is built from a data set $X \subset \mathbb{R}^D$, not necessarily finite. If the tree has k levels, then it partitions the space into 2^k cells. We define the *radius* of a cell $C \subset \mathbb{R}^D$ to be the smallest $r > 0$ such that $X \cap C \subset B(x, r)$ for some $x \in C$. Our first theorem gives an upper bound on the rate at which the radius of cells in an RPTree-Max decreases as one moves down the tree.

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

Our next theorem gives a result for the second type of RPTree. In this case, we are able to quantify the improvement per level, rather than amortized over levels. Recall that an RPTree-Mean has two different types of splits; let's call them splits *by distance* and splits *by projection*.

Theorem 4 *There are constants $0 < c_1, c_2, c_3 < 1$ with the following property. Suppose an RPTree-Mean is built using data set $X \subset \mathbb{R}^D$ of local covariance dimension (d, ϵ, r), where $\epsilon < c_1$. Consider any cell C of radius $\leq r$. Pick a point $x \in S \cap C$ at random, and let C' be the cell that contains it at the next level down.*

- *If C is split by distance then*
$$\mathbb{E}\left[\Delta(S \cap C')\right] \leq c_2 \Delta(S \cap C).$$
- *If C is split by projection, then*
$$\mathbb{E}\left[\Delta_A^2(S \cap C')\right] \leq \left(1 - \frac{c_3}{d}\right) \Delta_A^2(S \cap C).$$

In both cases, the expectation is over the randomization in splitting C and the choice of $x \in S \cap C$.

2.5 A lower bound for k-d trees

Finally, we remark that this property of automatically adapting to intrinsic dimension does not hold for k-d trees.

Theorem 5 *There is a data set $X \subset \mathbb{R}^D$ which has doubling dimension $\leq \log D$ and for which the k-d tree contains cells at level $D-1$ with radius $>$ radius$(X)/2$.*

This bad example is very simple, and applies to any variant of k-d trees that uses axis-aligned splits.

3 An RPTree-Max adapts to doubling dimension

In this section, we prove Theorem 3. The argument makes heavy use of the small ϵ-covers guaranteed by low doubling dimension, and the properties of random projections.

A very rough proof outline is as follows. Suppose an RP tree is built using data set $X \subset \mathbb{R}^D$ of doubling dimension d, and that C is some cell of the tree. If $X \cap C$ lies in a ball of radius Δ, then we need to show that after $O(d \log d)$ further levels of projection, each remaining cell is contained in a ball of radius $\leq \Delta/2$. To this end, we start by covering $X \cap C$ with balls $B_1, B_2, \ldots, B_N$ of radius $\Delta/\sqrt{d}$. The doubling dimension tells us we need $N = O(d^{d/2})$. We'll show that if two balls B_i and B_j are more than a distance $(\Delta/2) - (\Delta/\sqrt{d})$ apart, then a single random projection (with jittered split) has a constant probability of cleanly separating them, in the sense that B_i will lie entirely on one side of the split, and B_j on the other. There are at most N^2 such pairs i, j, and so after $\Theta(d \log d)$ projections every one of these pairs will have been split. In other words, $\Theta(d \log d)$ levels below C in the tree, each cell will only contain points from balls B_i which are within distance $(\Delta/2) - (\Delta/\sqrt{d})$ of each other. Hence the radius of these cells will be $\leq \Delta/2$.

Returning to the two balls B_i and B_j, we says that a split is *good* if it completely separates them. But there are also *bad* splits, in which the split point intersects *both* the balls. The remaining splits are *neutral* (Figure 2). Most of our proof consists in showing that the probability of a good split is strictly greater than that of a bad split.

For instance, to lower-bound the probability of a good split, we show that each of the following sequence of events occurs with constant probability.

- $\widetilde{B}_i$ and $\widetilde{B}_j$, the projections of B_i and B_j onto a random line, have a certain amount of empty space between them.
- The median of the projected data lies very close to this empty space.
- Picking a split point at random near the median will separate $\widetilde{B}_i$ from $\widetilde{B}_j$.

We begin by reviewing some relevant properties of random projection.

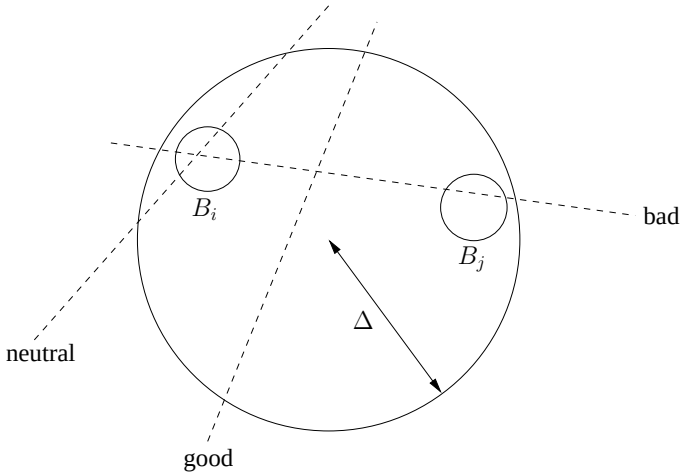


Figure 2: Cell C of the RP tree is contained in a ball of radius Δ. Balls B_i and B_j are part of a cover of this cell, and have radius $\Delta/\sqrt{d}$. From the point of view of this particular pair, there are three kinds of splits: good, bad, and neutral.

3.1 Properties of random projections

The most obvious way to pick a random projection from $\mathbb{R}^D$ to $\mathbb{R}$ is to choose a projection direction u uniformly at random from the surface of the unit sphere S^{D-1}, and to send $x \mapsto u \cdot x$.

Another common option is to select the projection vector from a multivariate Gaussian distribution, $u \sim N(0, (1/D)I_D)$. This gives almost the same distribution as before, and is slightly easier to work with in terms of the algorithm and analysis. We will therefore use this type of projection, bearing in mind that all proofs carry over to the other variety as well, with slight changes in constants.

The key property of a random projection from $\mathbb{R}^D$ to $\mathbb{R}$ is that it approximately preserves the lengths of vectors, modulo a scaling factor of $\sqrt{D}$. This is summarized in the lemma below.

Lemma 6 *Fix any $x \in \mathbb{R}^D$. Pick a random vector $U \sim N(0, (1/D)I_D)$. 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) $\mathbb{P}\left[|U \cdot x| \geq \beta \cdot \frac{\|x\|}{\sqrt{D}}\right] \leq \frac{2}{\beta}e^{-\beta^2/2}$

Proof. See Appendix. ∎

3.2 How does random projection affect the diameter of a set?

Recall that the split rule looks at a random projection of the data and then splits it *approximately* at the median. The perturbation added to the median depends on the diameter of the projected space.

Suppose set $X \subset \mathbb{R}^D$ has doubling dimension d. Let $\widetilde{X}$ denote its random projection into $\mathbb{R}$. How does $\text{diam}(\widetilde{X})$ compare to $\text{diam}(X)$? (Define the diameter of a set S in the obvious way, as $\sup_{x,y\in S}\|x-y\|$.) Obviously $\text{diam}(\widetilde{X}) \leq \text{diam}(X)$, but we would in fact expect it to be much smaller if $d \ll D$. One way to

get a rough upper bound is to consider a cover of X by balls of radius $\text{diam}(X)\sqrt{d/D}$; we know $(D/d)^{d/2}$ such balls are sufficient. Applying Lemma 6(b) to the distance between each pair of ball-centers and taking a union bound, we find that with constant probability:

$$\text{the projected distance between every pair of ball-centers is } \leq \text{ diam}(X) \cdot \sqrt{\frac{d \log D}{D}}$$

(give or take a factor of two). Thus this is also an upper bound on (half) the diameter of the projected set $\widetilde{X}$.

In fact, it is possible to get a slightly better bound which shaves off the $\log D$ factor:

$$\text{diam}(\widetilde{X}) \;\leq\; \text{diam}(X) \cdot O\left(\sqrt{\frac{d}{D}}\right).$$

To get this, we do not drop the radius from $\text{diam}(X)$ to $\text{diam}(X)\sqrt{d/D}$ in one go, but rather decrease it gradually, a factor of two at a time. This technique was demonstrated in Indyk and Naor (2005); the following uses an adaptation of their proof.

Lemma 7 *Suppose set $X \subset \mathbb{R}^D$ is contained in a ball $B(x_0, \Delta)$ and has doubling dimension d. Let $\widetilde{X}$ denote the random projection of X into $\mathbb{R}$. Then for any $0 < \delta < 1$, with probability $> 1 - \delta$ over the choice of projection, $\widetilde{X}$ is contained in an interval of radius $4 \cdot \frac{\Delta}{\sqrt{D}} \cdot \sqrt{2\left(d + \ln \frac{2}{\delta}\right)}$ centered at $\widetilde{x}_0$.*

Proof. See Appendix. ∎

3.3 Most projected points lie close to the median

Lemma 7 tells us that a set in $\mathbb{R}^D$ of radius Δ and doubling dimension d projects to an interval in $\mathbb{R}$ of radius at most $O(\Delta \cdot \sqrt{d/D})$. In fact, *most* of the projected points will be even closer together, in a *central interval* of size $O(\Delta/\sqrt{D})$. We will use this fact to argue that the median of the projected points also lies in this central interval.

Lemma 8 *Suppose $X \subset \mathbb{R}^D$ lies within some ball $B(x_0, \Delta)$. Pick any $0 < \delta, \epsilon \leq 1$ such that $\delta\epsilon \leq 1/e^2$. Let μ be any measure on X. Then with probability $> 1 - \delta$ over the choice of random projection onto $\mathbb{R}$, all but an ϵ fraction of $\widetilde{X}$ (measured according to μ) lies within distance $\sqrt{2 \ln \frac{1}{\delta\epsilon}} \cdot \frac{\Delta}{\sqrt{D}}$ of $\widetilde{x}_0$.*

Proof. See Appendix. ∎

As a corollary, the median of the projected points must lie near $\widetilde{x}_0$ with high probability.

