UNIVERSITY OF CALIFORNIA, SAN DIEGO

**Learning From Data With Low Intrinsic Dimension**

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Computer Science

by

Nakul Verma

Committee in charge:

Professor Sanjoy Dasgupta, Chair
Professor Charles Elkan
Professor Gert Lanckriet
Professor Justin Roberts
Professor Lawrence Saul

2012

The dissertation of Nakul Verma is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

Chair

University of California, San Diego

2012

DEDICATION

To my parents.

# EPIGRAPH

*If the facts don't fit the theory,*
*change the facts.*
—Albert Einstein

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

The journey of a Ph.D. student is typically considered a solo endeavor. However while taking this journey, you inevitabley meet people who have made a profound impact in your life. I was fortunate enough to meet several people who had a positive influence on me, and made my grad school experience enjoyable.

First and foremost, I would like to thank my advisor Professor Sanjoy Dasgupta for his support and unending patience with me over the years. Knowingly and unknowingly, he has taught me a tremendous amount and made me a better researcher.

I would like to thank my parents and my sister for their encouragement. I cannot imagine completing my Ph.D. without their support.

I would particularly like to thank my friends Boris Babenko, Eric Christiansen, Daniel Hsu, Ajay Iyengar, Mayank Kabra, Ming Kawaguchi, Bharath Kumar, Samory Kpotufe, Brian McFee, Aditya Menon, Brian Ryujin, Matus Telgarsky, Kai Wang and Ruixin Yang, who made grad school fun.

Portions of this dissertation are based on papers that have been published in various conferences, several of which I have co-authored with others. Listed below are the details along with my contributions.

Chapter 3, in full, is a reprint of material from the paper "A note on random projections for preserving paths on a manifold" by N. Verma in UC San Diego Tech. Report CS2011-0971. The dissertation author is the primary investigator.

Chapter 4, in full, is a reprint of material from the paper "Distance preserving embeddings for general $n$-dimensional manifolds" by N. Verma in Conference on Learning Theory. The dissertation author is the primary investigator.

Chapter 5, in full, is a reprint of material from the paper "Multiple instance learning wiht manifold bags" by B. Babenko, N. Verma, P. Dollár and S. Belongie in International Conference in Machine Learning. The dissertation author contributed towards the theoretical analysis and the writeup.

Chapters 6 and 7, in full, are a reprint of material from the papers "Learning the structure of manifolds using random projections" by Y. Freund, S. Dasgupta, M. Kabra and N. Verma in Neural Information Processing Systems, and

"Which spatial partition trees are adaptive to intrinsic dimension?" by N. Verma, S. Kpotufe and S. Dasgupta in Uncertainty in Artificial Intelligence. The dissertation author contributed towards the experiments and the writeup.

Chapter 8, in part, is unpublished work with S. Dasgupta. The dissertation author is the primary investigator.

## VITA

| | |
|---|---|
| 2004 | Bachelor of Science in Computer Science, *magna cum laude*, University of California, San Diego |
| 2008 | Master of Science in Computer Science, University of California, San Diego |
| 2012 | Doctor of Philosophy in Computer Science, University of California, San Diego |

## PUBLICATIONS

S. Dasgupta, D. Hsu, N. Verma. A concentration theorem for projections. *Twenty-Second Conference on Uncertainty in Artificial Intelligence (UAI)*, 2006.

Y. Freund, S. Dasgupta, M. Kabra, N. Verma. Learning the structure of manifolds using random projections. *Twenty-First Conference on Neural Information Processing Systems (NIPS)*, 2007.

N. Verma. Mathematical advances in manifold learning. *Survey*, 2008.

N. Verma, S. Kpotufe, S. Dasgupta. Which spatial trees are adaptive to intrinsic di- mension? *Twenty-Fifth Conference on Uncertainty in Artificial Intelligence (UAI)*, 2009.

N. Nikzad, C. Ziftci, P. Zappi, N. Quick, P. Aghera, N. Verma, B. Demchak, K. Patrick, H. Shacham, T. Rosing, I. Krueger, W. Griswold, S. Dasgupta. CitiSense – Adaptive Services for Community-Driven Behavioral and Environmental Monitoring to Induce Change. *UC San Diego Tech. Report CS2011-0961*, 2011.

B. Babenko, N. Verma, P. Dollár, S. Belongie. Multiple instance learning with manifold bags. *Twenty-Eighth International Conference on Machine Learning (ICML)*, 2011.

N. Verma, P. Zappi, T. Rosing. Latent variables based data estimation for sensing applications. *IEEE International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP)*, 2011.

N. Verma. A note on random projections for preserving paths on a manifold. *UC San Diego Tech. Report CS2011-0971*, 2011.

N. Verma, D. Mahajan, S. Sellamanickam, V. Nair. Learning hierarchical similarity metrics. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

N. Verma. Distance preserving embeddings for low-dimensional manifolds. *Conference on Learning Theory (COLT)*, 2012.

N. Nikzad, N. Verma, C. Ziftci, E. Bales, N. Quick, P. Zappi, K. Patrick, S. Dasgupta, I. Krueger, T. Rosing, W. Griswold CitiSense: Improving geospatial environmental assessment of air quality using a wireless personal exposure monitoring system. *Wireless Health*, 2012.

B. Milosevic, N. Verma, P. Zappi, E. Farella, T. Rosing, L. Benini. Efficient energy management and data recovery in sensor networks using latent variables based tensor factorization. In preparation, 2012.

## PATENTS

D. Krishnaswamy, N. Verma, V. Bychkovsky. Method and system for keyword correlation in a mobile environment. US Patent Application # 20090125517, May 2009.

D. Krishnaswamy, N. Verma, V. Bychkovsky. Method and system using keyword vectors and associated metrics for learning and prediction of user correlation of targeted content messages in a mobile environment. US Patent Application # 20090125462, May 2009.

D. Krishnaswamy, R. Daley, N. Verma. Delivery of targeted content related to a learned and predicted future behavior based on spatial, temporal, and user attributes and behavioral constraints. US Patent Application # 20110282964, November 2011.

ABSTRACT OF THE DISSERTATION

**Learning From Data With Low Intrinsic Dimension**

by

Nakul Verma

Doctor of Philosophy in Computer Science

University of California, San Diego, 2012

Professor Sanjoy Dasgupta, Chair

The information explosion of the past few decades has created tremendous opportunities for Machine Learning-based data analysis. Modern data typically possesses a large number of features. Consider the task of predicting the effectiveness of a particular treatment by analyzing a patient's genome. One hopes that by measuring several gene expression levels one can capture relevant information, leading to better predictions. However, the presence of a large number of irrelevant features adds to the statistical and computational complexity of the learning algorithm, without helping the practitioner to solve the task at hand. Indeed, conventional statistical wisdom suggests that in a general setting the learning task becomes significantly more difficult with an increase in the number of features,

making it especially difficult to design and analyze learning algorithms for modern, high-dimensional data.

This dissertation explores a specific way one can cope with this *curse of dimension*. The key observation is that while modern datasets are represented in high dimensions, they often adhere to some low-dimensional *intrinsic* structure. This intrinsic structure can manifest itself in several forms: some datasets such as text data have a sparse structure; other datasets such as speech and image articulation data follow a manifold structure. If this intrinsic structure is in fact low-dimensional (that is, has few degrees of freedom), then the complexity of learning algorithms should scale only with data's intrinsic dimensionality.

In the first part of this dissertation we study how the performance of learning algorithms is affected when the data have a low-dimensional manifold structure. We provide sharp bounds for unsupervised dimensionality reduction, and an improved PAC-learning framework for multiple instance learning in this setting.

The second part of this dissertation focuses on understanding and formalizing the general phenomenon of low intrinsic dimension. We explore a few notions that can effectively quantify low-dimensional geometric structure in data. We show that unlike traditional notions, some of the new notions are algorithmically verifiable. We can thus test a given dataset and guarantee a learning rate that scales only with its intrinsic dimension.

# Chapter 1

# Introduction

## 1.1 Nature of collected data

Lowering storage and communication costs has enabled us collect an unending source of rich multifaceted data. There is now a growing need to process, analyze and extract the relevant information from this data deluge. We are increasingly relying on automated machine learning based data analysis techniques to ease our processing burden. While such automated methods perform reasonably well when lots of data is available, the multifaceted high-dimensional nature of modern datasets has created unique challenges for the learning algorithms: we cannot provide good performance and quality guarantees for high-dimensional data. Part of the reason for poor scaling behavior of learning algorithms is because increase in dimensionality exponentially increases the volume of the representation space. As a result, even large amounts of data are not enough to yield salient patterns or trends. In order to obtain a statistically reliable result, the amount of data needed to fill this space often grows exponentially with the dimensionality.

To perform well on modern high-dimensional datasets, we need a way to cope with the effects of this *curse of dimensionality*. This dissertation explores a specific way of doing this. The key observation is that while modern datasets are represented in high dimensions, they often adhere to some low-dimensional *intrinsic* structure. This intrinsic structure can be manifested in data in several forms: some datasets such as text data have a sparse structure, other datasets

such as speech and image articulation data follow a manifold structure, while other datasets such as genomics data typically conform to a taxonomy or cluster structure. If we can determine that the intrinsic structure is in fact low-dimensional (that is, has few degrees of freedom), then we can expect that the complexity of learning algorithms should scale only with data's intrinsic dimensionality.

This dissertation focuses on theoretical and practical issues in designing effective learning algorithms (both unsupervised and supervised) when the given data does adhere to a *manifold* structure (Section 1.2). We will then discuss a way to understand and formalize the general phenomenon of low intrinsic dimension and its effects on the complexity of learning (Section 1.3).

## 1.2 Unsupervised and weakly-supervised methods for manifold structured data

Manifolds are an important geometric structure that have received significant attention in the past decade from the machine learning community [Tenebaum et al., 2000, Roweis and Saul, 2000, Belkin and Niyogi, 2003]. Their applicability in learning arises from the fact that several types of data can be well modeled as a manifold. Much of the existing work has focused on the practical applicability of this model. We study how the manifold assumption provides enhanced theoretical guarantees in some popular learning regimes.

**Unsupervised dimensionality reduction of manifolds.**

Dimensionality reduction is a popular preprocessing step to alleviate the curse of dimensionality. When the underlying data has low dimensional manifold structure, one expects that good information-preserving low dimension embeddings are possible. Formally, given an $n$-dimensional manifold in $\mathbb{R}^D$, the goal is to come up with an embedding of the data in $\mathbb{R}^d$ ($d \ll D$) that preserves some interesting property of the manifold, say, preserve interpoint geodesic distances. This problem was originally studied in differential geometry by Nash [1954] and Kuiper [1955]. Their results show that one can embed the $n$-dimensional manifold

in just $d = 2n+1$ dimensions while preserving all geodesic distances. In the context of Machine Learning, we want to develop algorithms that achieve something similar when we have access to only a *finite size* sample from the underlying manifold.

While several algorithms are suggested to solve this problem (see e.g. works by Tenebaum et al. [2000], Roweis and Saul [2000], Belkin and Niyogi [2003]), only a few provide any sort of guarantee on the quality of the embedding. Chapters 3 and 4 explore manifold embedding techniques that can provide good distance preserving guarantees. Chapter 3 shows that a random linear projection of a general $n$-dimensional manifold into $O(n/\epsilon^2)$ dimensions can preserve lengths of all paths on a manifold (not just the geodesics) within a factor of $(1 \pm \epsilon)$. This work provides an alternate technique to analyze random projections for manifolds, and improves upon the previous known results by Baraniuk and Wakin [2009] and Clarkson [2008].

Chapter 4 tackles the problem of minimizing the dependence on $\epsilon$ that is typically associated with random projection type results. Following the ideas discussed by Nash [1954], we can show that certain non-linear mappings can *completely remove* the dependence on $\epsilon$ from the embedding dimension. We derive an explicit algorithm for computing the target embedding based on samples from the underlying manifold and show that our embedding approximately maintains the geodesic distances between any pair of points from the underlying manifold (not just the input samples). This work can be viewed as an algorithmic realization of Nash's Embedding Theorem, and provides the sharpest known result for algorithmically embedding arbitrary $n$-dimensional manifolds in just $O(n)$ dimensions while maintaining the underlying geodesic distances.

These results on manifold dimensionality reduction validate our intuition that the complexity (in this case, the dimension of the embedding space) of learning depends only on the dimension of the intrinsic structure of the data (the dimension of the underlying manifold).

**Sample complexity of Multiple Instance Learning for manifold bags.**

Multiple Instance Learning (MIL) is a popular weakly-supervised learning paradigm in which training data is provided in the form of labeled *sets of instances* (or bags) rather than individual instances. Typically, a bag is labeled positive if some instance in the bag is positive. Chapter 5 reconciles the disconnect between how MIL is used in practice and its theoretical analysis that is existing in the current literature. We argue that in many applications of MIL the data is better modeled as low dimensional manifold bags in high dimensional feature space, and show that the geometric properties of such manifold bags are intimately related to its PAC-learnability. We also present a practical algorithm for MIL using manifold bags and show that it works well for real-world detection tasks in image and audio domains.

These results again corroborate our intuition that even though data may be represented in high dimensions, if the data distribution conforms to a low-dimensional manifold structure, then the learning complexity does in fact scale with the complexity of the underlying manifold and not the ambient space.

## 1.3 Formalizing low-dimensional intrinsic structure and its implications on learning

Since low-dimensional intrinsic structure is prevalent in modern datasets, one of the key problems is how can one quantify these intrinsic degrees of freedom of the underlying data distribution *without knowing the exact nature of the intrinsic structure*. Such a characterization should be (i) conducive to algorithmic analysis (that is, be able to provide guarantees for the learning algorithm), (ii) robust to noise, and (iii) empirically verifiable from a finite size sample. While certain notions of characterizing intrinsic complexity such as *doubling dimension* and *covering dimension* (see e.g. surveys by Cutler [1993] and Clarkson [2006] for definitions) do seem to satisfy requirement (i) and, with certain modifications to the definitions, (ii), the biggest hurdle seems to be how can one empirically verify that real-world

datasets do in fact have low, say, doubling dimension.

Chapter 6 studies the strengths and weaknesses of several notions of intrinsic dimension existing in the literature; it turns out many of these are not well suited for existing real-world datasets. It then explores an alternate notion of characterizing intrinsic complexity called the *local covariance dimension* (formalized in [Freund et al., 2007, Dasgupta and Freund, 2008]), and studies its applicability on modern datasets.

Chapter 7 explores how the learning complexity of tree based space partitioning algorithms is closely tied with this alternate notion. Extensive experimental analysis shows that the quality of regression estimates, nearest neighbor queries and quantization errors yielded by these space partition algorithms scales with the local covariance dimension of the underlying data distribution.

Developing on this theme, Chapter 8 shows that this *adaptivity to intrinsic dimension* is not necessarily tied to tree based regressors or with the notion of 'local covariance dimension'. We show that for several other regressor types, as long as we can exhibit some low-dimensional structure in data, we get similar results.

These results provide a holistic way to formalize and test the *intrinsic dimension hypothesis* in modern datasets. The accompanying sampling complexity results give good performance guarantees for these datasets that scale with the intrinsic dimension of the underlying dataset and not the ambient space.

# Chapter 2

# Notation and Preliminaries

Here we present our notation and review some concepts that will be useful throughout the text.

## 2.1 Manifolds: definition and properties

**Definition 2.1** *We say a function $f : U \mapsto V$ is a diffeomorphism, if it is smooth*[1] *and invertible with a smooth inverse.*

**Definition 2.2** *A topological space $M$ is said to be a smooth $n$-dimensional manifold, if $M$ is locally diffeomorphic to $\mathbb{R}^n$. That is, at each $p \in M$ we have an open neighborhood $U \subset M$ containing $p$ such that there exist a diffeomorphic map between $U$ and $\mathbb{R}^n$.*

It is always helpful to have a picture in mind. See Figure 2.1 for an example of 1-dimensional manifold in $\mathbb{R}^3$. Notice that locally any small segment of the manifold "looks like" an interval in $\mathbb{R}^1$.

**Definition 2.3** *A tangent space at a point $p \in M$, denoted by $T_pM$, is the vector space formed by collection of all vectors tangent to $M$ at $p$.*

[1] recall that a function is smooth if all its partial derivatives $\partial^n f / \partial x_{i_1} \dots \partial x_{i_n}$ exist and are continuous.

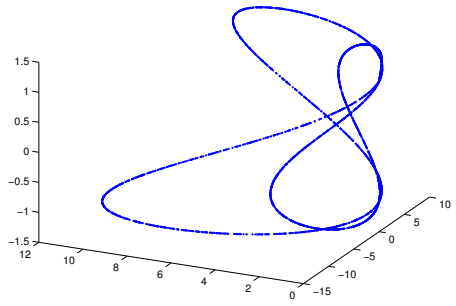



**Figure 2.1**: An example of 1-dimensional manifold in $\mathbb{R}^3$. Observe that a local enough region effectively looks like a line segment.

For the purposes of this dissertation we will restrict ourselves to the discussion of manifolds that are immersed in an ambient space $\mathbb{R}^D$, and whose tangent space at each point is equipped with an inner product structure that is inherited from the ambient space. Such manifolds are called *Riemannian* (sub-)manifolds and allow us to define various notions of length, angles, curvature, etc. on the manifold (see e.g. do Carmo [1992]).

Throughout the text we shall use $M$ to denote a smooth, $n$-dimensional compact Riemannian submanifold of $\mathbb{R}^D$. We will frequently refer to such a manifold as an $n$-manifold. Since we will be working with samples from $M$, we need to ensure certain amount of curvature regularity on $M$. Here we borrow the notation from Niyogi et al. [2008] about the condition number of $M$.

**Definition 2.4 (condition number [Niyogi et al., 2008])** *Let $M \subset \mathbb{R}^D$ be a compact Riemannian manifold. The condition number of $M$ is $\frac{1}{\tau}$, if $\tau$ is the largest number such that the normals of length $r < \tau$ at any two distinct points $p, q \in M$ don't intersect.*

The condition number of a manifold $M$ is an intuitive notion that captures the "complexity" of $M$ in terms of its curvature. Say $M$ has condition number $1/\tau$, then we can bound the directional curvature at any $p \in M$ by $\tau$. Figure 2.2 depicts the normals of a manifold. Notice that long non-intersecting normals are possible only if the manifold is relatively flat. Hence, the condition number of $M$

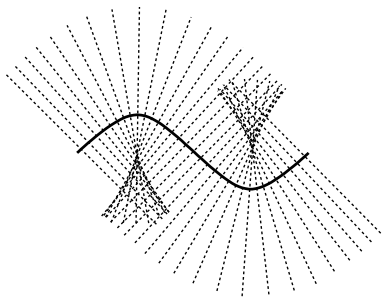


**Figure 2.2**: Tubular neighborhood of a manifold. Note that the normals (dotted lines) of a particular length incident at each point of the manifold (solid line) will intersect if the manifold is too curvy.

gives us a handle on how curvy can $M$ be. As a quick example, let's calculate the condition number of an $n$-dimensional sphere of radius $r$ (embedded in $\mathbb{R}^D$). Note that in this case one can have non-intersecting normals of length less than $r$ (since otherwise they will start intersecting at the center of the sphere). Thus, the condition number of such a sphere is $1/r$. Throughout the text we will assume that $M$ has a bounded condition number.

There are several useful properties of well-conditioned manifolds (that is, manifolds with bounded condition number). These are detailed in Appendix A.

We will use $D_G(p, q)$ to indicate the geodesic distance[2] between points $p$ and $q$ where the underlying manifold is understood from the context, and $\|p - q\|$ to indicate the Euclidean distance between points $p$ and $q$ where the ambient space is understood from the context.

[2]Geodesic distances and paths between two points are not necessarily unique. Here we are interested in the one that yields the shortest distance.

# Chapter 3

# Random Projections for Preserving Paths on Manifolds

Random projections have turned out to be a powerful tool for linear dimensionality reduction that approximately preserve Euclidean distances between pairs of points in a set $S \subset \mathbb{R}^D$. Their simplicity and universality stems from the fact the target embedding space is picked *without* looking at the individual samples from the set $S$. Interestingly, recent results by Baraniuk and Wakin [2009] and Clarkson [2008] show that even if the underlying set is a non-linear manifold (say of intrinsic dimensionality $n$), a random projection into a subspace of dimension $O(n)$ suffices to preserve interpoint Euclidean distances between the pairs of points.

It turns out that requiring Euclidean distances to be approximately preserved between pairs of points in a manifold is in a sense the strongest condition one can pose. This condition suffices to imply that the random projection will also preserve several other useful properties on manifolds. For instance, if one has a random projection that can approximately preserve the Euclidean distances, it will also approximately preserve the lengths of arbitrary curves on the manifold, and the curvature of the manifold.

Here we are interested in analyzing whether one can use random projections to reason *directly* about preserving the lengths of arbitrary paths on a manifold, without having to appeal to interpoint Euclidean distances. There is a two-fold reason for doing this: i) one can possibly get a sharper bound on the dimension

of target space by relaxing the Euclidean interpoint distance preservation requirement, and ii) since paths—unlike Euclidean distances—are inherently an intrinsic quantity, it should require a different technique to show path length preservation. Thus, giving us an alternate, direct proof.

Here we make progress on both fronts. We can remove the dependence on ambient dimension from the bound provided by Baraniuk and Wakin [2009], as well as simplify the bound provided by Clarkson [2008] by giving an explicit bound for all settings of the isometry parameter (and not just asymptotically small values). Our key lemma (Lemma 3.4) uses an elegant chaining argument on the coverings of vectors in tangent spaces providing an alternate proof technique.

## 3.1 Preliminaries

Given an $n$-manifold $M \subset \mathbb{R}^D$, recall that the length of any given curve $\gamma : [a, b] \to M$ is given by $\int_a^b \|\gamma'(s)\| ds$ (that is, length of a curve is an infinitesimal sum of the lengths of vectors tangent to points along the path). It thus suffices to bound the distortion induced by a random projection to the lengths of arbitrary vectors tangent to $M$. We shall assume that $M$ has condition number $1/\tau$ (cf. Definition 2.4)

Since we will be talking about random projections, or more formally, a function that maps data from $\mathbb{R}^D$ to a random subspace of dimension $\mathbb{R}^d$ via an orthogonal linear projection, we will frequently refer to the matrix form of this linear function as a *random projection matrix* or a *random orthoprojector*.

As a final piece of notation, we require a notion of covering on our manifold $M$. We define the *$\alpha$-geodesic covering number* of $M$ as the size of the smallest set $S \subset M$, with the property: for all $p \in M$, there exists $p' \in S$ such that $D_G(p, p') \leq \alpha$.

## 3.2 Main result

**Theorem 3.1** *Let $M$ be a smooth compact $n$-dimensional submanifold of $\mathbb{R}^D$ with condition number $1/\tau$. Let $G(M, \alpha)$ denote the $\alpha$-geodesic covering number of $M$. Pick any $0 < \epsilon < 1$ and $0 < \delta < 1$. Let $\phi$ be a random projection matrix that maps points from $\mathbb{R}^D$ into a random subspace of dimension $d$ ($d \leq D$) and define $\Phi := \sqrt{D/d}\phi$ as a scaled projection mapping.*

*If $d \geq \left\{ \frac{64}{\epsilon^2} \ln \frac{4G(M, \tau\epsilon^2/2^{18})}{\delta} + \frac{64n}{\epsilon^2} \ln \frac{117}{\epsilon} \right\}$, then with probability at least $1 - \delta$, for any path $\gamma$ in $M$, and its projection $\Phi(\gamma)$ in $\Phi(M) \subset \mathbb{R}^d$,*

$$(1 - \epsilon)L(\gamma) \leq L(\Phi(\gamma)) \leq (1 + \epsilon)L(\gamma),$$

*where $L(\beta)$ denotes the length of the path $\beta$.*

## 3.3 Proof

As discussed earlier, it suffices to uniformly bound the distortion induced by a random projection to the length of an arbitrary vector tangent to our manifold $M$. So we shall only focus on that. We start by stating a few useful lemmas that would help in our discussion.

**Lemma 3.2 (random projection of a fixed vector – see e.g. Lemma 2.2 of Dasgupta and Gupta [1999])** *Fix a vector $v \in \mathbb{R}^D$. Let $\phi$ be a random projection map that maps points from $\mathbb{R}^D$ to a random subspace of dimension $d$. Then,*

*i) For any $\beta \geq 1$,*

$$\mathbf{Pr}\left[ \|\phi(v)\|^2 \geq \beta \frac{d}{D} \|v\|^2 \right] \leq e^{(\beta - 1 - \ln \beta)(-d/2)}.$$

*ii) For any $0 < \epsilon < 1$, we have*

$$\mathbf{Pr}\left[ \|\phi(v)\|^2 \leq (1 - \epsilon)\frac{d}{D} \|v\|^2 \text{ or } \|\phi(v)\|^2 \geq (1 + \epsilon)\frac{d}{D} \|v\|^2 \right] \leq 2e^{-d\epsilon^2/4}.$$

**Lemma 3.3 (covering of a Euclidean unit-sphere – see e.g. Lemma 5.2 of Vershynin [2010])** *Let $S^{n-1}$ be an $(n-1)$-dimensional Euclidean unit sphere. Then, for all $\epsilon > 0$ there exists an $\epsilon$-cover of $S^{n-1}$ of size at most $(1+2/\epsilon)^n$. That is, there exists a set $C \subset S^{n-1}$, of size at most $(1+2/\epsilon)^n$, with the property: for any $x \in S^{n-1}$, exists $c \in C$ such that $\|x-c\| \leq \epsilon$.*

**Lemma 3.4 (random projection of a section of a manifold)** *Let $M \subset \mathbb{R}^D$ be a smooth compact $n$-dimensional manifold with condition number $1/\tau$. Pick any $0 < \epsilon < 1$. Fix some $p$ in $M$ and let $S := \{p' \in M : D_G(p,p') \leq \tau\epsilon^2/2^{18}\}$. Let $\phi$ be a random orthoprojector from $\mathbb{R}^D$ to $\mathbb{R}^d$. Then, if $d \geq 30n\ln 18$,*

$$\mathbf{Pr}\left[\exists p' \in S : \exists v' \in T_{p'}M : \|\phi v'\| \leq (1-\epsilon)\sqrt{\frac{d}{D}}\|v'\| \text{ or } \|\phi v'\| \geq (1+\epsilon)\sqrt{\frac{d}{D}}\|v'\|\right] \leq 4e^{n\ln(117/\epsilon)-(d\epsilon^2/64)}.$$

*Proof.* Note that the set $S$ is path-connected, and (see for instance Lemma A.4 in the Appendix) for any Euclidean ball $B(x,r)$ in $\mathbb{R}^D$, $S \cap B(x,r)$ can be covered by $9^n$ balls of half the radius. We will use this fact to create a hierarchy of covers of increasingly fine resolution. For each point in the hierarchy, we shall associate a covering of the tangent space at that point. We will inductively show that (with high probability) a random projection doesn't distort the lengths of the tangent vectors in the covering by too much. We will then conclude by showing that bounding the length distortion on tangent vectors in the covering implies a bound on the length distortion of all vectors in all the tangent spaces of all points in $S$. We now make this argument precise.

**Constructing a hierarchical cover of $S$:** Note that $S$ is contained in the Euclidean ball $B(p, \tau\epsilon^2/2^{18})$. We create a hierarchy of covers as follows. Pick a cover of $S \subset B(p,\tau\epsilon^2/2^{18})$ by $9^n$ balls of radius $\tau\epsilon^2/2^{19}$ (see Lemma A.4 in the Appendix). WLOG, we can assume that the centers of these balls lie in $S$ (see e.g. proof of Theorem 22 of Dasgupta and Freund [2008]). Each of these balls induces a subset of $S$, which in turn can then be covered by $9^n$ balls of radius $\tau\epsilon^2/2^{20}$. We

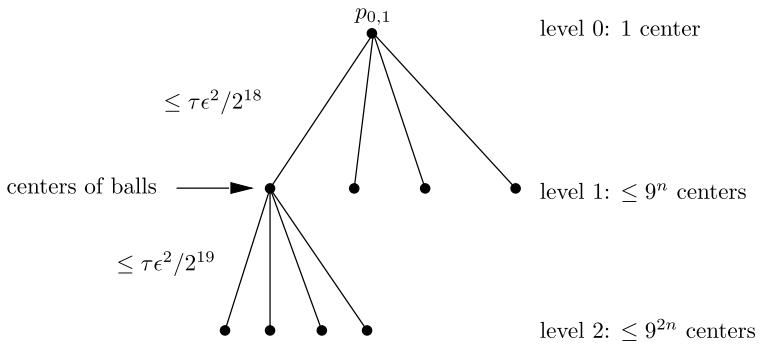



**Figure 3.1**: A hierarchy of covers of $S \subset B(p, \tau\epsilon^2/2^{18})$ for some point $p$ in an $n$-manifold $M$ with condition number $1/\tau$. Observe that at any level $i$, there are at most $9^{ni}$ points in the cover. Also note that the Euclidean distance between any point $p_{i,k}$ at level $i$ and its parent $p_{i-1,j}$ in the hierarchy is at most $\tau\epsilon^2/2^{17+i}$.

can continue this process to get an increasingly fine resolution such that at the end, any point of $S$ would have been arbitrarily well approximated by the center of some ball in the hierarchy. We will use the notation $p_{i,k}$ to denote the center of the $k^{\text{th}}$ ball at level $i$ of the hierarchy (note that with this notation $p_{0,1} = p$). (see Figure 3.1).

**A tangent space cover associated with each point in the hierarchy:** Associated with each $p_{i,k}$, we have a set $Q_{i,k} \subset T_{p_{i,k}}M$ of unit-length vectors tangent to $M$ at $p_{i,k}$ that forms a $(\epsilon/6)$-cover of the unit-vectors in $T_{p_{i,k}}M$ (that is, for all unit $v \in T_{p_{i,k}}M$, there exists $q \in Q_{i,k}$ where $\|q\| = 1$ such that $\|q - v\| \leq \epsilon/6$). We will define the individual vectors in $Q_{i,k}$ as follows. The set $Q_{0,1}$ is any $(\epsilon/6)$-cover of the unit-sphere in $T_{p_{0,1}}M$. Note that, by Lemma 3.3 and recalling that $\epsilon < 1$, we can assume that $|Q_{0,1}| =: L \leq e^{n\ln(13/\epsilon)}$. For levels $i = 1, 2, \ldots$, define $Q_{i,k}$ (associated with the point $p_{i,k}$) as the *parallel transport* (via the shortest geodesic path using the standard manifold connection, see Figure 3.2) of the vectors in $Q_{i-1,j}$ (associated with the point $p_{i-1,j}$) where $p_{i-1,j}$ is the parent of $p_{i,k}$ in the hierarchy. Note that parallel transporting a set of vectors preserves certain desirable properties – the dot product, for instance, between the vectors being transported is preserved (see, for instance, page 161 of Stoker [1969]). Thus, by construction, we have that $Q_{i,k}$ is a $(\epsilon/6)$-cover, since parallel transport doesn't change the lengths

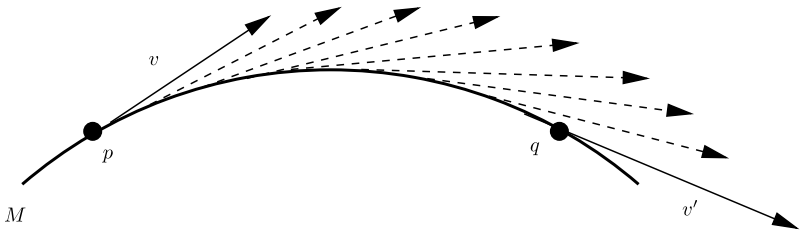



**Figure 3.2**: Parallel transport of the vector $v$ at point $p \in M$ to the point $q \in M$. The resulting transported vector is $v'$. Parallel transport is the translation of a (tangent) vector from one point to another while remaining tangent to the manifold. As the vector is transported infinitesimally along a path, it is also required to be parallel. Note that the resulting vector $v'$ is typically path-dependent: that is, for different paths from $p$ to $q$, the transport of $v$ is generally different. However, as expected, the transport does not change the length of the original vector. That is, $\|v\| = \|v'\|$.

or the mutual angles between the vectors being transported.

**A residual associated with each vector in the tangent space cover:** For $i \geq 1$, let $q_l^{i,k}$ be the $l^{\text{th}}$ vector in $Q_{i,k}$, which was formed by the transport of the vector $q_l^{i-1,j}$ in $Q_{i-1,j}$. We define the "residual" as $r_l^{i,k} := q_l^{i,k} - q_l^{i-1,j}$ (for $l = 1, \ldots, L$). Then we have that $\|r_l^{i,k}\|$ is bounded. In particular, since $\|q_l^{i-1,j}\| = \|q_l^{i,k}\| = 1$ (since the transport doesn't change vector lengths), and since $D_G(p_{i-1,j}, p_{i,k}) \leq 2\|p_{i-1,j} - p_{i,k}\| \leq \tau\epsilon^2/2^{16+i}$ (cf. Lemma A.1)

$$\|r_l^{i,k}\|^2 \leq 2D_G(p_{i-1,j}, p_{i,k})/\tau \leq \epsilon^2 2^{-i-15}.$$

**Effects of a random projection on the length of the residual:** Note that for a fixed $r_l^{i,k}$ (corresponding to a fixed point $p_{i,k}$ at level $i$ in the hierarchy and its parent $p_{i-1,j}$ in the hierarchy), using Lemma 3.2 (i), we have (for $\beta > 1$)

$$\mathbf{Pr}\left[\|\phi(r_l^{i,k})\|^2 \geq \beta\frac{d}{D}\|r_l^{i,k}\|^2\right] \leq e^{(\beta-1-\ln\beta)(-d/2)}. \tag{3.1}$$

By choosing $\beta = 2^{i/2}$ in Eq. (3.1), we have the following. For any fixed $i$ and $k$, with probability at least $1 - e^{n\ln(13/\epsilon)}e^{(2^{i/2}-1-\ln 2^{i/2})(-d/2)} \geq 1 - e^{n\ln(13/\epsilon)-di/30}$, we have $\|\phi(r_l^{i,k})\|^2 \leq \epsilon^2 2^{i/2}\frac{d}{D}\|r_l^{i,k}\|^2 \leq \epsilon^2 2^{-(i/2)-15}(d/D)$ (for $l = 1, \ldots, L$). By taking

a union bound over all edges in the hierarchy, (if $d \geq 30n \ln 18$)

$$\begin{aligned}\mathbf{Pr}\Big[\exists \text{ level } i : \exists \text{ ball } k \text{ at level } i \text{ with center } p_{i,k} : \exists \text{ residual } r_l^{i,k} : \\ \|\phi(r_l^{i,k})\|^2 \geq \epsilon^2 2^{-(i/2)-15}(d/D)\Big] \quad &\leq \sum_{i=1}^{\infty} e^{ni\ln 9} e^{n\ln(13/\epsilon)} e^{-di/30} \\ &= e^{n\ln(13/\epsilon)}\Big(\frac{1}{1-e^{n\ln 9 - d/30}} - 1\Big) \\ &\leq 2e^{n\ln(117/\epsilon)-(d/30)},\end{aligned}$$

where the equality is by noting that the geometric series converges (since $d \geq 30n \ln 18$), and the last inequality is by noting that $\frac{1}{1-s} \leq 1 + 2s$ for $0 \leq s \leq 1/2$.

**Effects of a random projection on the vectors in the tangent space cover:** We now use the (uniform) bound on $\|\phi(r_l^{i,k})\|^2$ to conclude inductively that $\phi$ doesn't distort the length of any vector $q_l^{i,k}$ too much (for any $i$, $k$, and $l$). In particular we will show that for all $i$, $k$ and $l$, we will have $(1 - \frac{\epsilon}{2})\frac{d}{D} \leq \|\phi(q_l^{i,k})\|^2 \leq (1 + \frac{\epsilon}{2})\frac{d}{D}$.

*Base case (level 0):* Since $|Q_{0,1}| \leq e^{n\ln(13/\epsilon)}$ we can apply Lemma 3.2 (ii), to conclude with probability at least $1 - 2e^{-d\epsilon^2/64 + n\ln(13/\epsilon)}$, for all $q \in Q_{0,1}$, $(1 - \frac{\epsilon}{4})\frac{d}{D} \leq \|\phi(q)\|^2 \leq (1 + \frac{\epsilon}{4})\frac{d}{D}$.

*Inductive hypothesis:* We assume that for all $q_l^{i,k} \in Q_{i,k}$ (for all $k$) at level $i$

$$\left(1 - \frac{\epsilon}{4} - \frac{\epsilon}{32}\sum_{j=1}^{i} 2^{-j/4}\right)\frac{d}{D} \leq \|\phi(q_l^{i,k})\|^2 \leq \left(1 + \frac{\epsilon}{4} + \frac{\epsilon}{32}\sum_{j=1}^{i} 2^{-j/4}\right)\frac{d}{D}. \tag{3.2}$$

*Inductive case:* Pick any $p_{i+1,k}$ at level $i+1$ in the hierarchy and let $p_{i,j}$ be its parent ($i \geq 0$). Then, for any $q_l^{i+1,k} \in Q_{i+1,k}$ (associated with the point $p_{i+1,k}$), let $q_l^{i,j} \in Q_{i,j}$ (associated with the point $p_{i,j}$) be the vector which after the parallel transport resulted in $q_l^{i+1,k}$. Then, we have:

$$\begin{aligned}\|\phi(q_l^{i+1,k})\|^2 &= \|\phi(q_l^{i,j}) + \phi(r_l^{i+1,k})\|^2 \\ &= \|\phi(q_l^{i,j})\|^2 + \|\phi(r_l^{i+1,k})\|^2 + 2\phi(q_l^{i,j}) \cdot \phi(r_l^{i+1,k})\end{aligned}$$

$$
\begin{aligned}
&\geq \|\phi(q_l^{i,j})\|^2 + \|\phi(r_l^{i+1,k})\|^2 - 2\|\phi(q_l^{i,j})\|\|\phi(r_l^{i+1,k})\| \\
&\geq \frac{d}{D}\left(1 - \frac{\epsilon}{4} - \frac{\epsilon}{32}\sum_{j=1}^{i} 2^{-j/4}\right) - 2\sqrt{\left(1+\frac{\epsilon}{2}\right)\frac{d}{D}}\sqrt{\frac{\epsilon^2 2^{-(i/2)-15.5} d}{D}} \\
&\geq \frac{d}{D}\left[\left(1 - \frac{\epsilon}{4} - \frac{\epsilon}{32}\sum_{j=1}^{i} 2^{-j/4}\right) \underbrace{-\epsilon\sqrt{2^{-(i/2)-12.5}}}_{\geq -\epsilon 2^{-(i/4)-(1/4)-5}}\right] \\
&\geq \frac{d}{D}\left(1 - \frac{\epsilon}{4} - \frac{\epsilon}{32}\sum_{j=1}^{i+1} 2^{-j/4}\right).
\end{aligned}
$$

Now, in the other direction we have

$$
\begin{aligned}
\|\phi(q_l^{i+1,k})\|^2 &= \|\phi(q_l^{i,j})\|^2 + \|\phi(r_l^{i+1,k})\|^2 + 2\|\phi(q_l^{i,j})\|\|\phi(r_l^{i+1,k})\| \\
&\leq \frac{d}{D}\left(1 + \frac{\epsilon}{4} + \frac{\epsilon}{32}\sum_{j=1}^{i} 2^{-j/4}\right) + \frac{\epsilon^2 2^{-(i/2)-15.5} d}{D} \\
&\quad + 2\sqrt{\left(1+\frac{\epsilon}{2}\right)\frac{d}{D}}\sqrt{\frac{\epsilon^2 2^{-(i/2)-15.5} d}{D}} \\
&\leq \frac{d}{D}\left[\left(1 + \frac{\epsilon}{4} + \frac{\epsilon}{32}\sum_{j=1}^{i} 2^{-j/4}\right) \underbrace{+\epsilon 2^{-(i/2)-15.5} + \epsilon\sqrt{2^{-(i/2)-12.5}}}_{\leq +\epsilon 2^{-(i/4)-(1/4)-5}}\right] \\
&\leq \frac{d}{D}\left(1 + \frac{\epsilon}{4} + \frac{\epsilon}{32}\sum_{j=1}^{i+1} 2^{-j/4}\right).
\end{aligned}
$$

So far we have shown that by picking $d \geq 30n \ln 18$, with probability at least $1 - 2(e^{n\ln(117/\epsilon)-(d/30)} + e^{n\ln(13/\epsilon)-(d\epsilon^2/64)}) \geq 1 - 4e^{n\ln(117/\epsilon)-(d\epsilon^2/64)}$, for all $i$, $k$ and $l$,

$$(1-\epsilon/2)(d/D) \leq \|\phi(q_l^{i,k})\|^2 \leq (1+\epsilon/2)(d/D).$$

**Effects of a random projection on any tangent vector at any point in the hierarchy:** Now, pick any point $p_{i,k}$ in the hierarchy and consider the corresponding set $Q_{i,k}$. We will show that for any unit vector $v \in T_{p_{i,k}}M$, $(1-\epsilon)\sqrt{d/D} \leq \|\phi(v)\| \leq (1+\epsilon)\sqrt{d/D}$.

