Random projection trees and low dimensional manifolds

Sanjoy Dasgupta
UC San Diego
dasgupta@cs.ucsd.edu

Yoav Freund
UC San Diego
yfreund@cs.ucsd.edu

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

A k-d tree [4] 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), and suppose that the dots are points in a database, while the cross is a query point q. The cell containing q, henceforth denoted 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. In *classification*, q is assigned the majority label in its cell, or the label of its nearest neighbor in the cell. In *regression*, q is assigned the average response value in its cell. In *vector quantization*, q is replaced by the mean of the data points in the cell. Naturally, the statistical theory around k-d trees is centered on *the rate at which the diameter of individual cells drops as you move down the tree*; for details, see page 320 of [8].

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 $\mathbb{R}^{1000}$, 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 machine learning has been the realization that a lot of data which superficially lie in a very high-dimensional space $\mathbb{R}^D$, actually have low *intrinsic* dimension, in the sense of lying close to a manifold of dimension $d \ll D$. There has been significant 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 standard methods will work well. This field is quite recent and yet the literature on it is already voluminous; early foundational work includes [24, 23, 3].

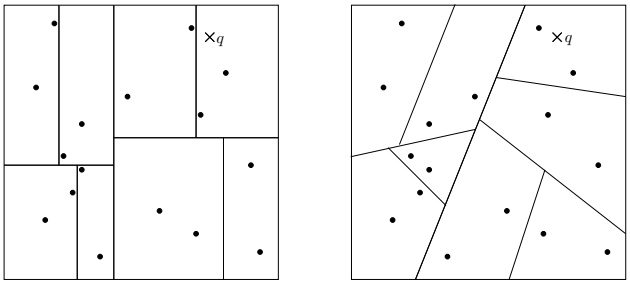


Figure 1: Left: A spatial partitioning of $\mathbb{R}^2$ induced by a k-d tree with three levels. The dots are data points; the cross marks a query point q. Right: Partitioning induced by an RP tree.

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: the bad example mentioned above has an intrinsic dimension of just $O(\log D)$. But we introduce a simple variant of k-d trees that does possess this property. Instead of splitting along coordinate directions at the median, 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 add a small amount of "jitter". We call these *random projection trees* (Figure 1, right), or RP trees for short, and we 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 on the extrinsic dimensionality (D) of the data.

STOC'08, May 17–20, 2008, Victoria, British Columbia, Canada.

2. DETAILED OVERVIEW

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

2.1 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. A common method for collecting motion data is to have a person wear a skin-tight suit with high contrast reference points printed on it. Video cameras are used to track the 3D trajectories of the reference points as the person is walking or running. In order to ensure good coverage, a typical suit has about $N = 100$ reference points. The position and posture of the body at a particular point of time is represented by a $(3N)$-dimensional vector. 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 positions of the reference points are more or less deterministic functions of these joint angles.

To take another example, a speech signal is commonly represented by a high-dimensional time series: the signal is broken into overlapping windows, and a variety of filters are applied within each window. Even richer representations can be obtained by using more filters, or by concatenating vectors corresponding to consecutive windows. Through all this, the intrinsic dimensionality remains small, because the system can be described by a few physical parameters describing the configuration of the speaker's vocal apparatus.

2.2 Intrinsic dimensionality

In this paper we explore three definitions of intrinsic dimension: Assouad dimension, manifold dimension, and local covariance dimension.

Assouad (or *doubling*) *dimension* appeared in [2].

DEFINITION 1. *For any point $x \in \mathbb{R}^D$ and any $r > 0$, let $B(x, r) = \{z : \|x - z\| \leq r\}$ denote the closed ball of radius r centered at x. The Assouad 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 $r/2$.*

This definition has proved fruitful in recent work on embeddings of metric spaces [2, 18, 17]. To relate it to manifolds, we show (Theorem 22) that the Assouad dimension of a d-dimensional Riemannian submanifold of $\mathbb{R}^D$ is $O(d)$, subject to a bound on the second fundamental form of the manifold.

Assouad dimension and manifold dimension have become common currency in the computer science literature. Yet they arose in contexts very different from data analysis, and it is not obvious that they are really the most appropriate quantities for capturing the intrinsic dimensionality of data. It is especially troubling that they seem quite resistant to empirical verification: given a sample of points drawn from an underlying distribution P, it is not easy to check whether P is concentrated near a low-dimensional manifold, or near a set of low Assouad dimension.

To address some of these qualms, we introduce a statistically motivated notion of dimension: we say that a set S has *local covariance dimension* (d, ϵ, r) if neighborhoods of radius 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 r has covariance matrix whose largest d eigenvalues satisfy $\sigma_1^2 + \cdots + \sigma_d^2 \geq (1 - \epsilon) \cdot (\sigma_1^2 + \cdots + \sigma_D^2)$.*

The intuitions behind this notion have informed some of the work on learning manifolds (for instance, [23]), but here we formalize it for the first time.

2.3 k-d trees and RP trees

Both k-d trees 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 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 $|S| < MinSize$ **return** $(Leaf)$
 $Rule \leftarrow$ 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 (typically the coordinate with largest spread) and then splits the data on its median value for that coordinate.

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

 choose a 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. Both are adaptive to intrinsic dimension, although the proofs are in different models and use 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 $x \in S$; let $y \in S$ be the farthest point from it
 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 [13]. 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 a critical respect: it occasionally performs a different kind of split, in which a cell is split into two pieces based on distance from the mean.