Corollary 9 *Under the hypotheses of Lemma 8, for any $0 < \delta < 2/e^2$, the following holds with probability at least $1 - \delta$ over the choice of projection:*

$$|\text{median}(\widetilde{X}) - \widetilde{x}_0| \;\leq\; \frac{\Delta}{\sqrt{D}} \cdot \sqrt{2 \ln \frac{2}{\delta}}.$$

Proof. Let μ be the uniform distribution over X and use $\epsilon = 1/2$. ∎

3.4 The chance that two faraway balls are cleanly separated by a single split

We now get to the main lemma, which gives a lower bound on the probability of a good split (recall Figure 2).

Lemma 10 *Suppose $X \subset B(x_0, \Delta)$ has doubling dimension $d \geq 1$. Pick any two balls $B = B(z, r)$ and $B' = B(z', r)$ such that*

- *their centers z and z' lie in $B(x_0, \Delta)$,*
- *the distance between these centers is $\|z - z'\| \geq \frac{1}{2}\Delta - r$, and*
- *the radius r is at most $\Delta/(512\sqrt{d})$.*

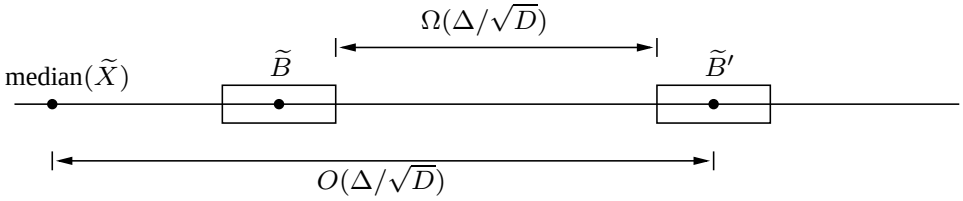
Now, suppose you pick a random projection U which sends X to (say) $\widetilde{X} \subset \mathbb{R}$, and then you pick a split point at random in the range $\text{median}(\widetilde{X}) \pm \frac{6\Delta}{\sqrt{D}}$. *Then with probability at least $1/192$ over the choice of U and the split point, $X \cap B$ and $X \cap B'$ will completely be contained in separate halves of the split.*

Proof. This is a sketch; full details are in the Appendix.

Let $\widetilde{B}$ and $\widetilde{B}'$ denote the projections of $X \cap B$ and $X \cap B'$, respectively. There is a reasonable chance the split point will cleanly separate $\widetilde{B}$ from $\widetilde{B}'$, *provided* the random projection U satisfies four critical properties:

1. $\widetilde{B}$ and $\widetilde{B}'$ are contained within intervals of radius at most $\Delta/(16\sqrt{D})$ around $\widetilde{z}$ and $\widetilde{z}'$, respectively.
2. $|\widetilde{z} - \widetilde{z}'| \geq \Delta/(4\sqrt{D})$.
3. $\widetilde{z}$ and $\widetilde{z}'$ both lie within distance $3\Delta/\sqrt{D}$ of $\widetilde{x}_0$.
4. The median of $\widetilde{X}$ lies within distance $3\Delta/\sqrt{D}$ of $\widetilde{x}_0$.

It follows from Lemmas 6 and 7 and Corollary 9 that with probability at least $1/2$, the projection U will satisfy all these properties. If this is the case, then we say U is "good", and the following picture of $\widetilde{X}$ is valid:



The "sweet spot" is the region between $\widetilde{B}$ and $\widetilde{B}'$; if the split point falls into it, then the two balls will be cleanly separated. By properties (1) and (2), the length of this sweet spot is at least $\Delta/(4\sqrt{D}) - 2\Delta/(16\sqrt{D}) = \Delta/(8\sqrt{D})$. Moreover, by properties (3) and (4), we know that its entirety must lie within distance $3\Delta/\sqrt{D}$ of $\widetilde{x}_0$ (since both $\widetilde{z}$ and $\widetilde{z}'$ do), and thus within distance $6\Delta/\sqrt{D}$ of median$(\widetilde{X})$. Thus, under the sampling strategy from the lemma statement, there will be a constant probability of hitting the sweet spot.

Putting it all together,

$$\mathbb{P}[B, B' \text{ cleanly separated}] \;\geq\; \mathbb{P}[U \text{ is good}] \cdot \mathbb{P}[B, B' \text{ separated} | U \text{ is good}] \;\geq\; \frac{1}{2} \cdot \frac{\Delta/(8\sqrt{D})}{12\Delta/\sqrt{D}} \;=\; \frac{1}{192},$$

as claimed. ∎

3.5 The chance that two faraway balls intersect the split point

We also upper bound the chance of a bad split (Figure 2). For the final qualitative result, all that matters is that this probability be strictly smaller than that of a good split.

Lemma 11 *Under the hypotheses of Lemma 10,*

$$\mathbb{P}[\widetilde{B}, \widetilde{B}' \text{ both intersect the split point}] \;<\; \frac{1}{384}.$$

Proof. See Appendix. ∎

3.6 Proof of Theorem 3

Finally, we complete the proof of Theorem 3.

Lemma 12 *Suppose $X \subset \mathbb{R}^D$ has doubling dimension d. Pick any cell C in the RP tree; suppose it is contained in a ball of radius Δ. Then the probability that there exists a descendant of C which is more than $\Omega(d \log d)$ levels below and yet has radius $> \Delta/2$ is at most $1/2$.*

Proof. Suppose $X \cap C \subset B(x_0, \Delta)$. Consider a cover of this set by balls of radius $r = \Delta/(512\sqrt{d})$. The doubling dimension tells us that at most $N = (O(d))^d$ balls are needed.

Fix any pair of balls B, B' from this cover whose centers are at distance at least $\Delta/2 - r$ from one another. For $k = 1, 2, \ldots$, let p_k be the probability that there is some cell k levels below C which contains points from both B and B'.

By Lemma 10, $p_1 \leq 191/192$. To express p_k in terms of p_{k-1}, think of the randomness in the subtree rooted at C as having two parts: the randomness in splitting cell C, and the rest of the randomness (for each of the two induced subtrees). Lemmas 10 and 11 then tell us that

$$\begin{aligned} p_k &\leq \mathbb{P}[\text{top split cleanly separates } B \text{ from } B'] \cdot 0 + \\ &\quad \mathbb{P}[\text{top split intersects both } B \text{ and } B'] \cdot 2p_{k-1} + \\ &\quad \mathbb{P}[\text{all other split configurations}] \cdot p_{k-1} \\ &\leq \frac{1}{192} \cdot 0 + \frac{1}{384} \cdot 2p_{k-1} + \left(1 - \frac{1}{192} - \frac{1}{384}\right) \cdot p_{k-1} \\ &= \left(1 - \frac{1}{384}\right) p_{k-1}. \end{aligned}$$

The three cases in the first inequality correspond respectively to good, bad, and neutral splits at C (Figure 2). It follows that for some constant c' and $k = c' d \log d$, we have $p_k \leq 1/N^2$.

To finish up, take a union bound over all pairs of balls from the cover which are at the prescribed minimum distance from each other. ∎

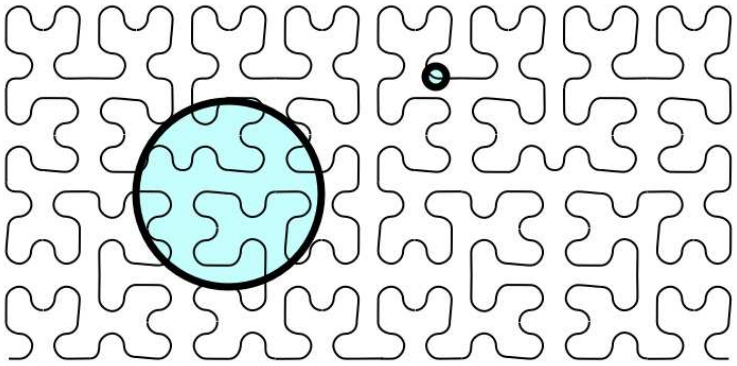


Figure 3: Hilbert's space filling curve: a 1-dimensional manifold that has doubling dimension 2, when the radius of balls is larger than the curvature of the manifold.

4 The doubling dimension of a smooth manifold

Suppose M is a d-dimensional Riemannian submanifold of $\mathbb{R}^D$. What is its doubling dimension? The simplest case is when M is a d-dimensional affine set, in which case it immediately has the same doubling dimension as $\mathbb{R}^d$, namely $O(d)$. We may expect that for more general M, *small enough neighborhoods* have this doubling dimension.

Consider a neighborhood $M \cap B(x, r)$ (recall that balls are defined with respect to Euclidean distance in $\mathbb{R}^D$ rather than geodesic distance on M). If this neighborhood has high *curvature* (speaking informally), then it could potentially have a large doubling dimension. For instance, $M \cap B(x, r)$ could be a 1-dimensional manifold and yet curve so much that $\Theta(2^D)$ balls of radius $r/2$ are needed to cover it (Figure 3). We will therefore limit attention to manifolds of bounded curvature, and to values of r small enough that the pieces of M in $B(x, r)$ are relatively flat.

To formalize things, we need to get a handle on how curved the manifold M is locally. This is a relationship between the Riemannian metric on M and that of the space $\mathbb{R}^D$ in which it is immersed, and is captured by the differential geometry concept of *second fundamental form*. For any point $p \in M$, this is a symmetric bilinear form $B : T_p \times T_p \to T_p^\perp$, where T_p denotes the tangent space at p and $T_p^\perp$ the normal space orthogonal to T_p. For details, see Chapter 6 of do Carmo (1992). Our assumption on curvature is the following.