Define $A := \max_{v \in T_{p_{i,k}}M, \|v\|=1} \|\phi(v)\|$ and let $v_0$ be a unit vector that attains this maximum. Let $q \in Q_{i,k}$ be such that $\|v_0 - q\| \leq \epsilon/6$. Now, if $\|v_0 - q\| = 0$,

then we get that $A = \|\phi(v_0)\| = \|\phi(q)\| \leq (1+\epsilon)\sqrt{d/D}$. Otherwise,

$$\begin{aligned} A = \|\phi(v_0)\| &\leq \|\phi(q)\| + \|\phi(v_0 - q)\| = \|\phi(q)\| + \|v_0 - q\| \left\| \phi\left( \frac{v_0 - q}{\|v_0 - q\|} \right) \right\| \\ &\leq (1 + \epsilon/2)\sqrt{d/D} + (\epsilon/6)(A). \end{aligned}$$

This yields that $A \leq (1+\epsilon)\sqrt{d/D}$, and thus for any unit $v \in T_{p_{i,k}}M$, $\|\phi(v)\| \leq \|\phi(v_0)\| = A \leq (1+\epsilon)\sqrt{d/D}$. Now, in the other direction, pick any unit $v \in T_{p_{i,k}}M$, and let $q \in Q_{i,k}$ be such that $\|v - q\| \leq \epsilon/6$. Again, if $\|v - q\| = 0$, then we have that $\|\phi(v)\| = \|\phi(q)\| \geq (1-\epsilon)\sqrt{d/D}$. Otherwise,

$$\begin{aligned} \|\phi(v)\| &\geq \|\phi(q)\| - \|\phi(v - q)\| = \|\phi(q)\| - \|v - q\| \left\| \phi\left( \frac{v - q}{\|v - q\|} \right) \right\| \\ &\geq (1 - \epsilon/2)\sqrt{d/D} - (\epsilon/6)(1+\epsilon)\sqrt{d/D} \geq (1-\epsilon)\sqrt{d/D}. \end{aligned}$$

Since $\phi$ is linear, it immediately follows that for *all* $v \in T_{p_{i,k}}M$ (not just unit-length $v$) we have

$$(1-\epsilon)\sqrt{d/D}\|v\| \leq \|\phi(v)\| \leq (1+\epsilon)\sqrt{d/D}\|v\|.$$

Observe that since the choice of the point $p_{i,k}$ was arbitrary, this holds true for *any* point in the hierarchy.

**Effects of a random projection on any tangent vector at any point in $S$:** We can finally give a bound on any tangent vector $v$ at any $p \in S$. Pick any $v$ tangent to $M$ at $p \in S$. Then, for any $\delta > 0$ such that $\delta \leq \tau/2$, we know that there exists some $p_{i,k}$ in the hierarchy such that $\|p - p_{i,k}\| \leq \delta$. Let $u$ be the parallel transport (via the shortest geodesic path) of $v$ from $p$ to $p_{i,k}$. Then, we know that $\|u\| = \|v\|$ and (cf. Lemma A.1) $\|\frac{u}{\|u\|} - \frac{v}{\|v\|}\| \leq \sqrt{4\delta/\tau}$. Thus,

$$\begin{aligned} \|\phi(v)\| &\leq \|\phi(u)\| + \|\phi(v - u)\| \leq (1+\epsilon)\sqrt{d/D}\|u\| + \|(v - u)\| \\ &\leq (1+\epsilon)\sqrt{d/D}\|v\| + 2\sqrt{\delta/\tau}. \end{aligned}$$

Similarly, in the other direction

$$\begin{aligned} \|\phi(v)\| &\geq \|\phi(u)\| - \|\phi(v - u)\| \geq (1-\epsilon)\sqrt{d/D}\|u\| - \|(v - u)\| \\ &\geq (1-\epsilon)\sqrt{d/D}\|v\| - 2\sqrt{\delta/\tau}. \end{aligned}$$

Note that since the choice of $\delta$ was arbitrary, by letting $\delta \to 0$ from above, we can conclude

$$(1-\epsilon)\sqrt{\frac{d}{D}}\|v\| \leq \|\phi(v)\| \leq (1+\epsilon)\sqrt{\frac{d}{D}}\|v\|.$$

∎

All the pieces are now in place to compute the distortion to tangent vectors induced by a random projection mapping. Let $C$ be a $(\tau\epsilon^2/2^{18})$-geodesic cover of $M$. Noting that one can have $C$ of size at most $G(M, \tau\epsilon^2/2^{18})$, we have (for $d > 30n \ln 9$)

$$\begin{aligned}\mathbf{Pr}\Bigg[\exists c \in C : &\exists p \text{ such that } D_G(c,p) \leq \tau\epsilon^2/2^{18} : \exists v \in T_pM : \\ &\|\phi(v)\| \leq (1-\epsilon)\sqrt{\frac{d}{D}}\|v\| \;\text{ or }\; \|\phi(v)\| \geq (1+\epsilon)\sqrt{\frac{d}{D}}\|v\|\Bigg] \\ &\leq\; 4G(M, \tau\epsilon^2/2^{18})\big(e^{n\ln(117/\epsilon)-(d\epsilon^2/64)}\big).\end{aligned}$$

The theorem follows by bounding this quantity by $\delta$.

## Acknowledgements

The contents of this chapter originally appeared in the following publication: N. Verma. A note on random projections for preserving paths on a manifold. *UC San Diego, Tech. Report CS2011-0971*, 2011.

# Chapter 4

# Sharper Bounds for Embedding General Manifolds

In the last chapter we saw that there exists a *linear* embedding that can preserve geodesic distances well. Here we are interested in studying how much can we reduce the embedding dimension without compromising the quality of geodesics. Recall that a random projection of a general $n$-manifold into $\tilde{O}(n/\epsilon^2)$ dimensions[1] guarantees $(1 \pm \epsilon)$-isometry. Observe that the $1/\epsilon^2$ dependence is troublesome: if we want an embedding with all distances within 99% of the original distances (i.e., $\epsilon = 0.01$), the bounds require the dimension of the target space to be at least 10,000!

In this chapter, we give two algorithms that achieve $(1 \pm \epsilon)$-isometry where the dimension of the target space is *independent* of the isometry constant $\epsilon$. As one expects, this dependency shows up in the sampling density (i.e. the size of $X$) required to compute the embedding. The first algorithm we propose is simple and easy to implement but embeds the given $n$-dimensional manifold in $\tilde{O}(2^{cn})$ dimensions(where $c$ is an absolute constant). The second algorithm, a variation on the first, focuses on minimizing the target dimension. It is computationally more involved and serves a more algorithmic purpose: it shows that one can embed the manifold in just $\tilde{O}(n)$ dimensions.

[1] $\tilde{O}(\cdot)$ notation suppresses the logarithmic dependence on parameters.

We would like to highlight that both our proposed algorithms work for a very general class of well-conditioned manifolds. There is no requirement that the underlying manifold is connected, or is globally isometric (or even globally diffeomorphic) to some subset of $\mathbb{R}^n$ as is frequently assumed by several manifold embedding algorithms. In addition, unlike spectrum-based embedding algorithms in the literature, our algorithms yield an explicit embedding that cleanly embeds out-of-sample data points, and provide isometry guarantees over the entire manifold (not just the input samples).

## 4.1 Isometrically embedding manifolds: Intuition

Given an underlying $n$-dimensional manifold $M \subset \mathbb{R}^D$, we shall use ideas from Nash's embedding [Nash, 1954] to develop our algorithms. To ease the burden of finding a $(1 \pm \epsilon)$-isometric embedding directly, our proposed algorithm will be divided in two stages. The first stage will embed $M$ in a lower dimensional space without having to worry about preserving any distances. Since interpoint distances will potentially be distorted by the first stage, the second stage will focus on adjusting these distances by applying a series of corrections. The combined effect of both stages is a distance preserving embedding of $M$ in lower dimensions. We now describe the stages in detail.

### 4.1.1 Embedding stage

We shall use the random projection result by Clarkson [2008] (with $\epsilon$ set to a constant) to embed $M$ into $d = \tilde{O}(n)$ dimensions. This gives an easy one-to-one low-dimensional embedding that doesn't collapse interpoint distances. Note that a projection does contract interpoint distances; by appropriately scaling the random projection, we can make sure that the distances are contracted by at most a constant amount, with high probability.

### 4.1.2 Correction stage

Since the random projection can contract different parts of the manifold by different amounts, we will apply several corrections—each corresponding to a different local region—to stretch-out and restore the local distances.

To understand a single correction better, we can consider its effect on a small section of the contracted manifold. Since manifolds are locally linear, the section effectively looks like a contracted $n$-dimensional affine space. Our correction map needs to restore distances over this $n$-flat.

For simplicity, let's temporarily assume $n = 1$ (this corresponds to a 1-dimensional manifold), and let $t \in [0, 1]$ parameterize a unit-length segment of the contracted 1-flat. Suppose we want to stretch the segment by a factor of $L \geq 1$ to restore the contracted distances. How can we accomplish this?

Perhaps the simplest thing to do is apply a linear correction mapping $\Psi : t \mapsto Lt$. While this mapping works well for individual local corrections, it turns out that this mapping makes it difficult to control interference between different corrections with overlapping localities.

We instead use extra coordinates and apply a *non-linear* map $\Psi : t \mapsto (t, \sin(Ct), \cos(Ct))$, where $C$ controls the stretch-size. Note that such a spiral map stretches the length of the tangent vectors by a factor of $\sqrt{1 + C^2}$, since $\|\Psi'\| = \|d\Psi/dt\| = \|(1, C\cos(Ct), -C\sin(Ct))\| = \sqrt{1 + C^2}$. Now since the geodesic distance between any two points $p$ and $q$ on a manifold is given by the expression $\int \|\gamma'(s)\| ds$, where $\gamma$ is a parameterization of the geodesic curve between points $p$ and $q$ (that is, length of a curve is infinitesimal sum of the length of tangent vectors along its path), $\Psi$ stretches the interpoint geodesic distances by a factor of $\sqrt{1 + C^2}$ on the resultant surface as well. Thus, to stretch the distances by a factor of $L$, we can set $C := \sqrt{L^2 - 1}$.

Now generalizing this to a local region for an arbitrary $n$-dimensional manifold, let $U := [u^1, \ldots, u^n]$ be a $d \times n$ matrix whose columns form an orthonormal basis for the (local) contracted $n$-flat in the embedded space $\mathbb{R}^d$ and let $\sigma^1, \ldots, \sigma^n$ be the corresponding shrinkages along the $n$ orthogonal directions. Then one can consider applying an $n$-dimensional analog of the spiral mapping:

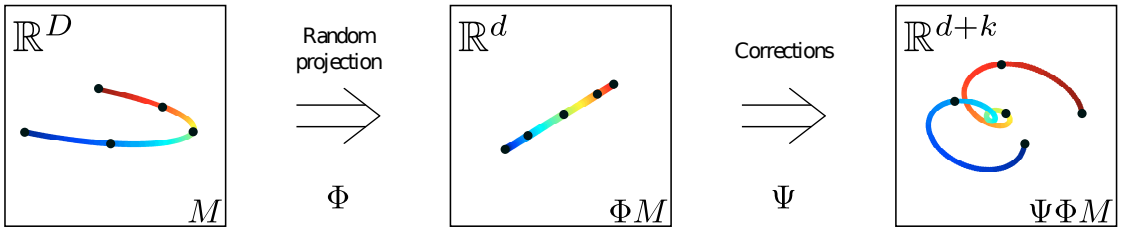



**Figure 4.1**: A simple example demonstrating our embedding technique on a 1-dimensional manifold. Left: The original 1-dimensional manifold in some high dimensional space. Middle: A low dimensional mapping of the original manifold via, say, a linear projection onto the vertical plane. Different parts of the manifold are contracted by different amounts – distances at the tail-ends are contracted more than the distances in the middle. Right: Final embedding after applying a series of spiralling corrections. Small size spirals are applied to regions with small distortion (middle), large spirals are applied to regions with large distortions (tail-ends). Resulting embedding is isometric (i.e., geodesic distance preserving) to the original manifold.

$\Psi : t \mapsto (t, \Psi_{\sin}(t), \Psi_{\cos}(t))$, where $t \in \mathbb{R}^d$

$$\Psi_{\sin}(t) := (\sin((Ct)_1), \ldots, \sin((Ct)_n)), \text{ and}$$
$$\Psi_{\cos}(t) := (\cos((Ct)_1), \ldots, \cos((Ct)_n)).$$

Here $C$ is an $n \times d$ "correction" matrix that encodes how much of the surface needs to stretch in the various orthogonal directions. It turns out that if one sets $C$ to be the matrix $SU^{\mathsf{T}}$, where $S$ is a diagonal matrix with entry $S_{ii} := \sqrt{(1/\sigma^i)^2 - 1}$ (recall that $\sigma^i$ was the shrinkage along direction $u^i$), then the correction $\Psi$ precisely restores the shrinkages along the $n$ orthonormal directions on the resultant surface (see Section 4.5.2 for a detailed derivation).

This takes care of the local regions individually. Now, globally, since different parts of the contracted manifold need to be stretched by different amounts, we localize the effect of the individual $\Psi$'s to a small enough neighborhood by applying a specific kind of kernel function known as the "bump" function in the analysis literature, given by (see also Figure 4.4 middle)

$$\lambda_x(t) := \mathbf{1}_{\{\|t-x\|<\rho\}} \cdot e^{-1/(1-(\|t-x\|/\rho)^2)}.$$

Applying different $\Psi$'s at different parts of the manifold has an aggregate effect of creating an approximate isometric embedding.

We now have a basic outline of our algorithm. Let $M$ be an $n$-dimensional manifold in $\mathbb{R}^D$. We first find a contraction of $M$ in $d = \tilde{O}(n)$ dimensions via a random projection. This embeds the manifold in low dimensions but distorts the interpoint geodesic distances. We estimate the distortion at different regions of the projected manifold by comparing a sample from $M$ (i.e. $X$) with its projection. We then perform a series of corrections—each applied locally—to adjust the lengths in the local neighborhoods. We will conclude that restoring the lengths in all neighborhoods yields a globally consistent approximately isometric embedding of $M$. See also Figure 4.1.

As briefly mentioned earlier, a key issue in preserving geodesic distances across points in different neighborhoods is reconciling the interference between different corrections with overlapping localities. Based on exactly *how* we apply these different local $\Psi$'s gives rise to our two algorithms. For the first algorithm, we shall allocate a fresh set of coordinates for each correction $\Psi$ so that the different corrections don't interfere with each other. Since a local region of an $n$-dimensional manifold can potentially have up to $O(2^{cn})$ overlapping regions, we shall require $O(2^{cn})$ additional coordinates to apply the corrections, making the final embedding dimension of $\tilde{O}(2^{cn})$ (where $c$ is an absolute constant). For the second algorithm, we will follow Nash's technique [Nash, 1954] more closely and apply $\Psi$ maps iteratively in the same embedding space without the use of extra coordinates. At each iteration we need to compute a pair of vectors *normal* to the embedded manifold. Since locally the manifold spreads across its tangent space, these normals indicate the locally empty regions in the embedded space. Applying the local $\Psi$ correction in the direction of these normals gives a way to mitigate the interference between different $\Psi$'s. Since we don't use extra coordinates, the final embedding dimension remains $\tilde{O}(n)$.

## 4.2 Preliminaries

Let $M$ be a smooth, $n$-dimensional compact Riemannian submanifold of $\mathbb{R}^D$ with condition number $1/\tau$ (cf. Definition 2.4).

To correctly estimate the distortion induced by the initial contraction mapping, we will additionally require a high-resolution covering of our manifold.

**Definition 4.1 (bounded manifold cover)** *Let $M \subset \mathbb{R}^D$ be a Riemannian $n$-manifold. We call $X \subset M$ an $\alpha$-bounded $(\rho, \delta)$-cover of $M$ if for all $p \in M$ and $\rho$-neighborhood $X_p := \{x \in X : \|x - p\| < \rho\}$ around $p$, we have*

- *exist points $x_0, \ldots, x_n \in X_p$ such that $\left|\frac{x_i - x_0}{\|x_i - x_0\|} \cdot \frac{x_j - x_0}{\|x_j - x_0\|}\right| \leq 1/2n$, for $i \neq j$. (covering criterion)*
- *$|X_p| \leq \alpha$. (local boundedness criterion)*
- *exists point $x \in X_p$ such that $\|x - p\| \leq \rho/2$. (point representation criterion)*
- *for any $n+1$ points in $X_p$ satisfying the covering criterion, let $\hat{T}_p$ denote the $n$-dimensional affine space passing through them (note that $\hat{T}_p$ does not necessarily pass through $p$). Then, for any unit vector $\hat{v}$ in $\hat{T}_p$, we have $\left|\hat{v} \cdot \frac{v}{\|v\|}\right| \geq 1 - \delta$, where $v$ is the projection of $\hat{v}$ onto the tangent space of $M$ at $p$. (tangent space approximation criterion)*

The above is an intuitive notion of manifold sampling that can estimate the local tangent spaces. Curiously, we haven't found such "tangent-space approximating" notions of manifold sampling in the literature. We do note in passing that our sampling criterion is similar in spirit to the $(\epsilon, \delta)$-sampling (also known as "tight" $\epsilon$-sampling) criterion popular in the Computational Geometry literature (see e.g. works by Dey et al. [2002] and Giesen and Wagner [2003]).

**Remark 4.2** *Given an $n$-manifold $M$ with condition number $1/\tau$, and some $0 < \delta \leq 1$. If $\rho \leq \tau\delta/3\sqrt{2}n$, then there exists a $2^{10n+1}$-bounded $(\rho, \delta)$-cover of $M$ (see Section 4.7 on how to ensure such a cover).*

We can now state our two algorithms.

## 4.3 The algorithms

**Inputs**

We assume the following quantities are given:

(i) $n$ – the intrinsic dimension of $M$.

(ii) $1/\tau$ – the condition number of $M$.

(iii) $X$ – an $\alpha$-bounded $(\rho, \delta)$-cover of $M$.

(iv) $\rho$ – the $\rho$ parameter of the cover.

**Notation**

Let $\phi$ be a random orthogonal projection map that maps points from $\mathbb{R}^D$ into a random subspace of dimension $d$ ($n \leq d \leq D$). We will have $d$ to be about $\tilde{O}(n)$. Set $\Phi := (2/3)(\sqrt{D/d})\phi$ as a scaled version of $\phi$. Since $\Phi$ is linear, $\Phi$ can also be represented as a $d \times D$ matrix. In our discussion below we will use the function notation and the matrix notation interchangeably, that is, for any $p \in \mathbb{R}^D$, we will use the notation $\Phi(p)$ (applying function $\Phi$ to $p$) and the notation $\Phi p$ (matrix-vector multiplication) interchangeably.

For any $x \in X$, let $x_0, \ldots, x_n$ be $n+1$ points from the set $\{x' \in X : \|x - x'\| < \rho\}$ such that $\left|\frac{x_i - x_0}{\|x_i - x_0\|} \cdot \frac{x_j - x_0}{\|x_j - x_0\|}\right| \leq 1/2n$, for $i \neq j$ (cf. Definition 4.1). Let $F_x$ be the $D \times n$ matrix whose column vectors form some orthonormal basis of the $n$-dimensional subspace spanned by the vectors $\{x_i - x_0\}_{i \in [n]}$. Note that $F_x$ serves as a good approximation to the tangent spaces at different points in the neighborhood of $x \in M \subset \mathbb{R}^D$.

**Estimating local contractions**

We estimate the contraction caused by $\Phi$ at a small enough neighborhood of $M$ containing the point $x \in X$, by computing the "thin" Singular Value Decomposition (SVD) $U_x \Sigma_x V_x^{\mathsf{T}}$ of the $d \times n$ matrix $\Phi F_x$ and representing the singular values in the conventional descending order. That is, $\Phi F_x = U_x \Sigma_x V_x^{\mathsf{T}}$, and since

$\Phi F_x$ is a tall matrix ($n \leq d$), we know that the bottom $d-n$ singular values are zero. Thus, we only consider the top $n$ (of $d$) left singular vectors in the SVD (so, $U_x$ is $d \times n$, $\Sigma_x$ is $n \times n$, and $V_x$ is $n \times n$) and $\sigma_x^1 \geq \sigma_x^2 \geq \ldots \geq \sigma_x^n$ where $\sigma_x^i$ is the $i^{\text{th}}$ largest singular value.

Observe that the singular values $\sigma_x^1, \ldots, \sigma_x^n$ are precisely the distortion amounts in the directions $u_x^1, \ldots, u_x^n$ at $\Phi(x) \in \mathbb{R}^d$ ($[u_x^1, \ldots, u_x^n] = U_x$) when we apply $\Phi$. To see this, consider the direction $w^i := F_x v_x^i$ in the column-span of $F_x$ ($[v_x^1, \ldots, v_x^n] = V_x$). Then $\Phi w^i = (\Phi F_x) v_x^i = \sigma_x^i u_x^i$, which can be interpreted as: $\Phi$ maps the vector $w^i$ in the subspace $F_x$ (in $\mathbb{R}^D$) to the vector $u_x^i$ (in $\mathbb{R}^d$) with the scaling of $\sigma_x^i$.

Note that if $0 < \sigma_x^i \leq 1$ (for all $x \in X$ and $1 \leq i \leq n$), we can define an $n \times d$ correction matrix (corresponding to each $x \in X$) $C^x := S_x U_x^\mathsf{T}$, where $S_x$ is a diagonal matrix with $(S_x)_{ii} := \sqrt{(1/\sigma_x^i)^2 - 1}$. We can also write $S_x$ as $(\Sigma_x^{-2} - I)^{1/2}$. The correction matrix $C^x$ will have an effect of stretching the direction $u_x^i$ by the amount $(S_x)_{ii}$ and killing any direction $v$ that is orthogonal to (column-span of) $U_x$.

We compute the corrections $C^x$'s as follows:

**Algorithm 4.1** Compute Corrections $C^x$'s

1: **for** $x \in X$ (in any order) **do**
2: Let $x_0, \ldots, x_n \in \{x' \in X : \|x' - x\| < \rho\}$ be such that $\left|\frac{x_i - x_0}{\|x_i - x_0\|} \cdot \frac{x_j - x_0}{\|x_j - x_0\|}\right| \leq 1/2n$ (for $i \neq j$).
3: Let $F_x$ be a $D \times n$ matrix whose columns form an orthonormal basis of the $n$-dimensional span of the vectors $\{x_i - x_0\}_{i \in [n]}$.
4: Let $U_x \Sigma_x V_x^\mathsf{T}$ be the "thin" SVD of $\Phi F_x$.
5: Set $C^x := (\Sigma_x^{-2} - I)^{1/2} U_x^\mathsf{T}$.
6: **end for**

**Algorithm 4.2** Embedding Technique I

**Preprocessing Stage:** Partition the given covering $X$ into disjoint subsets such that no subset contains points that are too close to each other. Let $x_1, \ldots, x_{|X|}$ be the points in $X$ in some arbitrary but fixed order. We can do the partition as follows:

1: Initialize $X^{(1)}, \ldots, X^{(K)}$ as empty sets.
2: **for** $x_i \in X$ (in any fixed order) **do**
3: Let $j$ be the smallest positive integer such that $x_i$ is not within distance $2\rho$ of any element in $X^{(j)}$. That is, the smallest $j$ such that for all $x \in X^{(j)}$, $\|x - x_i\| \geq 2\rho$.
4: $X^{(j)} \leftarrow X^{(j)} \cup \{x_i\}$.
5: **end for**

**The Embedding:** For *any* $p \in M \subset \mathbb{R}^D$, embed it in $\mathbb{R}^{d+2nK}$ as follows:

1: Let $t = \Phi(p)$.
2: Define $\Psi(t) := (t, \Psi_{1,\sin}(t), \Psi_{1,\cos}(t), \ldots, \Psi_{K,\sin}(t), \Psi_{K,\cos}(t))$ where

$$\Psi_{j,\sin}(t) := (\psi^1_{j,\sin}(t), \ldots, \psi^n_{j,\sin}(t)),$$
$$\Psi_{j,\cos}(t) := (\psi^1_{j,\cos}(t), \ldots, \psi^n_{j,\cos}(t)).$$

The individual terms are given by

$$\psi^i_{j,\sin}(t) := \textstyle\sum_{x \in X^{(j)}} \left(\sqrt{\Lambda_{\Phi(x)}(t)}/\omega\right) \sin(\omega(C^x t)_i) \qquad i = 1, \ldots, n;$$
$$\psi^i_{j,\cos}(t) := \textstyle\sum_{x \in X^{(j)}} \left(\sqrt{\Lambda_{\Phi(x)}(t)}/\omega\right) \cos(\omega(C^x t)_i) \qquad j = 1, \ldots, K$$

where $\Lambda_a(b) = \frac{\lambda_a(b)}{\sum_{q \in X} \lambda_{\Phi(q)}(b)}$.
3: **return** $\Psi(t)$ as the embedding of $p$ in $\mathbb{R}^{d+2nK}$.

A few remarks are in order.

**Remark 4.3** *The goal of the Preprocessing Stage is to identify samples from $X$ that can have overlapping ($\rho$-size) local neighborhoods. The partitioning procedure described above ensures that corrections associated with nearby neighborhoods are*

*applied in separate coordinates to minimize interference.*

**Remark 4.4** *If $\rho \leq \tau/4$, the number of subsets (i.e. $K$) produced by Embedding I is at most $\alpha 2^{cn}$ for an $\alpha$-bounded $(\rho, \delta)$ cover $X$ of $M$ (where $c \leq 4$). See Section 4.8 for details.*

**Remark 4.5** *The function $\Lambda$ acts as a (normalized) localizing kernel that helps in localizing the effects of the spiralling corrections (discussed in detail in Section 4.5.2).*

**Remark 4.6** *$\omega > 0$ is a free parameter that controls the interference due to overlapping local corrections.*

---

**Algorithm 4.3** Embedding Technique II

---

**The Embedding:** Let $x_1, \ldots, x_{|X|}$ be the points in $X$ in arbitrary but fixed order. For *any* $p \in M$, we embed it in $\mathbb{R}^{2d+3}$ by:

1: Let $t = \Phi(p)$.

2: Define $\Psi_{0,n}(t) := (t, \underbrace{0, \ldots, 0}_{d+3})$.

3: **for** $i = 1, \ldots, |X|$ **do**

4: Define $\Psi_{i,0} := \Psi_{i-1,n}$.

5: **for** $j = 1, \ldots, n$ **do**

6: Let $\eta_{i,j}(t)$ and $\nu_{i,j}(t)$ be two mutually orthogonal unit vectors normal to $\Psi_{i,j-1}(M)$ at $\Psi_{i,j-1}(t)$.

7: Define

$$\Psi_{i,j}(t) := \Psi_{i,j-1}(t) + \frac{\sqrt{\Lambda_{\Phi(x_i)}(t)}}{\omega_{i,j}} \Big( \eta_{i,j}(t) \sin(\omega_{i,j}(C^{x_i}t)_j) + \nu_{i,j}(t) \cos(\omega_{i,j}(C^{x_i}t)_j) \Big),$$

where $\Lambda_a(b) = \frac{\lambda_a(b)}{\sum_{q \in X} \lambda_{\Phi(q)}(b)}$.

8: **end for**

9: **end for**

10: **return** $\Psi_{|X|,n}(t)$ as the embedding of $p$ into $\mathbb{R}^{2d+3}$.

---

**Remark 4.7** *The function $\Lambda$, and the free parameters $\omega_{i,j}$ (one for each $i, j$ iteration) have roles similar to those in Embedding I.*

**Remark 4.8** *The success of Embedding II depends upon finding a pair of normal unit vectors $\eta$ and $\nu$ in each iteration; we discuss how to approximate these in Section 4.10.*

We shall see that for appropriate choice of $d$, $\rho$, $\delta$ and $\omega$ (or $\omega_{i,j}$), our algorithm yields an approximate isometric embedding of $M$.

## 4.4 Main result

**Theorem 4.9** *Let $M \subset \mathbb{R}^D$ be a compact $n$-manifold with volume $V$ and condition number $1/\tau$ (as above). Let $d = \Omega\left(n + \ln(V/\tau^n)\right)$ be the target dimension of the initial random projection mapping such that $d \leq D$. For any $0 < \epsilon \leq 1$, let $\rho \leq (\tau d/D)(\epsilon/350)^2$, $\delta \leq (d/D)(\epsilon/250)^2$, and let $X \subset M$ be an $\alpha$-bounded $(\rho, \delta)$-cover of $M$. Now, let*

*i.* $N_{\mathrm{I}} \subset \mathbb{R}^{d+2\alpha n 2^{cn}}$ *be the embedding of $M$ returned by Algorithm* I *(where $c \leq 4$),*

*ii.* $N_{\mathrm{II}} \subset \mathbb{R}^{2d+3}$ *be the embedding of $M$ returned by Algorithm* II.

*Then, with probability at least $1 - 1/poly(n)$ over the choice of the initial random projection, for all $p, q \in M$ and their corresponding mappings $p_{\mathrm{I}}, q_{\mathrm{I}} \in N_{\mathrm{I}}$ and $p_{\mathrm{II}}, q_{\mathrm{II}} \in N_{\mathrm{II}}$, we have*

*i.* $(1-\epsilon)D_G(p,q) \leq D_G(p_{\mathrm{I}}, q_{\mathrm{I}}) \leq (1+\epsilon)D_G(p,q)$,

*ii.* $(1-\epsilon)D_G(p,q) \leq D_G(p_{\mathrm{II}}, q_{\mathrm{II}}) \leq (1+\epsilon)D_G(p,q)$.

## 4.5 Proof

Our goal is to show that the two proposed embeddings approximately preserve the lengths of all geodesic curves. Now, since the length of any given curve

$\gamma : [a, b] \to M$ is given by $\int_a^b \|\gamma'(s)\| ds$, it is vital to study how our embeddings modify the length of the tangent vectors at any point $p \in M$.

In order to discuss tangent vectors, we need to introduce the notion of a tangent space $T_pM$ at a particular point $p \in M$. Consider any smooth curve $c : (-\epsilon, \epsilon) \to M$ such that $c(0) = p$, then we know that $c'(0)$ is the vector tangent to $c$ at $p$. The collection of all such vectors formed by all such curves is a well defined vector space (with origin at $p$), called the tangent space $T_pM$. In what follows, we will fix an arbitrary point $p \in M$ and a tangent vector $v \in T_pM$ and analyze how the various steps of the algorithm modify the length of $v$.

Let $\Phi$ be the initial (scaled) random projection map (from $\mathbb{R}^D$ to $\mathbb{R}^d$) that may contract distances on $M$ by various amounts, and let $\Psi$ be the subsequent correction map that attempts to restore these distances (as defined in Step 2 for Embedding I or as a sequence of maps in Step 7 for Embedding II). To get a firm footing for our analysis, we need to study how $\Phi$ and $\Psi$ modify the tangent vector $v$. It is well known from differential geometry that for any smooth map $F : M \to N$ that maps a manifold $M \subset \mathbb{R}^k$ to a manifold $N \subset \mathbb{R}^{k'}$, there exists a linear map $(DF)_p : T_pM \to T_{F(p)}N$, known as the derivative map or the pushforward (at $p$), that maps tangent vectors incident at $p$ in $M$ to tangent vectors incident at $F(p)$ in $N$. To see this, consider a vector $u$ tangent to $M$ at some point $p$. Then, there is some smooth curve $c : (-\epsilon, \epsilon) \to M$ such that $c(0) = p$ and $c'(0) = u$. By mapping the curve $c$ into $N$, i.e. $F(c(t))$, we see that $F(c(t))$ includes the point $F(p)$ at $t = 0$. Now, by calculus, we know that the derivative at this point, $\left.\frac{dF(c(t))}{dt}\right|_{t=0}$ is the directional derivative $(\nabla F)_p(u)$, where $(\nabla F)_p$ is a $k' \times k$ matrix called the gradient (at $p$). The quantity $(\nabla F)_p$ is precisely the matrix representation of this linear "pushforward" map that sends tangent vectors of $M$ (at $p$) to the corresponding tangent vectors of $N$ (at $F(p)$). Figure 4.2 depicts how these quantities are affected by applying $F$. Also note that if $F$ is linear, then $DF = F$.

Observe that since pushforward maps are linear, without loss of generality we can assume that $v$ has unit length.

**A quick roadmap for the proof.** In the next three sections, we take a brief

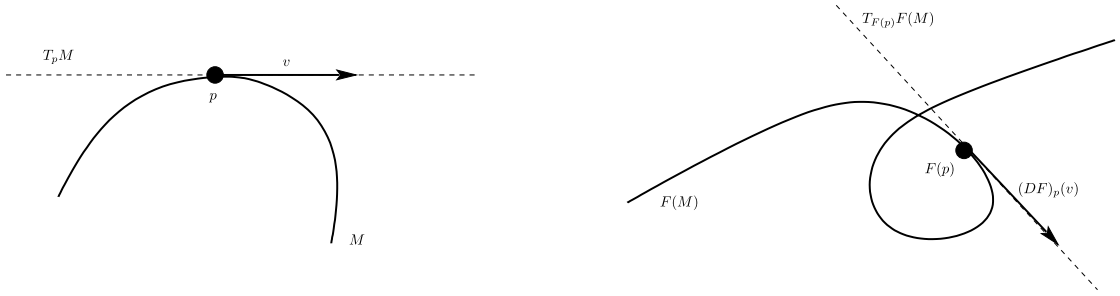



**Figure 4.2**: Effects of applying a smooth map $F$ on various quantities of interest. Left: A manifold $M$ containing point $p$. $v$ is a vector tangent to $M$ at $p$. Right: Mapping of $M$ under $F$. Point $p$ maps to $F(p)$, tangent vector $v$ maps to $(DF)_p(v)$.

detour to study the effects of applying $\Phi$, applying $\Psi$ for Algorithm I, and applying $\Psi$ for Algorithm II separately. This will give us the necessary tools to analyze the combined effect of applying $\Psi \circ \Phi$ on $v$ (Section 4.5.4). We will conclude by relating tangent vectors to lengths of curves, showing approximate isometry (Section 4.5.5). Figure 4.3 provides a quick sketch of our two stage mapping with the quantities of interest. We defer the proofs of all the supporting lemmas to Section 4.9.

### 4.5.1 Effects of applying $\Phi$

It is well known as an application of Sard's theorem from differential topology (see e.g. Milnor [1972]) that almost every smooth mapping of an $n$-dimensional manifold into $\mathbb{R}^{2n+1}$ is a differential structure preserving embedding of $M$. In particular, a projection onto a random subspace (of dimension $2n + 1$) constitutes such an embedding with probability 1.

This translates to stating that a random projection into $\mathbb{R}^{2n+1}$ is enough to guarantee that $\Phi$ doesn't collapse the lengths of non-zero tangent vectors almost surely. However, due to computational issues, we additionally require that the lengths are bounded away from zero (that is, a statement of the form $\|(D\Phi)_p(v)\| \geq \Omega(1)\|v\|$ for all $v$ tangent to $M$ at all points $p$).

We can thus appeal to the random projections result by Clarkson [2008] (with the isometry parameter set to a constant, say 1/4) to ensure this condition. In particular, the following holds.

**Lemma 4.10** *Let $M \subset \mathbb{R}^D$ be a smooth $n$-manifold (as defined above) with volume $V$ and condition number $1/\tau$. Let $R$ be a random projection matrix that maps points from $\mathbb{R}^D$ into a random subspace of dimension $d$ ($d \leq D$). Define $\Phi := (2/3)(\sqrt{D/d})R$ as a scaled projection mapping. If $d = \Omega(n + \ln(V/\tau^n))$, then with probability at least $1 - 1/poly(n)$ over the choice of the random projection matrix, we have*

*(a) For all $p \in M$ and all tangent vectors $v \in T_pM$, $(1/2)\|v\| \leq \|(D\Phi)_p(v)\| \leq (5/6)\|v\|$.*

*(b) For all $p, q \in M$, $(1/2)\|p - q\| \leq \|\Phi p - \Phi q\| \leq (5/6)\|p - q\|$.*

*(c) For all $x \in \mathbb{R}^D$, $\|\Phi x\| \leq (2/3)(\sqrt{D/d})\|x\|$.*

In what follows, we assume that $\Phi$ is such a scaled random projection map. Then, a bound on the length of tangent vectors also gives us a bound on the spectrum of $\Phi F_x$ (recall the definition of $F_x$ from Section 4.3).

**Corollary 4.11** *Let $\Phi$, $F_x$ and $n$ be as described above (recall that $x \in X$ that forms a bounded $(\rho, \delta)$-cover of $M$). Let $\sigma_x^i$ represent the $i^{th}$ largest singular value of the matrix $\Phi F_x$. Then, for $\delta \leq d/32D$, we have $1/4 \leq \sigma_x^n \leq \sigma_x^1 \leq 1$ (for all $x \in X$).*

We will be using these facts in our discussion below in Section 4.5.4.

### 4.5.2 Effects of applying $\Psi$ (Algorithm I)

As discussed in Section 4.1, the goal of $\Psi$ is to restore the contraction induced by $\Phi$ on $M$. To understand the action of $\Psi$ on a tangent vector better, we will first consider a simple case of flat manifolds (Section 4.5.2), and then develop the general case (Section 4.5.2).