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
$Rule(x) := x \cdot v \leq \text{median}(\{z \cdot v : z \in S\})$
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:

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

2.4 Main results

Suppose an RP tree is built from a data set $S \subset \mathbb{R}^D$, 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 $S \cap C \subset B(x, r)$ 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 $S \subset \mathbb{R}^D$. Pick any cell C in the RP tree; suppose that $S \cap C$ has Assouad 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 radius$(C') \leq$ radius$(C)/2$.*

Our next theorem gives a result for the second type of RP-Tree. 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 $S \subset \mathbb{R}^D$. Consider any cell C of radius r, such that $S \cap C$ has local covariance dimension (d, ϵ, r), where $\epsilon < c_1$. 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, $\mathbb{E}[\Delta(S \cap C')] \leq c_2 \Delta(S \cap C)$.*
- *If C is split by projection, then $\mathbb{E}\left[\Delta_A^2(S \cap C')\right] \leq (1 - (c_3/d))\, \Delta_A^2(S \cap C)$.*

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

Consider $S \subset \mathbb{R}^D$ made up of the coordinate axes between -1 and 1: $S = \bigcup_{i=1}^D \{te_i : -1 \leq t \leq 1\}$. Here $e_1, \ldots, e_D$ is the canonical basis of $\mathbb{R}^D$. There are many application domains, such as text, in which data is *sparse*; this example is an extreme case.

S lies within $B(0, 1)$ and can be covered by $2D$ balls of radius $1/2$. It is not hard to see that the Assouad dimension of S 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.

2.6 Connections to other work

Uses of k-d trees

As described in the introduction, the use of k-d trees for classification, regression, near neighbor search, and vector quantization leads to rates of convergence that depend on the rate at which the diameter of cells decreases down the tree. Based on our results, RP trees might considerably extend the scope of these methods, from data that is low dimensional to data that is just intrinsically low dimensional. [12] contains experimental results in this direction.

A related problem is *nearest* neighbor search, for which k-d trees are commonly used. Here, the criterion governing the efficacy of search is harder to make precise. Interestingly, some state-of-the-art practical work on tree-based nearest neighbor search [21] uses random projection as a preprocessing step. Another notable use of random projections in this context is locality-sensitive hashing [14]. Also relevant is work on other tree structures with complexity guarantees for nearest neighbor search [1, 20, 5]. It would be interesting if similar guarantees could be shown for a data structure as simple as ours.

Vector quantization

Vector quantization [16] is a basic building block of lossy data compression. Here, random vectors X are generated from some distribution P over $\mathbb{R}^D$, and the goal is to pick a finite codebook $C \subset \mathbb{R}^D$ and an encoding function $\alpha : \mathbb{R}^D \to C$ such that $\mathbb{E}_P \|X - \alpha(X)\|^2$ is small.

Ideally we'd let $\alpha(x)$ be the nearest neighbor of x in C, but often (in audio or video compression) the number of codewords is so enormous that this nearest neighbor computation cannot be performed in real time. A more efficient scheme is to have the codewords arranged in a tree [7]: there is a partition of space like a k-d tree, and each x is mapped to the mean value in cell(x). Our Theorem 4 shows that the vector quantization error of RP trees behaves like $e^{-O(r/d)}$, where r is the depth of the tree and d is the intrinsic dimension. There is a substantial body of work that obtains rates for vector quantization, and as one may expect, these turn out to be of the form $e^{-r/D}$ [15].

Compressed sensing

The field of compressed sensing has grown out of the surprising realization that high-dimensional sparse data can be accurately reconstructed from just a few random projections [6, 10]. The central premise of this research area is that the original data thus never even needs to be collected: all one ever sees are the random projections.

RP trees are similar in spirit and entirely compatible with this viewpoint. Theorem 4 holds even if the random pro-

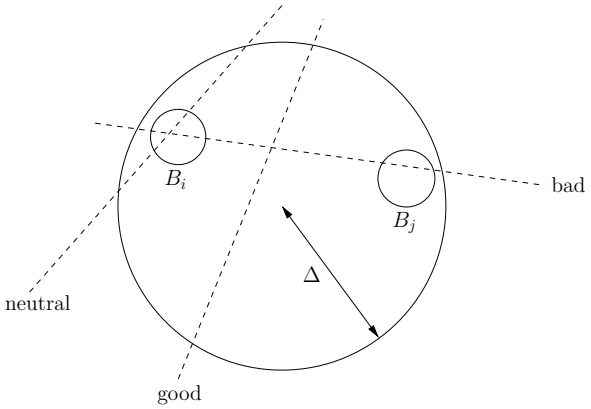


Figure 2: Cell C of the RP tree is contained in a ball of radius Δ. Balls B_i and B_j are part of a cover of this cell, and have radius $\Delta/\sqrt{d}$. For this pair of balls, there are three kinds of splits: good, bad, and neutral.

jections are forced to be the same across each entire level of the tree. For a tree of depth k, this means only k random projections are ever needed, and these can be computed beforehand (the split-by-distance can be reworked to operate in the projected space rather than the high-dimensional space). The data are not accessed in any other way.