> **Assumption.** The norm of the second fundamental form is uniformly bounded by some $\kappa \geq 0$; that is, for all $p \in M$ and unit norm $\eta \in T_p^\perp$ and $u \in T_p$, we have $\frac{\langle \eta, B(u,u) \rangle}{\langle u, u \rangle} \leq \kappa$.

We will henceforth limit attention to balls of radius $O(1/\kappa)$.

An additional, though minor, effect is that $M \cap B(x, r)$ may consist of several connected components, in which case we need to cover each of them. If there are N components, this would add a factor of $\log N$ to the doubling dimension, making it $O(d + \log N)$.

Almost all the technical details needed to bound the doubling dimension of manifolds appear in a separate context in a paper of Niyogi, Smale, and Weinberger (2006), henceforth NSW. Here we just put them together differently.

Theorem 13 *Suppose M is a d-dimensional Riemannian submanifold of $\mathbb{R}^D$ that satisfies the assumption*

above for some $\kappa \geq 0$. For any $x \in \mathbb{R}^D$ and $0 < r \leq 1/2\kappa$, the set $M \cap B(x,r)$ can be covered by $N \cdot 2^{O(d)}$ balls of radius $r/2$, where N is the number of connected components of $M \cap B(x,r)$.

Proof. We'll show that each connected component of $M \cap B(x,r)$ can be covered by $2^{O(d)}$ balls of radius $r/2$. To this end, fix one such component, and denote its restriction to $B(x,r)$ by M'.

Pick $p \in M'$, and let T_p be the tangent space at p. Now consider the projection of M' onto T_p; let f denote this projection map. We will make use of two facts, both of which are proved in NSW.

Fact 1 (Lemma 5.4 of NSW). The projection map $f : M' \to T_p$ is $1-1$.

Now, $f(M')$ is contained in a d-dimensional ball of radius $2r$ and can therefore be covered by $2^{O(d)}$ balls of radius $r/4$. We are almost done, as long as we can show that for any such ball $B \subset T_p$, the inverse image $f^{-1}(B)$ is contained in a D-dimensional ball of radius $r/2$. This follows immediately from the second fact.

Fact 2 (implicit in proof of Lemma 5.3 of NSW). For any $x, y \in M'$,

$$\|f(x) - f(y)\|^2 \geq \|x-y\|^2 \cdot (1 - r^2\kappa^2).$$

Thus the inverse image of the cover in the tangent space yields a cover of M'. ∎

5 An RPTree-Mean adapts to local covariance dimension

An RPTree-Mean has two types of splits. If a cell C has much larger diameter than average diameter, then it is split according to the distances of points from the mean. Otherwise, a random projection is used.

The first type of split is particularly easy to analyze.

5.1 Splitting by distance from the mean

This option is invoked when the points in the current cell, call them S, satisfy $\Delta^2(S) > c\Delta_A^2(S)$.

Lemma 14 *Suppose that $\Delta^2(S) > c\Delta_A^2(S)$. Let S_1 denote the points in S whose distance to* $\text{mean}(S)$ *is less than or equal to the median distance, and let S_2 be the remaining points. Then the expected squared diameter after the split is*

$$\frac{|S_1|}{|S|}\Delta^2(S_1) + \frac{|S_2|}{|S|}\Delta^2(S_2) \;\leq\; \left(\frac{1}{2} + \frac{2}{c}\right)\Delta^2(S).$$

Proof. See Appendix. ∎

5.2 Splitting by projection: overview

Suppose the current cell contains a set of points $S \subset \mathbb{R}^D$ for which $\Delta^2(S) \leq c\Delta_A^2(S)$. We will show that a split by projection has a constant probability of reducing the average squared diameter $\Delta_A^2(S)$ by $\Omega(\Delta_A^2(S)/d)$. Our proof has several parts:

1. First of all, the random projection has a predictable effect upon certain key properties of S, such as its average squared diameter. Moreover, it is possible to give a coarse characterization of the distribution of the projected points; we will use this to upper-bound the averages of various functions of the projected data.

2. The definition of local covariance dimension tells us that the data lie near a d-dimensional subspace. To make use of this, we operate in the eigenbasis of the covariance matrix of S, designating by H the principal d-dimensional subspace, and by L the remaining $D - d$ dimensions.

3. When picking a random projection $U \sim N(0, (1/D)I_D)$, we first consider just its component U_H in the principal subspace H. If we project onto U_H and split S into two parts S_1, S_2 at the median, then the average squared diameter decreases by a factor of $1/d$ with constant probability.

4. Finally, using U in place of U_H does not change this outcome significantly, even though the contribution of U_L to U is much larger than that of U_H.

5.3 Properties of the projected data

How does projection into one dimension affect the average squared diameter of a data set? Roughly, it gets divided by the original dimension. To see this, we start with the fact that when a data set with covariance A is projected onto a vector U, the projected data have variance $U^T AU$. We now show that for random U, such quadratic forms are concentrated about their expected values.

Lemma 15 *Suppose A is an $n \times n$ positive semidefinite matrix, and $U \sim N(0, (1/n)I_n)$. Then for any $\alpha, \beta > 0$:*

(a) $\mathbb{P}[U^T AU < \alpha \cdot \mathbb{E}[U^T AU]] \leq e^{-((1/2)-\alpha)/2}$*, and*

(b) $\mathbb{P}[U^T AU > \beta \cdot \mathbb{E}[U^T AU]] \leq e^{-(\beta-2)/4}$.

Proof. See Appendix. ∎

In our applications of this lemma, the dimension n will be either D or d.

Next, we examine the overall distribution of the projected points. When $S \subset \mathbb{R}^n$ has diameter Δ, its projection into the line can have diameter upto Δ, but as we saw in Lemma 8, most of it will lie within a central interval of size $O(\Delta/\sqrt{n})$. What can be said about points that fall outside this interval?

Lemma 16 *Suppose $S \subset B(0, \Delta) \subset \mathbb{R}^n$. Pick any $0 < \delta \leq 1/e^2$. Pick $U \sim N(0, (1/n)I_n)$. Then with probability at least $1 - \delta$ over the choice of U, the projection $S \cdot U = \{x \cdot U : x \in S\}$ satisfies the following property for all positive integers k.*

The fraction of points outside the interval $\left(-\frac{k\Delta}{\sqrt{n}}, +\frac{k\Delta}{\sqrt{n}}\right)$ is at most $\frac{2^k}{\delta} \cdot e^{-k^2/2}$.

Proof. This follows by applying Lemma 8 for each positive integer k (with corresponding failure probability $\delta/2^k$), and then taking a union bound. ∎

5.4 Properties of average diameter

The average squared diameter $\Delta_A^2(S)$ has certain reformulations and decompositions that make it particularly convenient to work with. These properties are consequences of the following two observations, the first of which the reader may recognize as a standard "bias-variance" decomposition of statistics.

Lemma 17 *Let X, Y be independent and identically distributed random variables in $\mathbb{R}^n$, and let $z \in \mathbb{R}^n$ be any fixed vector.*

(a) $\mathbb{E}\left[\|X - z\|^2\right] = \mathbb{E}\left[\|X - \mathbb{E}X\|^2\right] + \|z - \mathbb{E}X\|^2$.

(b) $\mathbb{E}\left[\|X - Y\|^2\right] = 2\,\mathbb{E}\left[\|X - \mathbb{E}X\|^2\right]$.

Proof. Part (a) is immediate when both sides are expanded. For (b), we use part (a) to assert that for any fixed y, we have $\mathbb{E}\left[\|X - y\|^2\right] = \mathbb{E}\left[\|X - \mathbb{E}X\|^2\right] + \|y - \mathbb{E}X\|^2$. We then take expectation over $Y = y$. ∎

Corollary 18 *The average squared diameter of a set S can also be written as:*

$$\Delta_A^2(S) = \frac{2}{|S|} \sum_{x \in S} \|x - \text{mean}(S)\|^2.$$

Proof. $\Delta_A^2(S)$ is simply $\mathbb{E}\left[\|X - Y\|^2\right]$, when X, Y are i.i.d. draws from the uniform distribution over S. ∎

In other words, $\Delta_A^2(S)$ is twice the average squared radius of S, that is, twice the average squared distance of points in S from their mean. If we define

$$\text{radius}_A^2(S, \mu) = \frac{1}{|S|} \sum_{x \in S} \|x - \mu\|^2$$

we see at once that $\Delta_A^2(S) = 2 \inf_\mu \text{radius}_A^2(S, \mu)$. In our proof, we will occasionally consider values of μ other than mean(S).

At each successive level of the tree, the current cell is split into two, either by a random projection or according to distance from the mean. Suppose the points in the current cell are S, and that they are split into sets S_1 and S_2. It is obvious that the expected diameter is nonincreasing:

$$\Delta(S) \;\geq\; \frac{|S_1|}{|S|}\Delta(S_1) + \frac{|S_2|}{|S|}\Delta(S_2).$$

This is also true of the expected average diameter. In fact, we can precisely characterize how much it decreases on account of the split.

Lemma 19 *Suppose set S is partitioned (in any manner) into S_1 and S_2. Then*

$$\Delta_A^2(S) - \left\{ \frac{|S_1|}{|S|}\Delta_A^2(S_1) + \frac{|S_2|}{|S|}\Delta_A^2(S_2) \right\} \;=\; \frac{2|S_1| \cdot |S_2|}{|S|^2} \|\text{mean}(S_1) - \text{mean}(S_2)\|^2.$$

Proof. See Appendix. ∎

5.5 The principal subspace

The local covariance dimension (d, ϵ, r) tells us that if S is sufficiently small (if it is contained in a ball of radius r), then its variance is concentrated near a d-dimensional subspace. Specifically, if its covariance matrix (henceforth denoted $\text{cov}(S)$) has ordered eigenvalues $\sigma_1^2 \geq \cdots \geq \sigma_D^2$, then the top d eigenvalues satisfy

$$\sigma_i^2 \;\geq\; \frac{1-\epsilon}{d} \cdot (\sigma_1^2 + \cdots + \sigma_D^2).$$

Let H denote the d-dimensional subspace spanned by the top d eigenvectors, and L its orthogonal complement. Without loss of generality, these are the first d and the last $D-d$ coordinates, respectively.