#### Warm-up: flat $M$

Let's first consider applying a simple one-dimensional spiral map $\bar{\Psi} : \mathbb{R} \to \mathbb{R}^3$ given by $t \mapsto (t, \sin(Ct), \cos(Ct))$, where $t \in I = (-\epsilon, \epsilon)$. Let $\bar{v}$ be a unit vector

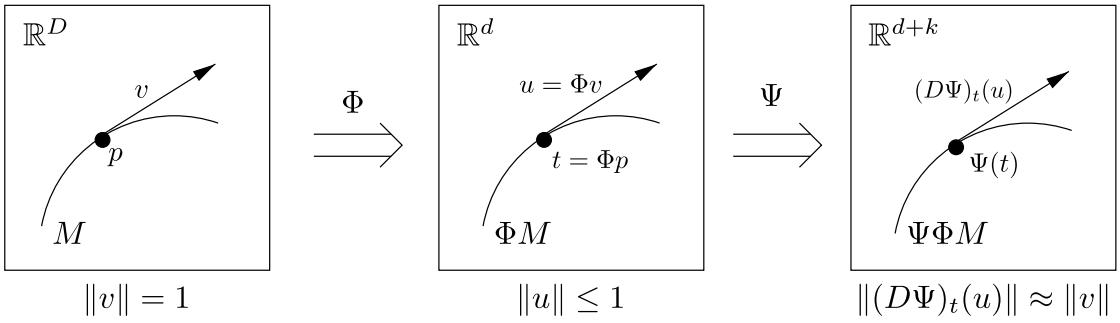



**Figure 4.3**: Two stage mapping of our embedding technique. Left: Underlying manifold $M \subset \mathbb{R}^D$ with the quantities of interest – a fixed point $p$ and a fixed unit-vector $v$ tangent to $M$ at $p$. Center: A (scaled) linear projection of $M$ into a random subspace of $d$ dimensions. The point $p$ maps to $\Phi p$ and the tangent vector $v$ maps to $u := (D\Phi)_p(v) = \Phi v$. The length of $v$ contracts to $\|u\|$. Right: Correction of $\Phi M$ via a non-linear mapping $\Psi$ into $\mathbb{R}^{d+k}$. We have $k = O(\alpha 2^{cn})$ for correction technique I, and $k = d + 3$ for correction technique II (see also Section 4.3). Our goal is to show that $\Psi$ stretches length of contracted $v$ (i.e. $u$) back to approximately its original length.

tangent to $I$ (at, say, 0). Then note that

$$(D\bar{\Psi})_{t=0}(\bar{v}) = \frac{d\bar{\Psi}}{dt}\Big|_{t=0} = (1, C\cos(Ct), -C\sin(Ct))\Big|_{t=0}.$$

Thus, applying $\bar{\Psi}$ stretches the length from 1 to $\left\|(1, C\cos(Ct), -C\sin(Ct))|_{t=0}\right\| = \sqrt{1+C^2}$. Notice the advantage of applying the spiral map in computing the lengths: the sine and cosine terms combine together to yield a simple expression for the size of the stretch. In particular, if we want to stretch the length of $\bar{v}$ from 1 to, say, $L \geq 1$, then we simply need $C = \sqrt{L^2 - 1}$ (notice the similarity between this expression and our expression for the diagonal component $S_x$ of the correction matrix $C^x$ in Section 4.3).

We can generalize this to the case of $n$-dimensional flat manifold (a section of an $n$-flat) by considering a map similar to $\bar{\Psi}$. For concreteness, let $F$ be a $D \times n$ matrix whose column vectors form some orthonormal basis of the $n$-flat manifold (in the original space $\mathbb{R}^D$). Let $U\Sigma V^\mathsf{T}$ be the "thin" SVD of $\Phi F$. Then $FV$ forms an orthonormal basis of the $n$-flat manifold (in $\mathbb{R}^D$) that maps to an orthogonal basis $U\Sigma$ of the projected $n$-flat manifold (in $\mathbb{R}^d$) via the contraction mapping $\Phi$. Define

the spiral map $\bar{\Psi} : \mathbb{R}^d \to \mathbb{R}^{d+2n}$ in this case as follows. $\bar{\Psi}(t) := (t, \bar{\Psi}_{\sin}(t), \bar{\Psi}_{\cos}(t))$, with $\bar{\Psi}_{\sin}(t) := (\bar{\psi}^1_{\sin}(t), \ldots, \bar{\psi}^n_{\sin}(t))$ and $\bar{\Psi}_{\cos}(t) := (\bar{\psi}^1_{\cos}(t), \ldots, \bar{\psi}^n_{\cos}(t))$. The individual terms are given as

$$\begin{array}{l} \bar{\psi}^i_{\sin}(t) := \sin((Ct)_i) \\ \bar{\psi}^i_{\cos}(t) := \cos((Ct)_i) \end{array} \quad i = 1, \ldots, n,$$

where $C$ is now an $n \times d$ correction matrix. It turns out that setting $C = (\Sigma^{-2} - I)^{1/2} U^{\mathsf{T}}$ precisely restores the contraction caused by $\Phi$ to the tangent vectors (notice the similarity between this expression with the correction matrix in the general case $C^x$ in Section 4.3 and our motivating intuition in Section 4.1). To see this, let $v$ be a vector tangent to the $n$-flat at some point $p$ (in $\mathbb{R}^D$). We will represent $v$ in the $FV$ basis (that is, $v = \sum_i \alpha_i (Fv^i)$ where $[Fv^1, \ldots, Fv^n] = FV$). Note that $\|\Phi v\|^2 = \|\sum_i \alpha_i \Phi F v^i\|^2 = \|\sum_i \alpha_i \sigma^i u^i\|^2 = \sum_i (\alpha_i \sigma^i)^2$ (where $\sigma^i$ are the individual singular values of $\Sigma$ and $u^i$ are the left singular vectors forming the columns of $U$). Now, let $w$ be the pushforward of $v$ (that is, $w = (D\Phi)_p(v) = \Phi v = \sum_i w_i e^i$, where $\{e^i\}_i$ forms the standard basis of $\mathbb{R}^d$). Now, since $D\bar{\Psi}$ is linear, we have $\|(D\bar{\Psi})_{\Phi(p)}(w)\|^2 = \|\sum_i w_i (D\bar{\Psi})_{\Phi(p)}(e^i)\|^2$, where $(D\bar{\Psi})_{\Phi(p)}(e^i) = \frac{d\bar{\Psi}}{dt^i}\big|_{t=\Phi(p)} = \left(\frac{dt}{dt^i}, \frac{d\bar{\Psi}_{\sin}(t)}{dt^i}, \frac{d\bar{\Psi}_{\cos}(t)}{dt^i}\right)\Big|_{t=\Phi(p)}$. The individual components are given by

$$\begin{array}{l} d\bar{\psi}^k_{\sin}(t)/dt^i = +\cos((Ct)_k) C_{k,i} \\ d\bar{\psi}^k_{\cos}(t)/dt^i = -\sin((Ct)_k) C_{k,i} \end{array} \quad k = 1, \ldots, n;\ i = 1, \ldots, d.$$

By algebra, we see that

$$\begin{aligned} \|(D(\bar{\Psi} \circ \Phi))_p(v)\|^2 &= \|(D\bar{\Psi})_{\Phi(p)}((D\Phi)_p(v))\|^2 = \|(D\bar{\Psi})_{\Phi(p)}(w)\|^2 \\ &= \sum_{k=1}^{d} w_k^2 + \sum_{k=1}^{n} \cos^2((C\Phi(p))_k)((C\Phi v)_k)^2 + \sum_{k=1}^{n} \sin^2((C\Phi(p))_k)((C\Phi v)_k)^2 \\ &= \sum_{k=1}^{d} w_k^2 + \sum_{k=1}^{n} ((C\Phi v)_k)^2 = \|\Phi v\|^2 + \|C\Phi v\|^2 = \|\Phi v\|^2 + (\Phi v)^{\mathsf{T}} C^{\mathsf{T}} C (\Phi v) \\ &= \|\Phi v\|^2 + (\sum_i \alpha_i \sigma^i u^i)^{\mathsf{T}} U (\Sigma^{-2} - I) U^{\mathsf{T}} (\sum_i \alpha_i \sigma^i u^i) \\ &= \|\Phi v\|^2 + [\alpha_1 \sigma^1, \ldots, \alpha_n \sigma^n](\Sigma^{-2} - I)[\alpha_1 \sigma^1, \ldots, \alpha_n \sigma^n]^{\mathsf{T}} \\ &= \|\Phi v\|^2 + (\sum_i \alpha_i^2 - \sum_i (\alpha_i \sigma^i)^2) = \|\Phi v\|^2 + \|v\|^2 - \|\Phi v\|^2 = \|v\|^2. \end{aligned}$$

In other words, our non-linear correction map $\bar{\Psi}$ can *exactly* restore the contraction caused by $\Phi$ for *any* vector tangent to an $n$-flat manifold.

In the fully general case, the situation gets slightly more complicated since we need to apply different spiral maps, each corresponding to a different size correction at different locations on the contracted manifold. Recall that we localize the effect of a correction by applying the so-called "bump" function (details below). These bump functions, although important for localization, have an undesirable effect on the stretched length of the tangent vector. Thus, to ameliorate their effect on the length of the resulting tangent vector, we control their contribution via a free parameter $\omega$.

**The general case**

More specifically, Embedding Technique I restores the contraction induced by $\Phi$ by applying a map $\Psi(t) := (t, \Psi_{1,\sin}(t), \Psi_{1,\cos}(t), \ldots, \Psi_{K,\sin}(t), \Psi_{K,\cos}(t))$ (recall that $K$ is the number of subsets we decompose $X$ into – cf. description in Embedding I in Section 4.3), with $\Psi_{j,\sin}(t) := (\psi^1_{j,\sin}(t), \ldots, \psi^n_{j,\sin}(t))$ and $\Psi_{j,\cos}(t) := (\psi^1_{j,\cos}(t), \ldots, \psi^n_{j,\cos}(t))$. The individual terms are given as

$$\begin{array}{l} \psi^i_{j,\sin}(t) := \sum_{x \in X^{(j)}} (\sqrt{\Lambda_{\Phi(x)}(t)}/\omega) \sin(\omega (C^x t)_i) \\ \psi^i_{j,\cos}(t) := \sum_{x \in X^{(j)}} (\sqrt{\Lambda_{\Phi(x)}(t)}/\omega) \cos(\omega (C^x t)_i) \end{array} \qquad i = 1, \ldots, n; j = 1, \ldots, K,$$

where $C^x$'s are the correction amounts for different locations $x$ on the manifold, $\omega > 0$ controls the frequency (cf. Section 4.3), and $\Lambda_{\Phi(x)}(t)$ is defined to be $\lambda_{\Phi(x)}(t) / \sum_{q \in X} \lambda_{\Phi(q)}(t)$, with

$$\lambda_{\Phi(x)}(t) := \begin{cases} \exp(-1/(1 - \|t - \Phi(x)\|^2/\rho^2)) & \text{if } \|t - \Phi(x)\| < \rho. \\ 0 & \text{otherwise.} \end{cases}$$

$\lambda$ is a classic example of a *bump function* (see Figure 4.4 middle). It is a smooth function with compact support. Its applicability arises from the fact that it can be made "to specifications". That is, it can be made to vanish outside any interval of our choice. Here we exploit this property to localize the effect of our corrections. The normalization of $\lambda$ (the function $\Lambda$) creates the so-called

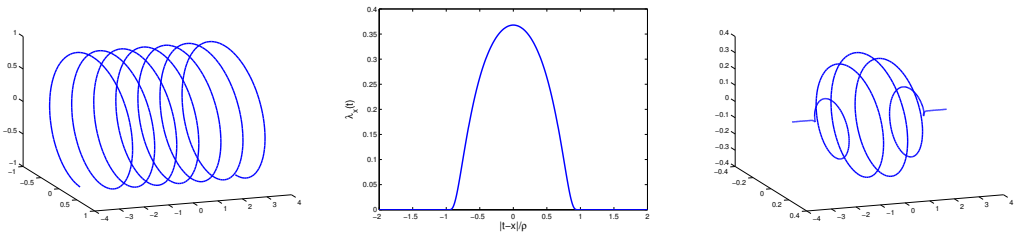


**Figure 4.4**: Effects of applying a bump function on a spiral mapping. Left: Spiral mapping $t \mapsto (t, \sin(t), \cos(t))$. Middle: Bump function $\lambda_x$: a smooth function with compact support. The parameter $x$ controls the location while $\rho$ controls the width. Right: The combined effect: $t \mapsto (t, \lambda_x(t)\sin(t), \lambda_x(t)\cos(t))$. Note that the effect of the spiral is localized while keeping the mapping smooth.

smooth partition of unity that helps to vary smoothly between the spirals applied at different regions of $M$.

Since any tangent vector in $\mathbb{R}^d$ can be expressed in terms of the basis vectors, it suffices to study how $D\Psi$ acts on the standard basis $\{e^i\}$. Note that

$$(D\Psi)_t(e^i) = \Big(\frac{dt}{dt^i}, \frac{d\Psi_{1,\sin}(t)}{dt^i}, \frac{d\Psi_{1,\cos}(t)}{dt^i}, \ldots, \frac{d\Psi_{K,\sin}(t)}{dt^i}, \frac{d\Psi_{K,\cos}(t)}{dt^i}\Big)\Big|_t,$$

where for $k \in [n]$, $i \in [d]$ and $j \in [K]$

$$\begin{aligned}
\frac{d\psi^k_{j,\sin}(t)}{dt^i} &= \sum_{x \in X^{(j)}} \frac{1}{\omega}\Big(\sin(\omega(C^x t)_k)\frac{d\Lambda^{1/2}_{\Phi(x)}(t)}{dt^i}\Big) + \sqrt{\Lambda_{\Phi(x)}(t)}\cos(\omega(C^x t)_k)C^x_{k,i} \\
\frac{d\psi^k_{j,\cos}(t)}{dt^i} &= \sum_{x \in X^{(j)}} \frac{1}{\omega}\Big(\cos(\omega(C^x t)_k)\frac{d\Lambda^{1/2}_{\Phi(x)}(t)}{dt^i}\Big) - \sqrt{\Lambda_{\Phi(x)}(t)}\sin(\omega(C^x t)_k)C^x_{k,i}
\end{aligned}$$

One can now observe the advantage of having the term $\omega$. By picking $\omega$ sufficiently large, we can make the first part of the expression sufficiently small. Now, for any tangent vector $u = \sum_i u_i e^i$ such that $\|u\| \le 1$, we have (by algebra)

$$\begin{aligned}
\|(D\Psi)_t(u)\|^2 &= \Big\|\sum u_i (D\Psi)_t(e^i)\Big\|^2 \\
&= \sum_{k=1}^{d} u_k^2 + \sum_{k=1}^{n}\sum_{j=1}^{K}\Big[\sum_{x \in X^{(j)}}\Big(\frac{A^{k,x}_{\sin}(t)}{\omega}\Big) + \sqrt{\Lambda_{\Phi(x)}(t)}\cos(\omega(C^x t)_k)(C^x u)_k\Big]^2 \\
&\qquad + \Big[\sum_{x \in X^{(j)}}\Big(\frac{A^{k,x}_{\cos}(t)}{\omega}\Big) - \sqrt{\Lambda_{\Phi(x)}(t)}\sin(\omega(C^x t)_k)(C^x u)_k\Big]^2, \qquad (4.1)
\end{aligned}$$

where

$$A^{k,x}_{\sin}(t) := \sum_i u_i \sin(\omega(C^x t)_k)(d\Lambda^{1/2}_{\Phi(x)}(t)/dt^i), \text{ and}$$
$$A^{k,x}_{\cos}(t) := \sum_i u_i \cos(\omega(C^x t)_k)(d\Lambda^{1/2}_{\Phi(x)}(t)/dt^i).$$

We can further simplify Eq. (4.1) and get

**Lemma 4.12** *Let $t$ be any point in $\Phi(M)$ and $u$ be any vector tagent to $\Phi(M)$ at $t$ such that $\|u\| \leq 1$. Let $\epsilon$ be the isometry parameter chosen in Theorem 4.9. Pick $\omega \geq \Omega(n\alpha^2 9^n \sqrt{d}/\rho\epsilon)$, then*

$$\|(D\Psi)_t(u)\|^2 = \|u\|^2 + \sum_{x\in X} \Lambda_{\Phi(x)}(t) \sum_{k=1}^{n} (C^x u)_k^2 + \zeta, \tag{4.2}$$

*where $|\zeta| \leq \epsilon/2$.*

We will use this derivation of $\|(D\Psi)_t(u)\|^2$ to study the combined effect of $\Psi \circ \Phi$ on $M$ in Section 4.5.4.

### 4.5.3 Effects of applying $\Psi$ (Algorithm II)

The goal of the second algorithm is to apply the spiralling corrections while using the coordinates more economically. We achieve this goal by applying them sequentially in the same embedding space (rather than simultaneously by making use of extra $2nK$ coordinates as done in the first algorithm), see also Nash [1954]. Since all the corrections will be sharing the same coordinate space, one needs to keep track of a pair of normal vectors in order to prevent interference among the different local corrections.

More specifically, $\Psi : \mathbb{R}^d \to \mathbb{R}^{2d+3}$ (in Algorithm II) is defined recursively as $\Psi := \Psi_{|X|,n}$ such that (see also Embedding II in Section 4.3)

$$\Psi_{i,j}(t) := \Psi_{i,j-1}(t) + \frac{\sqrt{\Lambda_{\Phi(x_i)}(t)}}{\omega_{i,j}} \Big( \eta_{i,j}(t) \sin(\omega_{i,j}(C^{x_i}t)_j) + \nu_{i,j}(t) \cos(\omega_{i,j}(C^{x_i}t)_j) \Big),$$

where $\Psi_{i,0}(t) := \Psi_{i-1,n}(t)$, and the base function $\Psi_{0,n}(t)$ is given as $t \mapsto (t, \overbrace{0, \ldots, 0}^{d+3})$. $\eta_{i,j}(t)$ and $\nu_{i,j}(t)$ are mutually orthogonal unit vectors that are approximately nor-

mal to $\Psi_{i,j-1}(\Phi M)$ at $\Psi_{i,j-1}(t)$. In this section we assume that the normals $\eta$ and $\nu$ have the following properties:

- $|\eta_{i,j}(t) \cdot v| \leq \epsilon_0$ and $|\nu_{i,j}(t) \cdot v| \leq \epsilon_0$ for all unit-length $v$ tangent to $\Psi_{i,j-1}(\Phi M)$ at $\Psi_{i,j-1}(t)$. (quality of normal approximation)

- For all $1 \leq l \leq d$, we have $\|d\eta_{i,j}(t)/dt^l\| \leq K_{i,j}$ and $\|d\nu_{i,j}(t)/dt^l\| \leq K_{i,j}$. (bounded directional derivatives)

We refer the reader to Section 4.10 for details on how to estimate such normals.

Now, as before, representing a tangent vector $u = \sum_l u_l e^l$ (such that $\|u\|^2 \leq 1$) in terms of its basis vectors, it suffices to study how $D\Psi$ acts on basis vectors. Observe that $(D\Psi_{i,j})_t(e^l) = \left(\frac{d\Psi_{i,j}(t)}{dt^l}\right)_{k=1}^{2d+3}\Big|_t$, with the $k^{\text{th}}$ component given as

$$\begin{aligned}
&\left(\frac{d\Psi_{i,j-1}(t)}{dt^l}\right)_k + (\eta_{i,j}(t))_k\sqrt{\Lambda_{\Phi(x_i)}(t)}C^{x_i}_{j,l}B^{i,j}_{\cos}(t) - (\nu_{i,j}(t))_k\sqrt{\Lambda_{\Phi(x_i)}(t)}C^{x_i}_{j,l}B^{i,j}_{\sin}(t)\\
&\qquad + \frac{1}{\omega_{i,j}}\Big[\Big(\frac{d\eta_{i,j}(t)}{dt^l}\Big)_k\sqrt{\Lambda_{\Phi(x_i)}(t)}B^{i,j}_{\sin}(t) + \Big(\frac{d\nu_{i,j}(t)}{dt^l}\Big)_k\sqrt{\Lambda_{\Phi(x_i)}(t)}B^{i,j}_{\cos}(t)\\
&\qquad\qquad + (\eta_{i,j}(t))_k\frac{d\Lambda^{1/2}_{\Phi(x_i)}(t)}{dt^l}B^{i,j}_{\sin}(t) + (\nu_{i,j}(t))_k\frac{d\Lambda^{1/2}_{\Phi(x_i)}(t)}{dt^l}B^{i,j}_{\cos}(t)\Big],
\end{aligned}$$

where $B^{i,j}_{\cos}(t) := \cos(\omega_{i,j}(C^{x_i}t)_j)$ and $B^{i,j}_{\sin}(t) := \sin(\omega_{i,j}(C^{x_i}t)_j)$. For ease of notation, let $R^{k,l}_{i,j}$ be the terms in the bracket (being multiplied to $1/\omega_{i,j}$) in the above expression. Then, we have (for any $i$, $j$)

$$\begin{aligned}
\|(D\Psi_{i,j})_t(u)\|^2 &= \Big\|\sum_l u_l(D\Psi_{i,j})_t(e^l)\Big\|^2\\
&= \sum_{k=1}^{2d+3}\Big[\underbrace{\sum_l u_l\left(\frac{d\Psi_{i,j-1}(t)}{dt^l}\right)_k}_{\zeta^{k,1}_{i,j}} + \underbrace{(\eta_{i,j}(t))_k\sqrt{\Lambda_{\Phi(x_i)}(t)}\cos(\omega_{i,j}(C^{x_i}t)_j)\sum_l C^{x_i}_{j,l}u_l}_{\zeta^{k,2}_{i,j}}\\
&\qquad \underbrace{-(\nu_{i,j}(t))_k\sqrt{\Lambda_{\Phi(x_i)}(t)}\sin(\omega_{i,j}(C^{x_i}t)_j)\sum_l C^{x_i}_{j,l}u_l}_{\zeta^{k,3}_{i,j}} + (1/\omega_{i,j})\underbrace{\sum_l u_l R^{k,l}_{i,j}}_{\zeta^{k,4}_{i,j}}\Big]^2\\
&= \underbrace{\|(D\Psi_{i,j-1})_t(u)\|^2}_{=\sum_k\left(\zeta^{k,1}_{i,j}\right)^2} + \underbrace{\Lambda_{\Phi(x_i)}(t)(C^{x_i}u)^2_j}_{=\sum_k\left(\zeta^{k,2}_{i,j}\right)^2+\left(\zeta^{k,3}_{i,j}\right)^2}
\end{aligned}$$

$$+\underbrace{\sum_k \Big[\big(\zeta_{i,j}^{k,4}/\omega_{i,j}\big)^2 + \big(2\zeta_{i,j}^{k,4}/\omega_{i,j}\big)\big(\zeta_{i,j}^{k,1}+\zeta_{i,j}^{k,2}+\zeta_{i,j}^{k,3}\big) + 2\big(\zeta_{i,j}^{k,1}\zeta_{i,j}^{k,2}+\zeta_{i,j}^{k,1}\zeta_{i,j}^{k,3}\big)\Big]}_{Z_{i,j}},$$

(4.3)

where the last equality is by expanding the square and by noting that $\sum_k \zeta_{i,j}^{k,2}\zeta_{i,j}^{k,3} = 0$ since $\eta$ and $\nu$ are orthogonal to each other. The base case $\|(D\Psi_{0,n})_t(u)\|^2$ equals $\|u\|^2$.

By picking $\omega_{i,j}$ sufficiently large, and by noting that the cross terms $\sum_k(\zeta_{i,j}^{k,1}\zeta_{i,j}^{k,2})$ and $\sum_k(\zeta_{i,j}^{k,1}\zeta_{i,j}^{k,3})$ are very close to zero since $\eta$ and $\nu$ are approximately normal to the tangent vector, we have

**Lemma 4.13** *Let $t$ be any point in $\Phi(M)$ and $u$ be any vector tagent to $\Phi(M)$ at $t$ such that $\|u\| \leq 1$. Let $\epsilon$ be the isometry parameter chosen in Theorem 4.9. Pick $\omega_{i,j} \geq \Omega\big((K_{i,j} + (\alpha 9^n/\rho))(nd|X|)^2/\epsilon\big)$ (recall that $K_{i,j}$ is the bound on the directional derivate of $\eta$ and $\nu$). If $\epsilon_0 \leq O\big(\epsilon/d(n|X|)^2\big)$ (recall that $\epsilon_0$ is the quality of approximation of the normals $\eta$ and $\nu$), then we have*

$$\|(D\Psi)_t(u)\|^2 = \|(D\Psi_{|X|,n})_t(u)\|^2 = \|u\|^2 + \sum_{i=1}^{|X|}\Lambda_{\Phi(x_i)}(t)\sum_{j=1}^{n}(C^{x_i}u)_j^2 + \zeta, \quad (4.4)$$

*where $|\zeta| \leq \epsilon/2$.*

### 4.5.4 Combined effect of $\Psi(\Phi(M))$

We can now analyze the aggregate effect of both our embeddings on the length of an arbitrary unit vector $v$ tangent to $M$ at $p$. Let $u := (D\Phi)_p(v) = \Phi v$ be the pushforward of $v$. Then $\|u\| \leq 1$ (cf. Lemma 4.10). See also Figure 4.3.

Now, recalling that $D(\Psi \circ \Phi) = D\Psi \circ D\Phi$, and noting that pushforward maps are linear, we have $\|(D(\Psi\circ\Phi))_p(v)\|^2 = \big\|(D\Psi)_{\Phi(p)}(u)\big\|^2$. Thus, representing $u$ as $\sum_i u_i e^i$ in ambient coordinates of $\mathbb{R}^d$, and using Eq. (4.2) (for Algorithm I) or Eq. (4.4) (for Algorithm II), we get

$$\big\|(D(\Psi\circ\Phi))_p(v)\big\|^2 \;=\; \big\|(D\Psi)_{\Phi(p)}(u)\big\|^2 = \|u\|^2 + \sum_{x\in X}\Lambda_{\Phi(x)}(\Phi(p))\|C^x u\|^2 + \zeta,$$

where $|\zeta| \leq \epsilon/2$. We can give simple lower and upper bounds for the above expression by noting that $\Lambda_{\Phi(x)}$ is a localization function. Define $N_p := \{x \in X : \|\Phi(x) - \Phi(p)\| < \rho\}$ as the neighborhood around $p$ ($\rho$ as per the theorem statement). Then only the points in $N_p$ contribute to above equation, since $\Lambda_{\Phi(x)}(\Phi(p)) = d\Lambda_{\Phi(x)}(\Phi(p))/dt^i = 0$ for $\|\Phi(x) - \Phi(p)\| \geq \rho$. Also note that for all $x \in N_p$, $\|x - p\| < 2\rho$ (cf. Lemma 4.10).

Let $x_M := \arg\max_{x \in N_p} \|C^x u\|^2$ and $x_m := \arg\min_{x \in N_p} \|C^x u\|^2$ are quantities that attain the maximum and the minimum respectively, then:

$$\|u\|^2 + \|C^{x_m} u\|^2 - \epsilon/2 \leq \|(D(\Psi \circ \Phi))_p(v)\|^2 \leq \|u\|^2 + \|C^{x_M} u\|^2 + \epsilon/2. \quad (4.5)$$

Notice that ideally we would like to have the correction factor "$C^p u$" in Eq. (4.5) since that would give the perfect stretch around the point $p$. But what about correction $C^x u$ for closeby $x$'s? The following lemma helps us continue in this situation.

**Lemma 4.14** *Let $p$, $v$, $u$ be as above. For any $x \in N_p \subset X$, let $C^x$ and $F_x$ also be as discussed above (recall that $\|p - x\| < 2\rho$, and $X \subset M$ forms a bounded $(\rho, \delta)$-cover of the fixed underlying manifold $M$ with condition number $1/\tau$). Define $\xi := (4\rho/\tau) + \delta + 4\sqrt{\rho\delta/\tau}$. If $\rho \leq \tau/4$ and $\delta \leq d/32D$, then*

$$1 - \|u\|^2 - 40 \cdot \max\left\{\sqrt{\xi D/d}, \xi D/d\right\} \leq \|C^x u\|^2 \leq 1 - \|u\|^2 + 51 \cdot \max\left\{\sqrt{\xi D/d}, \xi D/d\right\}.$$

Note that we chose $\rho \leq (\tau d/D)(\epsilon/350)^2$ and $\delta \leq (d/D)(\epsilon/250)^2$ (cf. theorem statement). Thus, combining Eq. (4.5) and Lemma 4.14, we get (recall $\|v\| = 1$)

$$(1 - \epsilon)\|v\|^2 \leq \|(D(\Psi \circ \Phi))_p(v)\|^2 \leq (1 + \epsilon)\|v\|^2.$$

So far we have shown that our embedding approximately preserves the length of a fixed tangent vector at a fixed point. Since the choice of the vector and the point was arbitrary, it follows that our embedding approximately preserves the tangent vector lengths throughout the embedded manifold uniformly. We will now show that preserving the tangent vector lengths implies preserving the geodesic curve lengths.

### 4.5.5 Preservation of the geodesic lengths

Pick any two (path-connected) points $p$ and $q$ in $M$, and let $\alpha$ be the geodesicpath between $p$ and $q$. Further let $\bar{p}$, $\bar{q}$ and $\bar{\alpha}$ be the images of $p$, $q$ and $\alpha$ under our embedding. Note that $\bar{\alpha}$ is not necessarily the geodesic path between $\bar{p}$ and $\bar{q}$, thus we need an extra piece of notation: let $\bar{\beta}$ be the geodesic path between $\bar{p}$ and $\bar{q}$ (under the embedded manifold) and $\beta$ be its inverse image in $M$. We need to show $(1-\epsilon)L(\alpha) \leq L(\bar{\beta}) \leq (1+\epsilon)L(\alpha)$, where $L(\cdot)$ denotes the length of the path $\cdot$ (end points are understood).

First recall that for any differentiable map $F$ and curve $\gamma$, $\bar{\gamma} = F(\gamma) \Rightarrow \bar{\gamma}' = (DF)(\gamma')$. By $(1 \pm \epsilon)$-isometry of tangent vectors, this immediately gives us $(1-\epsilon)L(\gamma) \leq L(\bar{\gamma}) \leq (1+\epsilon)L(\gamma)$ for any path $\gamma$ in $M$ and its image $\bar{\gamma}$ in embedding of $M$. So,

$$(1-\epsilon)D_G(p,q) = (1-\epsilon)L(\alpha) \leq (1-\epsilon)L(\beta) \leq L(\bar{\beta}) = D_G(\bar{p},\bar{q}).$$

Similarly,

$$D_G(\bar{p},\bar{q}) = L(\bar{\beta}) \leq L(\bar{\alpha}) \leq (1+\epsilon)L(\alpha) = (1+\epsilon)D_G(p,q).$$

## 4.6 Discussion

This work provides two algorithms for $(1\pm\epsilon)$-isometric embedding of generic $n$-dimensional manifolds. Our algorithms are similar in spirit to Nash's construction [Nash, 1954], and manage to remove the dependence on the isometry constant $\epsilon$ from the final embedding dimension. Note that this dependency does necessarily show up in the sampling density required to make the corrections.

The correction procedure discussed here can also be readily adapted to create isometric embeddings from any manifold embedding procedure (under some mild conditions). Take any off-the-shelf manifold embedding algorithm $\mathcal{A}$ (such as LLE, Laplacian Eigenmaps, etc.) that maps an $n$-dimensional manifold in, say, $d$ dimensions, but does not necessarily guarantee an approximate isometric embedding. Then as long as one can ensure that $\mathcal{A}$ is a one-to-one mapping that doesn't collapse interpoint distances, we can scale the output returned by $\mathcal{A}$ to

create a contraction. The scaled version of $\mathcal{A}$ acts as the Embedding Stage of our algorithm. We can thus apply the Corrections Stage (either the one discussed in Algorithm I or Algorithm II) to produce an approximate isometric embedding of the given manifold in slightly higher dimensions. In this sense, the correction procedure presented here serves as a *universal procedure* for approximate isometric manifold embeddings.

## Acknowledgements

The contents of this chapter originally appeared in the following publication: N. Verma. Distance preserving embeddings for general $n$-dimensional manifolds. *Conference on Learning Theory (COLT)*, 2012.

## 4.7 On constructing a bounded manifold cover

Given a compact $n$-manifold $M \subset \mathbb{R}^D$ with condition number $1/\tau$, and some $0 < \delta \leq 1$, we can construct an $\alpha$-bounded $(\rho, \delta)$ cover $X$ of $M$ (with $\alpha \leq 2^{10n+1}$ and $\rho \leq \tau\delta/3\sqrt{2}n$) as follows.

Set $\rho \leq \tau\delta/3\sqrt{2}n$ and pick a $(\rho/2)$-net $C$ of $M$ (that is $C \subset M$ such that, i. for $c, c' \in C$ such that $c \neq c'$, $\|c - c'\| \geq \rho/2$, ii. for all $p \in M$, exists $c \in C$ such that $\|c - p\| < \rho/2$). WLOG we shall assume that all points of $C$ are in the interior of $M$. Then, for each $c \in C$, define $M_{c,\rho/2} := \{p \in M : \|p - c\| \leq \rho/2\}$, and the orthogonal projection map $f_c : M_{c,\rho/2} \to T_cM$ that projects $M_{c,\rho/2}$ onto $T_cM$ (note that, cf. Lemma A.3(i), $f_c$ is one-to-one). Note that $T_cM$ can be identified with $\mathbb{R}^n$ with the $c$ as the origin. We will denote the origin as $x_0^{(c)}$, that is, $x_0^{(c)} = f_c(c)$.

Now, let $B_c$ be any $n$-dimensional closed ball centered at the origin $x_0^{(c)} \in T_cM$ of radius $r > 0$ that is completely contained in $f_c(M_{c,\rho/2})$ (that is, $B_c \subset f_c(M_{c,\rho/2})$). Pick a set of $n$ points $x_1^{(c)}, \ldots, x_n^{(c)}$ on the surface of the ball $B_c$ such that $(x_i^{(c)} - x_0^{(c)}) \cdot (x_j^{(c)} - x_0^{(c)}) = 0$ for $i \neq j$.

Define the bounded manifold cover as

$$X := \bigcup_{c \in C, i=0,\ldots,n} f_c^{-1}(x_i^{(c)}). \tag{4.6}$$

**Lemma 4.15** *Let $0 < \delta \leq 1$ and $\rho \leq \tau\delta/3\sqrt{2}n$. Let $C$ be a $(\rho/2)$-net of $M$ as described above, and $X$ be as in Eq.* (4.6)*. Then $X$ forms a $2^{10n+1}$-bounded $(\rho, \delta)$ cover of $M$.*

*Proof.* Pick any point $p \in M$ and define $X_p := \{x \in X : \|x - p\| < \rho\}$. Let $c \in C$ be such that $\|p - c\| < \rho/2$. Then $X_p$ has the following properties.

*Covering criterion:* For $0 \leq i \leq n$, since $\|f_c^{-1}(x_i^{(c)}) - c\| \leq \rho/2$ (by construction), we have $\|f_c^{-1}(x_i^{(c)}) - p\| < \rho$. Thus, $f_c^{-1}(x_i^{(c)}) \in X_p$ (for $0 \leq i \leq n$). Now, for $1 \leq i \leq n$, noting that $D_G(f_c^{-1}(x_i^{(c)}), f_c^{-1}(x_0^{(c)})) \leq 2\|f_c^{-1}(x_i^{(c)}) - f_c^{-1}(x_0^{(c)})\| \leq \rho$ (cf. Lemma A.2), we have that for the vector $\hat{v}_i^{(c)} := \frac{f_c^{-1}(x_i^{(c)}) - f_c^{-1}(x_0^{(c)})}{\|f_c^{-1}(x_i^{(c)}) - f_c^{-1}(x_0^{(c)})\|}$ and its (normalized) projection $v_i^{(c)} := \frac{x_i^{(c)} - x_0^{(c)}}{\|x_i^{(c)} - x_0^{(c)}\|}$ onto $T_cM$, $\left\|\hat{v}_i^{(c)} - v_i^{(c)}\right\| \leq \rho/\sqrt{2}\tau$ (cf. Lemma A.5). Thus, for $i \neq j$, we have (recall, by construction, we have $v_i^{(c)} \cdot v_j^{(c)} = 0$)

$$\begin{aligned}
|\hat{v}_i^{(c)} \cdot \hat{v}_j^{(c)}| &= |(\hat{v}_i^{(c)} - v_i^{(c)} + v_i^{(c)}) \cdot (\hat{v}_j^{(c)} - v_j^{(c)} + v_j^{(c)})| \\
&= |(\hat{v}_i^{(c)} - v_i^{(c)}) \cdot (\hat{v}_j^{(c)} - v_j^{(c)}) + v_i^{(c)} \cdot (\hat{v}_j^{(c)} - v_j^{(c)}) + (\hat{v}_i^{(c)} - v_i^{(c)}) \cdot v_j^{(c)}| \\
&\leq \|(\hat{v}_i^{(c)} - v_i^{(c)})\|\|(\hat{v}_j^{(c)} - v_j^{(c)})\| + \|\hat{v}_i^{(c)} - v_i^{(c)}\| + \|\hat{v}_j^{(c)} - v_j^{(c)}\| \\
&\leq 3\rho/\sqrt{2}\tau \leq 1/2n.
\end{aligned}$$

*Point representation criterion:* There exists $x \in X_p$, namely $f_c^{-1}(x_0^{(c)})$ $(= c)$, such that $\|p - x\| \leq \rho/2$.

*Local boundedness criterion:* Define $M_{p,3\rho/2} := \{q \in M : \|q - p\| < 3\rho/2\}$. Note that $X_p \subset \{f_c^{-1}(x_i^{(c)}) : c \in C \cap M_{p,3\rho/2}, 0 \leq i \leq n\}$. Now, using Lemma A.4 we have that there exists a cover $N \subset M_{p,3\rho/2}$ of size at most $9^{3n}$ such that for any point $q \in M_{p,3\rho/2}$, there exists $n' \in N$ such that $\|q - n'\| < \rho/4$. Note that, by construction of $C$, there cannot be an $n' \in N$ such that it is within distance $\rho/4$ of two (or more) distinct $c, c' \in C$ (since otherwise the distance $\|c - c'\|$ will be less than $\rho/2$, contradicting the packing of $C$). Thus, $|C \cap M_{p,3\rho/2}| \leq 9^{3n}$. It follows

that $|X_p| \leq (n+1)9^{3n} \leq 2^{10n+1}$.

*Tangent space approximation criterion:* Let $\hat{T}_p$ be the $n$-dimensional span of $\{\hat{v}_i^{(c)}\}_{i\in[n]}$ (note that $\hat{T}_p$ may not necessarily pass through $p$). Then, for any unit vector $\hat{u} \in \hat{T}_p$, we need to show that its projection $u_p$ onto $T_pM$ has the property $|\hat{u} \cdot \frac{u_p}{\|u_p\|}| \geq 1-\delta$. Let $\theta$ be the angle between vectors $\hat{u}$ and $u_p$. Let $u_c$ be the projection of $\hat{u}$ onto $T_cM$, and $\theta_1$ be the angle between vectors $\hat{u}$ and $u_c$, and let $\theta_2$ be the angle between vectors $u_c$ (at $c$) and its parallel transport along the geodesic path to $p$. WLOG we can assume that $\theta_1$ and $\theta_2$ are at most $\pi/2$. Then, $\theta \leq \theta_1+\theta_2 \leq \pi$. We get the bound on the individual angles as follows. By applying Lemma A.6, $\cos(\theta_1) \geq 1-\delta/4$, and by applying Lemma A.1, $\cos(\theta_2) \geq 1-\delta/4$. Finally, by using Lemma 4.16, we have $\left|\hat{u} \cdot \frac{u_p}{\|u_p\|}\right| = \cos(\theta) \geq \cos(\theta_1+\theta_2) \geq 1-\delta$. ∎

**Lemma 4.16** *Let $0 \leq \epsilon_1, \epsilon_2 \leq 1$. If $\cos\alpha \geq 1-\epsilon_1$ and $\cos\beta \geq 1-\epsilon_2$, then $\cos(\alpha+\beta) \geq 1-\epsilon_1-\epsilon_2-2\sqrt{\epsilon_1\epsilon_2}$.*

*Proof.* Applying the identity $\sin\theta = \sqrt{1-\cos^2\theta}$ immediately yields $\sin\alpha \leq \sqrt{2\epsilon_1}$ and $\sin\beta \leq \sqrt{2\epsilon_2}$. Now, $\cos(\alpha+\beta) = \cos\alpha\cos\beta - \sin\alpha\sin\beta \geq (1-\epsilon_1)(1-\epsilon_2) - 2\sqrt{\epsilon_1\epsilon_2} \geq 1-\epsilon_1-\epsilon_2-2\sqrt{\epsilon_1\epsilon_2}$. ∎

**Remark 4.17** *A dense enough sample from $M$ constitutes as a bounded cover. One can selectively prune the dense sampling to control the total number of points in each neighborhood, while still maintaining the cover properties.*

## 4.8 Bounding the number of subsets $K$ in Embedding I

By construction (see preprocessing stage of Embedding I), $K = \max_{x\in X} |X \cap B(x, 2\rho)|$ (where $B(x,r)$ denotes a Euclidean ball centered at $x$ of radius $r$). That is, $K$ is the largest number of $x$'s ($\in X$) that are within a $2\rho$ ball of some $x \in X$.