3. AN RPTREE-MAX ADAPTS TO ASSOUAD DIMENSION

In this section, we prove Theorem 3. A rough outline is as follows. Suppose an RP tree is built using data set $S \subset \mathbb{R}^D$ of Assouad dimension d, and that C is some cell of the tree. If $S \cap C$ lies in a ball of radius Δ, then we need to show that after $O(d \log d)$ further levels of splitting, each resulting cell is contained in a ball of radius $\leq \Delta/2$. To this end, we start by covering $S \cap C$ with balls $B_1, B_2, \ldots, B_N$ of radius $\Delta/\sqrt{d}$. The Assouad dimension tells us $N = O(d^{d/2})$ suffices. 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 and B_j will lie entirely on opposite sides of the split. There are at most N^2 such pairs i, j, so after $\Theta(d \log d)$ projections every one of these pairs will have been split. Thus, $\Theta(d \log d)$ 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 B_i and B_j, we say that a split is *good* if it completely separates them. 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 good splits are more likely than bad ones.

To lower-bound the probability of a good split, let $\widetilde{B}_i$ and $\widetilde{B}_j$ be the projections of B_i and B_j onto a random line. We show that with constant probability the following events occur: (1) $\widetilde{B}_i$ and $\widetilde{B}_j$ have a certain amount of space between them. (2) The median of the projected data lies very close to this space. (3) Picking a split point at random near the median will separate $\widetilde{B}_i$ from $\widetilde{B}_j$.

3.1 Gross statistics of projected data

We choose random projections from $\mathbb{R}^D$ to $\mathbb{R}$ by picking $U \sim N(0, (1/D)I_D)$ (multivariate Gaussian) and mapping $x \mapsto x \cdot U$. The key property of such a projection is that it approximately preserves the lengths of vectors, modulo a scaling factor of $\sqrt{D}$. This is summarized below.

LEMMA 5. *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) $\mathbb{P}\left[|U \cdot x| \leq \alpha \cdot \frac{\|x\|}{\sqrt{D}}\right] \leq \sqrt{\frac{2}{\pi}}\,\alpha$*; and*

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

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 space.

Suppose $S \subset \mathbb{R}^D$ has Assouad dimension d. Let $\widetilde{S} = S \cdot U$ be its random projection into $\mathbb{R}$. How does $\mathrm{diam}(\widetilde{S})$ compare to $\mathrm{diam}(S)$? (Here $\mathrm{diam}(S) = \sup_{x,y \in S} \|x - y\|$.) Clearly $\mathrm{diam}(\widetilde{S}) \leq \|U\| \cdot \mathrm{diam}(S)$, but we would in fact expect it to be much smaller if $d \ll D$. In fact, $\mathrm{diam}(\widetilde{S}) \leq \mathrm{diam}(S) \cdot O(\sqrt{d/D})$; the following is adapted from an argument due to [19].

LEMMA 6. *Suppose set $S \subset \mathbb{R}^D$ is contained in a ball $B(x_0, \Delta)$ and has Assouad dimension d. Let $\widetilde{S}$ denote the random projection of S into $\mathbb{R}$. Then for any $0 < \delta < 1$, with probability $> 1 - \delta$ over the choice of projection, $\widetilde{S}$ lies in an interval of radius $4 \cdot \frac{\Delta}{\sqrt{D}} \cdot \sqrt{2\left(d + \ln \frac{2}{\delta}\right)}$ centered at $\widetilde{x}_0$.*

Thus, S 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 $O(\Delta/\sqrt{D})$.

LEMMA 7. *Suppose $S \subset \mathbb{R}^D$ lies within 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 S. Then with probability $> 1 - \delta$ over the choice of random projection onto $\mathbb{R}$, all but an ϵ fraction of $\widetilde{S}$ (in μ-measure) lies within distance $\sqrt{2 \ln \frac{1}{\delta\epsilon}} \cdot \frac{\Delta}{\sqrt{D}}$ of $\widetilde{x}_0$.*

It follows that the median of the projected points also lies in this central interval; take μ to be the uniform distribution over S and use $\epsilon = 1/2$.

COROLLARY 8. *Under the hypotheses of Lemma 7, for any $0 < \delta < 2/e^2$, with probability at least $1 - \delta$ over the choice of projection: $|\mathrm{median}(\widetilde{S}) - \widetilde{x}_0| \leq \frac{\Delta}{\sqrt{D}} \cdot \sqrt{2 \ln \frac{2}{\delta}}$.*

3.2 The probability of good and bad splits

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

LEMMA 9. *Say $S \subset B(x_0, \Delta)$ has Assouad dimension $d \geq 1$. Pick balls $B = B(z, r)$ and $B' = B(z', r)$ such that*

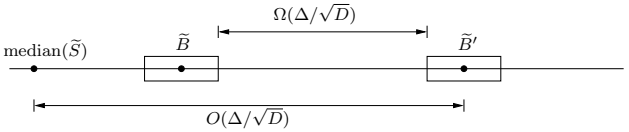
- *their centers z and z' lie in $B(x_0, \Delta)$,*
- *the distance between these centers is $\|z - z'\| \geq \frac{1}{2}\Delta - r$,*
- *and the radius r is at most $\Delta/(512\sqrt{d})$.*

Now pick a random projection U, that sends S to (say) $\widetilde{S} \subset \mathbb{R}$, and then pick a split point at random in the range median$(\widetilde{S}) \pm (6\Delta/\sqrt{D})$. With probability at least $1/192$ over the choice of U and the split point, $S \cap B$ and $S \cap B'$ will be contained in separate halves of the split.

PROOF. (Sketch.) Let $\widetilde{B}$ and $\widetilde{B}'$ be the projections of $S \cap B$ and $S \cap B'$. It follows from Lemmas 5 and 6 and Corollary 8 that with probability at least $1/2$, the random projection U will satisfy the following 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 $\widetilde{S}$ lies within distance $3\Delta/\sqrt{D}$ of $\widetilde{x}_0$.