Suppose we pick a random unit vector in subspace H and project S onto it. By the local covariance condition, we'd expect the projected data to have an average squared diameter of at least $(1-\epsilon)\Delta_A^2(S)/d$. We now formalize this intuition.

Lemma 20 *Pick $U_H \sim N(0, (1/d)I_d) \times \{0\}^{D-d}$. Then with probability at least $1/10$, the projection of S onto U_H has average squared diameter*

$$\Delta_A^2(S \cdot U_H) \geq \frac{(1-\epsilon)\Delta_A^2(S)}{4d}.$$

Proof. See Appendix. ∎

In what follows, we will assume for convenience that S has mean zero.

Suppose we project S onto a random vector U_H chosen from the principal subspace H. Splitting $S \cdot U_H$ at its median value, we get two subsets S_1 and S_2. The means of these subsets need not lie in the direction U_H, but we can examine their components $\mu_1, \mu_2 \in \mathbb{R}^D$ in this particular direction. In fact, we can lower bound the benefit of the split in terms of the distance between these two points.

To this end, define $\text{gain}(S_1, S_2)$ to be the reduction in average squared diameter occasioned by the split $S_1 \cup S_2 \to S_1, S_2$, that is,

$$\begin{aligned}
\text{gain}(S_1, S_2) &= \Delta_A^2(S_1 \cup S_2) - \frac{|S_1|}{|S_1|+|S_2|}\Delta_A^2(S_1) - \frac{|S_2|}{|S_1|+|S_2|}\Delta_A^2(S_2) \\
&= 2\text{radius}_A^2(S_1 \cup S_2) - \frac{2|S_1|}{|S_1|+|S_2|}\text{radius}_A^2(S_1) - \frac{2|S_2|}{|S_1|+|S_2|}\text{radius}_A^2(S_2).
\end{aligned}$$

Likewise, define $\text{gain}(S_1, S_2, \mu_1, \mu_2)$ to be (the lowerbound we obtain for) the gain if instead of using the optimal centers $\text{mean}(S_1)$ and $\text{mean}(S_2)$ for the two pieces of the split, we instead use μ_1 and μ_2. That is,

$$\text{gain}(S_1, S_2, \mu_1, \mu_2) \;=\; 2\text{radius}_A^2(S_1 \cup S_2) - \frac{2|S_1|}{|S_1|+|S_2|}\text{radius}_A^2(S_1, \mu_1) - \frac{2|S_2|}{|S_1|+|S_2|}\text{radius}_A^2(S_2, \mu_2).$$

Lemma 21 *Suppose $S \subset \mathbb{R}^D$ has mean zero. Choose any unit vector $W \in \mathbb{R}^D$ and split S into two parts S_1 and S_2 at the median of its projection onto W, that is,*

$$S_1 = \{x \in S : x \cdot W \geq \text{median}(S \cdot W)\}, \;\; S_2 = \{x \in S : x \cdot W < \text{median}(S \cdot W)\}.$$

Let μ_1 and μ_2 denote the (D-dimensional) components of $\text{mean}(S_1)$ *and* $\text{mean}(S_2)$ *in the direction W. Then the decrease in average diameter occasioned by this split is*

$$\text{gain}(S_1, S_2) \;\geq\; \text{gain}(S_1, S_2, \mu_1, \mu_2) \;=\; \frac{1}{2}\|\mu_1 - \mu_2\|^2.$$

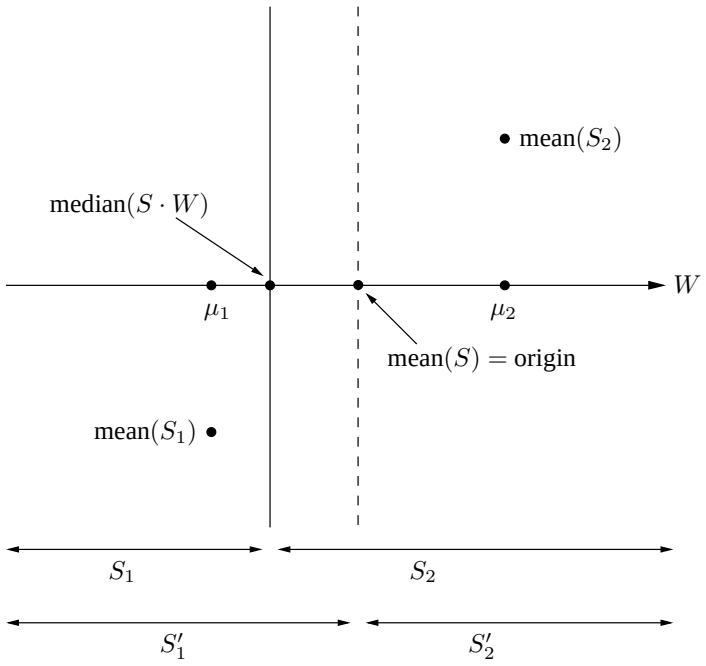


Figure 4: Points S have mean zero. They are projected onto direction W, and split at the median to get sets S_1, S_2. Instead of working with mean(S_1) and mean(S_2), we will work with their components in the direction W, which we denote by μ_1 and μ_2. Notice that $\mu_1 + \mu_2 = 0$. We will also consider the split at the mean: S_1', S_2'.

Proof. The first inequality is obvious. For the second, since S_1 and S_2 are each half the size of S,

$$\begin{aligned}\text{gain}(S_1, S_2, \mu_1, \mu_2) &= \frac{2}{|S|}\sum_{x\in S}\|x\|^2 - \frac{2}{|S|}\sum_{x\in S_1}\|x-\mu_1\|^2 - \frac{2}{|S|}\sum_{x\in S_2}\|x-\mu_2\|^2 \\ &= \frac{2}{|S|}\sum_{x\in S}(x\cdot W)^2 - \frac{2}{|S|}\sum_{x\in S_1}((x-\mu_1)\cdot W)^2 - \frac{2}{|S|}\sum_{x\in S_2}((x-\mu_2)\cdot W)^2\end{aligned}$$

where the second equality comes from noticing that μ_1 and μ_2 are nonzero only in the specific direction W. Now we invoke Lemma 19 to get $\text{gain}(S_1, S_2, \mu_1, \mu_2) = \frac{1}{2}(\mu_1 \cdot W - \mu_2 \cdot W)^2 = \frac{1}{2}\|\mu_1 - \mu_2\|^2$. ∎

A useful property of μ_1 and μ_2 is that their midpoint is zero (Figure 4). Thus, if we are using them as centers for the two halves of our partition, we might as well split S at the mean (the origin) rather than at the median.

Lemma 22 *Under the hypotheses of the previous lemma, consider the alternative partition*

$$S_1' = \{x \in S : x \cdot W \geq 0\}, \;\; S_2' = \{x \in S : x \cdot W < 0\}.$$

Then the decrease in average diameter occasioned by this split is

$$\text{gain}(S_1', S_2') \;\geq\; \text{gain}(S_1', S_2', \mu_1, \mu_2) \;\geq\; \frac{1}{2}\|\mu_1 - \mu_2\|^2.$$

Proof. The first inequality is obvious. The second follows from

$$\text{gain}(S_1', S_2', \mu_1, \mu_2) \;\geq\; \text{gain}(S_1, S_2, \mu_1, \mu_2).$$

This is because both partitions are based only on projections onto direction W. Of the two, the former partition is the optimal one with respect to μ_1, μ_2. We finish up by invoking the previous lemma. ∎

We now have all the ingredients we need to lower-bound the gain of a projection-and-split that is restricted to the principal subspace H.

Lemma 23 *Pick any $\delta > 0$. Pick $U_H \sim N(0, (1/d)I_d) \times \{0\}^{D-d}$ and split S into two halves S_1 and S_2 at the median of the projection onto direction $\widehat{U}_H = U_H/\|U_H\|$. Let μ_1 and μ_2 denote the (D-dimensional) components of* mean(S_1) *and* mean(S_2) *in direction $\widehat{U}_H$. Also define an alternative split*

$$S_1' = \{x \in S : x \cdot U_H \geq 0\}, \;\; S_2' = \{x \in S : x \cdot U_H < 0\}.$$

Then with probability at least $(1/10) - \delta$ over the choice of U_H, we have

$$\text{gain}(S_1', S_2', \mu_1, \mu_2) \;\geq\; \frac{1}{2}\|\mu_1 - \mu_2\|^2 \;\geq\; \frac{\Delta_A^2(S)}{d} \cdot \Omega\left(\frac{(1-\epsilon-c\delta)^2}{c\,\|U_H\|^2 \log(1/\delta)}\right).$$

Proof. Only the second inequality needs to be shown. To this end, consider the projection of S onto U_H (rather than $\widehat{U}_H$), and let the random variable $\widetilde{X}$ denote a uniform-random draw from the projected points. Lemma 20 tells us that with probability at least $1/10$ over the choice of U_H, the variance of $\widetilde{X}$ is at least $(1-\epsilon)\Delta_A^2(S)/(8d)$. We show that together with the tail bounds on $\widetilde{X}$ (Lemma 16), this means $\|\mu_1 - \mu_2\|^2$ cannot be too small. For details see the Appendix. ∎

5.6 The remaining coordinates

If the random projection were chosen specifically from subspace H, then Lemma 23 says that we would be in good shape. We now show that it doesn't hurt too much to pick the projection from $\mathbb{R}^D$. If the projection vector is U, we will use the split

$$S_1'' = \{x \in S : x \cdot U \geq 0\}, \;\; S_2'' = \{x \in S : x \cdot U < 0\}$$

instead of the S_1' and S_2' we had before. As a result of this, some points will change sides (Figure 5). However, the overall benefit of the partition will not decrease too much.