Now, pick any $x \in X$ and consider the set $M_x := M \cap B(x, 2\rho)$. Then, if $\rho \leq \tau/4$, $M_x$ can be covered by $2^{cn}$ balls of radius $\rho$ (see Lemma A.4). By recalling

that $X$ forms an $\alpha$-bounded $(\rho, \delta)$-cover, we have $|X \cap B(x, 2\rho)| = |X \cap M_x| \leq \alpha 2^{cn}$ (where $c \leq 4$).

## 4.9 Supporting proofs

### 4.9.1 Proof of Lemma 4.10

Since $R$ is a random orthoprojector from $\mathbb{R}^D$ to $\mathbb{R}^d$, it follows that

**Lemma 4.18 (random projection of $n$-manifolds – adapted from Theorem 1.5 of Clarkson [2008])** *Let $M$ be a smooth compact $n$-manifold with volume $V$ and condition number $1/\tau$. Let $\bar{R} := \sqrt{D/d}R$ be a scaling of $R$. Pick any $0 < \epsilon \leq 1$ and $0 < \delta \leq 1$. If $d = \Omega\big(\epsilon^{-2}\log(V/\tau^n) + \epsilon^{-2}n\log(1/\epsilon) + \ln(1/\delta)\big)$, then with probability at least $1 - \delta$, for all $p, q \in M$*

$$(1-\epsilon)\|p-q\| \leq \|\bar{R}p - \bar{R}q\| \leq (1+\epsilon)\|p-q\|.$$

We apply this result with $\epsilon = 1/4$. Then, for $d = \Omega(\log(V/\tau^n)+n)$, with probability at least $1 - 1/\text{poly}(n)$, $(3/4)\|p-q\| \leq \|\bar{R}p - \bar{R}q\| \leq (5/4)\|p-q\|$. Now let $\Phi : \mathbb{R}^D \to \mathbb{R}^d$ be defined as $\Phi x := (2/3)\bar{R}x = (2/3)(\sqrt{D/d})x$ (as per the lemma statement). Then we immediately get $(1/2)\|p-q\| \leq \|\Phi p - \Phi q\| \leq (5/6)\|p-q\|$.

Also note that for any $x \in \mathbb{R}^D$, we have $\|\Phi x\| = (2/3)(\sqrt{D/d})\|Rx\| \leq (2/3)(\sqrt{D/d})\|x\|$ (since $R$ is an orthoprojector).

Finally, for any point $p \in M$, a unit vector $u$ tangent to $M$ at $p$ can be approximated arbitrarily well by considering a sequence $\{p_i\}_i$ of points (in $M$) converging to $p$ (in $M$) such that $(p_i - p)/\|p_i - p\|$ converges to $u$. Since for all points $p_i$, $(1/2) \leq \|\Phi p_i - \Phi p\|/\|p_i - p\| \leq (5/6)$ (with high probability), it follows that $(1/2) \leq \|(D\Phi)_p(u)\| \leq (5/6)$.

### 4.9.2 Proof of Corollary 4.11

Let $v_x^1$ and $v_x^n$ ($\in \mathbb{R}^n$) be the right singular vectors corresponding to singular values $\sigma_x^1$ and $\sigma_x^n$ respectively of the matrix $\Phi F_x$. Then, quickly note that $\sigma_x^1 = \|\Phi F_x v^1\|$, and $\sigma_x^n = \|\Phi F_x v^n\|$. Note that since $F_x$ is orthonormal, we have that

$\|F_x v^1\| = \|F_x v^n\| = 1$. Now, since $F_x v^n$ is in the span of column vectors of $F_x$, by the sampling condition (cf. Definition 4.1), there exists a unit length vector $\bar{v}_x^n$ tangent to $M$ (at $x$) such that $|F_x v_x^n \cdot \bar{v}_x^n| \geq 1 - \delta$. Thus, decomposing $F_x v_x^n$ into two vectors $a_x^n$ and $b_x^n$ such that $a_x^n \perp b_x^n$ and $a_x^n := (F_x v_x^n \cdot \bar{v}_x^n)\bar{v}_x^n$, we have

$$\begin{aligned}
\sigma_x^n &= \|\Phi(F_x v^n)\| = \|\Phi((F_x v_x^n \cdot \bar{v}_x^n)\bar{v}_x^n) + \Phi b_x^n\| \\
&\geq (1-\delta)\,\|\Phi \bar{v}_x^n\| - \|\Phi b_x^n\| \\
&\geq (1-\delta)(1/2) - (2/3)\sqrt{2\delta D/d},
\end{aligned}$$

since $\|b_x^n\|^2 = \|F_x v_x^n\|^2 - \|a_x^n\|^2 \leq 1-(1-\delta)^2 \leq 2\delta$ and $\|\Phi b_x^n\| \leq (2/3)(\sqrt{D/d})\|b_x^n\| \leq (2/3)\sqrt{2\delta D/d}$. Similarly decomposing $F_x v_x^1$ into two vectors $a_x^1$ and $b_x^1$ such that $a_x^1 \perp b_x^1$ and $a_x^1 := (F_x v_x^1 \cdot \bar{v}_x^1)\bar{v}_x^1$, we have

$$\begin{aligned}
\sigma_x^1 &= \|\Phi(F_x v_x^1)\| = \|\Phi((F_x v_x^1 \cdot \bar{v}_x^1)\bar{v}_x^1) + \Phi b_x^1\| \\
&\leq \left\|\Phi \bar{v}_x^1\right\| + \|\Phi b_x^1\| \\
&\leq (5/6) + (2/3)\sqrt{2\delta D/d},
\end{aligned}$$

where the last inequality is by noting $\|\Phi b_x^1\| \leq (2/3)\sqrt{2\delta D/d}$. Now, by our choice of $\delta$ ($\leq d/32D$), and by noting that $d \leq D$, the corollary follows.

### 4.9.3 Proof of Lemma 4.12

We can simplify Eq. (4.1) by recalling how the subsets $X^{(j)}$ were constructed (see preprocessing stage of Embedding I). Note that for any fixed $t$, at most one term in the set $\{\Lambda_{\Phi(x)}(t)\}_{x \in X^{(j)}}$ is non-zero. Thus,

$$\begin{aligned}
\|(D\Psi)_t(u)\|^2 &= \sum_{k=1}^{d} u_k^2 + \sum_{k=1}^{n} \sum_{x \in X} \Lambda_{\Phi(x)}(t)(C^x u)_k^2(\cos^2(\omega(C^x t)_k) + \sin^2(\omega(C^x t)_k)) \\
&\quad + \frac{1}{\omega}\Bigg[ \underbrace{\left(\left(A_{\sin}^{k,x}(t)\right)^2 + \left(A_{\cos}^{k,x}(t)\right)^2\right)/\omega}_{\zeta_1} \\
&\quad\quad + \underbrace{2A_{\sin}^{k,x}(t)\sqrt{\Lambda_{\Phi(x)}(t)}\cos(\omega(C^x t)_k)(C^x u)_k}_{\zeta_2} \\
&\quad\quad \underbrace{-2A_{\cos}^{k,x}(t)\sqrt{\Lambda_{\Phi(x)}(t)}\sin(\omega(C^x t)_k)(C^x u)_k}_{\zeta_3} \Bigg]
\end{aligned}$$

$$= \|u\|^2 + \sum_{x\in X} \Lambda_{\Phi(x)}(t) \sum_{k=1}^{n} (C^x u)_k^2 + \zeta,$$

where $\zeta := (\zeta_1 + \zeta_2 + \zeta_3)/\omega$. Noting that i) the terms $|A^{k,x}_{\sin}(t)|$ and $|A^{k,x}_{\cos}(t)|$ are at most $O(\alpha 9^n\sqrt{d}/\rho)$ (see Lemma 4.19), ii) $|(C^x u)_k| \le 4$, and iii) $\sqrt{\Lambda_{\Phi(x)}(t)} \le 1$, we can pick $\omega$ sufficiently large (say, $\omega \ge \Omega(n\alpha^2 9^n\sqrt{d}/\rho\epsilon)$ such that $|\zeta| \le \epsilon/2$ (where $\epsilon$ is the isometry constant from our main theorem).

**Lemma 4.19** *For all $k$, $x$ and $t$, the terms $|A^{k,x}_{\sin}(t)|$ and $|A^{k,x}_{\cos}(t)|$ are at most $O(\alpha 9^n\sqrt{d}/\rho)$.*

*Proof.* We shall focus on bounding $|A^{k,x}_{\sin}(t)|$ (the steps for bounding $|A^{k,x}_{\cos}(t)|$ are similar). Note that

$$\begin{aligned} |A^{k,x}_{\sin}(t)| &= \Big|\sum_{i=1}^{d} u_i \sin(\omega (C^x t)_k) \frac{d\Lambda^{1/2}_{\Phi(x)}(t)}{dt^i}\Big| \\ &\le \sum_{i=1}^{d} |u_i| \cdot \Big|\frac{d\Lambda^{1/2}_{\Phi(x)}(t)}{dt^i}\Big| \le \sqrt{\sum_{i=1}^{d} \Big|\frac{d\Lambda^{1/2}_{\Phi(x)}(t)}{dt^i}\Big|^2}, \end{aligned}$$

since $\|u\| \le 1$. Thus, we can bound $|A^{k,x}_{\sin}(t)|$ by $O(\alpha 9^n\sqrt{d}/\rho)$ by noting the following lemma. ∎

**Lemma 4.20** *For all $i$, $x$ and $t$, $|d\Lambda^{1/2}_{\Phi(x)}(t)/dt^i| \le O(\alpha 9^n/\rho)$.*

*Proof.* Pick any $t \in \Phi(M)$, and let $p_0 \in M$ be (the unique element) such that $\Phi(p_0) = t$. Define $N_{p_0} := \{x \in X : \|\Phi(x) - \Phi(p_0)\| < \rho\}$ as the neighborhood around $p_0$. Fix an arbitrary $x_0 \in N_{p_0} \subset X$ (since if $x_0 \notin N_{p_0}$ then $d\Lambda^{1/2}_{\Phi(x_0)}(t)/dt^i = 0$), and consider the function

$$\Lambda^{1/2}_{\Phi(x_0)}(t) = \left(\frac{\lambda_{\Phi(x_0)}(t)}{\sum_{x\in N_{p_0}} \lambda_{\Phi(x)}(t)}\right)^{1/2} = \left(\frac{e^{-1/(1-(\|t-\Phi(x_0)\|^2/\rho^2))}}{\sum_{x\in N_{p_0}} e^{-1/(1-(\|t-\Phi(x)\|^2/\rho^2))}}\right)^{1/2}.$$

Pick an arbitrary coordinate $i_0 \in \{1, \ldots, d\}$ and consider the (directional) derivative of this function

$$\frac{d\Lambda^{1/2}_{\Phi(x_0)}(t)}{dt^{i_0}} = \frac{1}{2}\big(\Lambda^{-1/2}_{\Phi(x_0)}(t)\big)\Big(\frac{d\Lambda_{\Phi(x_0)}(t)}{dt^{i_0}}\Big)$$

$$
\begin{aligned}
&= \frac{\Big(\sum_{x\in N_{p_0}} e^{-A_t(x)}\Big)^{1/2}}{2\Big(e^{-A_t(x_0)}\Big)^{1/2}}\Bigg[\frac{\Big(\sum_{x\in N_{p_0}} e^{-A_t(x)}\Big)\Big(\frac{-2(t_{i_0}-\Phi(x_0)_{i_0})}{\rho^2}(A_t(x_0))^2\Big)\Big(e^{-A_t(x_0)}\Big)}{\Big(\sum_{x\in N_{p_0}} e^{-A_t(x)}\Big)^2} \\
&\qquad\qquad - \frac{\Big(e^{-A_t(x_0)}\Big)\Big(\sum_{x\in N_{p_0}} \frac{-2(t_{i_0}-\Phi(x)_{i_0})}{\rho^2}(A_t(x))^2 e^{-A_t(x)}\Big)}{\Big(\sum_{x\in N_{p_0}} e^{-A_t(x)}\Big)^2}\Bigg] \\
&= \frac{\Big(\sum_{x\in N_{p_0}} e^{-A_t(x)}\Big)\Big(\frac{-2(t_{i_0}-\Phi(x_0)_{i_0})}{\rho^2}(A_t(x_0))^2\Big)\Big(e^{-A_t(x_0)}\Big)^{1/2}}{2\Big(\sum_{x\in N_{p_0}} e^{-A_t(x)}\Big)^{1.5}} \\
&\qquad - \frac{\Big(e^{-A_t(x_0)}\Big)^{1/2}\Big(\sum_{x\in N_{p_0}} \frac{-2(t_{i_0}-\Phi(x)_{i_0})}{\rho^2}(A_t(x))^2 e^{-A_t(x)}\Big)}{2\Big(\sum_{x\in N_{p_0}} e^{-A_t(x)}\Big)^{1.5}},
\end{aligned}
$$

where $A_t(x) := 1/(1-(\|t-\Phi(x)\|^2/\rho^2))$. Observe that the domain of $A_t$ is $\{x \in X : \|t-\Phi(x)\| < \rho\}$ and the range is $[1,\infty)$. Recalling that for any $\beta \geq 1$, $|\beta^2 e^{-\beta}| \leq 1$ and $|\beta^2 e^{-\beta/2}| \leq 3$, we have that $|A_t(\cdot)^2 e^{-A_t(\cdot)}| \leq 1$ and $|A_t(\cdot)^2 e^{-A_t(\cdot)/2}| \leq 3$. Thus,

$$
\begin{aligned}
&\Big|\frac{d\Lambda^{1/2}_{\Phi(x_0)}(t)}{dt^{i_0}}\Big| \\
&\leq \frac{3\cdot\Big|\sum_{x\in N_{p_0}} e^{-A_t(x)}\Big|\cdot\Big|\frac{2(t_{i_0}-\Phi(x_0)_{i_0})}{\rho^2}\Big| + \Big|e^{-A_t(x_0)/2}\Big|\cdot\Big|\sum_{x\in N_{p_0}} \frac{2(t_{i_0}-\Phi(x)_{i_0})}{\rho^2}\Big|}{2\Big(\sum_{x\in N_{p_0}} e^{-A_t(x)}\Big)^{1.5}} \\
&\leq \frac{(3)(2/\rho)\Big|\sum_{x\in N_{p_0}} e^{-A_t(x)}\Big| + \Big|e^{-A_t(x_0)/2}\Big|\sum_{x\in N_{p_0}} (2/\rho)}{2\Big(\sum_{x\in N_{p_0}} e^{-A_t(x)}\Big)^{1.5}} \\
&\leq O(\alpha 9^n/\rho),
\end{aligned}
$$

where the last inequality is by noting: i) $|N_{p_0}| \leq \alpha 9^n$ (since for all $x \in N_{p_0}$, $\|x-p_0\| \leq 2\rho$ – cf. Lemma 4.10, $X$ is an $\alpha$-bounded cover, and by noting that

for $\rho \leq \tau/4$, a ball of radius $2\rho$ can be covered by $9^n$ balls of radius $\rho$ on the given $n$-manifold – cf. Lemma A.4), ii) $|e^{-A_t(x)}| \leq |e^{-A_t(x)/2}| \leq 1$ (for all $x$), and iii) $\sum_{x \in N_{p_0}} e^{-A_t(x)} \geq \Omega(1)$ (since our cover $X$ ensures that for any $p_0$, there exists $x \in N_{p_0} \subset X$ such that $\|p_0 - x\| \leq \rho/2$ – see also Remark 4.2, and hence $e^{-A_t(x)}$ is non-negligible for some $x \in N_{p_0}$). ∎

### 4.9.4 Proof of Lemma 4.13

Note that by definition, $\|(D\Psi)_t(u)\|^2 = \|(D\Psi_{|X|,n})_t(u)\|^2$. Thus, using Eq. (4.3) and expanding the recursion, we have

$$\begin{aligned}
\|(D\Psi)_t(u)\|^2 &= \|(D\Psi_{|X|,n})_t(u)\|^2 \\
&= \|(D\Psi_{|X|,n-1})_t(u)\|^2 + \Lambda_{\Phi(x_{|X|})}(t)(C^{x_{|X|}}u)_n^2 + Z_{|X|,n} \\
&\vdots \\
&= \|(D\Psi_{0,n})_t(u)\|^2 + \Big[\sum_{i=1}^{|X|} \Lambda_{\Phi(x_i)}(t) \sum_{j=1}^{n} (C^{x_i}u)_j^2\Big] + \sum_{i,j} Z_{i,j}.
\end{aligned}$$

Note that $(D\Psi_{i,0})_t(u) := (D\Psi_{i-1,n})_t(u)$. Now recalling that $\|(D\Psi_{0,n})_t(u)\|^2 = \|u\|^2$ (the base case of the recursion), all we need to show is that $|\sum_{i,j} Z_{i,j}| \leq \epsilon/2$. This follows directly from the lemma below.

**Lemma 4.21** *Let $\epsilon_0 \leq O\big(\epsilon/d(n|X|)^2\big)$, and for any $i$, $j$, let $\omega_{i,j} \geq \Omega\big((K_{i,j} + (\alpha 9^n/\rho))(nd|X|)^2/\epsilon\big)$ (as per the statement of Lemma 4.13). Then, for any $i$, $j$, $|Z_{i,j}| \leq \epsilon/2n|X|$.*

*Proof.* Recall that (cf. Eq. (4.3))

$$Z_{i,j} = \underbrace{\frac{1}{\omega_{i,j}^2} \sum_k \big(\zeta_{i,j}^{k,4}\big)^2}_{(a)} + \underbrace{2 \sum_k \frac{\zeta_{i,j}^{k,4}}{\omega_{i,j}} \big(\zeta_{i,j}^{k,1} + \zeta_{i,j}^{k,2} + \zeta_{i,j}^{k,3}\big)}_{(b)} + \underbrace{2 \sum_k \zeta_{i,j}^{k,1} \zeta_{i,j}^{k,2}}_{(c)} + \underbrace{2 \sum_k \zeta_{i,j}^{k,1} \zeta_{i,j}^{k,3}}_{(d)}.$$

*Term (a):* Note that $|\sum_k (\zeta_{i,j}^{k,4})^2| \leq O\big(d^3(K_{i,j} + (\alpha 9^n/\rho))^2\big)$ (cf. Lemma 4.22 (iv)). By our choice of $\omega_{i,j}$, we have term (a) at most $O(\epsilon/n|X|)$.

*Term (b):* Note that $\big|\zeta_{i,j}^{k,1} + \zeta_{i,j}^{k,2} + \zeta_{i,j}^{k,3}\big| \leq O(n|X| + (\epsilon/dn|X|))$ (by noting Lemma 4.22 (i)-(iii), recalling the choice of $\omega_{i,j}$, and summing over all $i'$, $j'$). Thus,

$\big|\sum_k \zeta_{i,j}^{k,4}(\zeta_{i,j}^{k,1}+\zeta_{i,j}^{k,2}+\zeta_{i,j}^{k,3})\big| \le O\big(\big(d^2(K_{i,j}+(\alpha 9^n/\rho))\big)\big(n|X|+(\epsilon/dn|X|)\big)\big)$. Again, by our choice of $\omega_{i,j}$, term (b) is at most $O(\epsilon/n|X|)$.

*Terms (c) and (d):* We focus on bounding term (c) (the steps for bounding term (d) are same). Note that $|\sum_k \zeta_{i,j}^{k,1}\zeta_{i,j}^{k,2}| \le 4|\sum_k \zeta_{i,j}^{k,1}(\eta_{i,j}(t))_k|$. Now, observe that $\big(\zeta_{i,j}^{k,1}\big)_{k=1,\ldots,2d+3}$ is a tangent vector with length at most $O(dn|X|+(\epsilon/dn|X|))$ (cf. Lemma 4.22 (i)). Thus, by noting that $\eta_{i,j}$ is almost normal (with quality of approximation $\epsilon_0$), we have term (c) at most $O(\epsilon/n|X|)$.

By choosing the constants in the order terms appropriately, we can get the lemma. ∎

**Lemma 4.22** *Let $\zeta_{i,j}^{k,1}$, $\zeta_{i,j}^{k,2}$, $\zeta_{i,j}^{k,3}$, and $\zeta_{i,j}^{k,4}$ be as defined in Eq.* (4.3)*. Then for all $1 \le i \le |X|$ and $1 \le j \le n$, we have*

*(i)* $|\zeta_{i,j}^{k,1}| \le 1 + 8n|X| + \sum_{i'=1}^{i}\sum_{j'=1}^{j-1} O(d(K_{i',j'}+(\alpha 9^n/\rho))/\omega_{i',j'})$,

*(ii)* $|\zeta_{i,j}^{k,2}| \le 4$,

*(iii)* $|\zeta_{i,j}^{k,3}| \le 4$,

*(iv)* $|\zeta_{i,j}^{k,4}| \le O(d(K_{i,j}+(\alpha 9^n/\rho)))$.

*Proof.* First note for any $\|u\| \le 1$ and for any $x_i \in X$, $1 \le j \le n$ and $1 \le l \le d$, we have $|\sum_l C_{j,l}^{x_i} u_l| = |(C^{x_i}u)_j| \le 4$ (cf. Lemma 4.24 (b) and Corollary 4.11).

Noting that for all $i$ and $j$, $\|\eta_{i,j}\| = \|\nu_{i,j}\| = 1$, we have $|\zeta_{i,j}^{2,k}| \le 4$ and $|\zeta_{i,j}^{3,k}| \le 4$.

Observe that $\zeta_{i,j}^{k,4} = \sum_l u_l R_{i,j}^{k,l}$. For all $i$, $j$, $k$ and $l$, note that i) $\|d\eta_{i,j}(t)/dt^l\| \le K_{i,j}$ and $\|d\nu_{i,j}(t)/dt^l\| \le K_{i,j}$ and ii) $|d\lambda_{\Phi(x_i)}^{1/2}(t)/dt^l| \le O(\alpha 9^n/\rho)$ (cf. Lemma 4.20). Thus we have $|\zeta_{i,j}^{k,4}| \le O(d(K_{i,j}+(\alpha 9^n/\rho)))$.

Now for any $i$, $j$, note that $\zeta_{i,j}^{k,1} = \sum_l u_l d\Psi_{i,j-1}(t)/dt^l$. Thus by recursively expanding, $|\zeta_{i,j}^{k,1}| \le 1 + 8n|X| + \sum_{i'=1}^{i}\sum_{j'=1}^{j-1} O(d(K_{i',j'}+(\alpha 9^n/\rho))/\omega_{i',j'})$. ∎

### 4.9.5 Proof of Lemma 4.14

We start by stating the following useful observations:

**Lemma 4.23** *Let $A$ be a linear operator such that $\max_{\|x\|=1} \|Ax\| \leq \delta_{\max}$. Let $u$ be a unit-length vector. If $\|Au\| \geq \delta_{\min} > 0$, then for any unit-length vector $v$ such that $|u \cdot v| \geq 1 - \epsilon$, we have*

$$1 - \frac{\delta_{\max}\sqrt{2\epsilon}}{\delta_{\min}} \leq \frac{\|Av\|}{\|Au\|} \leq 1 + \frac{\delta_{\max}\sqrt{2\epsilon}}{\delta_{\min}}.$$

*Proof.* Let $v' = v$ if $u \cdot v > 0$, otherwise let $v' = -v$. Quickly note that $\|u - v'\|^2 = \|u\|^2 + \|v'\|^2 - 2u \cdot v' = 2(1 - u \cdot v') \leq 2\epsilon$. Thus, we have,

i. $\|Av\| = \|Av'\| \leq \|Au\| + \|A(u - v')\| \leq \|Au\| + \delta_{\max}\sqrt{2\epsilon}$,

ii. $\|Av\| = \|Av'\| \geq \|Au\| - \|A(u - v')\| \geq \|Au\| - \delta_{\max}\sqrt{2\epsilon}$.

Noting that $\|Au\| \geq \delta_{\min}$ yields the result. ∎

**Lemma 4.24** *Let $x_1, \ldots, x_n \in \mathbb{R}^D$ be $n$ mutually orthonormal vectors, $F := [x_1, \ldots, x_n]$ be a $D \times n$ matrix and let $\Phi$ be a linear map from $\mathbb{R}^D$ to $\mathbb{R}^d$ ($n \leq d \leq D$) such that for all non-zero $a \in span(F)$ we have $0 < \|\Phi a\| \leq \|a\|$. Let $U\Sigma V^\mathsf{T}$ be the thin SVD of $\Phi F$. Define $C = (\Sigma^{-2} - I)^{1/2} U^\mathsf{T}$. Then,*

*(a) $\|C(\Phi a)\|^2 = \|a\|^2 - \|\Phi a\|^2$, for any $a \in span(F)$,*

*(b) $\|C\|^2 \leq (1/\sigma^n)^2$, where $\|\cdot\|$ denotes the spectral norm of a matrix and $\sigma^n$ is the $n^{th}$ largest singular value of $\Phi F$.*

*Proof.* Note that $FV$ forms an orthonormal basis for the subspace spanned by columns of $F$ that maps to $U\Sigma$ via the mapping $\Phi$. Thus, since $a \in \text{span}(F)$, let $y$ be such that $a = FVy$. Note that i) $\|a\|^2 = \|y\|^2$, ii) $\|\Phi a\|^2 = \|U\Sigma y\|^2 = y^\mathsf{T}\Sigma^2 y$. Now,

$$\begin{aligned}
\|C\Phi a\|^2 &= \|((\Sigma^{-2} - I)^{1/2} U^\mathsf{T})\Phi F V y\|^2 \\
&= \|(\Sigma^{-2} - I)^{1/2} U^\mathsf{T} U \Sigma V^\mathsf{T} V y\|^2 \\
&= \|(\Sigma^{-2} - I)^{1/2} \Sigma y\|^2 \\
&= y^\mathsf{T} y - y^\mathsf{T}\Sigma^2 y \\
&= \|a\|^2 - \|\Phi a\|^2.
\end{aligned}$$

Now, consider $\|C\|^2$.

$$\begin{aligned}
\|C\|^2 &\leq \|(\Sigma^{-2} - I)^{1/2}\|^2 \|U^\mathsf{T}\|^2 \\
&\leq \max_{\|x\|=1} \|(\Sigma^{-2} - I)^{1/2} x\|^2 \\
&\leq \max_{\|x\|=1} x^\mathsf{T} \Sigma^{-2} x \\
&= \max_{\|x\|=1} \sum_i x_i^2/(\sigma^i)^2 \\
&\leq (1/\sigma^n)^2,
\end{aligned}$$

where $\sigma^i$ are the (top $n$) singular values forming the diagonal matrix $\Sigma$. ∎

**Lemma 4.25** *Let $M \subset \mathbb{R}^D$ be a compact Riemannian $n$-manifold with condition number $1/\tau$. Pick any $x \in M$ and let $F_x$ be any $n$-dimensional affine space with the property: for any unit vector $v_x$ tangent to $M$ at $x$, and its projection $v_{xF}$ onto $F_x$, $\left|v_x \cdot \frac{v_{xF}}{\|v_{xF}\|}\right| \geq 1 - \delta$. Then for any $p \in M$ such that $\|x - p\| \leq \rho \leq \tau/2$, and any unit vector $v$ tangent to $M$ at $p$, ($\xi := (2\rho/\tau) + \delta + 2\sqrt{2\rho\delta/\tau}$)*

*i.* $\left|v \cdot \frac{v_F}{\|v_F\|}\right| \geq 1 - \xi$,

*ii.* $\|v_F\|^2 \geq 1 - 2\xi$,

*iii.* $\|v_r\|^2 \leq 2\xi$,

*where $v_F$ is the projection of $v$ onto $F_x$ and $v_r$ is the residual (i.e. $v = v_F + v_r$ and $v_F \perp v_r$).*

*Proof.* Let $\gamma$ be the angle between $v_F$ and $v$. We will bound this angle.

Let $v_x$ (at $x$) be the parallel transport of $v$ (at $p$) via the (shortest) geodesic path via the manifold connection. Let the angle between vectors $v$ and $v_x$ be $\alpha$. Let $v_{xF}$ be the projection of $v_x$ onto the subspace $F_x$, and let the angle between $v_x$ and $v_{xF}$ be $\beta$. WLOG, we can assume that the angles $\alpha$ and $\beta$ are acute. Then, since $\gamma \leq \alpha + \beta \leq \pi$, we have that $\left|v \cdot \frac{v_F}{\|v_F\|}\right| = \cos\gamma \geq \cos(\alpha + \beta)$. We bound the individual terms $\cos\alpha$ and $\cos\beta$ as follows.

Now, since $\|p - x\| \leq \rho$, using Lemmas A.1 and A.2, we have $\cos(\alpha) = |v \cdot v_x| \geq 1 - 2\rho/\tau$. We also have $\cos(\beta) = \left|v_x \cdot \frac{v_{xF}}{\|v_{xF}\|}\right| \geq 1 - \delta$. Then, using Lemma 4.16, we finally get $\left|v \cdot \frac{v_F}{\|v_F\|}\right| = |\cos(\gamma)| \geq 1 - 2\rho/\tau - \delta - 2\sqrt{2\rho\delta/\tau} = 1 - \xi$.

Also note since $1 = \|v\|^2 = (v \cdot \frac{v_F}{\|v_F\|})^2 \left\| \frac{v_F}{\|v_F\|} \right\|^2 + \|v_r\|^2$, we have $\|v_r\|^2 = 1 - \left( v \cdot \frac{v_F}{\|v_F\|} \right)^2 \leq 2\xi$, and $\|v_F\|^2 = 1 - \|v_r\|^2 \geq 1 - 2\xi$. ∎

Now we are in a position to prove Lemma 4.14. Let $v_F$ be the projection of the unit vector $v$ (at $p$) onto the subspace spanned by (the columns of) $F_x$ and $v_r$ be the residual (i.e. $v = v_F + v_r$ and $v_F \perp v_r$). Then, noting that $p$, $x$, $v$ and $F_x$ satisfy the conditions of Lemma 4.25 (with $\rho$ in the Lemma 4.25 replaced with $2\rho$ from the statement of Lemma 4.14), we have ($\xi := (4\rho/\tau) + \delta + 4\sqrt{\rho\delta/\tau}$)

a) $\left| v \cdot \frac{v_F}{\|v_F\|} \right| \geq 1 - \xi$,

b) $\|v_F\|^2 \geq 1 - 2\xi$,

c) $\|v_r\|^2 \leq 2\xi$.

We can now bound the required quantity $\|C^x u\|^2$. Note that

$$\begin{aligned} \|C^x u\|^2 &= \|C^x \Phi v\|^2 = \|C^x \Phi (v_F + v_r)\|^2 \\ &= \|C^x \Phi v_F\|^2 + \|C^x \Phi v_r\|^2 + 2C^x \Phi v_F \cdot C^x \Phi v_r \\ &= \underbrace{\|v_F\|^2 - \|\Phi v_F\|^2}_{(a)} + \underbrace{\|C^x \Phi v_r\|^2}_{(b)} + \underbrace{2C^x \Phi v_F \cdot C^x \Phi v_r}_{(c)} \end{aligned}$$

where the last equality is by observing $v_F$ is in the span of $F_x$ and applying Lemma 4.24 (a). We now bound the terms $(a)$,$(b)$, and $(c)$ individually.

**Term** $(a)$**:** Note that $1 - 2\xi \leq \|v_F\|^2 \leq 1$ and observing that $\Phi$ satisfies the conditions of Lemma 4.23 with $\delta_{\max} = (2/3)\sqrt{D/d}$, $\delta_{\min} = (1/2) \leq \|\Phi v\|$ (cf. Lemma 4.10) and $\left| v \cdot \frac{v_F}{\|v_F\|} \right| \geq 1 - \xi$, we have (recall $\|\Phi v\| = \|u\| \leq 1$)

$$\begin{aligned} \|v_F\|^2 - \|\Phi v_F\|^2 &\leq 1 - \|v_F\|^2 \left\| \Phi \frac{v_F}{\|v_F\|} \right\|^2 \\ &\leq 1 - (1 - 2\xi) \left\| \Phi \frac{v_F}{\|v_F\|} \right\|^2 \\ &\leq 1 + 2\xi - \left\| \Phi \frac{v_F}{\|v_F\|} \right\|^2 \\ &\leq 1 + 2\xi - \left(1 - (4/3)\sqrt{2\xi D/d}\right)^2 \|\Phi v\|^2 \\ &\leq 1 - \|u\|^2 + \left(2\xi + (8/3)\sqrt{2\xi D/d}\right), \end{aligned} \tag{4.7}$$

where the fourth inequality is by using Lemma 4.23. Similarly, in the other direction

$$\begin{aligned}
\|v_F\|^2 - \|\Phi v_F\|^2 &\geq 1 - 2\xi - \|v_F\|^2 \left\| \Phi \frac{v_F}{\|v_F\|} \right\|^2 \\
&\geq 1 - 2\xi - \left\| \Phi \frac{v_F}{\|v_F\|} \right\|^2 \\
&\geq 1 - 2\xi - \left(1 + (4/3)\sqrt{2\xi D/d}\right)^2 \|\Phi v\|^2 \\
&\geq 1 - \|u\|^2 - \left(2\xi + (32/9)\xi(D/d) + (8/3)\sqrt{2\xi D/d}\right). \quad (4.8)
\end{aligned}$$

**Term** $(b)$**:** Note that for any $x$, $\|\Phi x\| \leq (2/3)(\sqrt{D/d})\|x\|$. We can apply Lemma 4.24 (b) with $\sigma_x^n \geq 1/4$ (cf. Corollary 4.11) and noting that $\|v_r\|^2 \leq 2\xi$, we immediately get

$$0 \leq \|C^x \Phi v_r\|^2 \leq 4^2 \cdot (4/9)(D/d)\|v_r\|^2 \leq (128/9)(D/d)\xi. \quad (4.9)$$

**Term** $(c)$**:** Recall that for any $x$, $\|\Phi x\| \leq (2/3)(\sqrt{D/d})\|x\|$, and using Lemma 4.24 (b) we have that $\|C^x\|^2 \leq 16$ (since $\sigma_x^n \geq 1/4$ – cf. Corollary 4.11).

Now let $a := C^x \Phi v_F$ and $b := C^x \Phi v_r$. Then $\|a\| = \|C^x \Phi v_F\| \leq \|C^x\| \|\Phi v_F\| \leq 4$, and $\|b\| = \|C^x \Phi v_r\| \leq (8/3)\sqrt{2\xi D/d}$ (see Eq. (4.9)).

Thus, $|2a \cdot b| \leq 2\|a\|\|b\| \leq 2 \cdot 4 \cdot (8/3)\sqrt{2\xi D/d} = (64/3)\sqrt{2\xi D/d}$. Equivalently,

$$-(64/3)\sqrt{2\xi D/d} \leq 2C^x \Phi v_F \cdot C^x \Phi v_r \leq (64/3)\sqrt{2\xi D/d}. \quad (4.10)$$

Combining (4.7)-(4.10), and noting $d \leq D$, yields the lemma.

## 4.10 Computing the normal vectors

The success of the second embedding technique crucially depends upon finding (at each iteration step) a pair of mutually orthogonal unit vectors that are normal to the embedding of manifold $M$ (from the previous iteration step) at a given point $p$. At a first glance finding such normal vectors seems infeasible since we only have access to a finite size sample $X$ from $M$. The saving grace comes from noting that the corrections are applied to the $n$-dimensional manifold $\Phi(M)$

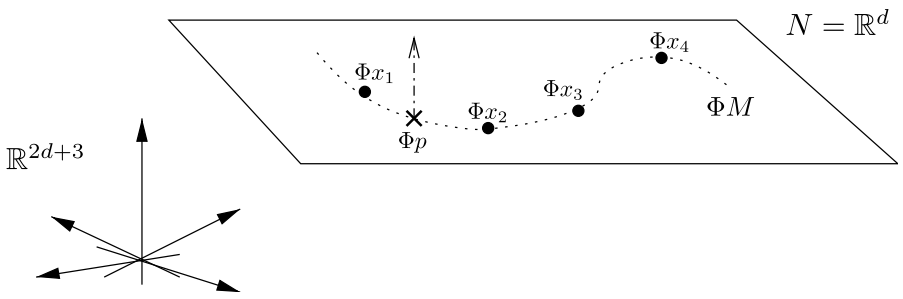



**Figure 4.5**: Basic setup for computing the normals to the underlying $n$-manifold $\Phi M$ at the point of interest $\Phi p$. Observe that even though it is difficult to find vectors normal to $\Phi M$ at $\Phi p$ within the containing space $\mathbb{R}^d$ (because we only have a finite-size sample from $\Phi M$, viz. $\Phi x_1$, $\Phi x_2$, etc.), we can treat the point $\Phi p$ as part of the bigger ambient manifold $N$ ($= \mathbb{R}^d$, that contains $\Phi M$) and compute the desired normals in a space that contains $N$ itself. Now, for each $i$, $j$ iteration of Algorithm II, $\Psi_{i,j}$ acts on the entire $N$, and since we have complete knowledge about $N$, we can compute the desired normals.

that is actually a *submanifold* of $d$-dimensional space $\mathbb{R}^d$. Let's denote this space $\mathbb{R}^d$ as a flat $d$-manifold $N$ (containing our manifold of interest $\Phi(M)$). Note that even though we only have partial information about $\Phi(M)$ (since we only have samples from it), we have full information about $N$ (since it is the entire space $\mathbb{R}^d$). What it means is that given some point of interest $\Phi p \in \Phi(M) \subset N$, finding a vector normal to $N$ (at $\Phi p$) automatically is a vector normal to $\Phi(M)$ (at $\Phi p$). Of course, to find two mutually orthogonal normals to a $d$-manifold $N$, $N$ itself needs to be embedded in a larger dimensional Euclidean space (although embedding into $d + 2$ should suffice, for computational reasons we will embed $N$ into Euclidean space of dimension $2d + 3$). This is precisely the first thing we do before applying any corrections (cf. Step 2 of Embedding II in Section 4.3). See Figure 4.5 for an illustration of the setup before finding any normals.

Now for every iteration of the algorithm, note that we have complete knowledge of $N$ and exactly what function (namely $\Psi_{i,j}$ for iteration $i$, $j$) is being applied to $N$. Thus with additional computation effort, one can compute the necessary normal vectors.

More specifically, we can estimate a pair of mutually orthogonal unit vectors that are normal to $\Psi_{i,j}(N)$ at $\Phi p$ (for any step $i$, $j$) as follows.

**Algorithm 4.4** Compute Normal Vectors

**Preprocessing Stage:**

1: Let $\eta_{i,j}^{\text{rand}}$ and $\nu_{i,j}^{\text{rand}}$ be vectors in $\mathbb{R}^{2d+3}$ drawn independently at random from the surface of the unit-sphere (for $1 \leq i \leq |X|$, $1 \leq j \leq n$).

**Compute Normals:** For any point of interest $p \in M$, let $t := \Phi p$ denote its projection into $\mathbb{R}^d$. Now, for any iteration $i, j$ (where $1 \leq i \leq |X|$, and $1 \leq j \leq n$), we shall assume that vectors $\eta$ and $\nu$ upto iterations $i, j-1$ are already given. Then we can compute the (approximated) normals $\eta_{i,j}(t)$ and $\nu_{i,j}(t)$ for the iteration $i$, $j$ as follows.