In this case, we say U is "good", and the following picture of $\widetilde{S}$ is valid:



The "sweet spot" is the region between $\widetilde{B}$ and $\widetilde{B}'$; if the split point falls in it, 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 (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$(\widetilde{S})$. Thus, under the sampling strategy from the lemma statement, there is a constant probability of hitting the sweet spot. Putting it all together,

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

as claimed. ∎

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 10. *Under the hypotheses of Lemma 9,*

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

3.3 Proof of Theorem 3

Finally, we complete the proof of Theorem 3.

LEMMA 11. *Suppose $S \subset \mathbb{R}^D$ has Assouad 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 $S \cap C \subset B(x_0, \Delta)$. Cover this set by balls of radius $r = \Delta/(512\sqrt{d})$; the Assouad dimension tells us that $N = (O(d))^d$ balls suffice. Now, fix any pair of balls B, B' from this cover whose centers are at distance at least $\Delta/2 - r$ from one another; and, 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 9, $p_1 \leq 191/192$. 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 9 and 10 then tell us that

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

The three cases in the first inequality correspond 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 faraway pairs of balls from the cover. ∎

4. 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 (that is, average interpoint distance), then it is split according to the distances of points from the mean. Otherwise, a random projection is used.

4.1 Splitting by distance from the mean

This is invoked when the points in the current cell, call them S, satisfy $\Delta^2(S) > c\Delta_A^2(S)$ (recall that $\Delta(S)$ is the diameter of S while $\Delta_A^2(S)$ is the average interpoint distance).

LEMMA 12. *Suppose that $\Delta^2(S) > c\Delta_A^2(S)$. Let S_1 be the points in S whose distance to 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 $\frac{|S_1|}{|S|}\Delta^2(S_1) + \frac{|S_2|}{|S|}\Delta^2(S_2) \leq \left(\frac{1}{2} + \frac{2}{c}\right)\Delta^2(S)$.*

4.2 Splitting by projection: proof outline

Suppose the current cell contains a set of points $S \subset \mathbb{R}^D$ with $\Delta^2(S) \leq c\Delta_A^2(S)$. We 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 three parts:

I. Suppose S is split into S_1 and S_2, with means μ_1 and μ_2. Then the reduction in average diameter can be expressed in a remarkably simple form, as a multiple of $\|\mu_1 - \mu_2\|^2$.

II. Next, we give a lower bound on the distance between the *projected* means, $(\widetilde{\mu}_1 - \widetilde{\mu}_2)^2$. We show that the distribution of the projected points is subgaussian with variance $O(\Delta_A^2(S)/D)$. This well-behavedness implies that $(\widetilde{\mu}_1 - \widetilde{\mu}_2)^2 = \Omega(\Delta_A^2(S)/D)$.

III. We finish by showing that, approximately, $\|\mu_1-\mu_2\|^2 \geq (D/d)(\widetilde{\mu}_1-\widetilde{\mu}_2)^2$. This is because $\mu_1-\mu_2$ lies close to the subspace spanned by the top d eigenvectors of the covariance matrix of S; and with high probability, *every* vector in this subspace shrinks by $O(\sqrt{d/D})$ when projected on a random line.

We now tackle these three parts of the proof in order.

4.3 Quantifying the reduction in average diameter

The average squared diameter $\Delta_A^2(S)$ has certain reformulations that make it 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 13. *Let X, Y be independent and identically distributed random variables in $\mathbb{R}^n$, and fix any $z \in \mathbb{R}^n$.*

(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 $Y = y$. ∎

This can be used to show that the averaged squared diameter, $\Delta_A^2(S)$, is twice the average squared distance of points in S from their mean.

COROLLARY 14. *The average squared diameter of a set S can also be written as: $\Delta_A^2(S) = \frac{2}{|S|}\sum_{x\in S}\|x - mean(S)\|^2$.*

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. ∎

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

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

This completes part I of the proof outline.

4.4 Properties of the projected data

Projection from $\mathbb{R}^D$ into $\mathbb{R}^1$ shrinks the average squared diameter of a data set by roughly D. 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^TAU. We then observe that for random U, such quadratic forms are concentrated about their expected values.

LEMMA 16. *Pick $U \sim N(0, (1/D)I_D)$. For any $S \subset \mathbb{R}^D$, with probability at least $1/10$, the projection of S onto U has average squared diameter $\Delta_A^2(S\cdot U) \geq \Delta_A^2(S)/(4D)$.*

PROOF. By Corollary 14,

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

where $\text{cov}(S)$ is the covariance of data set S. This quadratic term has expectation (over the choice of U) $\mathbb{E}[2U^T\text{cov}(S)U] = 2\sum_{i,j}\mathbb{E}[U_iU_j]\text{cov}(S)_{ij} = \frac{2}{D}\sum_i \text{cov}(S)_{ii} = \Delta_A^2(S)/D$. Lemma 23(a) then bounds the concentration of $U^T\text{cov}(S)U$ around its expected value. ∎

Next, we examine the overall distribution of the projected points. When $S \subset \mathbb{R}^D$ has diameter Δ, its projection into the line can have diameter upto Δ, but as we saw in Lemma 7, most of it will lie within a central interval of size $O(\Delta/\sqrt{D})$. Now we characterize the distribution more precisely.