To make the transition from the previous section a bit smoother, we will pick our random vector $U \sim N(0, (1/D)I_D)$ in two phases:

- Pick $U_H \sim N(0, \frac{1}{d}I_d) \times \{0\}^{D-d}$ and $U_L \sim \{0\}^d \times N(0, \frac{1}{D-d}I_{D-d})$.
- Set $U = \sqrt{\frac{d}{D}}\,U_H + \sqrt{\frac{D-d}{D}}\,U_L$

Lemma 24 *Pick $U \sim N(0, (1/D)I_D)$, and define a partition of S as follows.*

$$S_1'' = \{x \in S : x \cdot U \geq 0\}, \;\; S_2'' = \{x \in S : x \cdot U < 0\}.$$

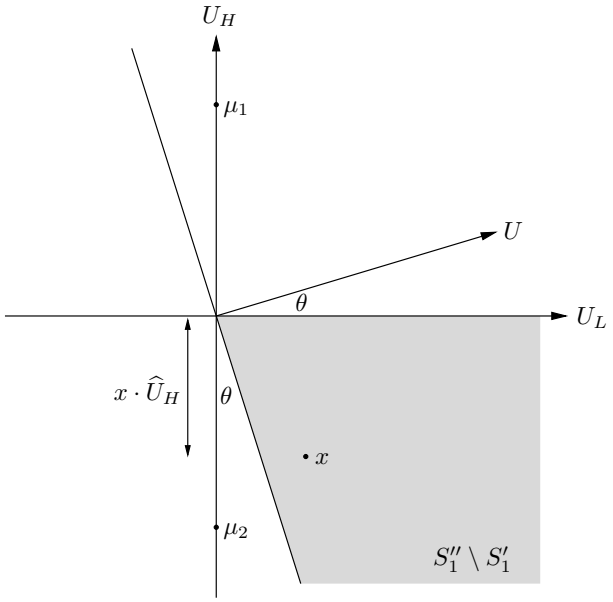


Figure 5: The plane determined by U_H and U_L.

Let μ_1, μ_2 be determined by U_H as in the statement of Lemma 23. Then

$$\text{gain}(S_1'', S_2'') \;\geq\; \frac{1}{2}\|\mu_1 - \mu_2\|^2 - \frac{4}{|S|}\|\mu_1 - \mu_2\| \sum_{x \in (S_1'' \setminus S_1') \cup (S_2'' \setminus S_2')} \left| x \cdot \widehat{U}_H \right|,$$

where $\widehat{U}_H = U_H / \|U_H\|$.

Proof. For each partition S_i'', we keep using the center μ_i from before. We can then compare the benefit of partition (S_1'', S_2'') to that of (S_1', S_2'), both relative to the centers μ_1, μ_2. The points which do not change sides (that is, points in S_1' which are also in S_1'') incur no additional charge. Thus

$$\begin{aligned} \text{gain}(S_1'', S_2'') &\geq \text{gain}(S_1'', S_2'', \mu_1, \mu_2) \\ &\geq \text{gain}(S_1', S_2', \mu_1, \mu_2) - \sum_{x \in (S_1'' \setminus S_1') \cup (S_2'' \setminus S_2')} \text{(additional charge for point } x\text{)}. \end{aligned}$$

The additional cost for a point in $S_1'' \setminus S_1'$ is

$$\frac{2}{|S|}\left(\|x - \mu_1\|^2 - \|x - \mu_2\|^2\right) \;=\; \frac{4x \cdot (\mu_2 - \mu_1)}{|S|} \;=\; \frac{4|x \cdot \widehat{U}_H| \|\mu_1 - \mu_2\|}{|S|}.$$

Likewise for points in $S_2'' \setminus S_2'$. The bound then follows immediately from Lemma 23. ∎

To show that the additional costs do not add up to very much, we divide the points $x \in (S_1'' \setminus S_1') \cup (S_2'' \setminus S_2')$ into two groups: those with $|x \cdot \widehat{U}_H| = O(\Delta_A(S)\sqrt{\epsilon/d})$, which is so small as to be insignificant; and the remainder, for which $|x \cdot \widehat{U}_H|$ could potentially be a lot larger. We'll show that on account of the local covariance condition, the overwhelming majority of points belong to the first group.

Lemma 25 *Pick any $\delta > 0$. Choose $U \sim N(0, (1/D)I_D)$. Then with probability at least $0.98 - \delta$, at most a δ fraction of points $x \in S$ satisfy:*

$$x \in (S_1'' \setminus S_1') \cup (S_2'' \setminus S_2') \quad \textit{and} \quad \left|x \cdot \widehat{U}_H\right| \;>\; \frac{\Delta_A(S)}{\sqrt{d}} \cdot \sqrt{\frac{9\epsilon}{\delta}} \cdot \frac{1}{\|U_H\|}.$$

Proof. We will show that because of the local covariance condition, all but δ fraction of the points have a relatively small projection onto $\widehat{U}_L$. Their projections onto $\widehat{U}_H$ could, of course, be very large, *but not if they are in* $(S_1'' \setminus S_1') \cup (S_2'' \setminus S_2')$. There are several parts of the proof.

(1) With probability at least $1 - e^{-4} \geq 0.98$, the direction U_L satisfies

$$\Delta_A^2(S \cdot U_L) \;\leq\; \frac{18\epsilon\Delta_A^2(S)}{D-d}.$$

Since $\Delta_A^2(S \cdot U_L) = 2(U_L)^T \text{cov}(S) U_L$, it has expectation $\mathbb{E}[\Delta_A^2(S \cdot U_L)] \leq \epsilon\Delta_A^2(S)/(D-d)$ by the local covariance condition. Lemma 15(b) then bounds the chance that it is much larger than its expected value.

(2) With probability at least $1 - \delta$, at most a δ fraction of points $x \in S$ have

$$|x \cdot U_L| \;>\; \sqrt{\frac{9\epsilon\Delta_A^2(S)}{\delta(D-d)}}.$$

This follows immediately from (1), by applying Chebyshev's inequality.

(3) For any point $x \in (S_1'' \setminus S_1') \cup (S_2'' \setminus S_2')$, we have

$$\left|x \cdot \widehat{U}_H\right| \;\leq\; \left|x \cdot \widehat{U}_L\right| \cdot \sqrt{\frac{D-d}{d}} \cdot \frac{\|U_L\|}{\|U_H\|}.$$

Consider the plane spanned by U_H and U_L. The projection of a point x into this plane completely determines its membership in the sets S_i' and S_i'' (Figure 5). We then have

$$\frac{|x \cdot \widehat{U}_L|}{|x \cdot \widehat{U}_H|} \;\geq\; \tan\theta \;=\; \frac{\|U_H\|\sqrt{d/D}}{\|U_L\|\sqrt{(D-d)/D}}$$

from which (3) follows easily.

Putting these together gives the lemma. ∎

This δ fraction of points could potentially have very high $|x \cdot \widehat{U}_H|$ values, but the tail bounds of Lemma 16 assure us that the total contribution of these points must be small.

Lemma 26 *Let T denote the δ fraction of S with the highest values of $|x \cdot \widehat{U}_H|$. Then*

$$\frac{1}{|S|}\sum_{x \in T}\left|x \cdot \widehat{U}_H\right| \;\leq\; \frac{\delta\Delta(S)}{\sqrt{d}\,\|U_H\|} \cdot O\left(\sqrt{\log\frac{1}{\delta}}\right).$$

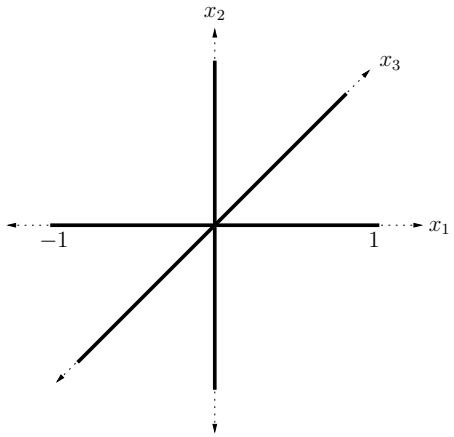


Figure 6: A bad case for k-d trees. Here the intrinsic dimension is just $O(\log D)$, and yet k-d trees need D levels in order to halve the diameter.

Putting it all together, we have that with constant probability, the split reduces the expected average diameter $\Delta_A^2(S)$ by $\Omega(\Delta_A^2(S)/d)$.

Lemma 27 *Suppose set S has $\Delta^2(S) \leq c\Delta_A^2(S)$ and has local covariance dimension (d, ϵ, r), where $r > \Delta(S)$. Then there are constants $c_1, c_2, c_3 > 0$ such that if $\epsilon < c_1$, then with probability at least c_2, a random projection U produces a partition S_1'', S_2'' with*

$$\text{gain}(S_1'', S_2'') \;\geq\; c_3 \cdot \frac{\Delta_A^2(S)}{cd}.$$

Proof. This follows directly from Lemmas 23, 24, 25, and 26. The only additional step is to argue that $\|U_H\|$ is likely to be close to 1. ∎

6 k-d trees do not adapt to intrinsic dimension

Consider a set $X \subset \mathbb{R}^D$ made up of the coordinate axes between -1 and $+1$ (Figure 6):

$$X \;=\; \bigcup_{i=1}^{D} \{te_i : -1 \leq t \leq +1\}.$$

Here $e_1, \ldots, e_D$ is the canonical basis of $\mathbb{R}^D$.

X lies within $B(0, 1)$ and can be covered by $2D$ balls of radius $1/2$. It is not hard to see that the doubling dimension of X is $d = \log 2D$. On the other hand, a k-d tree would clearly need D levels before halving the diameter of its cells. Thus k-d trees cannot be said to adapt to the intrinsic dimensionality of data.