1: Let $\Delta > 0$ be the quality of approximation.

2: **for** $k = 1, \ldots, d$ **do**

3: Approximate the $k^{\text{th}}$ tangent vector as

$$T^k := \frac{\Psi_{i,j-1}(t + \Delta e^k) - \Psi_{i,j-1}(t)}{\Delta},$$

where $\Psi_{i,j-1}$ is as defined in Section 4.5.3, and $e^k$ is the $k^{\text{th}}$ standard vector.

4: **end for**

5: Let $\eta = \eta_{i,j}^{\text{rand}}$, and $\nu = \nu_{i,j}^{\text{rand}}$.

6: Use Gram-Schmidt orthogonalization process to extract $\hat{\eta}$ (from $\eta$) that is orthogonal to vectors $\{T^1, \ldots, T^d\}$.

7: Use Gram-Schmidt orthogonalization process to extract $\hat{\nu}$ (from $\nu$) that is orthogonal to vectors $\{T^1, \ldots, T^d, \hat{\eta}\}$.

8: **return** $\hat{\eta}/\|\hat{\eta}\|$ and $\hat{\nu}/\|\hat{\nu}\|$ as mutually orthogonal unit vectors that are approximately normal to $\Psi_{i,j-1}(\Phi M)$ at $\Psi_{i,j-1}(t)$.

A few remarks are in order.

**Remark 4.26** *The choice of target dimension of size* $2d + 3$ *(instead of* $d + 2$*) ensures that a pair of random unit-vectors* $\eta$ *and* $\nu$ *are not parallel to any vector in the tangent bundle of* $\Psi_{i,j-1}(N)$ *with probability* 1*. This follows from Sard's theorem (see e.g. Milnor [1972]), and is the key observation in reducing the embedding size in Whitney's embedding [Whitney, 1936]. This also ensures that our orthogonalization process (Steps 6 and 7) will not result in a null vector.*

**Remark 4.27** *By picking $\Delta$ sufficiently small, we can approximate the normals $\eta$ and $\nu$ arbitrarily well by approximating the tangents $T^1, \ldots, T^d$ well.*

**Remark 4.28** *For each iteration $i, j$, the vectors $\hat{\eta}/\|\hat{\eta}\|$ and $\hat{\nu}/\|\hat{\nu}\|$ that are returned (in Step 8) are a smooth modification to the starting vectors $\eta_{i,j}^{\text{rand}}$ and $\nu_{i,j}^{\text{rand}}$ respectively. Now, since we use the same starting vectors $\eta_{i,j}^{\text{rand}}$ and $\nu_{i,j}^{\text{rand}}$ regardless of the point of application ($t = \Phi p$), it follows that the respective directional derivates of the returned vectors are bounded as well.*

By noting Remarks 4.27 and 4.28, the approximate normals we return satisfy the conditions needed for Embedding II (see our discussion in Section 4.5.3).

# Chapter 5

# Multiple Instance Learning for Manifold Bags

Traditional supervised learning requires example/label pairs during training. However, in many domains labeling every single instance of data is either tedious or impossible. The Multiple Instance Learning (MIL) framework, introduced by Dietterich et al. [1997], provides a general paradigm for a more relaxed form of supervised learning: instead of receiving example/label pairs, the learner gets unordered sets of instances, or *bags*, and labels are provided for each bag, rather than for each instance. A bag is labeled positive if it contains at least one positive instance. In recent years MIL has received significant attention in terms of both algorithm design and applications [Maron and Lozano-Perez, 1998, Andrews et al., 2002, Zhang and Goldman, 2002, Viola et al., 2005].

Theoretical PAC-style analysis of MIL problems has also seen progress in the last decade [Auer et al., 1997, Blum and Kalai, 1998, Long and Tan, 1998, Sabato and Tishby, 2009, Sabato et al., 2010]. Typical analysis formulates the MIL problem as follows: a fixed number of instances, $r$, is drawn from an instance space $\mathcal{I}$ to form a bag. The sample complexity for bag classification is then analyzed in terms of the bag size ($r$). Most of the theory work has focused on reducing the dependence on $r$ under various settings. For example, Blum and Kalai [1998] showed that if one has access to a noise tolerant learner and the bags are formed by drawing $r$ independent samples from a fixed distribution over $\mathcal{I}$,

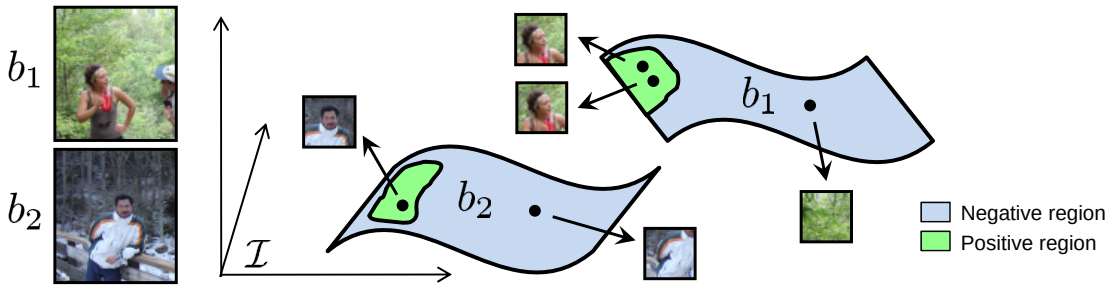



**Figure 5.1**: Better data modeling with manifold bags. In this example the task is to predict whether an image contains a face. Each bag is an image, and individual instances are image patches of a fixed size. Examples of two positive bags $b_1$ and $b_2$ (left), and a visualization of the instance space $\mathcal{I}$ (right) are shown. The two bags trace out low-dimensional manifolds in $\mathcal{I}$; in this case the manifold dimension is two since there are two degrees of freedom (the $x$ and $y$ location of the image patch). The green regions on the manifolds indicate the portion of the bags that is positive.

then the sample complexity grows linearly with $r$. Recently, Sabato and Tishby [2009] showed that if one can minimize the empirical error on bags, then even if the instances in a bag have arbitrary statistical dependence, sample complexity grows only logarithmically with $r$.

The above line of work is rather restrictive. Any dependence on $r$ makes it impossible to apply these generalization bounds to problems where bags have infinitely many instances – a typical case in practice. Consider the following motivating example: we would like to predict whether an image contains a face (as in Viola et al. [2005]). Putting this in the MIL framework, a bag is an entire image, which is labeled positive if and only if there is a face somewhere in the image. The individual instances are image patches. Notice that in this scenario the instances collectively form (a discrete approximation to) a low-dimensional manifold; see Figure 5.1. Here we expect the sample complexity to scale with the geometric properties of the underlying manifold bag rather than the number of instances per bag.

This situation arises in many other MIL applications where some type of sliding window is used to break up an object into many overlapping pieces: images [Andrews et al., 2002, Viola et al., 2005], video [Ali and Shah, 2008, Buehler et al.,

2009], audio [Saul et al., 2001, Mandel and Ellis, 2008], and sensor data [Stikic and Schiele, 2009]. Consider also the original molecule classification task that motivated Dietterich et al. [1997] to develop MIL, where a bag corresponds to a molecule, and instances are different shapes that molecule can assume. Even in this application, "as the molecule changes its shape, it traces out a manifold through [feature] space" [Maron and Lozano-Perez, 1998]. Thus, manifold structure is an integral aspect of these problems that needs to be taken into account in MIL analysis and algorithm design.

Here we analyze the MIL framework for bags containing potentially infinite instances. In this setting a bag is drawn from a bag distribution, and is labeled positive if it contains at least one positive instance. In order to have a tractable analysis, we impose a structural constraint on the bags: we assume that bags are low dimensional manifolds in the instance space, as discussed above. We show that the geometric structure of such bags is intimately related to the PAC-learnability of MIL problems. We investigate how learning is affected if we have have access to only a limited number of instances per manifold bag. We then discuss how existing MIL algorithms, that are designed for finite sized bags, can be adapted to learn from manifold bags efficiently using an iterative querying heuristic. Our experiments on real-world data (image and audio) validate the intuition of our analysis and show that our querying heuristic works well in practice.

## 5.1 Problem formulation and analysis

Let $\mathcal{I}$ be the domain of instances (for the purposes of our discussion we assume it to be $\mathbb{R}^D$ for some large $D$), and let $\mathcal{B}$ be the domain of bags. Here we impose a structural constraint on $\mathcal{B}$: each bag from $\mathcal{B}$ is a low dimensional manifold over the instances of $\mathcal{I}$. More formally, each bag $b \in \mathcal{B}$ is a $n$-dimensional submanifold of ambient instance space $\mathcal{I}$ ($n \ll D$). The geometric properties of such bags are integral to our analysis. We shall use $\textsc{dim}(b)$, $\textsc{vol}(b)$, and $\textsc{cond}(b)$ to denote the intrinsic dimension, volume, and the condition number (cf. Definition 2.4) of the manifold bag $b$ respectively.

With this formalism, we can define a structured family of bag spaces.

**Definition 5.1** *We say that a bag space $\mathcal{B}$ belongs to class $(V, n, \tau)$, if for every $b \in \mathcal{B}$, we have that* $\textsc{dim}(b) = n$, $\textsc{vol}(b) \leq V$, *and* $\textsc{cond}(b) \leq 1/\tau$.

In what follows, we will assume that $\mathcal{B}$ belongs to class $(V, n, \tau)$. We now provide our main results, with all the supporting proofs in Section 5.3.

### 5.1.1 Learning with manifold bags

Since we are interested in PAC-style analysis, we will be working with a fixed hypothesis class $\mathcal{H}$ over the instance space $\mathcal{I}$ (that is, each $h \in \mathcal{H}$ is of the form $h : \mathcal{I} \to \{0, 1\}$). The corresponding *bag* hypothesis class $\overline{\mathcal{H}}$ over the bag space $\mathcal{B}$ (where each $\bar{h} \in \overline{\mathcal{H}}$ is of the form $\bar{h} : \mathcal{B} \to \{0, 1\}$) is defined as the set of classifiers $\{\bar{h} : h \in \mathcal{H}\}$ where, for any $b \in \mathcal{B}$, $\bar{h}(b) := \max_{\alpha \in b} h(\alpha)$. We assume that there is some unknown instance classification rule $h^* : \mathcal{I} \to \{0, 1\}$ that gives the true labels for all instances.

The learner gets access to $m$ bag/label pairs $(b_i, y_i)_{i=1}^m$, where each bag $b_i$ is drawn independently from an unknown but fixed distribution $\mathcal{D}_{\mathcal{B}}$ over $\mathcal{B}$, and is labeled according to the MIL rule $y_i := \max_{\alpha \in b_i} h^*(\alpha)$. We denote a sample of size $m$ as $S_m$.

Our learner should ideally return the hypothesis $\bar{h}$ that achieves the lowest bag generalization[1] error: $\text{err}(\bar{h}) := \mathbb{E}_{b \sim \mathcal{D}_{\mathcal{B}}}[\bar{h}(b) \neq y]$. This, of course, is not possible as the learner typically does not have access to the underlying data distribution $\mathcal{D}_{\mathcal{B}}$. Instead, the learner has access to the sample $S_m$, and can minimize the *empirical* error: $\widehat{\text{err}}(\bar{h}, S_m) := \frac{1}{m} \sum_{i=1}^m \mathbf{1}\{\bar{h}(b_i) \neq y_i\}$. Various PAC results relate these two quantities in terms of the properties of $\overline{\mathcal{H}}$.

Perhaps the most obvious way to bound $\text{err}(\bar{h})$ in terms of $\widehat{\text{err}}(\bar{h}, S_m)$ is by analyzing the VC-dimension of the bag hypotheses, $\text{VC}(\overline{\mathcal{H}})$, and applying the standard VC-bounds (see e.g. Vapnik and Chervonenkis [1971]). While finding the VC-dimension of the bag hypothesis class is non-trivial, the VC-dimension of

[1] One can also talk about the generalization error over instances. As noted in previous work (e.g. Sabato and Tishby [2009]), PAC analysis of the instance error typically requires stronger assumptions.

the corresponding *instance* hypotheses, $\mathrm{VC}(\mathcal{H})$, is well known for many popular choices of $\mathcal{H}$. Sabato and Tishby [2009] showed that for finite sized bags the VC-dimension of *bag* hypotheses (and thus the generalization error) can be bounded in terms of the VC-dimension of the underlying *instance* hypotheses. Although one might hope that this analysis could be carried over to bags of infinite size that are well structured, this turns out to not be the case.

### $\mathrm{VC}(\overline{\mathcal{H}})$ is unbounded for arbitrarily smooth manifold bags

We begin with a surprising result which goes against our intuition that requiring bag smoothness should suffice in bounding $\mathrm{VC}(\overline{\mathcal{H}})$. We demonstrate that requiring the bags to be low-dimensional, arbitrarily flat manifolds with fixed volume is not enough to get a handle on generalization error even for one of the simplest instance hypothesis classes (set of hyperplanes in $\mathbb{R}^D$). In particular,

**Theorem 5.2** *For any $V > 0$, $n \geq 1$, $\tau < \infty$, let $\mathcal{B}$ contain all manifolds $M$ such that $\dim(M) = n$, $\mathrm{VOL}(M) \leq V$, and $\mathrm{COND}(M) \leq 1/\tau$ (i.e. $\mathcal{B}$ is the largest member of class $(V, n, \tau)$). Let $\mathcal{H}$ be the set of hyperplanes in $\mathbb{R}^D$ ($D > n$). Then for any $m \geq 1$, there exists a set of $m$ bags $b_1, \ldots, b_m \in \mathcal{B}$, such that the corresponding bag hypothesis class $\overline{\mathcal{H}}$ (over the bag space $\mathcal{B}$) realizes all possible $2^m$ labelings.*

Thus, $\mathrm{VC}(\overline{\mathcal{H}})$ is unbounded making PAC-learnability seemingly impossible. To build intuition for this apparent richness of $\overline{\mathcal{H}}$, and possible alternatives to bound the generalization error, let us take a quick look at the case of one-dimensional manifolds in $\mathbb{R}^2$ with halfspaces as our $\mathcal{H}$. For any $m$, we can place a set of $m$ manifold bags in such a way that all labelings are realizable by $\overline{\mathcal{H}}$ (see Fig. 5.2 for an example where $m = 3$; see Section 5.3.1 for a detailed construction).

The key observation is that in order to label a bag positive, the instance hypothesis needs to label just a *single* instance in that bag positive. Considering that our bags have an infinite number of points, the positive region can occupy an arbitrarily small fraction of a positively labeled bag. This gives our bag hypotheses immense flexibility even when the underlying instance hypotheses are quite simple.

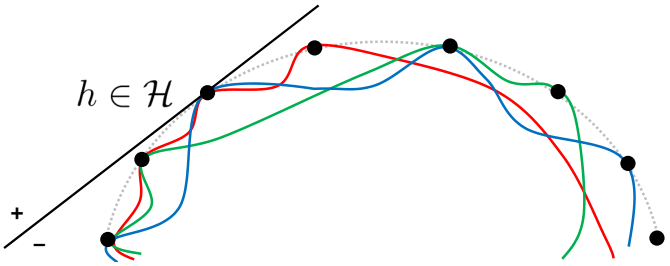



**Figure 5.2**: Bag hypotheses over manifold bags have unbounded VC-dimension. Three bags (colored blue, green and red) go around the eight anchor points (shown as black dots) that are arranged along a section of a circle. Notice that the hyperplanes tangent to the anchor points achieve all possible bag labelings. The hypothesis $h$ shown above, for example, labels the red and blue bags positive, and the green bag negative.

It seems that to bound $\text{err}(\bar{h})$ we must ensure that a non-negligible portion of a positive bag be labeled positive. A natural way of accomplishing this is to use a real-valued version of the instance hypothesis class (i.e., classifiers of the form $h_r : \mathcal{I} \to [0, 1]$, and labels determined by thresholding), and requiring that functions in this class (a) be *smooth*, and (b) label a positive bag with a certain *margin*. To understand why these properties are needed, consider three ways that $h_r$ can label the instances of a positive bag $b$ as one varies the latent parameter $\alpha$:

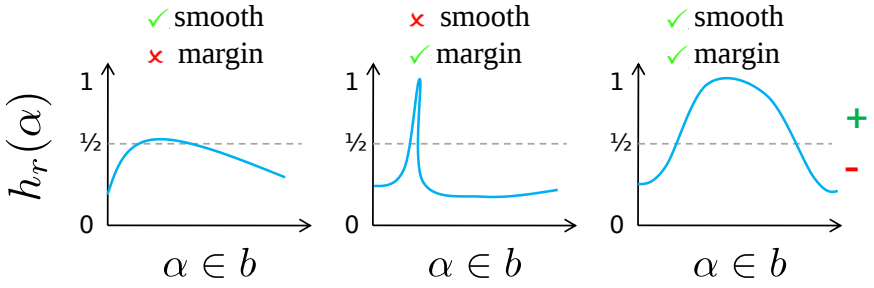



**Figure 5.3**: Potential ways classifier $h_r$ labels instances in a bag $b$. Ideally we want a non-negligible portion of the bag $b$ labeled positive (right figure).

In both the left and center panels, $h_r$ labels only a tiny portion of the bag positive: in the first case $h_r$ barely labels any instance above the threshold of $1/2$, resulting in a small margin; in the second case, although the margin is large, $h_r$

changes rapidly along the bag. Finally, in the right panel, when both the margin and smoothness conditions are met, a non-negligible portion of $b$ is labeled positive. We shall thus study how to bound the generalization error in this setting.

**Learning with a margin**

Let $\mathcal{H}_r$ be the real-valued relaxation of $\mathcal{H}$ (i.e. each $h_r \in \mathcal{H}_r$ is now of the form $h_r : \mathcal{I} \to [0,1]$). In order to ensure smoothness we impose a $\lambda$-Lipschitz constraint on the instance hypotheses: $\forall h_r \in \mathcal{H}_r,\ x, x' \in \mathcal{I},\ |h_r(x) - h_r(x')| \leq \lambda \|x - x'\|_2$. We denote the corresponding bag hypothesis class as $\overline{\mathcal{H}}_r$. Note that the true bag labels are still binary in this setting (i.e. determined by $h^*$).

Similar to the VC-dimension, the "fat-shattering dimension" of a real-valued bag hypothesis class, $\text{FAT}_\gamma(\overline{\mathcal{H}}_r)$, relates the generalization error to the empirical error at margin $\gamma$ (see for example Anthony and Bartlett [1999]):

$$\widehat{\text{err}}_\gamma(\bar{h}_r, S_m) := \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}\{\text{MARGIN}(\bar{h}_r(b_i), y_i) < \gamma\}, \tag{5.1}$$

where $\text{MARGIN}(x, y) := \begin{cases} x - 1/2 & y = 1 \\ 1/2 - x & \text{otherwise} \end{cases}$.

Recall that it was not possible to bound generalization error in terms of the instance hypotheses using VC dimension. However, analogous to Sabato & Tishby's analysis of finite size bags [2009], we *can* bound generalization error for manifold bags in terms of the fat-shattering dimension of instance hypotheses, $\text{FAT}_\gamma(\mathcal{H})$. In particular, we have the following:

**Theorem 5.3** *Let $\mathcal{B}$ belong to class $(V, n, \tau)$. Let $\mathcal{H}_r$ be $\lambda$-Lipschitz smooth (w.r.t. $\ell_2$-norm), and $\overline{\mathcal{H}}_r$ be the corresponding bag hypotheses over $\mathcal{B}$. Pick any $0 < \gamma < 1$ and $m \geq \text{FAT}_{\gamma/16}(\mathcal{H}_r) \geq 1$. For any $0 < \delta < 1$, we have with probability at least $1 - \delta$ over an i.i.d. sample $S_m$ (of size $m$), for every $\bar{h}_r \in \overline{\mathcal{H}}_r$:*

$$err(\bar{h}_r) \ \leq \ \widehat{err}_\gamma(\bar{h}_r, S_m) + O\left(\sqrt{\frac{n^2 \text{FAT}_{\frac{\gamma}{16}}(\mathcal{H}_r)}{m} \log^2\left(\frac{Vm}{\gamma^2 \tau_0^n}\right) + \frac{1}{m} \ln \frac{1}{\delta}}\right), \tag{5.2}$$

*where $\tau_0 = \min\{\frac{\tau}{2}, \frac{\gamma}{8}, \frac{\gamma}{8\lambda}\}$.*

Observe that the complexity term in Eq. (5.2) is independent of the "bag size"; it has instead been replaced by the volume and other geometric properties of the manifold bags. The other term captures the sample error for individual hypotheses at margin $\gamma$. Thus a natural strategy for a learner is to return a hypothesis that minimizes the empirical error while maximizing the margin.

### 5.1.2 Learning from queried instances

So far we have analyzed the MIL learner as a black box entity, which can minimize the empirical bag error by somehow accessing the bags. Since the individual bags in our case are low-dimensional manifolds (with an infinite number of instances), we must also consider *how* these bags are accessed by the learner. Perhaps the simplest approach is to query $\rho$ instances uniformly from each bag, thereby "reducing" the problem to standard MIL (with finite size bags) for which there are algorithms readily available (e.g. works by Maron and Lozano-Perez [1998], Andrews et al. [2002], Zhang and Goldman [2002], Viola et al. [2005]). More formally, for a bag sample $S_m$, let $p_1^i, \ldots, p_\rho^i$ be $\rho$ independent instance samples drawn uniformly from the (image of) bag $b_i \in S_m$, and let $S_{m,\rho} := \bigcup_{i,j} p_j^i$ be the set of all instances. Assuming that our manifold bags have well-conditioned boundaries, the following theorem relates the empirical error of sampled bags, $\widehat{\mathrm{err}}_\gamma(\bar{h}_r, S_{m,\rho}) := \frac{1}{m}\sum_{i=1}^m \mathbf{1}\{\textsc{margin}(\max_{j\in[\rho]} h(p_j^i), y_i) < \gamma\}$, to the generalization error.

**Theorem 5.4** *Let $\mathcal{B}$ belong to class $(V, n, \tau)$. Let $\mathcal{H}_r$ be $\lambda$-Lipschitz smooth (w.r.t. $\ell_2$-norm), and $\overline{\mathcal{H}}_r$ be the corresponding bag hypotheses over $\mathcal{B}$. Pick any $0 < \delta_1, \delta_2 < 1$, then with probability at least $1 - \delta_1 - \delta_2$, over the draw of $m$ bags ($S_m$) and $\rho$ instances per bag ($S_{m,\rho}$), for all $\bar{h}_r \in \overline{\mathcal{H}}_r$ we have the following:*

*Let $\frac{1}{\kappa} := \max_{b_i \in S_m}\{\textsc{cond}(\partial b_i)\}$ (where $\partial b_i$ is the boundary of the manifold bag $b_i$)[2] and set $\tau_1 = \min\{\frac{\tau}{32}, \frac{\kappa}{8}, \frac{\gamma}{9\lambda}, \frac{\gamma}{9}\}$. If*

$$\rho \geq \Omega\left(\left(V/\tau_1^{c_0 n}\right)\left(n + \ln\left(\frac{mV}{\tau_1^n \delta_2}\right)\right)\right),$$

[2]If $\partial b_i$ is empty, then define $\textsc{cond}(\partial b_i) = 0$.

*then*

$$err(\bar{h}_r) \quad \leq \quad \widehat{err}_{2\gamma}(\bar{h}_r, S_{m,\rho}) + O\left(\sqrt{\frac{n^2 \mathrm{FAT}_{\frac{\gamma}{16}}(\mathcal{H}_r)}{m} \log^2\left(\frac{Vm}{\gamma^2 \tau_0^n}\right) + \frac{1}{m} \ln \frac{1}{\delta_1}}\right),$$

*where* $\tau_0 = \min\{\frac{\tau}{2}, \frac{\gamma}{8}, \frac{\gamma}{8\lambda}\}$ *and* $\mathrm{c}_0$ *is an absolute constant.*

Notice the effect of the two key parameters in the above theorem: the number of training bags, $m$, and the number of queried instances per bag, $\rho$. Increasing either quantity improves generalization – increasing $m$ drives down the error (via the complexity term), while increasing $\rho$ helps improve the confidence (via $\delta_2$). While ideally we would like both quantities to be large, increasing these parameters is, of course, computationally burdensome for a standard MIL learner. Note, however, the difference between $m$ and $\rho$: increasing $m$ comes at an *additional cost* of obtaining extra labels, whereas increasing $\rho$ does not. We would therefore like an algorithm that can take advantage of using a large $\rho$ while avoiding computational costs.

**Iterative querying heuristic**

As we saw in the previous section, we would ideally like to train with a large number of queried instances, $\rho$, per training bag. However, this may be impractical in terms of both speed and memory constraints. Suppose we have access to a black box MIL algorithm $\mathcal{A}$ that can only train with $\hat{\rho} < \rho$ instances per bag at once. We propose a procedure called Iterative Querying Heuristic (IQH), described in detail in Algorithm 5.1 (the main steps are highlighted in blue).

Notice that IQH uses a total of $T\hat{\rho}$ instances per bag for training ($T$ iterations times $\hat{\rho}$ instances per iteration). Thus, setting $T \approx \rho/\hat{\rho}$ should achieve performance comparable to using $\rho$ instances at once. The free parameter $\omega$ controls how many new instances are considered in each iteration.

The intuition behind IQH is as follows. For positive bags, we want to ensure that at least one of the queried instances is positive; hence we use the current estimate of the classifier to select the most positive instances. For negative bags, we know all instances are negative. In this case we select the instances that are

**Algorithm 5.1** Iterative Querying Heuristic (IQH)

**Input:** Training bags $(b_1, \ldots, b_m)$, labels $(y_1, \ldots, y_m)$, parameters $T$, $\omega$ and $\hat{\rho}$

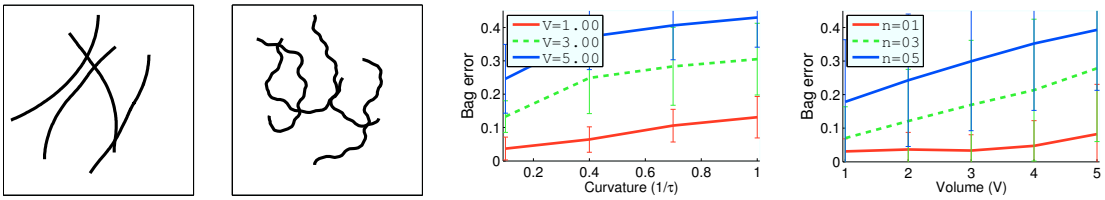
1: Initialize $I_i^0 = \emptyset$, $h_r^0$ as any classifier in $\mathcal{H}_r$.
2: **for** $t = 1, \ldots, T$ **do**
3: **Query $\omega$ new candidate instances per bag**:
$Z_i^t := I_i^{t-1} \cup \{p_1^i, \ldots, p_\omega^i\}$ where $p_j^i \sim b_i$, $\forall i$.
4: **Keep $\hat{\rho}$ highest scoring inst. using $h_r^{t-1}$**:
$I_i^t \subset Z_i^t$ s.t. $|I_i^t| = \hat{\rho}$ and $h_r^{t-1}(p) \geq h_r^{t-1}(p')$
for all $p \in I_i^t, p' \in Z_i^t \setminus I_i^t$.
5: **Train $\bar{h}_r^t$ with the selected instances**:
$\bar{h}_r^t \leftarrow \mathcal{A}(\{I_1^t \ldots I_m^t\}, \{y_1 \ldots y_m\})$.
6: **end for**
7: **Return** $h_r^T$ and the corresponding $\bar{h}_r^T$

closest to the decision boundary of our current classifier (corresponding to the most difficult negative instances); the motivation for this is similar to bootstrapping negative examples [Felzenszwalb et al., 2009] and some active learning techniques [Cohn et al., 1994]. We then use these selected instances to find a better classifier.

Thus one expects IQH to take advantage of a large number of instances per bag, without actually having to train with all of them at one time.

## 5.2 Experiments

Recall that we have shown that the generalization error is bounded in terms of key geometric properties of the manifold bags, such as curvature $(1/\tau)$ and volume $(V)$. Here we will experimentally validate that generalization error does indeed scale with these quantities, providing an empirical lower bound. Additionally, we study how the choice of $\rho$ affects the error, and show that our Iterative Heuristic (IQH) is effective in reducing the number of instances needed to train in each iteration. In all our experiments we use a boosting algorithm for MIL called MILBoost




(A) (B) (C) (D)

**Figure 5.4**: Results on synthetic data. Examples of four synthetically generated bags in $\mathbb{R}^2$ with (A) low curvature and (B) high curvature. (C) and (D): Test error scales with the manifold parameters: volume ($V$), curvature ($\frac{1}{\tau}$), and dimension ($n$).

[Viola et al., 2005] as the black box $\mathcal{A}$. We expect the same for any other choice of $\mathcal{A}$. Note that we use IQH only where specified.

### 5.2.1 Synthetic data

We begin with a carefully designed synthetic dataset, where we have complete control over the manifold curvature, volume and dimension, and study its effects on the generalization. The details on how we generate the dataset are provided in Section 5.4; see Figure 5.4 (A) and (B) for examples of the generated manifolds.

For the first set of experiments, we study the interplay between the volume and curvature while keeping the manifold dimension fixed. Here we generated one-dimensional curves of specified volume ($V$) and curvature ($1/\tau$) in $\mathbb{R}^2$. We set $h^*$ to be a vertical hyperplane and labeled the samples accordingly (see Section 5.4). For training, we used 10 positive and 10 negative bags with 500 queried instances per bag (forming a good cover); for testing we used 100 bags. Figure 5.4 (C) shows the test error, averaged over 50 trials, as we vary these parameters. Observe that for a fixed $V$, as we increase $1/\tau$ (making the manifolds more curvy) generalization error goes up.

For the next set of experiments, we want to understand how manifold dimensionality affects the error. Here we set the ambient dimension to 10 and varied the manifold dimension (with all other experiment settings as before). Figure 5.4

(D) shows how the test error scales for different dimensional bags as we vary the volume ($1/\tau$ set to 1).

These results corroborate the general intuition of our analysis, and give an empirical verification that the error indeed scales with the geometric properties of a manifold bag.

### 5.2.2 Real data

In this section we present results on image and audio datasets. We will see that the generalization behavior is consistent with our analysis across these different domains. We also study the effects of varying $\rho$ on generalization error, and see how using IQH helps achieve similar error rates with less instances per iteration.


**Figure 5.5**: INRIA Heads dataset. For our experiments we have labeled the heads in the INRIA Pedestrian Dataset [Dalal and Triggs, 2005]. We can construct bags of different volume by padding the head region. The above figure shows positive bags for two different amounts of padding.

**INRIA Heads.** For these experiments we chose the task of head detection (e.g. positive bags are images which contain at least one head). We used the INRIA Pedestrian Dataset Dalal and Triggs [2005], which contains both pedestrian and non-pedestrian images, to create an INRIA Heads dataset as follows. We manually labeled the location of the head in the pedestrian images. The images were resized such that the size of the head is roughly $24 \times 24$ pixels; therefore, instances in this

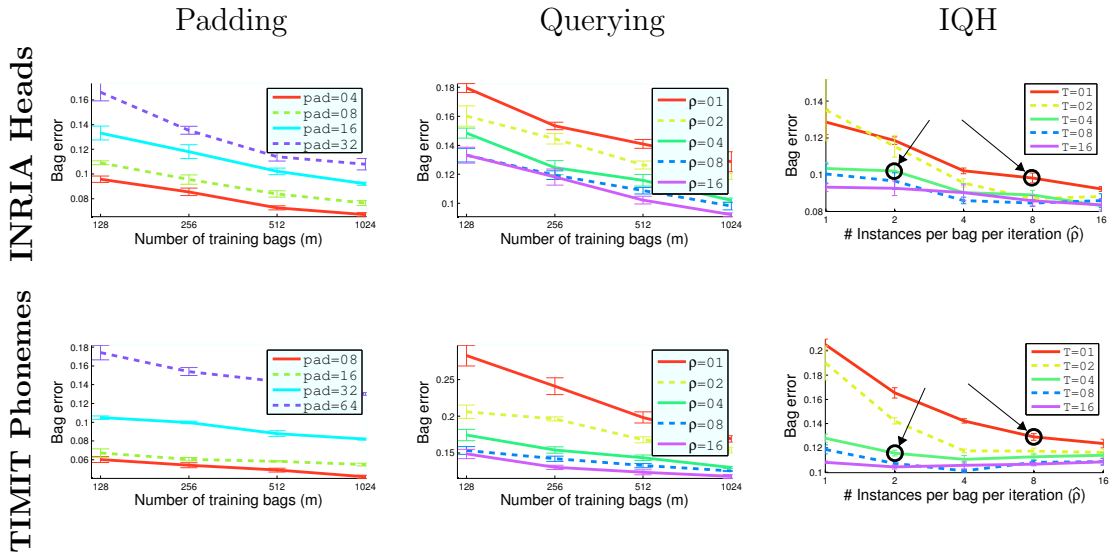



**Figure 5.6**: Results on image and audio datasets. Three different experiments (columns) – varying padding (volume), number of queried instances, and number of IQH iterations – on two different datasets (rows); see text for details. Note that x-axes are in logarithmic scale. All reported results are averages over 5 trials.

experiment are image patches of that size. For each image patch we computed Haar-like features on various channels as in Dollár et al. [2009], which corresponds to our instance space $\mathcal{I}$.

Using the ground truth labels, we generated 2472 positive bags by cropping out the head region with different amounts of padding (see Figure 5.5), which corresponds to changing the volume of the manifold bags. For example, padding by 6 pixels results in a bag that is a $30 \times 30$ pixel image. To generate negative bags we cropped 2000 random patches from the non-pedestrian images, as well as non-head regions from the pedestrian images. Unless otherwise specified, padding was set to 16.

**TIMIT Phonemes.** Our other application is in the audio domain, and is analogous to the image data described above. The task here was to detect whether a particular phoneme is spoken in an audio clip (we arbitrarily chose the phoneme "s" to be the positive class). We used the TIMIT dataset [Garofolo et al., 1993], which contains recordings of over 600 speakers reading text; the dataset also contains phoneme annotations. Bags in this experiment are audio clips, and instances

are audio pieces of length 0.2 seconds (i.e. this is the size of our sliding window). As in the image experiments, we had ground truth annotation for instances, and generated bags of various volumes/lengths by padding. We computed features as follows: we split each sliding window into 25 millisecond pieces, computed Mel-frequency cepstral coefficients (MFCC) [Davis and Mermelstein, 1980, Ellis, 2005] for each piece, and concatenated them to form a 104 dimensional feature vector for each instance. The reported padding amounts are in terms of a 5 millisecond step size (i.e., padding of 8 corresponds to 40 milliseconds of concatenation). Unless otherwise specified, padding was set to 64.

**Results.** Our first set of experiments involved sweeping over the amount of padding (corresponding to varying the volume of bags). We train with a fixed number of instances per bag, $\rho = 4$. Results for different training set sizes ($m$) are shown in the first column of Figure 5.6. As observed in the synthetic experiments, we see that increasing the padding (volume) leads to poorer generalization for both datasets. This corroborates our basic intuition that learning becomes more difficult with manifolds of larger volume.

In our second set of experiments, the goal was to see how generalization error is affected by varying the number of queried instances per bag, which compliments Theorem 5.4. Results are shown in the middle column of Figure 5.6. Observe the interplay between $m$ and $\rho$: increasing either, while keeping the other fixed, drives the error down. Recall, however, that increasing $m$ also requires additional labels while querying more instances per bag does not. The number of instances indeed has a significant impact on generalization – for example, in the audio domain, querying more instances per bag can improve the error by up to 15%. As per our analysis, these results suggest that to fully leverage the training data, we must query many instances per bag. Since training with a large number of instances can become computationally prohibitive, this further justifies the Iterative Querying Heuristic (IQH) described in Section 5.1.2.

Our final set of experiments evaluates the proposed IQH method (see Algorithm 5.1). The number of training bags, $m$, was fixed to 1024, and the number of candidate instances per iteration, $\omega$, was fixed to 32 for both datasets. Note

that $T = 1$ corresponds to querying instances and training MILBoost once (i.e. no iterative querying). Results are shown in the right column of Figure 5.6. These results show that our heuristic works quite well. Consider the highlighted points in both plots: using IQH with $T = 4$ and just 2 instances per bag during training we are able to achieve comparable test error to the naive method (i.e. $T = 1$) with 8 instances per bag. Thus, using IQH, we can obtain a good classifier while needing to use less memory and computational resources per iteration.

## Acknowledgments

The contents of this chapter originally appeared in the following publication: B. Babenko, N. Verma, P. Dollár and S. Belongie. Multiple instance learning with manifold bags. *International Conference on Machine Learning (ICML)*, 2011.

## 5.3 Supporting proofs

### 5.3.1 Proof of Theorem 5.2

We will show this for $n = 1$ and $D = 2$ (the generalization to $D > n \geq 1$ is immediate). We first construct $2^m$ anchor points $p_0, \ldots, p_{2^m-1}$ on a section of a circle in $\mathbb{R}^2$ that will serve as a guide on how to place $m$ bags $b_1, \ldots, b_m$ of dimension $n = 1$, volume[3] $\leq V$, and condition number $\leq 1/\tau$ in $\mathbb{R}^2$. We will then show that the class of hyperplanes in $\mathbb{R}^2$ can realize all possible $2^m$ labelings of these $m$ bags.

Let $V_0 := \min(V/2, \pi)$. Define anchor points $p_i := 2\tau\big(\cos(\frac{V_0 i}{2\tau 2^m}), \sin(\frac{V_0 i}{2\tau 2^m})\big)$ for $0 \leq i \leq 2^m - 1$. Observe that the points $p_i$ are on a circle centered at the origin of radius $2\tau$ in $\mathbb{R}^2$.

We use points $p_0, \ldots, p_{2^m-1}$ as guides to place $m$ bags $b_1, \ldots, b_m$ in $\mathbb{R}^2$ that are contained entirely in the disc of radius $2\tau$ centered at the origin and pass through the anchor points as follows. Let $k_m^i \ldots k_1^i$ represent the binary representation of the number $i$ ($0 \leq i \leq 2^m - 1$). Place bag $b_j$ such that $b_j$ passes through

[3]Volume of a 1-dimensional manifold is its length.

the anchor point $p_i$, if and only if $k_j^i = 1$. (see figure below for a visual example for 3 bags and 8 anchor points). Note that since, by construction, the arc (the dotted line in the figure) containing the anchor points has condition number at most $1/2\tau$ with volume strictly less than $V$, bags $b_j$ can be made to have condition number at most $1/\tau$ with volume at most $V$.

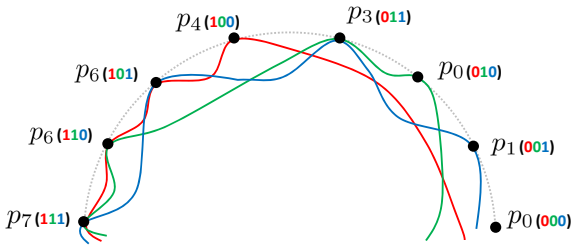



**Figure 5.7**: Placement of arbitrarily smooth bags along a section of a disk. Three bags (colored blue, green and red) go around the eight anchor points $p_0, \ldots, p_7$ in such a way that the hypothesis class of hyperplanes can realize all possible bag labelings.

It is clear that hyperplanes in $\mathbb{R}^2$ can realize any possible labeling of these $m$ bags. Say, we want some arbitrary labeling $(+1, +1, 0, \ldots, +1)$. We look at the number $i$ with the same bit representation. Then a hyperplane that is tangent to the circle (centered at the origin and radius $2\tau$) at the anchor point $p_i$, labels $p_i$ positive, and all other $p_k$'s negative. Note that this hypothesis will also label exactly those bags $b_j$ positive that are passing through the point $p_i$, and rest of the bags labeled negative. Thus, realizing the arbitrary labeling.