LEMMA 17. *Suppose $S \subset B(0,\Delta) \subset \mathbb{R}^D$. Pick $\delta > 0$ and $U \sim N(0,(1/D)I_D)$. With probability $\geq 1-\delta$, $S\cdot U = \{x\cdot U : x \in S\}$ satisfies the following property for all positive integers k: The fraction of points outside the interval $\left(-k\Delta/\sqrt{D}, +k\Delta/\sqrt{D}\right)$ is at most $(2^k/\delta)\cdot e^{-k^2/2}$.*

PROOF. Apply Lemma 7 for each k (with failure probability $\delta/2^k$) and take a union bound. ∎

Finally, we examine what happens when a d-dimensional linear subspace of $\mathbb{R}^D$ is projected into $\mathbb{R}^1$. We show a uniform bound over all vectors in the subspace.

LEMMA 18. *There exists $\kappa > 0$ with the following property. Fix any $\delta > 0$ and any d-dimensional subspace $H \subset \mathbb{R}^D$. Pick $U \sim N(0,(1/D)I_D)$. Then with probability at least $1-\delta$ over the choice of U,*

$$\sup_{x\in H}\frac{|x\cdot U|^2}{\|x\|^2} \;\leq\; \kappa\cdot\frac{d+\ln 1/\delta}{D}.$$

PROOF. Apply Lemma 6 to the intersection of H with the surface of the unit sphere in $\mathbb{R}^D$. This set has Assouad dimension $O(d)$. ∎

4.5 Distance between projected means

We are dealing with the case when $\Delta^2(S) \leq c\cdot\Delta_A^2(S)$, that is, the diameter of set S is at most a constant factor times the average interpoint distance. If S is projected onto a random direction, the projected points will have variance about $\Delta_A^2(S)/D$, by Lemma 16; and by Lemma 17, it isn't too far from the truth to think of these points as having roughly a Gaussian distribution. Thus, if the projected points are split into two groups at the mean, we would expect the means of these two groups to be separated by a distance of about $\Delta_A(S)/\sqrt{D}$. Indeed, this is the case. The same holds if we split at the median, which isn't all that different from the mean for close-to-Gaussian distributions.

LEMMA 19. *There is a constant κ_2 for which the following holds. Pick any $0 < \delta < 1/16c$. Pick $U \sim N(0, (1/D)I_D)$ and split S into two pieces:*

$$S_1 = \{x \in S : x \cdot U < s\} \quad \text{and} \quad S_2 = \{x \in S : x \cdot U \geq s\},$$

where s is either mean$(S \cdot U)$ or median$(S \cdot U)$. Write $p = |S_1|/|S|$, and let $\widetilde{\mu}_1$ and $\widetilde{\mu}_2$ denote the means of $S_1 \cdot U$ and $S_2 \cdot U$, respectively. Then with probability at least $1/10 - \delta$,

$$(\widetilde{\mu}_2 - \widetilde{\mu}_1)^2 \;\geq\; \kappa_2 \cdot \frac{1}{(p(1-p))^2} \cdot \frac{\Delta_A^2(S)}{D} \cdot \frac{1}{c\,\log(1/\delta)}.$$

PROOF. Let the r.v. $\widetilde{X}$ be a uniform-random draw from the projected points $S \cdot U$. Without loss of generality S has mean 0, so $\mathbb{E}\widetilde{X} = 0$ and thus $p\widetilde{\mu}_1 + (1-p)\widetilde{\mu}_2 = 0$. Rearranging, $\widetilde{\mu}_1 = -(1-p)(\widetilde{\mu}_2 - \widetilde{\mu}_1)$ and $\widetilde{\mu}_2 = p(\widetilde{\mu}_2 - \widetilde{\mu}_1)$.

We already know from Lemma 16 (and Corollary 14) that with probability at least $1/10$, the variance of the projected points is significant: $\text{var}(\widetilde{X}) \geq \Delta_A^2(S)/8D$. We'll show this implies a similar lower bound on $(\widetilde{\mu}_2 - \widetilde{\mu}_1)^2$.

Using $\mathbf{1}(\cdot)$ to denote $0-1$ indicator variables, for any $t > 0$,

$$\begin{aligned}\text{var}(\widetilde{X}) &\leq \mathbb{E}[(\widetilde{X} - s)^2] \\ &\leq \mathbb{E}[2t|\widetilde{X} - s| + (|\widetilde{X} - s| - t)^2 \cdot \mathbf{1}(|\widetilde{X} - s| \geq t)]\end{aligned}$$

This is convenient since the linear term gives us $\widetilde{\mu}_2 - \widetilde{\mu}_1$:

$$\begin{aligned}\mathbb{E}[2t|\widetilde{X} - s|] &= 2t(p(s - \widetilde{\mu}_1) + (1-p)(\widetilde{\mu}_2 - s)) \\ &= 4t \cdot p(1-p) \cdot (\widetilde{\mu}_2 - \widetilde{\mu}_1) + 2ts(2p-1).\end{aligned}$$

The last term vanishes since the split is either at the mean of the projected points, in which case $s = 0$, or at the median, in which case $p = 1/2$.