What makes k-d trees fail in this case is simply that the coordinate directions are not helpful for splitting the data. If other directions are allowed, it is possible to do much better. For instance, the direction $(1, 1, \ldots, 1)/\sqrt{D}$ immediately divides X into two groups of diameter $\sqrt{2}$: the points with all-positive coordinates, and those with all-negative coordinates. Subsequently, it is useful to choose directions corresponding to subsets of coordinates (value 1 on those coordinates, 0 on the others), and it is easy to pick $\log D$ such directions that will reduce the diameter of each cell to 1.

Thus there exist $d = 1 + \log D$ directions such that median splits along these directions yield cells of diameter 1. And as we have seen, if we increase the number of levels slightly, to $O(d \log d)$, even randomly chosen directions will work well.

References

P. Assouad (1983). Plongements lipschitziens dans $\mathbb{R}^n$. *Bull. Soc. Math. France*, 111(4):429-448.

M. Belkin and P. Niyogi (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373-1396.

W. Boothby (2003). *An Introduction to Differentiable Manifolds and Riemannian Geometry.* Academic Press.

L. Devroye, L. Gyorfi, and G. Lugosi (1996). *A Probabilistic Theory of Pattern Recognition*. Springer.

M. do Carmo (1992). *Riemannian Geometry*. Birkhauser.

R. Durrett (1995). *Probability: Theory and Examples*. Second edition, Duxbury.

J. Bentley (1975). Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 18(9):509–517.

H. Fuchs, Z. Kedem, and B. Naylor (1980). On visible surface generation by a priori tree structures. *Computer Graphics*, 14(3):124–133.

A. Gupta, R. Krauthgamer, and J. Lee (2003). Bounded geometries, fractals, and low-distortion embeddings. *IEEE Symposium on Foundations of Computer Science*, 534–544.

J. Heinonen (2001). *Lectures on Analysis on Metric Spaces*. Springer.

P. Indyk and A. Naor (2005). Nearest neighbor preserving embeddings. *ACM Transactions on Algorithms*, to appear.

P. Niyogi, S. Smale, and S. Weinberger (2006). Finding the homology of submanifolds with high confidence from random samples. *Discrete and Computational Geometry*, to appear.

S. Roweis and L. Saul (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science* 290:2323-2326.

J. Tenenbaum, V. de Silva, and J. Langford (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

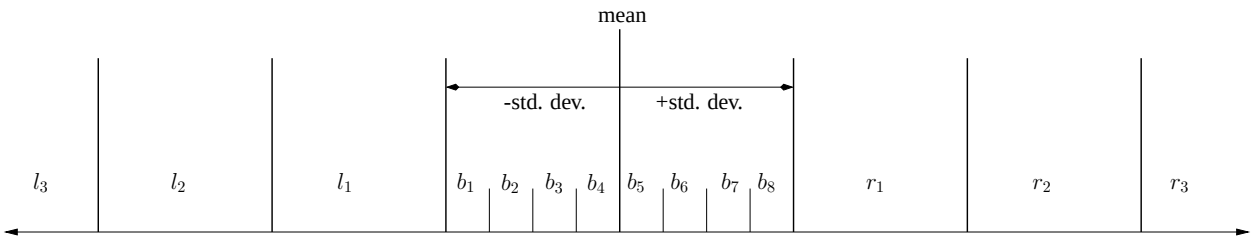


Figure 7: For each of these bins, its frequency and mean value are maintained.

7 Appendix I: a streaming version of the RP tree algorithm

In many large-scale applications, it is desirable to process data in a *streaming* manner. Formally, this is a model in which the data arrive one at a time, and are read but not stored in their entirety. Despite this constraint, the goal is to achieve (roughly) the same result as if all data were available in memory.

RP trees can handle streaming because the random projections can be chosen in advance and fixed thereafter. However, the split points themselves are based on statistics of the data, and some care must be taken to maintain these accurately.

In the streaming model, we build the RP tree gradually, starting with a single node which is split in two once enough data has been seen. This same process is then applied to the two children.

Let's focus on a single node. Since the projection vector associated with the node is randomly chosen and fixed beforehand, the data arriving at the node is effectively a stream of scalar values (each a dot product of a data point with the projection vector). Let's call this sequence $x_1, x_2, \ldots$. The node then goes through three phases.

Phase 1. The mean and variance of the x_i are calculated. Call these μ and σ^2.

This phase lasts during the first N_1 data points.

Phase 2. The data are partitioned into $k + 20$ bins, corresponding to the following intervals (Figure 7):

- ten left-hand intervals $r_1 = [\mu - \sigma, \mu - 2\sigma)$, $r_2 = [\mu - 2\sigma, \mu - 3\sigma)$, ... $r_10 = [\mu - 10\sigma, -\infty)$;
- k middle intervals $b_1, \ldots, b_k$ of equal width, spanning $(\mu - \sigma, \mu + \sigma)$;
- ten right-hand intervals $r_1 = [\mu + \sigma, \mu + 2\sigma)$, $r_2 = [\mu + 2\sigma, \mu + 3\sigma)$, ... $r_10 = [\mu + 10\sigma, \infty)$.

For each bin, two statistics are maintained: the fraction of points that fall in that interval, and their mean. We'll use $p(\cdot)$ to denote the former, and $m(\cdot)$ the latter.

This phase occupies the next N_2 data points.

Phase 3. Based on the statistics collected in phase 2, a split point is chosen at one of the interval boundaries. The node is then split into two children.

This is just a decision-making phase which consumes no additional data.

For Phase 3, we use the following rule.

procedure CHOOSERULE($BinStatistics, k, C, \epsilon$)

for $i = 2, \ldots, 10$ { **if** $\sum_{j=i}^{10} [p(l_j) + p(r_j)] > \max(\epsilon, C(1 - \text{erf}(i/\sqrt{2})))$ **then return** $(Rule := (|x - \mu| < i\sigma))$

{ Choose the boundary b between $b_1, \ldots, b_k$ that maximizes:
$p(\text{left side})p(\text{right side})(m(\text{left side}) - m(\text{right side}))^2$
return $(Rule := (x < b))$

In the simplest implementation the node does not update statistics after Phase 2, and just serves as a data splitter. A better implementation would continue to update the node, albeit slowly. New examples x_t (for $t > N_1 + N_2$) could lead to updates of the form:

$$\begin{aligned} \mu_t &\leftarrow (1 - \alpha_1)\mu_{t-1} + \alpha_1 x_t \\ \sigma_t^2 &\leftarrow (1 - \alpha_2)\sigma_{t-1}^2 + \alpha_2 (x - \mu_t)^2 \end{aligned}$$

where $\alpha_1 = 1/2N_1$, $\alpha_2 = 1/2N_2$.

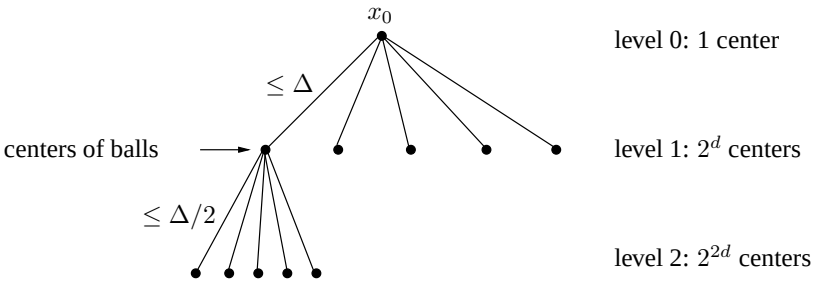


Figure 8: A hierarchy of covers. At level k, there are 2^{kd} points in the cover. Each of them has distance $\leq \Delta/2^k$ to its children (which constitute the cover at level $k+1$). At the leaves are individual points of X.

8 Appendix II: proofs

8.1 Proof of Lemma 6

Since U has a Gaussian distribution, and any linear combination of independent Gaussians is a Gaussian, it follows that the projection $U \cdot x$ is also Gaussian. Its mean and variance are easily seen to be zero and $\|x\|^2/D$, respectively. Therefore, writing

$$Z \;=\; \frac{\sqrt{D}}{\|x\|}\,(U \cdot x)$$

we have that $Z \sim N(0,1)$. The bounds stated in the lemma now follow from properties of the standard normal. In particular, $N(0,1)$ is roughly flat in the range $[-1,1]$ and then drops off rapidly; the two cases in the lemma statement correspond to these two regimes.

The highest density achieved by the standard normal is $1/\sqrt{2\pi}$. Thus the probability mass it assigns to the interval $[-\alpha,\alpha]$ is at most $2\alpha/\sqrt{2\pi}$; this takes care of (a). For (b), we use a standard tail bound for the normal, $\mathbb{P}(|Z| \geq \beta) \leq (2/\beta)e^{-\beta^2/2}$; see, for instance, page 7 of Durrett (1995).

8.2 Proof of Lemma 7

Pick a cover of $X \subset B(x_0,\Delta)$ by 2^d balls of radius $\Delta/2$. Without loss of generality, we can assume the centers of these balls lie in $B(x_0,\Delta)$. Each such ball B induces a subset $X \cap B$; cover each such subset by 2^d smaller balls of radius $\Delta/4$, once again with centers in B. Continuing this process, the final result is a hierarchy of covers, at increasingly finer granularities (Figure 8).