### 5.3.2 Proof of Theorem 5.3

For any domain $X$, real-valued hypothesis class $H \subset [0,1]^X$, margin $\gamma > 0$ and a sample $S \subset X$, define

$$\mathrm{cov}_\gamma(H, S) \;:=\; \{C \subset H : \forall h \in H, \exists h' \in C, \max_{s \in S} |h(s) - h'(s)| \leq \gamma\}$$

as a set of $\gamma$-covers of $S$ by $H$. Let $\gamma$-covering number of $H$ for any integer $m > 0$ be defined as

$$\mathcal{N}_\infty(\gamma, H, m) := \max_{S \subset X : |S| = m} \; \min_{C \in \mathrm{cov}_\gamma(H,S)} |C|.$$

We will first relate the covering numbers of $\mathcal{H}_r$ and $\overline{\mathcal{H}}_r$ with the fat-shattering dimension in the following two lemmas.

**Lemma 5.5 (relating hypothesis cover to the fat-shattering dimension – see Theorem 12.8 Anthony and Bartlett [1999])** *Let $H$ be a set of real functions from a domain $X$ to the interval $[0,1]$. Let $\gamma > 0$. Then for $m \geq \mathrm{FAT}_{\gamma/4}(H)$,*

$$\mathcal{N}_\infty(\gamma, H, m) < 2\big(4m/\gamma^2\big)^{\mathrm{FAT}_{\gamma/4}(H)\log\frac{4em}{\mathrm{FAT}_{\gamma/4}(H)\gamma}}.$$

**Lemma 5.6 (adapted from Lemma 17 of Sabato and Tishby [2009])** *Let $\mathcal{H}_r$ be an instance hypothesis class such that each $h_r \in \mathcal{H}_r$ is $\lambda$-lipschitz (w.r.t. $\ell_2$-norm), and let $\overline{\mathcal{H}}_r$ be the corresponding bag hypothesis class over $\mathcal{B}$ that belongs to the class $(V, n, \tau)$. For any $\gamma > 0$ and $m \geq 1$, we have*

$$\mathcal{N}_\infty(2\gamma, \overline{\mathcal{H}}_r, m) \leq \mathcal{N}_\infty(\gamma, \mathcal{H}_r, m2^{\mathrm{c}_0 n}(V/\epsilon^n)),$$

*where $\epsilon = \min\{\frac{\tau}{2}, \frac{\gamma}{2}, \frac{\gamma}{2\lambda}\}$, and $\mathrm{c}_0$ is an absolute constant.*

*Proof.* Let $S = \{b_1, \ldots, b_m\}$ be a set of $m$ manifold bags. Set $\epsilon = \min\{\frac{\tau}{2}, \frac{\gamma}{2}, \frac{\gamma}{2\lambda}\}$. For each bag $b_i \in S$, let $C_i$ be the smallest $\epsilon$-cover of (the image of) $b_i$ (by Lemma A.7, we know that $|C_i| \leq 2^{\mathrm{c}_0 n}(V/\epsilon^n)$ for some absolute constant $\mathrm{c}_0$).

Define $S^\cup := \cup_i C_i$ and let $R \in \mathrm{cov}_\gamma(\mathcal{H}_r, S^\cup)$ be some $\gamma$-cover of $S^\cup$. Now, for any $h_r \in \mathcal{H}_r$, let $\bar{h}_r \in \overline{\mathcal{H}}_r$ denote the corresponding bag classifier, and define $\widetilde{h}_r(C_i) := \max_{c \in C_i} h_r(c)$ as the maximum attained by $h_r$ on the sample $C_i$. Then, since $h_r$ is $\lambda$-lipschitz (w.r.t. $\ell_2$-norm), we have for any bag $b_i$ and its corresponding $\epsilon$-cover $C_i$,

$$|\bar{h}_r(b_i) - \widetilde{h}_r(C_i)| \leq \lambda\epsilon.$$

It follows that $\forall x \in S^\cup$: for any $h_r \in \mathcal{H}_r$ and $h'_r \in R$ such that $|h_r(x) - h'_r(x)| \leq \gamma$ (and the corresponding bag classifiers $\bar{h}_r$ and $\bar{h}'_r$ in $\overline{\mathcal{H}}_r$),

$$\begin{aligned}\max_{i\in[m]} |\bar{h}_r(b_i) - \bar{h}'_r(b_i)| &= \max_{i\in[m]} |\bar{h}_r(b_i) - \widetilde{h}_r(C_i) + \widetilde{h}_r(C_i) - \widetilde{h}'_r(C_i) + \widetilde{h}'_r(C_i) - \bar{h}'_r(b_i)| \\ &\leq 2\lambda\epsilon + \gamma \leq 2\gamma.\end{aligned}$$

Also, note that for any $h_r \in \mathcal{H}_r$ and $h'_r \in R$ such that $|h_r(x) - h'_r(x)| \leq \gamma$ $(x \in S^{\cup})$, we have $\bar{h}'_r \in \{\bar{h}_r | h_r \in R\} := \bar{R}$. It follows that for any $R \in \mathrm{cov}_\gamma(\mathcal{H}_r, S^{\cup})$, $\{\bar{h}_r | h_r \in R\} \in \mathrm{cov}_{2\gamma}(\overline{\mathcal{H}}_r, S)$. Thus,

$$\{\overline{H}_r | H_r \in \mathrm{cov}_\gamma(\mathcal{H}_r, S^{\cup})\} \subset \mathrm{cov}_{2\gamma}(\overline{\mathcal{H}}_r, S).$$

Hence, we have

$$\begin{aligned}
\mathcal{N}_\infty(2\gamma, \overline{\mathcal{H}}_r, m) &= \max_{S\subset\mathcal{B}:|S|=m} \min_{\bar{R}\in\mathrm{cov}_{2\gamma}(\overline{\mathcal{H}}_r,S)} |\bar{R}| \\
&\leq \max_{S\subset\mathcal{B}:|S|=m} \min_{\bar{R}\in\{\bar{H}_r | H_r\in\mathrm{cov}_\gamma(\mathcal{H}_r,S^{\cup})\}} |\bar{R}| \\
&= \max_{S\subset\mathcal{B}:|S|=m} \min_{R\in\mathrm{cov}_\gamma(\mathcal{H}_r,S^{\cup})} |R| \\
&= \max_{S\subset\mathcal{I}:|S|=|S^{\cup}|m} \min_{R\in\mathrm{cov}_\gamma(\mathcal{H}_r,S)} |R| \\
&\leq \max_{S\subset\mathcal{I}:|S|=m2^{\mathrm{c}_0 n}(V/\epsilon^n)} \min_{R\in\mathrm{cov}_\gamma(\mathcal{H}_r,S)} |R| \\
&= \mathcal{N}_\infty(\gamma, \mathcal{H}_r, m2^{\mathrm{c}_0 n}(V/\epsilon^n)),
\end{aligned}$$

where $\mathrm{c}_0$ is an absolute constant. ∎

Now we can relate the empirical error with generalization error by noting the following lemma.

**Lemma 5.7 (generalization error bound for real-valued functions – Theorem 10.1 of Anthony and Bartlett [1999])** *Suppose that $F$ is a set of real-valued functions defined on the domain $X$. Let $\mathcal{D}$ be any probability distribution on $Z = X \times \{0, 1\}$, $0 \leq \epsilon \leq 1$, real $\gamma > 0$ and integer $m \geq 1$. Then,*

$$\mathbf{Pr}_{S_m\sim\mathcal{D}}\big[\exists f \in F : err(f) \geq \widehat{err}_\gamma(f, S_m) + \epsilon\big] \;\leq\; 2\mathcal{N}_\infty\Big(\frac{\gamma}{2}, F, 2m\Big)e^{-\epsilon^2 m/8},$$

*where $S_m$ is an i.i.d. sample of size $m$ from $\mathcal{D}$, $err(f)$ is the error of $f$ with respect to $\mathcal{D}$, and $\widehat{err}_\gamma(f, S_m)$ is the empirical error of $f$ with respect to $S_m$ at margin $\gamma$.*

Combining Lemmas 5.7, 5.6 and 5.5, we have (for $m \geq \mathrm{FAT}_{\frac{\gamma}{16}}(\mathcal{H}_r)$):

$$\begin{aligned}
\mathbf{Pr}_{S_m\sim\mathcal{D}_\mathcal{B}}\big[\exists\bar{h}_r \in \overline{\mathcal{H}}_r : \mathrm{err}(\bar{h}_r) \geq \widehat{\mathrm{err}}_\gamma(\bar{h}_r, S_m) + \epsilon\big] &\leq 2\,\mathcal{N}_\infty\Big(\frac{\gamma}{2}, \overline{\mathcal{H}}_r, 2m\Big)\, e^{-\epsilon^2 m/8} \\
&\leq 2\,\mathcal{N}_\infty\Big(\frac{\gamma}{4}, \mathcal{H}_r, m2^{\mathrm{c}_0 n}(V/\tau_0^n)\Big)\, e^{-\epsilon^2 m/8} \\
&\leq 4\Big(\frac{64\cdot 2^{\mathrm{c}_0 n}Vm}{\gamma^2\tau_0^n}\Big)^{d\log\left(\frac{16e2^{\mathrm{c}_0 n}Vm}{\tau_0^n d\gamma}\right)}\, e^{-\epsilon^2 m/8},
\end{aligned}$$

where $c_0$ is an absolute constant, $d := \text{FAT}_{\frac{\gamma}{16}}(\mathcal{H}_r)$, and $\tau_0 = \min\{\frac{\tau}{2}, \frac{\gamma}{8}, \frac{\gamma}{8\lambda}\}$. For $d \geq 1$, the theorem follows.

### 5.3.3 Proof of Theorem 5.4

We start with the following useful observations that will help in our proof. Notation: for any two points $p$ and $q$ on a Riemannian manifold $M$,

- let $D_G(p, q)$ denote the geodesic distance between points $p$ and $q$.
- $B_G(p, \epsilon) := \{p' \in M | D_G(p', p) \leq \epsilon\}$ denote the geodesic ball centered at $p$ of radius $\epsilon$.

**Lemma 5.8** *Let $M \subset \mathbb{R}^D$ be a compact $n$-dimensional manifold with $\text{VOL}(M) \leq V$ and $\text{COND}(M) \leq 1/\tau$. Let $\mu(M)$ denote the uniform probability measure over $M$. Define $\mathcal{F}(M, \epsilon) := \{B_G(p, \epsilon) : p \in M$ and $B_G(p, \epsilon)$ contains no points from the boundary of $M\}$, that is, the set of all geodesic balls of radius $\epsilon$ that are contained entirely in the interior of $M$. Let $\tau_0 \leq \tau$ and $\rho \geq 1$. Let $p_1, \ldots, p_\rho$ be $\rho$ independent draws from $\mu(M)$. Then,*

$$\mathbf{Pr}_{p_1,\ldots,p_\rho \sim \mu(M)}\big[\exists F \in \mathcal{F}(M, \tau_0) : \forall i, p_i \notin F\big] \;\leq\; 2^{c_0 n}(V/\tau_0^n)e^{-\rho(\tau_0^{c_0 n}/V)},$$

*where $c_0$ is an absolute constant.*

*Proof.* Let $M^\circ$ denote the interior of $M$ (i.e., it contains all points of $M$ that are not at the boundary). Let $q_0 \in M$ be any fixed point such that $B_G(q_0, \frac{\tau_0}{2}) \subset M^\circ$. Then, by Lemmas A.8 and A.2 we know that $\text{VOL}(B_G(q_0, \frac{\tau_0}{2})) \geq \tau_0^{c_0 n}$. Observing that $M$ has volume at most $V$, we immediately get that $B_G(q_0, \frac{\tau_0}{2})$ occupies at least $\tau_0^{c_0 n}/V$ fraction of $M$. Thus

$$\mathbf{Pr}_{p_1,\ldots,p_\rho \sim \mu(M)}\Big[\forall i, p_i \notin B_G\big(q_0, \frac{\tau_0}{2}\big)\Big] \;\leq\; \Big(1 - \frac{\tau_0^{c_0 n}}{V}\Big)^\rho.$$

Now, let $C \subset M$ be a $(\frac{\tau_0}{2})$-geodesic covering of $M$. Using Lemmas A.7 and A.2, we can have $|C| \leq 2^{c_1 n}(V/\tau_0^n)$ (where $c_1$ is an absolute constant). Define

$C' \subset C$ as the set $\{c \in C : B_G(c, \frac{\tau_0}{2}) \subset M^\circ\}$. Then by union bounding over points in $C'$, we have

$$\mathbf{Pr}_{p_1,\ldots,p_\rho \sim \mu(M)}\Big[\exists c \in C' : \forall i, p_i \notin B_G\big(c, \frac{\tau_0}{2}\big)\Big] \;\; \leq \;\; |C'|\Big(1 - \frac{\tau_0^{\mathrm{c}_0 n}}{V}\Big)^\rho.$$

Equivalently we can say that, with probability at least $1 - |C'|e^{-\tau_0^{\mathrm{c}_0 n}\rho/V}$, for all $c' \in C'$, there exists $p_i \in \{p_1, \ldots, p_\rho\}$ such that $p_i \in B_G(c', \frac{\tau_0}{2})$.

Now, pick *any* $F \in \mathcal{F}(M, \tau_0)$, and let $q \in M$ denote its center (i.e., $q$ such that $B_G(q, \tau_0) = F$). Then since $C$ is a $(\frac{\tau_0}{2})$-geodesic cover of $M$, there exists $c \in C$ such that $D_G(q, c) \leq \tau_0/2$. Also, note that $c$ belongs to the set $C'$, since $B_G(c, \tau_0/2) \subset B_G(q, \tau_0) = F \subset M^\circ$. Thus with probability $\geq 1 - |C'|e^{-\tau_0^{\mathrm{c}_0 n}\rho/V}$, there exists $p_i$ such that

$$p_i \in B_G(c, \tau_0/2) \subset B_G(q, \tau_0) = F.$$

Observe that since the choice of $F$ was arbitrary, we have that for any $F \in \mathcal{F}$ (uniformly), there exists $p_i \in \{p_1, \ldots, p_\rho\}$ such that $p_i \in F$. The lemma follows. ∎

**Lemma 5.9** *Let $\mathcal{B}$ belong to class $(V, n, \tau)$. Fix a sample of size $m$ $\{b_1, \ldots, b_m\} := S_m \subset \mathcal{B}$, and let $\partial b_i$ denote the boundary of the manifold bag $b_i \in S_m$. Define $\frac{1}{\kappa} := \max_{b_i \in S_m}\{\textsc{cond}(\partial b_i)\}$. Now let $p_1^i, \ldots, p_\rho^i$ be the $\rho$ independent instances drawn uniformly from (the image of) $b_i$. Let $\mathcal{H}_r$ be a $\lambda$-lipschitz (w.r.t. $\ell_2$-norm) hypothesis class. Then, for any $\epsilon \leq \min\{\frac{\tau}{32}, \frac{\kappa}{8}\}$,*

$$\mathbf{Pr}\Big[\exists h_r \in \mathcal{H}_r, \exists b_i \in S_m : |\bar{h}_r(b_i) - \max_{j \in [\rho]} h_r(b_i(p_j^i))| > 9\epsilon\lambda\Big] \leq m 2^{\mathrm{c}_0 n}(V/\epsilon^n)e^{-\rho\epsilon^{\mathrm{c}_0 n}/V},$$

*where $\mathrm{c}_0$ is an absolute constant.*

*Proof.* Fix a bag $b_i \in S_m$, and let $M$ denote the manifold $b_i$. Quickly note that $\textsc{cond}(M) \leq 1/\tau$.

Define $M_{2\epsilon} := \{p \in M : \min_{q \in \partial M} D_G(p, q) \geq 2\epsilon\}$. By recalling that $\textsc{cond}(\partial M) \leq \frac{1}{\kappa}$ and $\epsilon \leq \min\{\frac{\tau}{32}, \frac{\kappa}{8}\}$, it follows that i) $M_{2\epsilon}$ is non-empty, ii) $\forall x \in M \setminus M_{2\epsilon}$, $\min_{y \in M_{2\epsilon}} D_G(x, y) \leq 8\epsilon$.

Observe that for all $p \in M_{2\epsilon}$, $B_G(p, \epsilon)$ is in the interior of $M$. Thus by applying Lemma 5.8, we have:

$$\mathbf{Pr}_{p_1,\ldots,p_\rho \sim \mu(M)}\big[\exists p \in M_{2\epsilon} : \forall i, p_i \notin B_G(p, \epsilon)\big] \;\; \leq \;\; 2^{\mathrm{c}_0 n}(V/\epsilon^n)e^{-\rho(\epsilon^{\mathrm{c}_0 n}/V)},$$

where $\mu(M)$ denotes the uniform probability measure on $M$.

Now for *any* $h_r \in \mathcal{H}_r$, let $x^* := \arg\max_{p\in M} h_r(p)$. Then with the same failure probability, we have that there exists some $p_i \in \{p_1, \ldots, p_\rho\}$ such that $D_G(p_i, x^*) \leq 9\epsilon$. To see this, consider:

if $x^* \in M_{2\epsilon}$, $D_G(x^*, p_i) \leq \epsilon$ (for some $p_i \in \{p_1, \ldots, p_\rho\}$), otherwise if $x^* \in M \setminus M_{2\epsilon}$, then exists $q \in M_{2\epsilon}$ such that $D_G(x^*, q) \leq 8\epsilon$.

Noting that $h_r$ is $\lambda$-Lipschitz, and union bounding over $m$ bags, the lemma follows. ∎

By Theorem 5.3 we have for any $0 < \gamma < 1$, with probability at least $1 - \delta_1$ over the sample $S_m$, for every $\bar{h}_r \in \overline{\mathcal{H}}_r$:

$$\mathrm{err}(\bar{h}_r) \;\leq\; \widehat{\mathrm{err}}_\gamma(\bar{h}_r, S_m) + O\left(\sqrt{\frac{n^2 \mathrm{FAT}_{\frac{\gamma}{16}}(\mathcal{H}_r)}{m} \log^2\left(\frac{Vm}{\gamma^2 \tau_0^n}\right) + \frac{1}{m}\ln\frac{1}{\delta_1}}\right),$$

where $\tau_0 = \min\{\frac{\tau}{2}, \frac{\gamma}{8}, \frac{\gamma}{8\lambda}\}$.

By applying Lemma 5.9 (with $\epsilon$ set to $\tau_1 = \min\{\frac{\tau}{32}, \frac{\kappa}{8}, \frac{\gamma}{9\lambda}, \frac{\gamma}{9}\}$), it follows that if $\rho \geq \Omega\big((V/\tau_1^{c_0 n})\big(n + \ln\big(\frac{mV}{\tau_1^n \delta_2}\big)\big)\big)$, then with probability $1 - \delta_2$: $\widehat{\mathrm{err}}_\gamma(\bar{h}_r, S_m) \leq \widehat{\mathrm{err}}_{2\gamma}(\bar{h}_r, S_{m,\rho})$, yielding the theorem.

## 5.4 Synthetic dataset generation

We generate a 1-dimensional manifold (in $\mathbb{R}^2$) of curvature $1/\tau$ and volume (length) $V$ as follows (see also Figure 5.8).

1. Pick a circle with radius $\tau$, a point $p$ on the circle, and a random angle $\theta$ (less than $\pi$).

2. Choose a direction (either clockwise or counterclockwise) and trace out an arc of length $\theta\tau$ starting at $p$ and ending at point, say, $q$.

3. Now pick another circle of the same radius $\tau$ that is tangent to the original circle at the point $q$.

4. Repeat the process of tracing out another arc on the new circle, starting at point $q$ and going in the reverse direction.

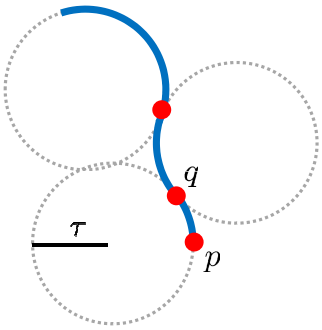



**Figure 5.8**: Synthetic bags. An example of a synthetic 1-dimensional manifold of a specified volume (length) $V$ and curvature $1/\tau$ generated by our procedure.

5. Terminate this process once we have a manifold of volume $V$.

Notice that this procedure can potentially result in a curve that intersects itself or has the condition number less than $1/\tau$. If this happens, we simply reject such a curve and generate another random curve until we have a well-conditioned manifold.

To generate a higher dimensional manifold, we extend our 1-dimensional manifold ($M \subset \mathbb{R}^2$) in the extra dimensions by taking a Cartesian product with a cube: $M \times [0,1]^{n-1}$. Notice that the "cube"-extension does not alter the condition number (i.e. it remains $1/\tau$). Since the resulting manifold fills up only $n+1$ dimensions, we randomly rotate it the ambient space.

Now, to label the generated manifolds positive and negative, we first fix $h^*$ to be a vertical hyperplane (w.r.t. the first coordinate) in $\mathbb{R}^D$. To label a manifold $b$ negative, we translate it such that the entire manifold lies in the negative region induced by $h^*$. And to label it positive, we translate it such that a part of $b$ lies in the positive region induced by $h^*$.

# Chapter 6

# Formalizing Intrinsic Dimension

So far we have discussed how the complexity of various popular learning algorithms scale with a *known* intrinsic structure, namely manifold structure. However, as discussed in Chapter 1, there are several other intrinsic structures of interest. Consider for instance a sparse or a cluster structure. How do we guarantee good rates for those structures? More importantly, what if our data conforms to some low dimensional *irregular* structure? Ideally we would like to provide good performance guarantees *without* having knowledge of the exact low-dimensional intrinsic structure.

Here we survey some popular notions of intrinsic dimension that have been inspired by data geometry. We evaluate strengths and weaknesses of each notion and introduce a more statistically verifiable notion.

## 6.1 Intrinsic dimension

Let $\mathcal{X}$ denote the space in which data lie. Here we assume $\mathcal{X}$ is a subset of $\mathbb{R}^D$, and that the metric of interest is Euclidean ($L_2$) distance. How can we characterize the intrinsic dimension of $\mathcal{X}$? This question has aroused keen interest in many different scientific communities, and has given rise to a variety of definitions. Here are four of the most successful such notions, arranged in decreasing order of generality:

- Covering dimension
- Assouad dimension
- Manifold dimension
- Affine dimension

The most general is the *covering dimension*: the smallest $d$ for which there is a constant $C > 0$ such that for any $\epsilon > 0$, $\mathcal{X}$ has an $\epsilon$-cover of size $C(1/\epsilon)^d$. This notion lies at the heart of much of empirical process theory. Although it permits many kinds of analysis and is wonderfully general, for our purposes it falls short on one count: for nonparametric estimators, we need small covering numbers for $\mathcal{X}$, but also for individual *neighborhoods* of $\mathcal{X}$. Thus we would like this same covering condition (with the same constant $C$) to hold for all $L_2$-balls in $\mathcal{X}$. This additional stipulation yields the *Assouad dimension*, which is defined as the smallest $d$ such that for any (Euclidean) ball $B \subset \mathbb{R}^D$, $X \cap B$ can be covered by $2^d$ balls of half the radius.

At the bottom end of the spectrum is the *affine dimension*, which is simply the smallest $d$ such that $\mathcal{X}$ is contained in a $d$-dimensional affine subspace of $\mathbb{R}^D$. It is a tall order to expect this to be smaller than $D$, although we may hope that $\mathcal{X}$ lies close to such a subspace. A more general hope is that $\mathcal{X}$ lies on (or close to) a $d$-dimensional Riemannian submanifold of $\mathbb{R}^D$. This notion makes a lot of intuitive sense, but in order for it to be useful either in algorithmic analysis or in estimating dimension, it is necessary to place conditions on the curvature of the manifold (as discussed in Chapter 2; see also works by Amenta and Bern [1998] and Niyogi et al. [2008]).

In what sense is our list arranged by decreasing generality? If $\mathcal{X}$ has an affine dimension of $d$, it certainly has manifold dimension at most $d$ (whatever the restriction on curvature). Similarly, low Assouad dimension implies small covering numbers. The only nontrivial containment result is that if $\mathcal{X}$ is a $d$-dimensional Riemannian submanifold with bounded curvature, then sufficiently small neighborhoods of $\mathcal{X}$ (where this neighborhood radius depends on the curvature) have Assouad dimension $O(d)$. This result is formalized and proved in Dasgupta and

Freund [2008]. The containment is strict: there is a substantial gap between manifolds of bounded curvature and sets of low Assouad dimension, on account of the smoothness properties of the former. This divide is not just a technicality but has important algorithmic implications. For instance, a variant of the Johnson Lindenstrauss lemma states that when a $d$-dimensional manifold (of bounded curvature) is projected onto a random subspace of dimension $O(d/\epsilon^2)$, then all interpoint Euclidean distances are preserved within $1 \pm \epsilon$ (see e.g. Baraniuk and Wakin [2009] and Clarkson [2008]). This does not hold for sets of Assouad dimension $d$ [Indyk and Naor, 2007].

None of these four notions arose in the context of data analysis, and it is not clear that any of them is well-suited to the dual purpose of (i) capturing a type of intrinsic structure that holds (verifiably) for many data sets and (ii) providing a formalism in which to analyze statistical procedures. In addition, they all describe sets, whereas in statistical contexts we are more interested in characterizing the dimension of a probability distribution. The recent machine learning literature, while appealing to the manifold idea for intuition, seems gradually to be moving towards a notion of "local flatness". Dasgupta and Freund [2008] formalized this notion and called it the *local covariance dimension*.

## 6.2 Local covariance dimension

**Definition 6.1** *Let $\mu$ be any measure over $\mathbb{R}^D$ and let $S$ be its covariance matrix. We say that $\mu$ has covariance dimension $(d, \epsilon)$ if the largest $d$ eigenvalues of $S$ account for $(1-\epsilon)$ fraction of its trace. That is, if the eigenvalues of $S$ are $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_D$, then*

$$\lambda_1 + \cdots + \lambda_d \geq (1-\epsilon)(\lambda_1 + \cdots + \lambda_D).$$

A distribution has covariance dimension $(d, \epsilon)$ if all but an $\epsilon$ fraction of its variance is concentrated in a $d$-dimensional affine subspace. Equivalently, the projection of the distribution onto this subspace leads to at most an $\epsilon$ total loss in squared distances. It is, in general, too much to hope that an entire data distribution

would have low covariance dimension. But we might hope that this property holds *locally*; or more precisely, that all (or most) sufficiently-small neighborhoods have low covariance dimension. At this stage, we could make this definition more complicated by quantifying the "most" or "sufficiently small" (as Dasgupta and Freund [2008] did to some extent), but it will turn out that we don't need to do this in order to state our theorems, so we leave things as they are.

Intuitively, the local covariance condition lies somewhere between manifold dimension and Assouad dimension, although it is more general in that merely requires points to be close to a locally flat set, rather than exactly on it.

## 6.3 Experiments with dimension

Covariance dimension is an intuitive notion, and recalls standard constructs in statistics such as mixtures of factor analyzers. It is instructive to see how it might be estimated, and whether there is evidence that many data sets do exhibit low covariance dimension.

First let's set our expectations properly. Even if data truly lies near a low-dimensional manifold, this property would only be apparent at a certain *scale*, that is, when considering neighborhoods whose radii lie within an appropriate range. For larger neighborhoods, the data set might seem slightly higher dimensional: the union of a slew of local low-dimensional subspaces. And for smaller neighborhoods, all we would see is pure noise, and the data set would seem full-dimensional.

Thus we will empirically estimate covariance dimension at different resolutions. First, we determine the diameter $\Delta$ of the dataset $\mathbf{X}$ by computing the maximum interpoint distance, and we choose multiple values $r \in [0, \Delta]$ as our different scales (radii). For each such radius $r$, and each data point point $x \in \mathbf{X}$, we compute the covariance matrix of the data points lying in the ball $B(x, r)$, and we determine (using a standard eigenvalue computation) how many dimensions suffice for capturing a $(1 - \epsilon)$ fraction of the variance. In our experiments, we try $\epsilon = 0.1$ and 0.01. We then take the dimension at scale $r$ (call it $d(r)$) to be average of all these values (over $x$).

If the balls $B(x, r)$ are so small as to contain very few data points, then the estimate $d(r)$ is not reliable. Thus we also keep track of $n(r)$, the average number of data points within the balls $B(x, r)$ (averaged over $x$). Roughly, we can expect $d(r)$ to be a reliable estimate if $n(r)$ is an order of magnitude larger than $d(r)$.

Figure 6.1 plots $d(r)$ against $r$ for several data sets. The numerical annotations on each curve represent the values $n(r)$. The larger the ratio $n(r)/d(r)$, the higher our confidence in the estimate.

The top figure shows dimensionality estimates for a noisy version of the ever-popular "swiss roll". In small neighborhoods, it is noise that dominates, and thus the data appear full-dimensional. In larger neighborhoods, the two-dimensional structure emerges: notice that the neighborhoods have very large numbers of points, so that we can feel very confident about the estimate of the local covariances. In even larger neighborhoods, we capture a significant chunk of the swiss roll and again revert to three dimensions.

The middle figure is for a data set consisting of images of a rotating teapot, each $30 \times 50$ pixels in size. Thus the ambient dimension is 1500, although the points lie close to a one-dimensional manifold (a circle describing the rotation). There is clear low-dimensional structure at a small scale, although in the figure, these $d(r)$ values seem to be 3 or 4 rather than 1.

The figure on the bottom is for a data set of noisy measurements from 12 sensors placed on a robotic arm with two joints. Thus the ambient dimension is 12, but there are only two underlying degrees of freedom.

In the next chapter we shall relate the learning rates with data's local covariance dimension using partition based trees datastructures.

# Acknowledgements

The contents of this chapter originally appeared in the following publication: N. Verma, S. Kpotufe and S. Dasgupta. Which spatial partition trees are adaptive to intrinsic dimension? *Uncertainty in Artificial Intelligence (UAI)*, 2009.

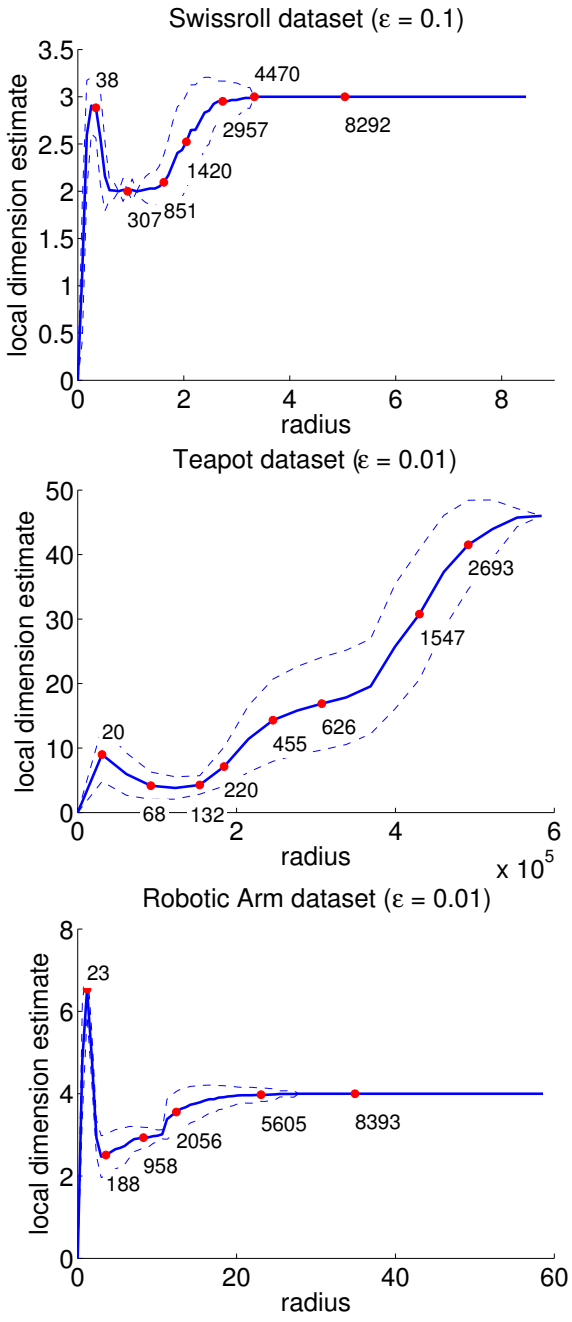



**Figure 6.1**: Local covariance dimension estimates for various datasets. The bold line shows the dimension estimate, with dashed confidence bands giving standard deviations over the different balls of each radius. The numeric annotations are average numbers of datapoints falling in balls of the specified radius. Left: Noisy swissroll (ambient dimension 3). Middle: Rotating teapot dataset (ambient dimension 1500). Right: Sensors on a robotic arm (ambient dimension 12).

# Chapter 7

# Learning Rates with Partition Trees

A *spatial partitioning tree* recursively divides space into increasing fine partitions. The most popular such data structure is probably the $k$-d tree, which splits the input space into two cells, then four, then eight, and so on, all with axis-parallel cuts. Each resulting partitioning has cells that are axis-aligned hyperrectangles (see Figure 7.1, middle). Once such a hierarchical partitioning is built from a data set, it can be used for standard statistical tasks. When a new query point $q$ arrives, that point can quickly be moved down the tree to a leaf cell (call it $C$). For classification, the majority label of the data points in $C$ can be returned. For regression, it will be the average of the response values in $C$. For nearest neighbor search, the closest point to $q$ in $C$ can be returned; of course, this might not be $q$'s nearest neighbor overall, but if the cell $C$ is sufficiently small, then it will at any rate be a point close enough to $q$ to have similar properties.

There are different ways to build a $k$-d tree, depending on which split coordinate is chosen at each stage. There are also many other types of spatial partitioning trees, such as dyadic trees (Figure 7.1, left) and PCA trees. We are interested in understanding the relative merits of these different data structures, to help choose between them. A natural first step, therefore, is to look at the underlying statistical theory. This theory, nicely summarized in Chapter 20 of Devroye et al. [1996], says that the convergence properties of tree-based estimators can be characterized

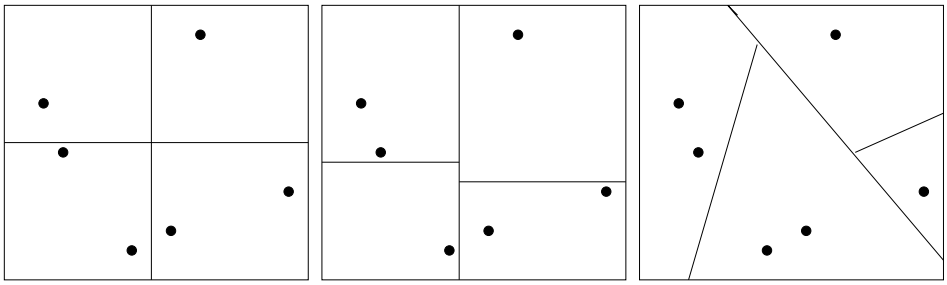


**Figure 7.1**: Some examples of spatial trees. Left: dyadic tree – cycles through coordinates and splits the data at the mid point. Middle: $k$-d tree – picks the coordinate direction with maximum spread and splits the data at the median value. Right: RP tree – picks a random direction from the unit sphere and split the data at the median value.

in terms of the rate at which cell diameters shrink as you move down the tree. The more rapidly these cells shrink, the better.

For $k$-d trees, these cell diameters can shrink very slowly when the data is high-dimensional. For $D$-dimensional data, it may require $D$ levels of the tree (and thus at least $2^D$ data points) to just *halve* the diameter. Thus $k$-d trees suffer from the same curse of dimensionality as other nonparametric statistical methods. But what if the data has low intrinsic dimension; for instance, if it lies close to a low-dimensional manifold? We are interested in understanding the behavior of spatial partitioning trees in such situations.

Some recent work Dasgupta and Freund [2008] provides new insights into this problem. It begins by observing that there is more than one way to define a cell's diameter. The statistical theory has generally considered the diameter to be the distance between the furthest pair of points on its boundary (if it is convex, then this is the distance between the furthest pair of vertices of the cell). It is very difficult to get bounds on this diameter unless the cells are of highly regular shape, such as hyperrectangles. A different, more flexible, notion of diameter looks at the furthest pair of *data points* within the cell, or even better, the typical interpoint distance of data within the cell (see Figure 7.2). It turns out that rates of convergence for statistical estimators can be given in terms of these kinds of *data diameter* (specifically, in terms of the rate at which these diameters decrease down the tree). Moreover, these data diameters can be bounded even if the cells are of

unusual shapes. This immediately opens the door to analyzing spatial partitioning trees that produce non-rectangular cells.

Dasgupta and Freund [2008] introduced *random projection trees*—in which the split at each stage is at the median along a direction chosen at random from the surface of the unit sphere (Figure 7.1, right)—and showed that the data diameter of the cells decreases at a rate that depends only on the *intrinsic dimension* of the data, not $D$:

> Let $d$ be the intrinsic dimension of data falling in a particular cell $C$ of an RP tree. Then all cells $O(d)$ levels below $C$ have at most half the data diameter of $C$.

(There is no dependence on the ambient dimension $D$.) They proved this for two notions of dimension: *Assouad dimension*, which is standard in the literature on analysis on metric spaces, and *local covariance dimension* (see Chapter 6 for details).

We are interested in exploring these phenomena more broadly, and for other types of trees. We start by examining the notion of local covariance dimension, and contrast it with other notions of dimension through a series of inclusion results. To get more intuition, we then investigate a variety of data sets and examine the extent to which these data verifiably have low local covariance dimension. The results suggest that this notion is quite reasonable and is of practical use. We then consider a variety of spatial partition trees: (i) $k$-d trees (of two types), (ii) dyadic trees, (iii) random projection trees, (iv) PCA trees, and (v) 2-means trees. We give upper and lower bounds on the diameter decrease rates achieved by these trees, as a function of local covariance dimension. Our strongest upper bounds on these rates are for PCA trees and 2-means trees, followed by RP trees. On the other hand, dyadic trees and $k$-d trees are weaker in their adaptivity. Our next step is to examine these effects experimentally, again on a range of data sets. We also investigate how the diameter decrease rate is correlated with performance in standard statistical tasks like regression and nearest neighbor search.

## 7.1 Spatial partition trees

Spatial partition trees conform to a simple template:

**Algorithm 7.1** PartitionTree(dataset $A \subset \mathcal{X}$)

**if** $|A| \leq$ MinSize **then**
  **return** leaf
**else**
  $(A_{\text{left}}, A_{\text{right}}) \leftarrow$ SplitAccordingToSomeRule($A$)
  LeftTree $\leftarrow$ PartitionTree($A_{\text{left}}$)
  RightTree $\leftarrow$ PartitionTree($A_{\text{right}}$)
**end if**
**return** (LeftTree, RightTree)

Different types of trees are distinguished by their splitting criteria. Here are some common varieties:

- **Dyadic tree:** Pick a coordinate direction and splits the data at the midpoint along that direction. One generally cycles through all the coordinates as one moves down the tree.
- **$k$-D tree:** Pick a coordinate direction and splits the data at the median along that direction. One often chooses the coordinate with largest spread.
- **Random Projection (RP) tree:** Split the data at the median along a random direction chosen from the surface of the unit sphere.
- **Principal Direction (PD) tree:** Split at the median along the principal eigenvector of the covariance matrix.
- **Two Means (2M) tree:** Pick the direction spanned by the centroids of the 2-means solution, and split the data as per the cluster assignment.

### 7.1.1 Notions of diameter

The generalization behavior of a spatial partitioning has traditionally been analyzed in terms of the physical diameter of the individual cells (see, for in-

stance, Devroye et al. [1996], Scott and Nowark [2006]). But this kind of diameter is hard to analyze for general convex cells. Instead we consider more flexible notions that measure the diameter of *data within the cell*. It has recently been shown that such measures are sufficient for giving generalization bounds (see Kpotufe and Dasgupta [2012] for the case of regression).