Next, we'll choose $t = t_o(\Delta(S)/\sqrt{D}) \cdot \sqrt{\log(1/\delta)}$ for some suitable constant t_o, so that the quadratic term in $\text{var}(\widetilde{X})$ can be bounded using Lemma 17 and Corollary 8: with probability at least $1 - \delta$, $E[(|\widetilde{X}| - t)^2 \cdot \mathbf{1}(|\widetilde{X}| \geq t)] \;\leq\; \delta \cdot (\Delta^2(S)/D)$ (a simple integration). Putting things together,

$$\frac{\Delta_A^2(S)}{8D} \;\leq\; \text{var}(\widetilde{X}) \;\leq\; 4t \cdot p(1-p) \cdot (\widetilde{\mu}_2 - \widetilde{\mu}_1) + \delta \cdot \frac{\Delta^2(S)}{D}.$$

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

4.6 Distance between high-dimensional means

Split S into two pieces as in the setting of Lemma 19, and let μ_1 and μ_2 denote the means of S_1 and S_2, respectively. We already have a lower bound on the distance between the projected means, $\widetilde{\mu}_2 - \widetilde{\mu}_1$; we will now show that $\|\mu_2 - \mu_1\|$ is larger than this by a factor of about $\sqrt{D/d}$. The main technical difficulty here is the dependence between the μ_i and the projection U. Incidentally, this is the only part of the entire argument that exploits intrinsic dimensionality.

LEMMA 20. *There is a constant κ_3 with the following property. Suppose set $S \subset \mathbb{R}^D$ is such that the top d eigenvalues of cov(S) account for more than $1 - \epsilon$ of its trace. Pick a random vector $U \sim N(0, (1/D)I_D)$, and split S into two pieces, S_1 and S_2, in any fashion (which may depend upon U). Let $p = |S_1|/|S|$. Let μ_1 and μ_2 be the means of S_1 and S_2, and $\widetilde{\mu}_1$ and $\widetilde{\mu}_2$ the means of $S_1 \cdot U$ and $S_2 \cdot U$. Then for any $\delta > 0$, with probability $\geq 1 - \delta$ over the choice of U,*

$$\|\mu_2 - \mu_1\|^2 \;\geq\; \frac{\kappa_3 D}{d + \ln 1/\delta}\left((\widetilde{\mu}_2 - \widetilde{\mu}_1)^2 - \frac{4}{p(1-p)}\frac{\epsilon\Delta_A^2(S)}{\delta D}\right).$$

PROOF. Assume without loss of generality that S has zero mean. Let H be the subspace spanned by the top d eigenvectors of cov(S), and let $H^\perp$ be its orthogonal subspace. Write any point $x \in \mathbb{R}^D$ as $x_H + x_\perp$, where each component is a vector in $\mathbb{R}^D$ that lies in the respective subspace.

Pick the random vector U; with probability $\geq 1 - \delta$ it satisfies the following two properties.

Property 1: For some constant $\kappa' > 0$, for every $x \in \mathbb{R}^D$

$$|x_H \cdot U|^2 \;\leq\; \|x_H\|^2 \cdot \kappa' \cdot \frac{d + \ln 1/\delta}{D} \;\leq\; \|x\|^2 \cdot \kappa' \cdot \frac{d + \ln 1/\delta}{D}.$$

This holds (with probability $1 - \delta/2$) by Lemma 18.

Property 2: Letting X be a uniform-random draw from S,

$$\begin{aligned}\mathbb{E}_X[(X_\perp \cdot U)^2] &\leq \frac{2}{\delta} \cdot \mathbb{E}_U\mathbb{E}_X[(X_\perp \cdot U)^2] \\ &= \frac{2}{\delta} \cdot \mathbb{E}_X\mathbb{E}_U[(X_\perp \cdot U)^2] \\ &= \frac{2}{\delta D} \cdot \mathbb{E}_X[\|X_\perp\|^2] \;\leq\; \frac{\epsilon\Delta_A^2(S)}{\delta D}.\end{aligned}$$

The first step is Markov's inequality, and holds with probability $1 - \delta/2$. The last inequality comes from the local covariance condition.

So assume the two properties hold. Writing $\mu_2 - \mu_1$ as $(\mu_{2H} - \mu_{1H}) + (\mu_{2\perp} - \mu_{1\perp})$,

$$\begin{aligned}(\widetilde{\mu}_2 - \widetilde{\mu}_1)^2 &= ((\mu_{2H} - \mu_{1H}) \cdot U + (\mu_{2\perp} - \mu_{1\perp}) \cdot U)^2 \\ &\leq 2((\mu_{2H} - \mu_{1H}) \cdot U)^2 + 2((\mu_{2\perp} - \mu_{1\perp}) \cdot U)^2.\end{aligned}$$

The first term can be bounded by Property 1:

$$((\mu_{2H} - \mu_{1H}) \cdot U)^2 \;\leq\; \|\mu_2 - \mu_1\|^2 \cdot \kappa' \cdot \frac{d + \ln 1/\delta}{D}.$$

For the second term, let $\mathbb{E}_X$ denote expectation over X chosen uniformly at random from S. Then

$$\begin{aligned}&((\mu_{2\perp} - \mu_{1\perp}) \cdot U)^2 \\ &\leq 2(\mu_{2\perp} \cdot U)^2 + 2(\mu_{1\perp} \cdot U)^2 \\ &= 2(\mathbb{E}_X[X_\perp \cdot U \mid X \in S_2])^2 + 2(\mathbb{E}_X[X_\perp \cdot U \mid X \in S_1])^2 \\ &\leq 2\mathbb{E}_X[(X_\perp \cdot U)^2 \mid X \in S_2] + 2\mathbb{E}_X[(X_\perp \cdot U)^2 \mid X \in S_1] \\ &\leq \frac{2}{1-p} \cdot \mathbb{E}_X[(X_\perp \cdot U)^2] + \frac{2}{p} \cdot \mathbb{E}_X[(X_\perp \cdot U)^2] \\ &= \frac{2}{p(1-p)}\mathbb{E}_X[(X_\perp \cdot U)^2] \;\leq\; \frac{2}{p(1-p)} \cdot \frac{\epsilon\Delta_A^2(S)}{\delta D}.\end{aligned}$$

by Property 2. The lemma follows by putting the various pieces together. ∎

We can now finish off the proof of Theorem 4.