Pick any center u at level k of the tree, along with one of its children v at level $k+1$. Then $\|u-v\| \leq \Delta/2^k$. Letting $\widetilde{u},\widetilde{v}$ denote the projections of these two points, we have from Lemma 6(b) that

$$\mathbb{P}\left[|\widetilde{u}-\widetilde{v}| \geq \beta\cdot\frac{\Delta}{\sqrt{D}}\cdot\left(\frac{3}{4}\right)^k\right] \;\leq\; \mathbb{P}\left[|\widetilde{u}-\widetilde{v}| \geq \beta\left(\frac{3}{2}\right)^k\cdot\frac{\|u-v\|}{\sqrt{D}}\right] \;\leq\; \frac{2}{\beta}\left(\frac{2}{3}\right)^k\exp\left(-\frac{\beta^2}{2}\cdot\left(\frac{3}{2}\right)^{2k}\right).$$

Choosing $\beta = \sqrt{2(d+\ln(2/\delta))}$, and using the relation $(3/2)^{2k} \geq k+1$ (for all $k \geq 0$), we have

$$\mathbb{P}\left[|\widetilde{u}-\widetilde{v}| \geq \beta\cdot\frac{\Delta}{\sqrt{D}}\cdot\left(\frac{3}{4}\right)^k\right] \;\leq\; \frac{\delta}{\beta}\left(\frac{\delta}{3}\right)^k e^{-(k+1)d}.$$

Now take a union bound over all edges (u, v) in the tree. There are $2^{(k+1)d}$ edges between levels k and $k+1$, so

$$\begin{aligned}\mathbb{P}\left[\exists k : \exists u \text{ in level } k \text{ with child } v \text{ in level } k+1 : |\widetilde{u} - \widetilde{v}| \geq \beta \cdot \frac{\Delta}{\sqrt{D}} \cdot \left(\frac{3}{4}\right)^k\right] \\ \leq \sum_{k=0}^{\infty} 2^{(k+1)d} \cdot \frac{\delta}{\beta}\left(\frac{\delta}{3}\right)^k e^{-(k+1)d} \\ \leq \frac{\delta}{\beta} \cdot \frac{1}{1-(\delta/3)} \leq \delta\end{aligned}$$

where for the last step we observe $\beta \geq 3/2$ whenever $d \geq 1$.

So with probability at least $1-\delta$, for all k, every edge between levels k and $k+1$ in the tree has projected length at most $\beta \cdot (3/4)^k \cdot \Delta/\sqrt{D}$. Thus every projected point in $\widetilde{X}$ has a distance from $\widetilde{x}_0$ of at most

$$\beta \cdot \frac{\Delta}{\sqrt{D}} \cdot \left(1 + \frac{3}{4} + \left(\frac{3}{4}\right)^2 + \cdots\right) = \frac{4\beta\Delta}{\sqrt{D}}.$$

Plugging in the value of β then yields the lemma.

8.3 Proof of Lemma 8

Set $c = \sqrt{2 \ln 1/(\delta\epsilon)} \geq 2$.

Fix any point x, and randomly choose a projection U. What is the chance that $\widetilde{x}$ lands far from $\widetilde{x}_0$? Define the bad event to be $F_x = \mathbf{1}(|\widetilde{x} - \widetilde{x}_0| \geq c\Delta/\sqrt{D})$. By Lemma 6(b), we have

$$\mathbb{E}_U[F_x] \leq \mathbb{P}_U\left[|\widetilde{x} - \widetilde{x}_0| \geq c \cdot \frac{\|x - x_0\|}{\sqrt{D}}\right] \leq \frac{2}{c} e^{-c^2/2} \leq \delta\epsilon.$$

Since this holds for any $x \in X$, it also holds in expectation over x drawn from μ. We are interested in bounding the probability (over the choice of U) that more than an ϵ fraction of μ falls far from $\widetilde{x}_0$. Using Markov's inequality and then Fubini's theorem, we have

$$\mathbb{P}_U\left[\mathbb{E}_\mu[F_x] \geq \epsilon\right] \leq \frac{\mathbb{E}_U[\mathbb{E}_\mu[F_x]]}{\epsilon} = \frac{\mathbb{E}_\mu[\mathbb{E}_U[F_x]]}{\epsilon} \leq \delta,$$

as claimed.

8.4 Proof of Lemma 10

It will help to define the failure probabilities $\delta_1 = 2/e^{31}$ and $\delta_2 = 1/20$.

Let $\widetilde{B}$ and $\widetilde{B}'$ denote the projections of $X \cap B$ and $X \cap B'$, respectively. There is a reasonable chance the split point will cleanly separate $\widetilde{B}$ from $\widetilde{B}'$, *provided* the random projection U satisfies four critical properties.

> **Property 1:** $\widetilde{B}$ and $\widetilde{B}'$ are contained within intervals of radius at most $\Delta/(16\sqrt{D})$ around $\widetilde{z}$ and $\widetilde{z}'$, respectively.

This follows by applying Lemma 7 to each ball in turn. For B, we have that with probability at least $1-\delta_1$, $\widetilde{B}$ is within radius

$$\frac{4r}{\sqrt{D}} \cdot \sqrt{2\left(d + \ln\frac{2}{\delta_1}\right)} \;\leq\; \frac{\Delta}{128\sqrt{D}} \cdot \sqrt{2\ln\frac{2e}{\delta_1}} \;=\; \frac{\Delta}{16\sqrt{D}}$$

of $\widetilde{z}$. Similarly with B', so this property holds with probability at least $1-2\delta_1$.

Property 2: $|\widetilde{z} - \widetilde{z}'| \geq \Delta/(4\sqrt{D})$.

By Lemma 6(a), this fails with probability at most $4\alpha/5$ for $\alpha = 1/(2-(4r/\Delta)) \leq 128/255$.

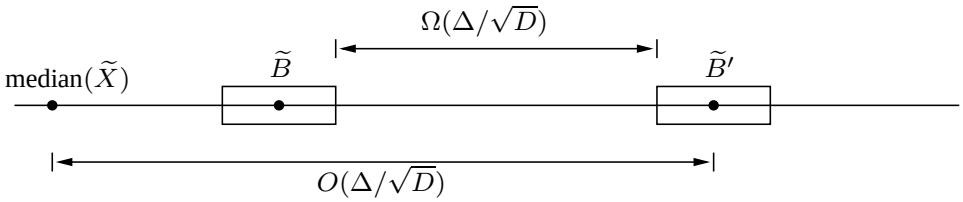
Property 3: $\widetilde{z}$ and $\widetilde{z}'$ both lie within distance $3\Delta/\sqrt{D}$ of $\widetilde{x}_0$.

By Lemma 6(b), this happens with probability at least $1 - 2\delta_2/\sqrt{2\ln(2/\delta_2)}$ (in that lemma, use $\beta = \sqrt{2\ln(2/\delta_2)}$).

Property 4: The median of $\widetilde{X}$ lies within distance $3\Delta/\sqrt{D}$ of $\widetilde{x}_0$.

By Corollary 9, this happens with probability at least $1-\delta_2$.

In summary, with probability at least $1/2$, all four properties hold, in which case we say the projection U is "good". If this is the case, then the following picture of $\widetilde{X}$ is valid:



The "sweet spot" is the region between $\widetilde{B}$ and $\widetilde{B}'$; if the split point falls into it, then the two balls will be cleanly separated. By properties (1) and (2), the length of this sweet spot is at least $\Delta/(4\sqrt{D}) - 2\Delta/(16\sqrt{D}) = \Delta/(8\sqrt{D})$. Moreover, by properties (3) and (4), we know that its entirety must lie within distance $3\Delta/\sqrt{D}$ of $\widetilde{x}_0$ (since both $\widetilde{z}$ and $\widetilde{z}'$ do), and thus within distance $6\Delta/\sqrt{D}$ of median($\widetilde{X}$). Thus, under the sampling strategy from the lemma statement, there will be a constant probability of hitting the sweet spot.

Putting it all together,

$$\mathbb{P}[B,\, B' \text{ cleanly separated}] \;\geq\; \mathbb{P}[U \text{ is good}] \cdot \mathbb{P}[B,\, B' \text{ separated} | U \text{ is good}] \;\geq\; \frac{1}{2} \cdot \frac{\Delta/(8\sqrt{D})}{12\Delta/\sqrt{D}} \;=\; \frac{1}{192},$$

as claimed.

8.5 Proof of Lemma 11

Define δ_1 as in the previous proof. As before (property (1)), with probability at least $1-2\delta_1$, the projections $\widetilde{B}$ and $\widetilde{B}'$ lie within radii $\leq \Delta/(16\sqrt{D})$ of their respective $\widetilde{z}, \widetilde{z}'$.

In order for $\widetilde{B}$ and $\widetilde{B}'$ to both intersect the split point, two unlikely events need to happen: first, $\widetilde{B}$ needs to intersect $\widetilde{B}'$; second, the split point needs to intersect $\widetilde{B}$. These are independent events (one involves the projection and the other involves the split point), so we will bound them in turn.

First of all, using Lemma 6(a),

$$\begin{aligned}\mathbb{P}[\widetilde{B} \text{ intersects } \widetilde{B}'] &\leq \mathbb{P}[|\widetilde{z} - \widetilde{z}'| \leq \Delta/(8\sqrt{D})] \\ &\leq \sqrt{\frac{2}{\pi}} \cdot \frac{\Delta/(8\sqrt{D})}{(1/\sqrt{D}) \cdot ((\Delta/2) - r)} \leq \sqrt{\frac{2}{\pi}} \cdot \frac{64}{255}\end{aligned}$$

under the conditions on r. Second,

$$\mathbb{P}[\text{split point intersects } \widetilde{B}] \leq \frac{\Delta/(8\sqrt{D})}{12\Delta/\sqrt{D}} = \frac{1}{96}.$$

It follows that

$$\begin{aligned}\mathbb{P}[\widetilde{B}, \widetilde{B}' \text{ both intersect split point}] &\leq 2\delta_1 + \mathbb{P}[\widetilde{B}, \widetilde{B}' \text{ touch}] \cdot \mathbb{P}[\text{split point touches } \widetilde{B}] \\ &\leq 2\delta_1 + \sqrt{\frac{2}{\pi}} \cdot \frac{64}{255} \cdot \frac{1}{96} < \frac{1}{384},\end{aligned}$$

as claimed.