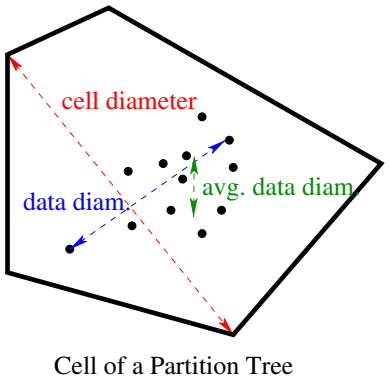



**Figure 7.2**: Various notions of diameter.

For any cell $A$, we will use two types of data diameter: the maximum distance between data points in $A$, denoted $\Delta(A)$, and the average interpoint distance among data in $A$, denoted $\Delta_a(A)$ (Figure 7.2).

## 7.2 Theoretical guarantees

Let $\mathbf{X} = \{X_1, \ldots, X_n\}$ be a data set drawn from underlying space $\mathcal{X}$, and let $\mu$ be the empirical distribution that assigns each weight to each of these points. Consider a partition of $\mathcal{X}$ into a collection of cells $\mathbf{A}$. For each such cell $A \in \mathbf{A}$, we can look at its maximum (data) diameter as well as its average (data) diameter; these are, respectively,

$$\Delta(A) := \max_{x,x' \in A \cap \mathbf{X}} \|x - x'\|$$

$$\Delta_a(A) := \frac{1}{(n\mu(A))} \left( \sum_{x,x' \in A \cap \mathbf{X}} \|x - x'\|^2 \right)^{1/2}$$

(for the latter it turns out to be a big convenience to use squared Euclidean distance.) We can also average these quantities all over cells $A \in \mathbf{A}$:

$$\Delta(\mathbf{A}) := \left(\frac{\sum_{A\in\mathbf{A}} \mu(A)\Delta^2(A)}{\sum_{A\in\mathbf{A}} \mu(A)}\right)^{1/2}$$
$$\Delta_a(\mathbf{A}) := \left(\frac{\sum_{A\in\mathbf{A}} \mu(A)\Delta_a^2(A)}{\sum_{A\in\mathbf{A}} \mu(A)}\right)^{1/2}$$

### 7.2.1 Irregular splitting rules

This section considers the RPTree, PDtree, and 2Mtree splitting rules. The nonrectangular partitions created by these trees turn out to be adaptive to the local dimension of the data: the decrease in average diameter resulting from a given split depends just on the eigenspectrum of the data in the local neighborhood, irrespective of the ambient dimension.

For the analysis, we consider a slight variant of these trees, in which an alternative type of split is used whenever the data in the cell has outliers (here, points that are much farther away from the mean than the typical distance-from-mean).

**Algorithm 7.2** split(region $A \subset \mathcal{X}$)

**if** $\Delta^2(A) \geq c \cdot \Delta_a^2(A)$ **then**
  {//SPLIT BY DISTANCE: remove outliers.}
  $A_{\text{left}} \leftarrow \{x \in A, \|x - \text{mean}(A)\| \leq \text{median}\{\|z - \text{mean}(A)\| : z \in \mathbf{X} \cap A\}\}$
**else**
  {//SPLIT BY PROJECTION: no outliers.}
  Choose a unit direction $v \in \mathbb{R}^D$ and a threshold $t \in \mathbb{R}$.
  $A_{\text{left}} \leftarrow \{x \in A,\, x \cdot v \leq t\}$
**end if**
$A_{\text{right}} \leftarrow A \setminus A_{\text{left}}$

The *distance split* is common to all three rules, and serves to remove outliers. It is guaranteed to reduce maximum data diameter by a constant fraction:

**Lemma 7.1 (Lemma 12 of Dasgupta and Freund [2008])** *Suppose* $\Delta^2(A) > c \cdot \Delta_a^2(A)$*, so that* $A$ *is split by* distance *under any instantiation of procedure* `split`*. Let* $\mathbf{A} = \{A_1, A_2\}$ *be the resulting split. We have*

$$\Delta^2(\mathbf{A}) \leq \left(\frac{1}{2} + \frac{2}{c}\right) \Delta^2(A).$$

We consider the three instantiations of procedure `split` in the following three sections, and we bound the decrease in diameter after a single split in terms of the local spectrum of the data.

### RPtree

For RPtree, the direction $v$ is picked randomly, and the threshold $t$ is the median of the projected data.

The diameter decrease after a split depends just on the parameter $d$ of the local covariance dimension, for $\epsilon$ sufficiently small.

**Theorem 7.2 (Theorem 4 of Dasgupta and Freund [2008])** *There exist constants* $0 < c_1, c_2 < 1$ *with the following property. Suppose* $\Delta^2(A) \leq c \cdot \Delta_a^2(A)$*, so that* $A$ *is split by* projection *into* $\mathbf{A} = \{A_1, A_2\}$ *using the RPtree split. If* $A \cap \mathbf{X}$ *has covariance dimension* $(d, c_1)$*, then*

$$\mathbb{E}\left[\Delta_a^2(\mathbf{A})\right] < (1 - c_2/d)\Delta_a^2(A),$$

*where the expectation is over the choice of direction.*

### PDtree

For PDtree, the direction $v$ is chosen to be the principal eigenvector of the covariance matrix of the data, and the threshold $t$ is the median of the projected data.

The diameter decrease after a split depends on the local spectrum of the data. Let $A$ be the current cell being split, and suppose the covariance matrix of the data in $A$ has eigenvalues $\lambda_1 \geq \cdots \geq \lambda_D$. If the covariance dimension of $A$ is $(d, \epsilon)$, define

$$k := \frac{1}{\lambda_1} \sum_{i=1}^{d} \lambda_i, \tag{7.1}$$

By definition, $k \leq d$.

The diameter decrease after the split depends on $k^2$, the worst case being when the data distribution in the cell has heavy tails. In the absence of heavy tails (condition (7.2)), we obtain a faster diameter decrease rate that depends just on $k$. This condition holds for any logconcave distribution (such as a Gaussian or uniform distribution), for instance. The decrease rate of $k$ could be much better than $d$ in situations where the first eigenvalue is dominant; and thus in such situations PD trees could do a lot better than RP trees.

**Theorem 7.3** *There exists constant $0 < c_1, c_2 < 1$ with the following property. Suppose $\Delta^2(A) \leq c \cdot \Delta_a^2(A)$, so that $A$ is split by* projection *into $\mathbf{A} = \{A_1, A_2\}$ using the PDtree split. If $A \cap \mathbf{X}$ has covariance dimension $(d, c_1)$, then*

$$\Delta_a^2(\mathbf{A}) < (1 - c_2/k^2)\Delta_a^2(A),$$

*where $k$ is as defined in (7.1).*

*If in addition the empirical distribution on $A \cap \mathbf{X}$ satisfies (for any $s \in \mathbb{R}$ and some $c_0 \geq 1$)*

$$\mathbb{E}_A[(X \cdot v - s)^2] \leq c_0 \left(\mathbb{E}_A[X \cdot v - s]\right)^2 \tag{7.2}$$

*we obtain a faster decrease where*

$$\Delta_a^2(\mathbf{A}) < (1 - c_2/k)\Delta_a^2(A).$$

*Proof.* The argument is based on the following fact which holds for any bi-partiton $\mathbf{A} = \{A_1, A_2\}$ of $A$ (see Lemma 15 of Dasgupta and Freund [2008]):

$$\Delta_a^2(A) - \Delta_a^2(\mathbf{A}) \;=\; 2\mu(A_1) \cdot \mu(A_2) \left\|\text{mean}(A_1) - \text{mean}(A_2)\right\|^2. \tag{7.3}$$

We start with the first part of the statement with no assumption on the data distribution. Let $\widetilde{x} \in \mathbb{R}$ be the projection of $x \in A \cap \mathbf{X}$ to the principal direction. WLOG assume that the median on the principal direction is 0. Notice

$$\begin{aligned} \left\|\text{mean}(A_1) - \text{mean}(A_2)\right\| &\geq \mathbb{E}\left[\widetilde{x} \,|\widetilde{x} > 0\right] - \mathbb{E}\left[\widetilde{x} \,|\widetilde{x} \leq 0\right] \\ &\geq \max\left\{\mathbb{E}\left[\widetilde{x} \,|\widetilde{x} > 0\right], -\mathbb{E}\left[\widetilde{x} \,|\widetilde{x} \leq 0\right]\right\}, \end{aligned}$$

where the expectation is over $x$ chosen uniformly at random from $A \cap \mathbf{X}$. The claim is therefore shown by bounding the r.h.s. below by $O(\Delta_a(A)/k$ and applying Equation (7.3).

We have $\mathbb{E}\left[\widetilde{x}^2\right] \geq \lambda_1$, so either $\mathbb{E}\left[\widetilde{x}^2 | \widetilde{x} > 0\right]$ or $\mathbb{E}\left[\widetilde{x}^2 | \widetilde{x} \leq 0\right]$ is greater than $\lambda_1$. Assume WLOG that it is the former. Let $\widetilde{m} = \max\{\widetilde{x} > 0\}$. We have that $\lambda_1 \leq \mathbb{E}\left[\widetilde{x}^2 | \widetilde{x} > 0\right] \leq \mathbb{E}\left[\widetilde{x} \,|\widetilde{x} > 0\right] \widetilde{m}$, and since $\widetilde{m}^2 \leq c\Delta_a^2\left(A\right)$, we get $\mathbb{E}\left[\widetilde{x} \,|\widetilde{x} > 0\right] \geq \frac{\lambda_1}{\Delta_a(A)\sqrt{c}}$. Now, by the assumption on covariance dimension,

$$\lambda_1 = \frac{\sum_{i=1}^{d} \lambda_i}{k} \geq (1 - c_1)\frac{\sum_{i=1}^{D} \lambda_i}{k} = (1 - c_1)\frac{\Delta_a^2\left(A\right)}{2k}.$$

We therefore have (for appropriate choice of $c_1$) that $\mathbb{E}\left[\widetilde{x} \,|\widetilde{x} > 0\right] \geq \Delta_a(A)/4k\sqrt{c}$, which concludes the argument for the first part.

For the second part, assumption (7.2) yields

$$\mathbb{E}\left[\widetilde{x} \,|\widetilde{x} > 0\right] - \mathbb{E}\left[\widetilde{x} \,|\widetilde{x} \leq 0\right] \;=\; 2\mathbb{E}\left|\widetilde{x}\right| \geq 2\sqrt{\frac{\mathbb{E}\left|\widetilde{x}\right|^2}{c_0}} \geq 2\sqrt{\frac{\lambda_1}{c_0}} \;=\; 2\sqrt{\frac{\Delta_a^2\left(A\right)}{4c_0k}}.$$

We finish up by appealing to Equation (7.3). ∎

**2Mtree**

For 2Mtree, the direction $v = \text{mean}(A_1) - \text{mean}(A_2)$ where $A = \{A_1, A_2\}$ is the bisection of $A$ that minimizes the 2-means cost. The threshold $t$ is the half point between the two means.

The 2-means cost can be written as

$$\sum_{i \in [2]} \sum_{x \in A_i \cap \mathbf{X}} \|x - \text{mean}(A_i)\|^2 = \frac{n}{2}\Delta_a^2\left(\mathbf{A}\right).$$

Thus, the 2Mtree (assuming an exact solver) minimizes $\Delta_a^2\left(\mathbf{A}\right)$. In other words, it decreases diameter at least as fast as RPtree and PDtree. Note however that, since these are greedy procedures, the decrease in diameter over multiple levels may not be superior to the decrease attained with the other procedures.

**Theorem 7.4** *Suppose* $\Delta^2\left(A\right) \leq c \cdot \Delta_a^2\left(A\right)$*, so that* $A$ *is split by* projection *into* $\mathbf{A} = \{A_1, A_2\}$ *using the RPtree split. There exists constants* $0 < c_1, c_2 < 1$ *with*

*the following property. Assume $A \cap \mathbf{X}$ has covariance dimension $(d, c_1)$. We then have*

$$\Delta_a^2(\mathbf{A}) < (1 - c_2/d')\Delta_a^2(A),$$

*where $d' \leq \min\{d, k^2\}$ for general distributions, and $d'$ is at most $k$ for distributions satisfying (7.2).*

**Diameter decrease over multiple levels**

The diameter decrease parameters $d, k^2, k, d'$ in Theorems 7.2, 7.3, 7.4 above are a function of the covariance dimension of the data in the cell $A$ being split. The covariance dimensions of the cells may vary over the course of the splits implying that the decrease rates may vary. However, we can bound the overall diameter decrease rate over multiple levels of the tree in terms of the worst case rate attained over levels.

**Lemma 7.5 (diameter decrease over multiple levels)** *Suppose a partition tree is built by calling* `split` *recursively (under any instantiation). Assume furthermore that every node $A \subset \mathcal{X}$ of the tree satisfies the following: let $\mathbf{A} = \{A_1, A_2\}$ represent the child nodes of $A$, we have for some constants $0 < c_1, c_2 < 1$ and $\kappa \leq D$ that*

*(i) If $A$ is split by distance, $\Delta^2(\mathbf{A}) < c_1\Delta^2(A)$.*

*(ii) If $A$ is split by projection, $\mathbb{E}[\Delta_a^2(\mathbf{A})] < (1 - c_2/\kappa)\Delta_a^2(A)$.*

*Then, there exists a constant $C$ such that the following holds: let $\mathbf{A}_l$ be the partition of $\mathcal{X}$ defined by the nodes at level $l$, we have*

$$\mathbb{E}\left[\Delta_a^2(\mathbf{A}_l)\right] \leq \mathbb{E}\left[\Delta^2(\mathbf{A}_l)\right] \leq \frac{1}{2^{\lfloor l/C\kappa \rfloor}}\Delta^2(\mathcal{X}),$$

*where the expectation is over the randomness in the algorithm for $\mathbf{X}$ fixed.*

So if every split decreases average diameter at a rate controlled by $\kappa$ as defined above, then it takes at most $O(\kappa \log(1/\epsilon))$ levels to decrease average diameter down to an $\epsilon$ fraction of the original diameter of the data. Combined with Theorems 7.2, 7.3, 7.4, we see that the three rules considered will decrease diameter at a fast rate whenever the covariance dimensions in local regions are small.

### 7.2.2 Axis parallel splitting rules

It was shown by Dasgupta and Freund [2008] that axis-parallel splitting rules do not always adapt to data that is intrinsically low-dimensional. They exhibit a data set in $\mathbb{R}^D$ that has low Assouad dimension $O(\log D)$, and where $k$-d trees (and also, it can be shown, dyadic trees) require $D$ levels to halve the data diameter.

The adaptivity of axis-parallel rules to covariance dimension is unclear. But they *are* guaranteed to decrease diameter at a rate depending on $D$. The following result states that it takes at most $O(D(\log D)\log(1/\epsilon))$ levels to decrease average diameter to an $\epsilon$ fraction of the original data diameter.

**Theorem 7.6** *Suppose a partition tree is built using either $k$-d tree or dyadic tree by cycling through the coordinates. Let $\mathbf{A}_l$ be the partition of $\mathcal{X}$ defined by the nodes at level $l$. Then we have*

$$\Delta_a^2(\mathbf{A}_l) \leq \Delta^2(\mathbf{A}_l) \leq \frac{D}{2^{\lfloor l/D \rfloor}} \Delta^2(\mathcal{X}).$$

## 7.3 Experiments

To highlight the adaptivity of these spatial trees, we need to resort to synthetic datasets where we can fully control the intrinsic and the ambient space. We will vary the ambient dimension while keeping the intrinsic dimension fixed and empirically calculate the rate at which the diameter decreases for various trees. We will then evaluate their performance on some common learning tasks on typical realworld datasets to see how these trees fair in practice.

**Spatial trees – versions used:** As discussed earlier, many versions are available for different spatial trees. Here we restrict our attention to the following variants: dyadic trees – fix a permutation and cycle through the coordinates, $k$-D trees – determine the spread over each coordinate by computing the coordinate vise diameter and picking the coordinate with maximum diameter, RP trees – pick the direction that results in the largest diameter decrease from a bag of 20 random

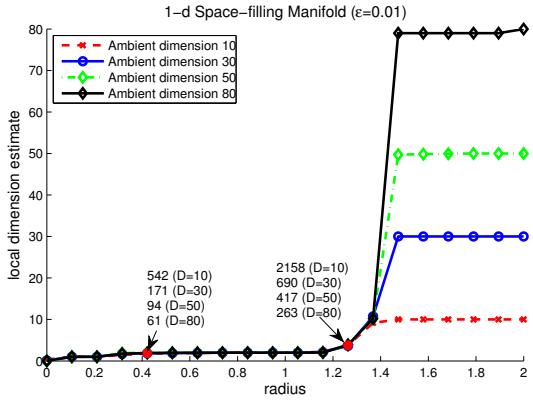



**Figure 7.3**: Local covariance dimension estimate of a space-filling 1-d manifold.

directions, PD trees – pick the principal direction in accordance to the data falling in each node of the tree, 2M trees – solve 2-means via the Lloyd's method and pick the direction spanned by the centroids of the 2-means solution.

### 7.3.1 Synthetic dataset: Space-filling manifold

To create a well behaved low dimensional manifold that shows adaptivity, one needs to take care of the following. As we vary the ambient dimension (keeping intrinsic dimension fixed) we want that: i) the curvature of the manifold shouldn't change by too much, ii) diameter of the manifold should remain constant, iii) the manifold should fill up the ambient space $D$ (it doesn't reside in some affine subspace of $D$).

We thus resort to a continuous 1-dimensional manifold on the surface of a $D-1$ dimensional sphere constructed via the sinusoidal basis. Data is generated by by sampling $20,000$ points uniformly at random from the interval $[0, 2\pi]$ and applying the map $M : t \mapsto \sqrt{\frac{2}{D}}\left(\sin(t), \cos(t), \sin(2t), \cos(2t) \ldots, \sin(\frac{Dt}{2}), \cos(\frac{Dt}{2})\right)$.

Figure 7.3 shows the local covariance dimension estimate for this space-filling 1-manifold (embedded in ambient space of dimension 10, 30, 50 and 80).

To show that these trees are *adaptive* to the intrinsic dimension of this space-filling manifold, we need to show that the number of levels needed to reduce the diameter by a certain factor eventually becomes constant (regardless of the

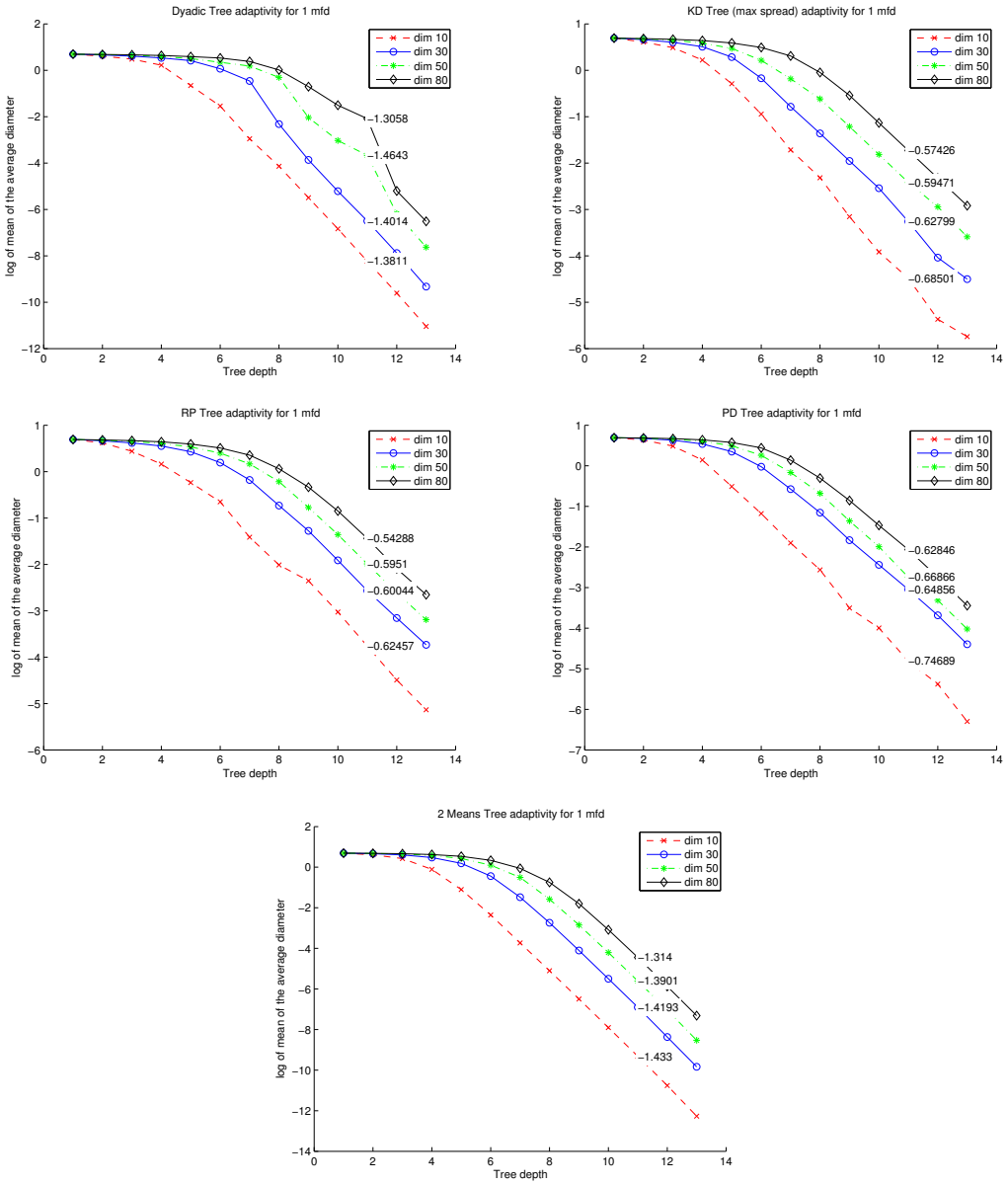



**Figure 7.4**: Adaptivity plots for various spatial trees on synthetic space-filling curve dataset. Note that the *slope* of the plotted curve shows the decrease rate (cf. (7.4)). Parallel lines highlight that the diameter decrease rates eventually become *independent* of the ambient dimension adapting to the low dimensional intrinsic structure of the manifold.

ambient dimension). More formally, since the average diameter decrease rate is given by

$$\frac{\text{change in log avg. diameter}}{\text{change in levels}}, \tag{7.4}$$

we plot of log of the average diameter (y-axis) against the tree depth (x-axis) for different spatial trees (see Figure 7.4). Notice that the quantity of interest – Equation (7.4) is the *slope* of the plotted curves. The annotated number on the curves is the average slope value of the last five measurements.

Observe, as one expects, that it takes a few levels to get down to the low dimensional manifold structure in the data. Notice that for RP, PD and 2M trees, the plots for various ambient dimensions essentially become parallel, highlighting that the decrease rates all converge to a single stable number *regardless of the size of the ambient dimension*, showing adaptivity to the data's intrinsic dimension. Notice that for dyadic trees the rate estimates (slopes) for high ambient dimension are not consistent with the low ambient dimension indicating that even after about 12 levels (dividing the space into 4096 partitions), showing lack of adaptivity. Note that the version of $k$-D tree that explicitly minimizes the diameter decrease criterion performs remarkably well. Note, however, that at each step the $k$-D has to compute this diameter decrease calculation, making it an expensive operation.

### 7.3.2 Real-world datasets

We now compare the performance of different spatial trees for typical learning tasks on some real-world dataset clusters. To exhibit a wide range of applicability, we choose the 'digit 1' cluster from the MNIST OCR dataset of handwritten digits, 'love' cluster from Australian Sign Language time-series dataset from UCI Machine Learning Repository [Kadous, 2002], and 'aw' phoneme from MFCC TIMIT dataset.

Experiments have been set as follows. For each cluster, we first estimate its local covariance dimension (as discussed in Sections 6.2 and 6.3). See Figure 7.5. We then do a 10-fold cross validation. For each fold, we use the training data to build the partition tree, and for each test point we compute the vector quantization error and the closest neighbor as it trickles down the tree.

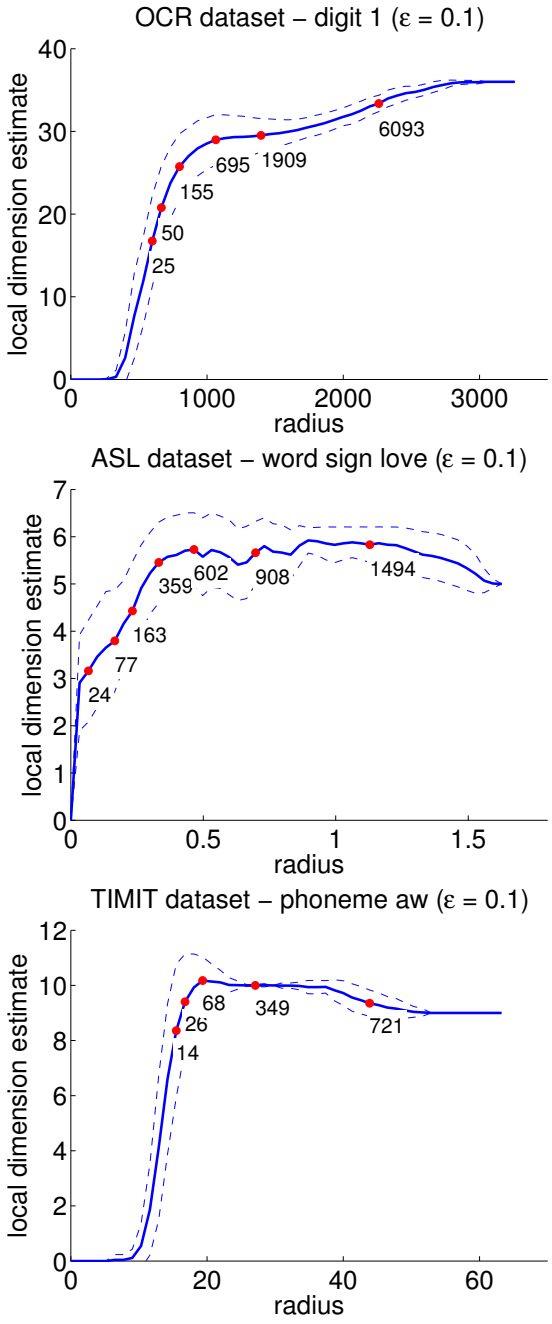



**Figure 7.5**: Local covariance dimension estimates for some real-world dataset clusters.

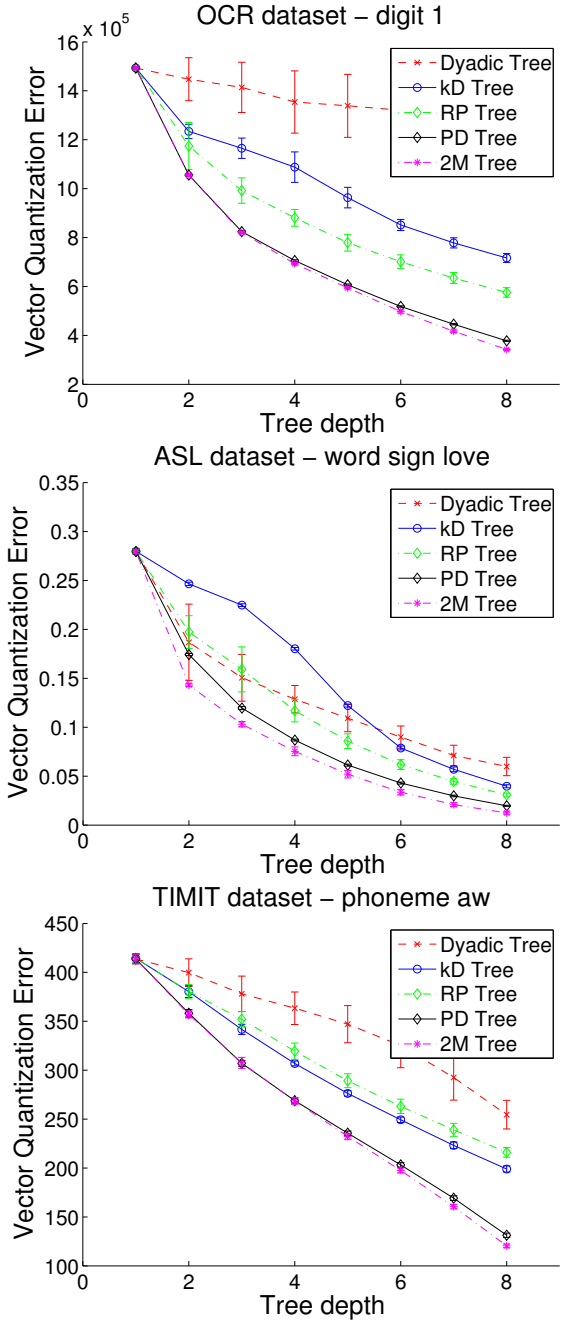



**Figure 7.6**: Average vector quantization error induced by different spatial trees on the datasets.

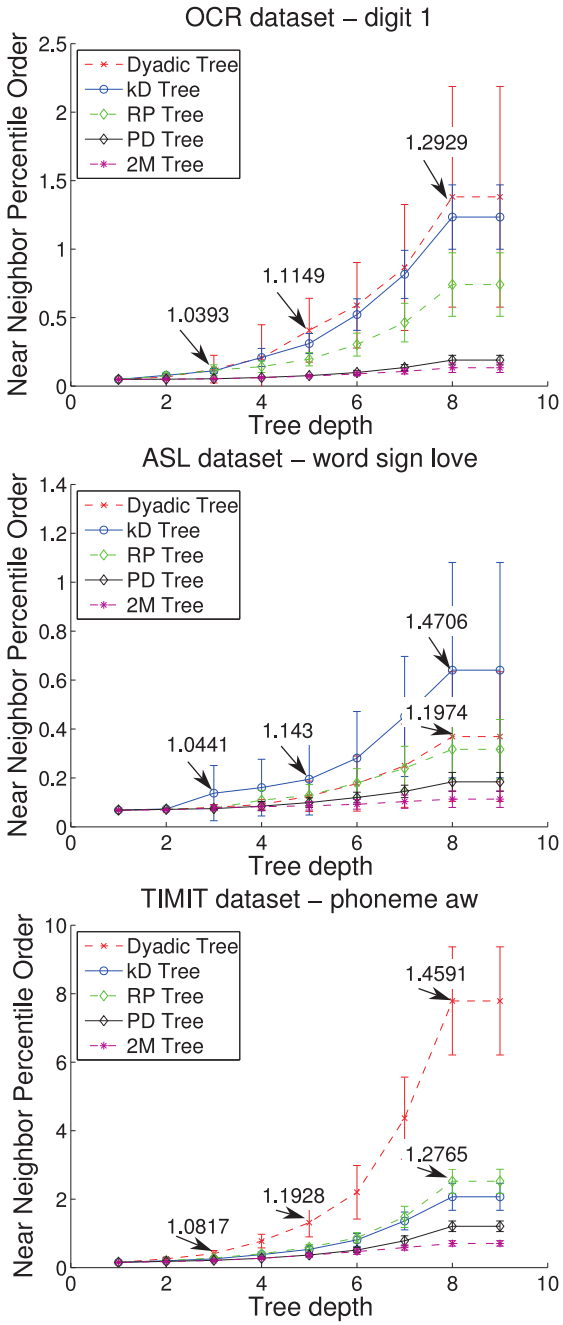



**Figure 7.7**: Results for near neighbor query. The annotated number shows the average ratio of the distance between the query point and found neighbor to the distance between the query point and the true nearest neighbor.

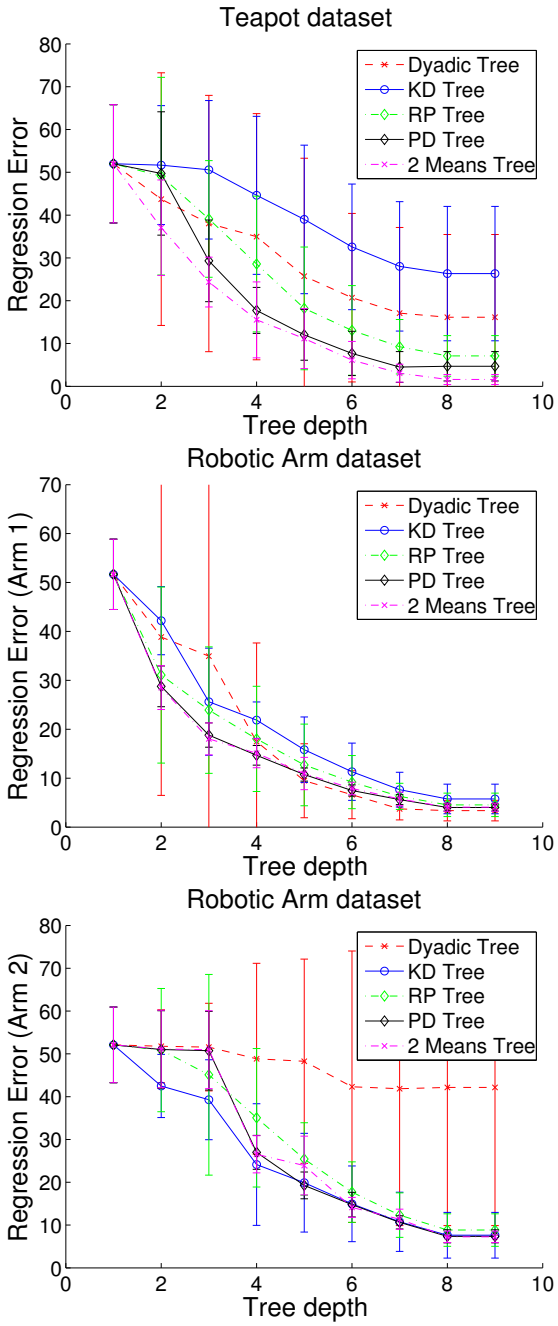



**Figure 7.8**: Plots showing relative performance of spatial trees on a typical regression task.

We report the average quantization error at different tree levels for various datasets (Figure 7.6). Notice that the PD and 2M trees consistently produce better quantization results than other trees.

Figure 7.7 shows the result of a near neighbor search. The plot shows the true percentile order of the found neighbor to the query point at different tree levels. The annotated numbers show the ratio of the distance between the query point and found neighbor to the distance between the query point and its true nearest neighbor. This helps in gauging the quality of the found neighbor in terms of distances.

As before, 2M and PD trees consistently yield better near neighbors to the query point. We should remark that the apparent good results of dyadic trees on the ASL dataset (middle row, middle column) should be taken in context with the number of datapoints falling in a particular node. For dyadic trees it is common to have unbalanced splits resulting in high number of datapoints falling in an individual cell. This significantly increases the chance of finding a close neighbor but also significantly increases the computational cost of finding that close neighbor.

Diverting our attention to the task of regression, we use the rotating teapot dataset (to predict the angle of rotation) and the robotic arm dataset (to predict the angular positions of the first and the second arm) (see Section 6.3 for description of the datasets). Figure 7.8 shows the relative performance of different spatial trees.

## Acknowledgements

The contents of this chapter originally appeared in the following publications: Y. Freund, S. Dasgupta, M. Kabra and N. Verma. Learning the structure of manifolds using random projections. *Neural Information Processing Systems (NIPS)*, 2007, and N. Verma, S. Kpotufe and S. Dasgupta. Which spatial partition trees are adaptive to intrinsic dimension? *Uncertainty in Artificial Intelligence (UAI)*, 2009.

## 7.4 Supporting proofs

### 7.4.1 Proof of Lemma 7.5

Fix $\mathbf{X}$. Consider the random variable $X$ drawn uniformly from $\mathbf{X}$. Let the r.v's $A_i = A_i(X), i = 0, \ldots, l$ denote the cell to which $X$ belongs at level $i$ in the tree. Define $I(A_i) := \mathbf{1}\{\Delta^2(A_i) \leq c\Delta_a^2(A_i)\}$.

Let $\mathbf{A}_l$ be the partition of $\mathcal{X}$ defined by the nodes at level $l$, we'll first show that $\mathbb{E}[\Delta^2(\mathbf{A}_l)] \leq \frac{1}{2}\Delta^2(\mathcal{X})$ for $l = C\kappa$ for some constant $C$. We point out that $\mathbb{E}[\Delta^2(\mathbf{A}_l)] = \mathbb{E}[\Delta^2(A_l)]$ where the last expectation is over the randomness in the algorithm and the choice of $X$.

To bound $\mathbb{E}[\Delta^2(A_l)]$, note that one of the following events must hold:

(a) $\exists\ 0 \leq i_1 < \cdots < i_m < l,\ m \geq \frac{l}{2}, I\left(A_{i_j}\right) = 0$

(b) $\exists\ 0 \leq i_1 < \cdots < i_m < l,\ m \geq \frac{l}{2}, I\left(A_{i_j}\right) = 1$

Let's first condition on event (a). We have

$$\mathbb{E}\left[\Delta^2(A_l)\right] \leq \mathbb{E}\left[\Delta^2(A_{i_m+1})\right] = \mathbb{E}\left[\mathbb{E}\left[\Delta^2(A_{i_m+1})\,|A_{i_m}\right]\right],$$

and since by the assumption, $\mathbb{E}[\Delta^2(A_{i_m+1})\,|A_{i_m}] \leq c_1\Delta^2(A_{i_m})$, we get $\mathbb{E}[\Delta^2(A_l)]$ $\leq c_1\mathbb{E}[\Delta^2(A_{i_m})]$. Applying the same argument recursively on $i_j$, $j = m, (m-1)$, $\ldots, 1$, we obtain $\mathbb{E}[\Delta^2(A_l)] \leq c_1^m \cdot \mathbb{E}[\Delta^2(A_{i_1})] \leq c_1^{l/2}\Delta^2(\mathcal{X})$.

Now condition on event (b). Using the fact that $\mathbb{E}[\Delta_a^2(A_i)]$ is non-increasing in $i$ (see Dasgupta and Freund [2008]), we can apply a similar recursive argument as above to obtain that $\mathbb{E}[\Delta_a^2(A_{i_m})] \leq (1 - c_2/d)^{m-1}\,\mathbb{E}[\Delta_a^2(A_{i_1})]$. It follows that $\mathbb{E}[\Delta^2(A_l)] \leq \mathbb{E}[\Delta^2(A_m)] \leq c\mathbb{E}[\Delta_a^2(A_m)] \leq c\left(1 - \frac{c_2}{d}\right)^{l/2-1}\Delta^2(\mathcal{X})$.

Thus, in either case we have

$$\mathbb{E}\left[\Delta^2(A_l)\right] \leq \max\left\{c_1^{l/2},\, c\,(1 - c_2/d)^{l/2-1}\right\} \cdot \Delta^2(\mathcal{X}),$$

and we can verify that there exists $C$ such that the r.h.s. above is at most $\frac{1}{2}\Delta^2(\mathcal{X})$ for $l \leq C\kappa$. Thus, we can repeat the argument over every $C\kappa$ levels to obtain the statement of the lemma.

### 7.4.2 Proof of Theorem 7.6

We assume that the procedure builds the tree by cycling through the coordinates (a single coordinate is used at each level).

Suppose a cell $A$ is split into $\mathbf{A} = \{A_1, A_2\}$ along some coordinate $e_i$. Then the average diameter along coordinate $i$ decreases under either split

$$\frac{1}{\mu(A)}\left(\mu(A_1)\Delta^2\left(A_1 \cdot e_i\right) + \mu(A_2)\Delta^2\left(A_2 \cdot e_i\right)\right) \;\leq\; \frac{1}{2}\Delta^2\left(A \cdot e_i\right).$$

To see this, notice that the masses of the resulting cells are halved under $k$-$d$ tree splits (we assume that $n$ is a power of 2), while the diameters are halved under the dyadic tree splits.