THEOREM 21. *Fix any $\epsilon \leq O(1/c)$. Suppose set $S \subset \mathbb{R}^D$ has the property that the top d eigenvalues of cov(S) account for more than $1 - \epsilon$ of its trace. Pick a random vector $U \sim N(0, (1/D)I_D)$ and split S into two parts,*

$$S_1 = \{x \in S : x \cdot U < s\} \quad \text{and} \quad S_2 = \{x \in S : x \cdot U \geq s\},$$

where s is either mean$(S \cdot U)$ or median$(S \cdot U)$. Then with probability $\Omega(1)$, the expected average diameter shrinks by $\Omega(\Delta_A^2(S)/cd)$.

PROOF. By Lemma 15, the reduction in expected average diameter is $2p(1-p)\|\mu_1 - \mu_2\|^2$, in the language of Lemmas 19 and 20. The rest follows from those lemmas. ∎

Acknowledgements

Dasgupta acknowledges the support of the National Science Foundation under grants IIS-0347646 and IIS-0713540.

5. REFERENCES

[1] S. Arya, D. Mount, N. Netanyahu, R. Silverman, and A. Wu. An optimal algorithm for approximate nearest neighbor searching. *Journal of the ACM*, 45:891–923, 1998.

[2] P. Assouad. Plongements lipschitziens dans r^n. *Bull. Soc. Math. France*, 111(4):429–448, 1983.

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

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

[5] A. Beygelzimer, S. Kakade, and J. Langford. Cover trees for nearest neighbor. In *23rd International Conference on Machine Learning*, 2006.

[6] E. Candes and T. Tao. Near optimal signal recovery from random projections: universal encoding strategies? *IEEE Transactions on Information Theory*, 52(12):5406–5425, 2006.

[7] P. Chou, T. Lookabaugh, and R. Gray. Optimal pruning with applications to tree-structured source coding and modeling. *IEEE Transactions on Information Theory*, 35(2):299–315, 1989.

[8] L. Devroye, L. Gyorfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

[9] M. do Carmo. *Riemannian Geometry*. Birkhauser, 1992.

[10] D. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

[11] R. Durrett. *Probability: Theory and Examples*. Duxbury, second edition, 1995.

[12] Y. Freund, S. Dasgupta, M. Kabra, and N. Verma. Learning the structure of manifolds using random projections. In *Neural Information Processing Systems*, 2007.

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

[14] A. Gionis, P. Indyk, and R. Motwani. Similarity search in high dimensions via hashing. In *25th International Conference on Very Large Databases*, 1999.

[15] S. Graf and H. Luschgy. *Foundations of quantization for probability distributions*. Springer, 2000.

[16] R. Gray and D. Neuhoff. Quantization. *IEEE Transactions on Information Theory*, 44(6):2325–2383, 1998.

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

[18] J. Heinonen. *Lectures on Analysis on Metric Spaces*. Springer, 2001.

[19] P. Indyk and A. Naor. Nearest neighbor preserving embeddings. *ACM Transactions on Algorithms*, 3(3), 2007.

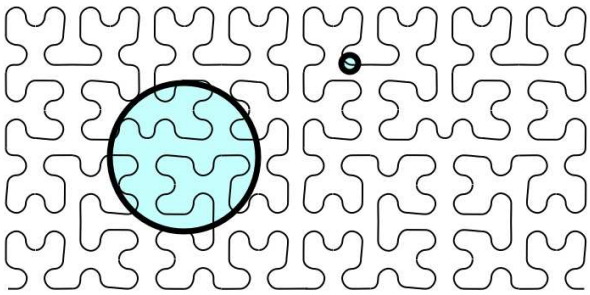


Figure 3: Hilbert's space filling curve: a 1D manifold that has Assouad dimension 2, when the radius of balls is larger than the curvature of the manifold.

[20] R. Krauthgamer and J. Lee. Navigating nets: simple algorithms for proximity search. In *ACM-SIAM Symposium on Discrete Algorithms*, 2004.

[21] T. Liu, A. Moore, A. Gray, and K. Yang. An investigation of practical approximate nearest neighbor algorithms. In *Neural Information Processing Systems*, 2004.

[22] P. Niyogi, S. Smale, and S. Weinberger. Finding the homology of submanifolds with high confidence from random samples. *Discrete and Computational Geometry*, 2006.

[23] S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, (290):2323–2326, 2000.

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

APPENDIX

A. ASSOUAD DIMENSION OF A SMOOTH MANIFOLD

If M is a d-dimensional Riemannian submanifold of $\mathbb{R}^D$, what is its Assouad dimension? An easy case is when M is an affine set, in which case it has the same Assouad dimension as $\mathbb{R}^d$, namely $O(d)$. We may expect that for more general M, the same holds true of *small enough neighborhoods*.

Recall that we define balls with respect to Euclidean distance in $\mathbb{R}^D$ rather than geodesic distance on M. If a neighborhood $M \cap B(x, r)$ has high *curvature* (speaking informally), then it could potentially have large Assouad dimension. For instance, it 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 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 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 *second fundamental form* (chapter 6 of [9]). 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. 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 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 can add a factor of $\log N$ to the Assouad dimension, making it $O(d + \log N)$.