8.6 Proof of Lemma 14

Let random variable X be distributed uniformly over S. Then

$$\mathbb{P}\left[\|X - \mathbb{E}X\|^2 \geq \text{median}(\|X - \mathbb{E}X\|^2)\right] \geq \frac{1}{2}$$

by definition of median, so $\mathbb{E}\left[\|X - \mathbb{E}X\|^2\right] \geq \text{median}(\|X - \mathbb{E}X\|^2)/2$. It follows from Corollary 18 that

$$\text{median}(\|X - \mathbb{E}X\|^2) \leq 2\mathbb{E}\left[\|X - \mathbb{E}X\|^2\right] = \Delta_A^2(S).$$

Set S_1 has squared diameter $\Delta^2(S_1) \leq (2\,\text{median}(\|X - \mathbb{E}X\|))^2 \leq 4\Delta_A^2(S)$. Meanwhile, S_2 has squared diameter at most $\Delta^2(S)$. Therefore,

$$\frac{|S_1|}{|S|}\Delta^2(S_1) + \frac{|S_2|}{|S|}\Delta^2(S_2) \leq \frac{1}{2} \cdot 4\Delta_A^2(S) + \frac{1}{2}\Delta^2(S)$$

and the lemma follows by using $\Delta^2(S) > c\Delta_A^2(S)$.

8.7 Proof of Lemma 15

This follows by examining the moment-generating function of $U^T AU$. Since the distribution of U is spherically symmetric, we can work in the eigenbasis of A and assume without loss of generality that $A = \text{diag}(a_1, \ldots, a_n)$, where $a_1, \ldots, a_n$ are the eigenvalues. Moreover, for convenience we take $\sum a_i = 1$.

Let $U_1, \ldots, U_n$ denote the individual coordinates of U. We can rewrite them as $U_i = Z_i/\sqrt{n}$, where $Z_1, \ldots, Z_n$ are i.i.d. standard normal random variables. Thus

$$U^T AU = \sum_i a_i U_i^2 = \frac{1}{n}\sum_i a_i Z_i^2.$$

This tells us immediately that $\mathbb{E}[U^T AU] = 1/n$.

We use Chernoff's bounding method for both parts. For (a), for any $t > 0$,

$$\begin{aligned}
\mathbb{P}\left[U^T AU < \alpha \cdot \mathbb{E}[U^T AU]\right] &= \mathbb{P}\left[\sum_i a_i Z_i^2 < \alpha\right] = \mathbb{P}\left[e^{-t\sum_i a_i Z_i^2} > e^{-t\alpha}\right] \\
&\leq \frac{\mathbb{E}\left[e^{-t\sum_i a_i Z_i^2}\right]}{e^{-t\alpha}} = e^{t\alpha}\prod_i \mathbb{E}\left[e^{-ta_i Z_i^2}\right] \\
&= e^{t\alpha}\prod_i \left(\frac{1}{1+2ta_i}\right)^{1/2}
\end{aligned}$$

and the rest follows by using $t = 1/2$ along with the inequality $1/(1+x) \leq e^{-x/2}$ for $0 < x \leq 1$. Similarly for (b), for $0 < t < 1/2$,

$$\begin{aligned}
\mathbb{P}\left[U^T AU > \beta \cdot \mathbb{E}[U^T AU]\right] &= \mathbb{P}\left[\sum_i a_i Z_i^2 > \beta\right] = \mathbb{P}\left[e^{t\sum_i a_i Z_i^2} > e^{t\beta}\right] \\
&\leq \frac{\mathbb{E}\left[e^{t\sum_i a_i Z_i^2}\right]}{e^{t\beta}} = e^{-t\beta}\prod_i \mathbb{E}\left[e^{ta_i Z_i^2}\right] \\
&= e^{-t\beta}\prod_i \left(\frac{1}{1-2ta_i}\right)^{1/2}
\end{aligned}$$

and it is adequate to choose $t = 1/4$ and invoke the inequality $1/(1-x) \leq e^{2x}$ for $0 < x \leq 1/2$.

8.8 Proof of Lemma 19

Let μ, μ_1, μ_2 denote the means of S, S_1, and S_2. Using Corollary 18 and Lemma 17(a), we have

$$\begin{aligned}
&\Delta_A^2(S) - \frac{|S_1|}{|S|}\Delta_A^2(S_1) - \frac{|S_2|}{|S|}\Delta_A^2(S_2) \\
&= \frac{2}{|S|}\sum_S \|x-\mu\|^2 - \frac{|S_1|}{|S|}\cdot\frac{2}{|S_1|}\sum_{S_1}\|x-\mu_1\|^2 - \frac{|S_2|}{|S|}\cdot\frac{2}{|S_2|}\sum_{S_2}\|x-\mu_2\|^2 \\
&= \frac{2}{|S|}\left\{\sum_{S_1}\left(\|x-\mu\|^2 - \|x-\mu_1\|^2\right) + \sum_{S_2}\left(\|x-\mu\|^2 - \|x-\mu_2\|^2\right)\right\} \\
&= \frac{2|S_1|}{|S|}\|\mu_1-\mu\|^2 + \frac{2|S_2|}{|S|}\|\mu_2-\mu\|^2.
\end{aligned}$$

Writing μ as a weighted average of μ_1 and μ_2 then completes the proof.

8.9 Proof of Lemma 20

By Corollary 18,

$$\Delta_A^2(S \cdot U_H) = \frac{2}{|S|}\sum_{x\in S}((x - \text{mean}(S)) \cdot U_H)^2 = 2(U_H)^T \text{cov}(S) U_H.$$

where $\text{cov}(S)$ is the covariance of data set S. This quadratic term has expectation

$$\mathbb{E}[2(U_H)^T \text{cov}(S) U_H] \;=\; \frac{2(\sigma_1^2 + \cdots + \sigma_d^2)}{d} \;\geq\; \frac{2(1-\epsilon)(\sigma_1^2 + \cdots + \sigma_D^2)}{d} \;=\; \frac{(1-\epsilon)\Delta_A^2(S)}{d}.$$

Lemma 15(a) then bounds the probability that it is much smaller than its expected value.

8.10 Proof of Lemma 23

Only the second inequality needs to be shown. To this end, consider the projection of S onto U_H (rather than $\widehat{U}_H$). Let s be the median of $S \cdot U_H$, and let the random variable $\widetilde{X}$ denote a uniform-random draw from the projected points. Lemma 20 tells us that with probability at least $1/10$ over the choice of U_H, the variance of $\widetilde{X}$ is at least $(1-\epsilon)\Delta_A^2(S)/(8d)$. We'll show that together with the tail bounds on $\widetilde{X}$ (Lemma 16), this means $\|\mu_1 - \mu_2\|^2$ cannot be too small.

Writing $Y = \widetilde{X} - s$, and using $\mathbf{1}(\cdot)$ to denote a $0-1$ indicator variable,

$$\begin{aligned}
\text{var}(\widetilde{X}) \;&\leq\; \mathbb{E}[(\widetilde{X}-s)^2] \;=\; \mathbb{E}Y^2 \;=\; \mathbb{E}[Y^2 \cdot \mathbf{1}(Y \geq 0)] + \mathbb{E}[Y^2 \cdot \mathbf{1}(Y < 0)] \\
&\leq\; \mathbb{E}[2tY \cdot \mathbf{1}(Y \geq 0)] + E[(Y-t)^2 \cdot \mathbf{1}(Y \geq t)] + \\
&\quad\; \mathbb{E}[-2tY \cdot \mathbf{1}(Y < 0)] + E[(Y+t)^2 \cdot \mathbf{1}(Y < -t)]
\end{aligned}$$

for any $t > 0$. The last inequality comes from noticing that

$$\begin{aligned}
Y^2 \cdot \mathbf{1}(Y \geq 0) \;&\leq\; 2tY \cdot \mathbf{1}(Y \geq 0) + (Y-t)^2 \cdot \mathbf{1}(Y \geq t) \\
Y^2 \cdot \mathbf{1}(Y < 0) \;&\leq\; -2tY \cdot \mathbf{1}(Y < 0) + (Y+t)^2 \cdot \mathbf{1}(Y < -t)
\end{aligned}$$

and taking expectations of both sides. This is a convenient formulation since the linear terms give us $\mu_1 - \mu_2$:

$$\mathbb{E}[2tY \cdot \mathbf{1}(Y \geq 0)] + \mathbb{E}[-2tY \cdot \mathbf{1}(Y < 0)] = t((\mu_1 \cdot U_H) - s) + t(s - (\mu_2 \cdot U_H)) = t \cdot \|U_H\| \cdot \|\mu_1 - \mu_2\|.$$

We'll use

$$t \;=\; \frac{\Delta(S)}{\sqrt{d}} \cdot \Theta\left(\sqrt{\log \frac{1}{\delta}}\right),$$

as a result of which we can be sure (with probability $1 - O(\delta)$) that the absolute value of the median is $|s| \leq t/2$ (Corollary 9). Thereafter, the two other terms can be bounded by Lemma 16:

$$E[(Y-t)^2 \cdot \mathbf{1}(Y \geq t)] + E[(Y+t)^2 \cdot \mathbf{1}(Y < -t)] \;=\; O\left(\delta \cdot \frac{\Delta^2(S)}{d}\right)$$

(with probability $1 - \delta$), whereupon

$$\frac{(1-\epsilon)\Delta_A^2(S)}{8d} \;\leq\; \text{var}(\widetilde{X}) \;\leq\; t \cdot \|U_H\| \cdot \|\mu_1 - \mu_2\| + O\left(\delta \cdot \frac{\Delta^2(S)}{d}\right).$$

The lemma now follows immediately by algebraic manipulation, using the relation $\Delta^2(S) \leq c\Delta_A^2(S)$.