Almost all the technical details needed to bound the Assouad dimension of manifolds appear in a separate context in [22]. Here we just put them together differently.

THEOREM 22. *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 make use of two facts proved in [22].

Fact 1 (Lemma 5.4 of [22]). 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 from

Fact 2 (implicit in proof of Lemma 5.3 of [22]). 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'. ∎

B. VARIOUS PROOFS

B.1 Proof of Lemma 5

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 [11].

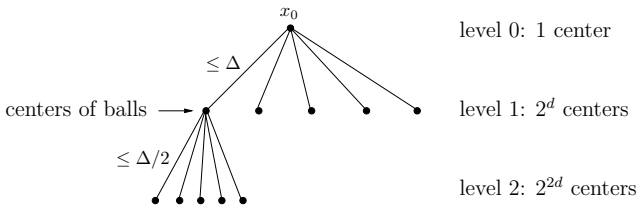


Figure 4: A hierarchy of covers. At level k, there are 2^{kd} points in the cover. Each of them has distance $\leq \Delta/2^k$ to its children (which constitute the cover at level $k+1$). At the leaves are individual points of S.

B.2 Proof of Lemma 6

Pick a cover of $S \subset B(x_0, \Delta)$ 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 $S \cap B$; 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 4).

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 5(b) that

$$\begin{aligned}
&\mathbb{P}\left[|\widetilde{u} - \widetilde{v}| \geq \beta \cdot \frac{\Delta}{\sqrt{D}} \cdot \left(\frac{3}{4}\right)^k\right] \\
&\leq \mathbb{P}\left[|\widetilde{u} - \widetilde{v}| \geq \beta \left(\frac{3}{2}\right)^k \cdot \frac{\|u-v\|}{\sqrt{D}}\right] \\
&\leq \frac{2}{\beta}\left(\frac{2}{3}\right)^k \exp\left(-\frac{\beta^2}{2} \cdot \left(\frac{3}{2}\right)^{2k}\right) \leq \frac{\delta}{\beta}\left(\frac{\delta}{3}\right)^k e^{-(k+1)d}
\end{aligned}$$

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

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

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 $\widetilde{S}$ has a distance from $\widetilde{x}_0$ of at most $\beta \cdot \frac{\Delta}{\sqrt{D}} \cdot \left(1 + \frac{3}{4} + (\frac{3}{4})^2 + \cdots\right) = \frac{4\beta\Delta}{\sqrt{D}}$. Plugging in the value of β then yields the lemma.

B.3 Proof of Lemma 7

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 5(b),

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

Since this holds for any $x \in S$, 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

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

B.4 Proof of Lemma 9

It will help to define the failure probabilities $\delta_1 = 2/e^{31}$ and $\delta_2 = 1/20$. In the proof sketch above, we defined four properties that make a projection U "good". We now verify that they all hold with probability at least 1/2.

Property (1) follows by applying Lemma 6 to each ball in turn. For B, we have that with probability at least $1-\delta_1$, $\widetilde{B}$ is within radius $(4r/\sqrt{D})\cdot\sqrt{2(d+\ln(2/\delta_1))} \;\leq\; (\Delta/128\sqrt{D})\cdot\sqrt{2\ln(2e/\delta_1)} \;=\; \Delta/(16\sqrt{D})$ of $\widetilde{z}$. Similarly with B', so this property holds with probability at least $1-2\delta_1$.

(2) follows from Lemma 5(a); specifically, it fails with probability at most $4\alpha/5$ for $\alpha = 1/(2-(4r/\Delta)) \leq 128/255$.

Property (3) is from Lemma 5(b), with probability at least $1-2\delta_2/\sqrt{2\ln(2/\delta_2)}$ (in that lemma, use $\beta = \sqrt{2\ln(2/\delta_2)}$).

Finally, (4) holds with probability $\geq 1-\delta_2$ by Corollary 8.

B.5 Proof of Lemma 10

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 occur: first, $\widetilde{B}$ must intersect $\widetilde{B}'$; second, the split point must 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.

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

by Lemma 5(a) and the conditions on r.

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

So the probability $\widetilde{B}, \widetilde{B}'$ both intersect the split point is at most $2\delta_1 + \mathbb{P}[\widetilde{B}, \widetilde{B}' \text{ touch}]\cdot\mathbb{P}[\text{split point touches } \widetilde{B}] < 1/384$.

B.6 Proof of Lemma 12

Let random variable X be distributed uniformly over S. $\mathbb{P}\left[\|X-\mathbb{E}X\|^2 \geq \text{median}(\|X-\mathbb{E}X\|^2)\right] \geq 1/2$ 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 14 that

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

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 $\Delta^2(S) > c\Delta_A^2(S)$.

B.7 Proof of Lemma 15

Let μ, μ_1, μ_2 denote the means of S, S_1, and S_2. Using Corollary 14 and Lemma 13(a), we have

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

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

B.8 Concentration of quadratic forms

LEMMA 23. *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) $\mathbb{P}[U^TAU < \alpha\cdot\mathbb{E}[U^TAU]] \leq e^{-((1/2)-\alpha)/2}$, *and*

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

PROOF. This follows by examining the moment-generating function of U^TAU. 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. 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 $U^TAU = \sum_i a_iU_i^2 = (1/n)\sum_i a_iZ_i^2$, and $\mathbb{E}[U^TAU] = 1/n$.

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

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

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

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

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