We can derive an upper bound on the diameter decrease rate over multiple levels as follows. Let $X \sim \mathcal{U}(\mathbf{X})$, and let $A_l$ be the cell to which $\widetilde{X}$ belongs at level $l \geq 0$ in the tree (built by either procedure). Let $l \geq 1$, if we condition on the event that the split at level $l-1$ is along coordinate $i$, we have by the above argument that $\mathbb{E}_X\left[\Delta^2\left(A_l \cdot e_i\right)\right] \leq \frac{1}{2}\mathbb{E}_X\left[\Delta^2\left(A_{l-1} \cdot e_i\right)\right]$. No matter the coordinate used for the previous split, we always have $\mathbb{E}_X\left[\Delta^2\left(A_l \cdot e_i\right)\right] \leq \mathbb{E}_X\left[\Delta^2\left(A_{l-1} \cdot e_i\right)\right]$, and it follows that after a multiple of $D$ levels we have

$$\mathbb{E}_X\left[\Delta^2\left(A_l \cdot e_i\right)\right] \leq \frac{1}{2^{l/D}}\Delta^2\left(\mathcal{X} \cdot e_i\right),$$

for all $i \in [D]$. Summing over all coordinates, we then get

$$\mathbb{E}_X\left[\Delta^2\left(A_l\right)\right] \leq \mathbb{E}_X\left[\sum_i^D \Delta^2\left(A_l \cdot e_i\right)\right] \leq \frac{D}{2^{l/D}}\Delta(\mathcal{X}).$$

To conclude, notice that $\mathbb{E}_X\left[\Delta^2\left(A_l\right)\right]$ is exactly $\Delta^2\left(\mathbf{A}_l\right)$ where $\mathbf{A}_l$ is the partition defined by the nodes at level $l$.

## 7.5 Empirical and distributional covariance dimensions

Theorems 7.2, 7.3, 7.4 concern the case when the data itself has low covariance dimension in local regions. For the sake of completeness, one may be

interested in what happens if the data is drawn from a distribution with low covariance dimension in local regions. Would the data also have low covariance dimension in most neighborhoods? This depends on whether enough points fall into a neighborhood, and also on the amount of outliers in the region since spectral quantities are very sensitive to outliers.

In this section we'll distinguish between the empirical measure and the underlying distribution. We denote the underlying distribution by $\mu$ while the empirical measure is denoted $\mu_n$. We have the following result.

**Lemma 7.7 (convergence of covariance dimension)** *Consider a collection $\mathcal{C}$ of subsets of $\mathcal{X}$, where $\mathcal{C}$ has VC dimension $\mathcal{V}$. The following holds simultaneously for all $A \subset \mathcal{C}$, with probability at least $1 - 2\delta$ over the sampling of $\mathbf{X}$.*

*Suppose $A \subset \mathcal{C}$ has covariance dimension $(d, \epsilon)_\mu$, and*

$$\mu_n(A) \geq 28672 \left( \frac{\Delta(A)}{\Delta_{n,a}(A)} \right)^4 \frac{\mathcal{V} \log n + \log(12/\delta)}{n\epsilon^2}.$$

*Then $A$ has empirical covariance dimension $(d, 2\epsilon)_n$.*

The sets of interest are the cells obtained by the partitioning procedures. By the nature of the splitting rules, these cells are intersections of at most $O(\log n)$ hyperplanes since the trees would typically be grown to a height of at most $O(\log n)$. Thus, the cells belong to a collection $\mathcal{C}$ of VC dimension at most $O(D \log n)$ using standard arguments on composition of VC classes and the fact that the class of half spaces has VC dimension $D + 1$.

The rest of the section gives an overview of the proof of Lemma 7.7. We will require additional notation to distinguish between empirical and distributional quantities. Notation will therefore be introduced where needed.

**Lemma 7.8 (relative VC bounds)** *Consider a class $\mathcal{C}$ of VC dimension $\mathcal{V}$. With probability at least $1 - \delta/3$ over the choice of the sample $\mathbf{X}$, we have for all $A \in \mathcal{C}'$ that*

$$\mu(A) \leq \mu_n(A) + 2\sqrt{\mu_n(A) \frac{\mathcal{V} \ln(2n) + \ln \frac{12}{\delta}}{n}} + 4 \frac{\mathcal{V} \ln(2n) + \ln \frac{12}{\delta}}{n}, \quad (7.5)$$

$$\mu_n(A) \leq \mu(A) + 2\sqrt{\mu(A) \frac{\mathcal{V} \ln(2n) + \ln \frac{12}{\delta}}{n}} + 4 \frac{\mathcal{V} \ln(2n) + \ln \frac{12}{\delta}}{n}. \quad (7.6)$$

**Lemma 7.9** *Let $\mathcal{C}$ be a class of subsets of $\mathbb{R}^D$ of VC-dimension $\mathcal{V}$. Consider a mapping which associates an orthonormal projection matrix $P_A \in \mathbb{R}^{D\times D}$ to $A \in \mathcal{C}$. Let $\mathbb{E}_{n,A}$ and $\mathbb{E}_A$ denote expectations taken with respect to $\mu_n$ and $\mu$ conditioned on $x \in A$. Also, let $\nu_A$ and $\nu_{n,A}$ denote the mean of $A$ under $\mu$ and the empirical mean. With probability at least $1-\delta$ over the sample $\mathbf{X}$, the following holds for all $A \in \mathcal{C}$ satisfying $\mu(A) \geq \frac{\mathcal{V}\ln(2n)+\ln(12/\delta)}{n}$:*

$$\begin{aligned}&\left|\mathbb{E}_{n,A}\left\|P_A(x-\nu_{n,A})\right\|^2 - \mathbb{E}_A\left\|P_A(x-\nu_A)\right\|^2\right| \\ &\quad\leq\ 8\frac{\Delta^2(A)}{\mu_n(A)}\sqrt{\mu(A)\frac{\mathcal{V}\log n+\log(12/\delta)}{n}} + 64\Delta^2(A)\cdot\mu(A)\frac{\mathcal{V}\log n+\log(12/\delta)}{n(\mu_n(A))^2}.\end{aligned}$$

*Proof.* [Proof Idea] It can be verified that

$$\begin{aligned}&\left|\mathbb{E}_{n,A}\left\|P_A(x-\nu_{n,A})\right\|^2 - \mathbb{E}_A\left\|P_A(x-\nu_A)\right\|^2\right| \\ &\quad\leq\ \left|\mathbb{E}_{n,A}\ \left\|P_A(x-\nu_A)\right\|^2 - \mathbb{E}_A\left\|P_A(x-\nu_A)\right\|^2\right| + \left\|\nu_{n,A}-\nu_A\right\|^2.\end{aligned}$$

The proof proceeds with standard symmetrization arguments (using Rademacher random variables) to bound the two terms on the r.h.s. separately. ∎

*Proof.* [Proof of Lemma 7.7] Let $A \in \mathcal{C}$ have covariance dimension $(d.\epsilon)_\mu$ and satisfy $\mu_n(A) \geq 28672\left(\frac{\Delta(A)}{\Delta_{n,a}(A)}\right)^4\frac{\mathcal{V}\log n+\log(12/\delta)}{n\epsilon^2}$. We'll show that $A$ has empirical covariance dimension $(d, 2\epsilon)$ with probability at least $1-2\delta$ by applying Lemma 7.9.

Assume Equations (7.5) and (7.6) hold (as is assumed in Lemma 7.9). Since $\mu_n(A) \geq 7\frac{\mathcal{V}'\ln(2n)+\ln(12/\delta)}{n}$, we have by Equation (7.6),that $\mu(A) \geq \frac{\mathcal{V}\ln(2n)+\ln(12/\delta)}{n}$. By Equation (7.5), $\mu(A) \leq 7\mu(A)$. The following then holds with probability at least $1-2\delta$ by Lemma 7.9:

Let $V = [v_{d+1}, \ldots, v_D]$ be the eigenvectors of the covariance of $A \cap \mathbf{X}$ corresponding to the smallest $D-d$ eigenvalues $\{\lambda_i\}_{d+1}^D$, and let $\{\lambda_{n,i}\}_1^D$ be the eigenvalues of the empirical covariance on $A$. Now define $\epsilon_0 := 8\sqrt{7\frac{\mathcal{V}\log n+\log(12/\delta)}{n\mu_n(A)}}$.

We set $P_A$ to $I_D$, then to $VV^\top$, to obtain the two inequalities below:

$$\sum_{i=1}^D \lambda_i \leq \sum_{i=1}^D \lambda_{n,i} + \Delta^2(A)(\epsilon_0+\epsilon_0^2), \text{ and}$$

$$\mathbb{E}_{n,A}\left\|VV^{\top}(x-\nu_{n,A})\right\|^2 \leq \sum_{i=d+1}^{D}\lambda_i + \Delta^2(A)(\epsilon_0+\epsilon_0^2).$$

Write $c := (\Delta^2(A)/\Delta^2(A))$, and note that $\Delta^2(A) \leq c\cdot\Delta_a^2(A) = 2c\cdot\sum_{i=1}^{D}\lambda_{n,i}$. It follows that

$$\begin{aligned}\sum_{i=d+1}^{D}\lambda_{n,i} &\leq \mathbb{E}_{n,A}\left\|VV^{\top}(x-\nu_{n,A})\right\|^2 \leq \sum_{i=d+1}^{D}\lambda_i + 2\epsilon_0\Delta^2(A)\\ &\leq \epsilon\left(\sum_{i=1}^{D}\lambda_i\right) + 2\epsilon_0\Delta^2(A) \leq \epsilon\left(\sum_{i=1}^{D}\lambda_{n,i} + 2\epsilon_0\Delta^2(A)\right) + 2\epsilon_0\Delta^2(A)\\ &\leq (\epsilon + 8c\cdot\epsilon_0)\sum_{i=1}^{D}\lambda_{n,i} \leq 2\epsilon\cdot\sum_{i=1}^{D}\lambda_{n,i},\end{aligned}$$

where the last inequality follows from the setting of $\mu_n(A)$. ∎

# Chapter 8

# Regression Rates with Other Low-dimensional Structures

Chapters 6 and 7 discussed how tree based regressors adapt well to a specific statistical notion of intrinsic dimension. This adaptivity yielded good nearest neighbor and vector quantization performance along with the regression performance. Here we show the universality of the phenomenon of *adaptivity to intrinsic dimension* in non-parametric regression.

It turns out that as long as one can exhibit some sort of *low-dimensional* organized structure on a given dataset (perhaps by showing some kind of small covering, or by some easily parameterizable sophisticated data-structure), a standard space partition based regressor induced by the organized structure yields good regression rates.

## 8.1 Partition based non-parametric regression

Given an i.i.d. sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ of size $n$, from an underlying distribution on $\mathcal{X} \times \mathcal{Y}$, we want to study what kinds of regression rates one can get by using a piecewise constant non-parametric regressor using a space partitioning method on $\mathcal{X}$. Here we assume that $\mathcal{X} \subseteq \mathbb{R}^D$ and $\mathcal{Y} \subseteq \mathbb{R}^m$.

Let $\mathbf{X}$ denote the set of $X_i$'s and $\mathbf{Y}$ denote the $Y_i$'s from the sample. Define

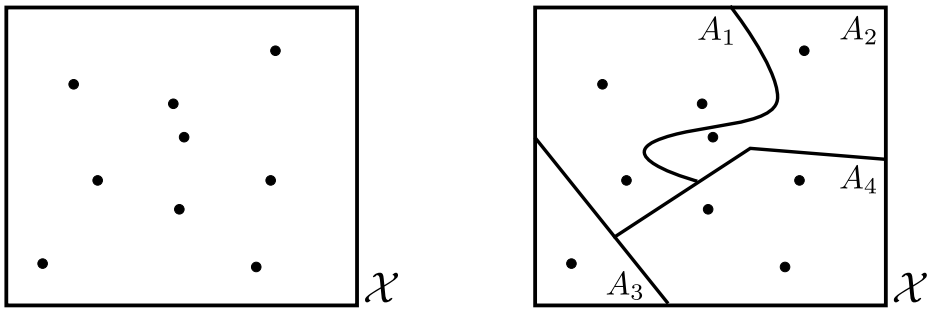



**Figure 8.1**: An example space partitioning of $\mathcal{X}$. Left: an example space $\mathcal{X}$ with a i.i.d. samples. Right: A partitioning of $\mathcal{X}$ in four cells $(A_1, \ldots A_4)$.

$f(X)$ as the *regression function*, that is,

$$f(X) := \mathbb{E}[Y|X].$$

Let $\mathbf{A}$ be some partitioning of $\mathcal{X}$, and for some $X \in \mathcal{X}$, let $A(X)$ denote the cell $A \in \mathbf{A}$ such that $X \in A$ (see Figure 8.1). For simplicity we shall assume that each cell in the partitioning contains at least one sample. We define the piecewise constant regressor w.r.t. the partition $\mathbf{A}$ as

$$f_{n,\mathbf{A}}(X) := \frac{1}{|A(X) \cap \mathbf{X}|} \sum_i Y_i \, \mathbf{1}[X_i \in A(X)]. \tag{8.1}$$

We are interested in bounding the excess "integrated" risk. Let $\Delta^2(S)$ denote the squared diameter of a set $S$, that is, $\Delta^2(S) := \sup_{x,y \in S} \|x - y\|^2$, and let $\Delta_a^2(S)$ denote the average squared diameter of $S$, that is, $\Delta_a^2(S) := \frac{1}{|S|^2} \sum_{x,y \in S} \|x - y\|^2$. Then we have the following (see Section 8.4 for all the supporting proofs).

**Theorem 8.1 (excess integrated risk in the fixed design setting)** *Let $(X_1, Y_1), \ldots, (X_n, Y_n) =: (\mathbf{X}, \mathbf{Y})$ be an i.i.d. sample from the underlying distribution on $\mathcal{X} \times \mathcal{Y}$. Let $\mathbf{A}$ be any partition of $\mathcal{X}$ that is independent of $\mathbf{Y}$, and $f_{n,\mathbf{A}}$ be a piecewise constant regressor w.r.t. $\mathbf{A}$ (as defined above). If the* regression function *$f$ is $\lambda$-Lipschitz (w.r.t. $L_2$-norm)*[1]*, then with probability at least $1 - \delta$, the excess*

[1]In light of Theorem 3.1 of Györfi et al. [2002], a Lipschitz-type smoothness assumption on $f$ is necessary to get non-trivial bounds.

*integrated risk (in the fixed design setting) is*

$$\frac{1}{n}\sum_{i=1}^{n}\|f_{n,\mathbf{A}}(X_i)-f(X_i)\|^2 \leq 4\Delta^2(\mathcal{Y})\frac{|\mathbf{A}|(\ln(|\mathbf{A}|/\delta)+2)}{n}+\lambda^2\sum_{A\in\mathbf{A}}\frac{|A\cap\mathbf{X}|}{n}\Delta_a^2(A\cap\mathbf{X}).$$

**Theorem 8.2 (excess integrated risk in the random design setting)** *Let* $(X_1,Y_1),\ldots,(X_n,Y_n)=:(\mathbf{X},\mathbf{Y})$ *be an i.i.d. sample from the underlying distribution on* $\mathcal{X}\times\mathcal{Y}$. *Let* $\mathbf{A}$ *be any partition of* $\mathcal{X}\subset\mathbb{R}^D$ *that is independent of* $\mathbf{Y}$, *and whose cells come from a fixed collection* $\mathcal{A}$, *and* $f_{n,\mathbf{A}}$ *be a piecewise constant regressor w.r.t.* $\mathbf{A}$ *(as defined above). If VC-dimension of* $\mathcal{A}$ *is at most* $\mathcal{V}<\infty$, *and the* regression function $f$ *is* $\lambda$*-Lipschitz (w.r.t.* $L_2$*-norm), then with probability at least* $1-\delta$, *the excess integrated risk (in the random design setting) is*

$$\int\|f_{n,\mathbf{A}}(x)-f(x)\|^2\ d\mu(x) \leq 14\lambda^2\left[\sum_{A\in\mathbf{A}}\frac{|A\cap\mathbf{X}|}{n}\Delta^2(A\cap\mathbf{X})\right]+42\Delta^2(\mathcal{Y})|\mathbf{A}|\frac{8+3\mathcal{V}\ln(6n)+\ln(|\mathbf{A}|/\delta^2)}{n}.$$

Theorems 8.1 and 8.2 provide a powerful relationship between the underlying regression function and our piecewise constant estimate via the partitioning. It turns out that arranging our data in an organized structure that induces a good partition[2] can yield good regression rates. Let's study some interesting data arrangements.

## 8.2 Organized structures that adapt to intrinsic dimension

### 8.2.1 Spatial trees

As discussed in Chapter 7 (see e.g. Lemma 7.5) several tree based partitioning procedures guarantee a diameter decrease rate of $2^{-l/d(\mathcal{X})}$ for going $l$ levels

[2] A good partition is typically the one that produces a good balance between the total number of cells and the data diameter in each cell.

down the tree (here $d(\mathcal{X})$ is some deterministic function of the geometry of the underlying data $\mathcal{X}$, such as doubling dimension or local covariance dimension of $\mathcal{X}$). A quick calculation yields the following rate.

Let $(\mathbf{X}, \mathbf{Y})$ be a size $n$ i.i.d. sample. Pick any $0 < c < 1/2$, and using the sample $\mathcal{X}$, grow a spatial tree to depth $c\log(n)$ (see Section 7.1 for example trees). Since each node of the spatial tree partitions the space in two cells, the leaves of a depth $c\log(n)$ tree induces a partition $\mathbf{A}$ of size $2^{c\log(n)} = n^c$. Assuming $c\log n \geq \log(\log(n^c/\delta) + 2)$, with probability at least $1 - \delta$, the piecewise constant regressor estimate $f_{n,\mathbf{A}}$ (cf. Eq. (8.1)) induced by the partition $\mathbf{A}$ gets the regression rate (cf. Theorem 8.1):

$$4\Delta^2(\mathcal{Y})n^{2c-1} + \lambda^2\Delta^2(\mathcal{X})n^{-c/d(\mathcal{X})}.$$

Optimizing for $c$ by setting it to $\frac{d(\mathcal{X})}{2d(\mathcal{X})+1}$ gives the rate of

$$\big(4\Delta^2(\mathcal{Y}) + \lambda^2\Delta^2(\mathcal{X})\big)n^{-1/1+2d(\mathcal{X})}.$$

It is instructive to contrast this rate to the standard $Cn^{-O(1/D)}$ rate induced by a "cube"-partitioning of $\mathcal{X} \subset \mathbb{R}^D$ (see for instance Theorem 4.3 of Györfi et al. [2002]). Our derivation signifies that the *choice* of partition is crucial to achieve good rates: a clever organization of data (an organization into partitions induced by spatial trees, in this case) can yield rates that are dependent only on the intrinsic geometry ($d(\mathcal{X})$) and not the ambient dimension ($D$).

### 8.2.2 Covering with balls

Suppose for all $r > 0$, we can exhibit an $r$-cover of size at most $(\Delta(\mathcal{X})/r)^{d(\mathcal{X})}$ (here $d(\mathcal{X})$ is again a deterministic function of the geometry of the underlying data $\mathcal{X}$). Recall from Section 6.1 that for $\mathcal{X}$ with covering dimension or with manifold dimension at most $b$, $d(\mathcal{X}) = O(b)$.

Let $(\mathbf{X}, \mathbf{Y})$ be a size $n$ i.i.d. sample and let $C$ be an $r$-cover of $\mathbf{X}$ (of size at most $(\Delta(\mathcal{X})/r)^{d(\mathcal{X})}$). Let $\mathbf{A}$ be any partition of $\mathcal{X}$ induced[3] by $C$. Assuming

[3]Given a covering $C$ of samples $\mathbf{X}$, we can induce a partition of our space $\mathcal{X}$ as follows. Let $C_1, \ldots, C_p$ be the covering elements in some arbitrary but fixed order. Define cell $A_i$ of the partition as $C_i \setminus \cup_{i'=1}^{i-1} C_{i'}$ (for $1 < i \leq p$). Finally define cell $A_1$ as $C_1 \cup (\cup_i C_i)^{\mathsf{C}}$. Define the partitioning $\mathbf{A} := \cup_i A_i$.

that $d(\mathcal{X})\log(\Delta(\mathcal{X})/r) \geq \log(\log((\Delta(\mathcal{X})/r)^{d(\mathcal{X})}/\delta) + 2)$, with probability at least $1-\delta$, the piecewise constant regressor estimate $f_{n,\mathbf{A}}$ yields the regression rate (cf. Theorem 8.1):

$$4\Delta^2(\mathcal{Y})(\Delta(\mathcal{X})/r)^{2d(\mathcal{X})}/n + 4\lambda^2 r^2.$$

By picking cover radius $r = (1/n)^{1/2+2d(\mathcal{X})}$, gives the rate

$$\left(4\Delta^2(\mathcal{Y})\Delta(\mathcal{X})^{2d(\mathcal{X})} + 4\lambda^2\right)n^{-1/1+d(\mathcal{X})}.$$

Our derived rate can be compared directly with Bickel and Li's [2006] manifold dimension adaptive rate of $Cn^{-1/O(d)}$ using a kernel based regressor (here $d$ is the dimension of the underlying manifold). The cover radius $r$ above has a role similar to the bandwidth parameter in Bickel and Li's work. Both results show adaptivity to data's intrinsic dimension. It is interesting to note that the structure induced by using localizing kernels (such as Gaussian kernel or box kernel) is similar to that induced by our partition $\mathbf{A}$ resulting in similar rates.

### 8.2.3 Covering with $k$-flats

Instead of an arbitrary $r$-covering with balls (as done above), suppose we can exhibit a more *structured* covering[4]: an $\epsilon$-covering of our samples with $k$-flats. That is, given an i.i.d. sample $(\mathbf{X}, \mathbf{Y})$ of size $n$, we construct a collection of $k$-dimensional flats $\mathbf{F}$ such that for every $X \in \mathbf{X}$, there exists a flat $F \in \mathbf{F}$, such that $\|X - \pi_F(X)\| \leq \epsilon$ (here $\pi_F(\cdot)$ denotes the projection of $\cdot$ into $F$). Let $N$ be the size of this collection, i.e. $N = |\,\mathbf{F}\,|$.

We can create a partition of $\mathcal{X}$ using our $k$-flat $\epsilon$-cover $\mathbf{F}$ as follows. For each flat $F$ in the cover, let $X_F \subset \mathbf{X}$ be the samples closest to $F$ (break ties arbitrarily). Let $C_F$ be a regular $r$-cover of $\pi_F(X_F)$ in $F$ (note that we can do it in such a way that $|C_F| \leq C_0(\Delta(\mathcal{X})/r)^k$, for some constant $C_0$ independent of $r$). Note that $C_F$ is a collection of $k$-dimensional balls in our $k$-flat $F$. We create the corresponding cover in $\mathcal{X}$ by extending $k$-dimensional balls in each $C_F$ in the ambient space by

[4] An advantage of such a structured covering is that we don't need to understand the underlying geometry of $\mathcal{X}$ as needed for an arbitrary cover.

extending it by $\epsilon$ amount in orthogonal complement of $F$. This gives us a cover of $\mathbf{X}$, which can now be made into a proper partition in a standard way (as done in previous section). We call this partition $\mathbf{A}$ as the partition induced by $k$-flat covering.

By our construction, we have a partition with total number of cells at most $NC_0(\Delta(\mathcal{X})/r)^k$. Again, by assuming that $NC_0(\Delta(\mathcal{X})/r)^k \geq \log(NC_0\Delta^k(\mathcal{X})/r^k\delta)+2$, with probability at least $1-\delta$, the piecewise constant regressor estimate $f_{n,\mathbf{A}}$ yields the following regression rate (cf. Theorem 8.1):

$$4\Delta^2(\mathcal{Y})N^2C_0^2(\Delta(\mathcal{X})/r)^{2k}/n + 8\lambda^2(r^2+\epsilon^2).$$

Optimizing for $r$ by setting it to $(N^2/n)^{1/2+2k}$, yields the rate of

$$\left(4C_0^2\Delta^2(\mathcal{Y})\Delta^{2k}(\mathcal{X}) + 8\lambda^2\right)N^{2/1+k}n^{-1/1+k} + 8\lambda^2\epsilon^2.$$

It is instructive to note that as long as $N = o(\sqrt{n})$, the l.h.s. converges to zero as we increase the number of samples $n$. This provides an interesting trade-off: (i) how well does the collection $\mathbf{F}$ fits the sample $\mathbf{X}$ (via $\epsilon$ in r.h.s.), and (ii) how "complex" is this collection (via $N = |\,\mathbf{F}\,|$ in l.h.s.). A balance between the two terms yields a good, ambient dimension independent rate.

### 8.2.4 Compression schemes

Suppose we have a compression scheme $(\phi, \psi, d, \epsilon, \alpha, \beta)$ on $\mathcal{X}$. That is,

- a compressor $\phi : \mathcal{X} \to \mathbb{R}^d$ such that $\Delta(\phi(\mathcal{X}))$ is bounded by $\alpha$,
- a decompressor $\psi : \mathbb{R}^d \to \mathcal{X}$ such that $\|X - \psi(\phi(X))\| \leq \epsilon$ for all $X \in \mathcal{X}$,
- $\|\psi(Y) - \psi(Y')\| \leq \beta\|Y - Y'\|$.

We can create a partition of $\mathcal{X}$ from this compression as follows. Let $(\mathbf{X}, \mathbf{Y})$ be a size $n$ i.i.d. sample. Let $C$ be an $r$-cover of $\phi(\mathbf{X})$ in $\mathbb{R}^d$. Note that $|C| \leq C_0(\alpha/r)^d$. Let $\mathbf{P}$ be a partition induced by $C$ of $\phi(\mathcal{X})$ in $\mathbb{R}^d$. Now consider any $\mathbf{P}$-respecting partition $\mathbf{A}$ of $\mathcal{X}$, that is, for any $\phi(X), \phi(X')$ in the same cell $P$ in $\mathbf{P}$, $X, X'$ is in the same cell in $\mathbf{A}$.

Assuming that $C_0(\alpha/r)^d \geq \log(C_0(\alpha/r)^d/\delta + 2)$, with probability at least $1-\delta$, the piecewise constant regressor estimate $f_{n,\mathbf{A}}$ yields the regression rate of:

$$4\Delta^2(\mathcal{Y})C_0^2(\alpha/r)^{2d}/n + 2\lambda^2(\beta^2 r^2 + 4\epsilon^2).$$

Optimizing for $r$ by setting it to $(\alpha^{2d}/\beta^2 n)^{1/2+2d}$, yields the rate

$$\left(4C_0^2\Delta^2(\mathcal{Y}) + 2\lambda^2\right)(\alpha\beta)^{2d/1+d}n^{-1/1+d} + 8\lambda^2\epsilon^2.$$

$\alpha$ and $\beta$ are complementary scaling parameters for the compression and decompression stages. A good compression scheme will typically have $\alpha\beta \approx 1$ and the reconstruction error $\epsilon \approx 0$. Thus, if one can find a good compression scheme of our data into $\mathbb{R}^d$, then the regression rates depend only on $d$, independent of the ambient dimension. As a quick example, consider data sampled uniformly from a patch parameterized manifold $M \subset \mathbb{R}^D$. That is, $M = f(H)$ for some $d$-dimensional patch $H \subset \mathbb{R}^d$. Then, the piecewise constant regression estimate derived from the partition induced by $(f^{-1}, f, d, 0, \alpha, \beta)$-compression scheme yields regression rates that scale with only with the manifold dimension $d$.

## 8.3 Discussion

Our results from previous section show that if one can find a clever way to arrange high-dimensional data in an organized structure, then the complexity of learning scales only with the intrinsic property of the structure, thus escaping the traditional curse of dimensionality.

## Acknowledgements

The contents of this chapter are based on unpublished work by N. Verma and S. Dasgupta.

## 8.4 Supporting proofs

### 8.4.1 Proof of Theorem 8.1

Consider the excess integrated risk (in the fixed design setting):

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\|f_{n,\mathbf{A}}(X_i)-f(X_i)\|^2 &= \frac{1}{n}\sum_{A\in\mathbf{A}}\sum_{X_i\in A}\|f_{n,\mathbf{A}}(X_i)-f(X_i)\|^2 \\
&\leq \frac{1}{n}\sum_{A\in\mathbf{A}}\sum_{X_i\in A} 2\Bigg[\underbrace{\Big\|\frac{1}{|A\cap\mathbf{X}|}\sum_{j:X_j\in A\cap\mathbf{X}}(Y_j-f(X_i))\Big\|^2}_{\text{variance}} \\
&\qquad + \underbrace{\Big\|\Big(\frac{1}{|A\cap\mathbf{X}|}\sum_{j:X_j\in A\cap\mathbf{X}}f(X_j)\Big)-f(X_i)\Big\|^2}_{\text{sq. bias}}\Bigg] \\
&\leq 4\Delta^2(\mathcal{Y})\frac{|\mathbf{A}|(\ln(|\mathbf{A}|/\delta)+2)}{n} + \lambda^2\underbrace{\sum_{A\in\mathbf{A}}\frac{|A\cap\mathbf{X}|}{n}\Delta_a^2(A\cap\mathbf{X})}_{\text{avg. sq. data diameter}},
\end{aligned}$$

where the last inequality is by noting Lemma 8.4 (for variance) and Lemma 8.3 for squared bias.

**Lemma 8.3 (bounding squared bias)** *Let $f$ be the $\lambda$-Lipschitz regression function, and $\mathbf{X}$ be the sample (as described above), and let $A$ be some cell in the partition of $\mathcal{X}$. Then,*

$$\frac{1}{|A\cap\mathbf{X}|}\sum_{X_i\in A\cap\mathbf{X}}\Big\|\Big(\frac{1}{|A\cap\mathbf{X}|}\sum_{X_j\in A\cap\mathbf{X}}f(X_j)\Big)-f(X_i)\Big\|^2 \;\leq\; \frac{\lambda^2}{2}\Delta_a^2(A\cap\mathbf{X}).$$

*Proof.* Observe that

$$\begin{aligned}
&\frac{1}{|A\cap\mathbf{X}|}\sum_{X_i\in A\cap\mathbf{X}}\Big\|\Big(\frac{1}{|A\cap\mathbf{X}|}\sum_{X_j\in A\cap\mathbf{X}}f(X_j)\Big)-f(X_i)\Big\|^2 \\
&\quad= \frac{1}{2}\frac{1}{|A\cap\mathbf{X}|^2}\sum_{X,X'\in A\cap\mathbf{X}}\|f(X)-f(X')\|^2 \\
&\quad\leq \frac{\lambda^2}{2|A\cap\mathbf{X}|^2}\sum_{X,X'\in A\cap\mathbf{X}}\|X-X'\|^2 \\
&\quad= \frac{\lambda^2}{2}\Delta_a^2(A\cap\mathbf{X}),
\end{aligned}$$

where the first equality is by noting Lemma 8.5. ∎

**Lemma 8.4 (bounding variance)** *Let $f$ be the regression function, $(X_1, Y_1), \ldots, (X_n, Y_n) =: (\mathbf{X}, \mathbf{Y})$ be an i.i.d. sample, and let $\mathbf{A}$ be some partition of $\mathcal{X}$ that is independent of $\mathbf{Y}$ (as described above). Pick any $\delta > 0$, then for all cells $A$ in the partition $\mathbf{A}$ of $\mathcal{X}$, we have*

$$\frac{1}{|A \cap \mathbf{X}|} \sum_{X_i \in A \cap \mathbf{X}} \Big\| \frac{1}{|A \cap \mathbf{X}|} \sum_{j : X_j \in A \cap \mathbf{X}} (Y_j - f(X_i)) \Big\|^2 \;\leq\; 2\Delta^2(\mathcal{Y}) \frac{2 + \ln(|\mathbf{A}|/\delta)}{|A \cap \mathbf{X}|}.$$

*Proof.* For each cell $A$ in $\mathbf{A}$, let $\mathbf{Y}_A := \{Y_i \in \mathbf{Y} : X_i \in A\} = (Y_{A,1}, \ldots, Y_{A,|A \cap \mathbf{X}|})$. Define the function $g_A : (Y_{A,1}, \ldots, Y_{A,|A \cap \mathbf{X}|}) \mapsto \| \frac{1}{|A \cap \mathbf{X}|} \sum_{Y_{A,j} \in \mathbf{Y}_A} (Y_{A,j} - f(X_i)) \|$ (for any $X_i \in A$). Then noting that changing a single $Y_{A,j}$ changes $g_A$ by only $\Delta(\mathcal{Y})/|A \cap \mathbf{X}|$ amount, we can apply Lemma 8.6 and get: with probability at least $1 - \delta/|\mathbf{A}|$,

$$|g_A - \mathbb{E} g_A| \leq \Delta(\mathcal{Y}) \sqrt{\frac{\ln(|\mathbf{A}|/\delta)}{2|A \cap \mathbf{X}|}}. \tag{8.2}$$

Also note that,

$$\begin{aligned} \mathbb{E}_{\mathbf{Y}_A} \Big\| \frac{1}{|A \cap \mathbf{X}|} \sum_{Y_{A,j} \in \mathbf{Y}_A} Y_{A,j} - f(X_i) \Big\| &\leq \frac{1}{|A \cap \mathbf{X}|} \Big( \mathbb{E}_{\mathbf{Y}_A} \Big\| \sum_{Y_{A,j} \in \mathbf{Y}_A} Y_{A,j} - f(X_i) \Big\|^2 \Big)^{\frac{1}{2}} \\ &= \frac{1}{|A \cap \mathbf{X}|} \Big( \sum_{Y_{A,j}} \mathbb{E}_{Y_{A,j}} \|Y_{A,j} - f(X_i)\|^2 \Big)^{\frac{1}{2}} \\ &\leq \frac{\Delta(\mathcal{Y})}{\sqrt{|A \cap \mathbf{X}|}}. \end{aligned} \tag{8.3}$$

Union bounding over all the $g_A$'s, and by combining Eqs. (8.2) and (8.3), we have with probability at least $1 - \delta$, for all $g_A$

$$\sum_{X_i \in A \cap \mathbf{X}} g_A^2 \;\leq\; 2 \sum_{X_i \in A \cap \mathbf{X}} \left[ \frac{\Delta^2(\mathcal{Y})}{|A \cap \mathbf{X}|} + \frac{\Delta^2(\mathcal{Y}) \ln(|\mathbf{A}|/\delta)}{2|A \cap \mathbf{X}|} \right] = 2\Delta^2(\mathcal{Y}) \frac{2 + \ln(|\mathbf{A}|/\delta)}{|A \cap \mathbf{X}|}.$$

∎

**Lemma 8.5 (Corollary 14 of Dasgupta and Freund [2008])** *For any set $S$ of numbers, let $\mu(S) := (1/|S|) \sum_{x \in S} x$, then*

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

**Lemma 8.6 (McDiarmid's concentration inequality [1989])** *Let $X_1, \ldots, X_n$ be $n$ independent random variables taking values in a set $A$. Let $g : A^n \to \mathbb{R}$ be a function such that for any $1 \leq i \leq n$,*

$$\sup_{x_1,\ldots,x_n,\hat{x}_i} |g(x_1, \ldots, x_n) - g(x_1, \ldots, x_{i-1}, \hat{x}_i, x_{i+1}, \ldots, x_n)| \leq c.$$

*Pick any $\delta > 0$, then with probability at least $1 - \delta$,*

$$g(X_1, \ldots, X_n) - \mathbb{E}g(X_1, \ldots, X_n) \leq \sqrt{\frac{n}{2c^2} \ln \frac{1}{\delta}}.$$

### 8.4.2 Proof of Theorem 8.2

Given the partition $\mathbf{A}$ of $\mathcal{X}$ consider a *refinement* of $\mathbf{A}$ as follows. For each $A \in \mathbf{A}$ we partition, let $B_A$ be the smallest enclosing ball, centered at one of the sample points from $A \cap \mathbf{X}$, that contains all the sample points from $A \cap \mathbf{X}$ (if $A$ contains no samples then set $B_A = \phi$). Define a refined partition $\mathbf{A}'$ as $\{A \cap B_A\}_{A \in \mathbf{A}} \cup \{A \cap B_A^{\mathsf{C}}\}_{A \in \mathbf{A}}$. Observe that the cells $A'$ of this refined partition come from the concept class: $\mathcal{A}' := \{A \cap B_A\}_{A \in \mathcal{A}} \cup \{A \cap B_A^{\mathsf{C}}\}_{A \in \mathcal{A}}$. Quickly note that the shatter coefficient $S(\mathcal{A}', 2n, n)$ (see Definition 8.7) is at most $d := 2(2ne/\mathcal{V})^{\mathcal{V}}(2n)^2 = 8n^2(2ne/\mathcal{V})^{\mathcal{V}}$. This follows from noting (i) Lemmas 8.10 and 8.9, and (ii) the concept class of balls centered at data has shatter coefficient at most $(2n)^2$.

Then the excess integrated risk (in the random design setting) becomes:

$$\int \|f_{n,\mathbf{A}}(x) - f(x)\|^2 \; d\mu(x) = \sum_{A' \in \mathbf{A}'} \int_{A'} \|f_{n,\mathbf{A}}(x) - f(x)\|^2 \; d\mu(x)$$

Now let $\mathbf{A}'_> := \{A' \in \mathbf{A}' : \mu_n(A') \geq (\ln(d) + \ln(8/\delta))/n\}$ and $\mathbf{A}'_< := \mathbf{A}' \setminus \mathbf{A}'_>$. Observe that (cf. Lemma 8.8) for every $A'_> \in \mathbf{A}'_>$, $\mu(A'_>) \leq 7\mu_n(A'_>)$, and for every $A'_< \in \mathbf{A}'_<$, $\mu(A'_<) \leq 7(\ln(d) + \ln(8/\delta))/n$. Thus, with probability at least $1 - \delta/2$

$$\begin{aligned} \sum_{A'_< \in \mathbf{A}'_<} \int_{A'_<} \|f_{n,\mathbf{A}}(x) - f(x)\|^2 \; d\mu(x) &\leq \sum_{A'_< \in \mathbf{A}'_<} \Delta^2(\mathcal{Y}) \mu(A'_<) \\ &\leq 14\Delta^2(\mathcal{Y})|\mathbf{A}| \frac{\ln(d) + \ln(8/\delta)}{n}. \quad (8.4) \end{aligned}$$

Also, with probability at least $1-\delta/2$

$$\begin{aligned}
&\sum_{A'_>\in\mathbf{A}'_>}\int_{A'_>}\|f_{n,\mathbf{A}}(x)-f(x)\|^2\ d\mu(x)=\sum_{A'_>\in\mathbf{A}'_>}\int_{A'_>}\|f_{n,\mathbf{A}'}(x)-f(x)\|^2\ d\mu(x)\\
&\leq \sum_{A'_>\in\mathbf{A}'_>}2\lambda^2\Delta^2(A'_>\cap\mathbf{X})\mu(A'_>)+\sum_{A'_>\in\mathbf{A}'_>}2\Delta^2(\mathcal{Y})\frac{2+\ln(4|\mathbf{A}|/\delta)}{|A'_>\cap\mathbf{X}|}\mu(A'_>)\\
&\leq 14\lambda^2\Big[\sum_{A_>\in\mathbf{A}_>}\frac{|A_>\cap\mathbf{X}|}{n}\Delta^2(A_>\cap\mathbf{X})\Big]+28\Delta^2(\mathcal{Y})|\mathbf{A}|\frac{2+\ln(4|\mathbf{A}|/\delta)}{n}. \quad (8.5)
\end{aligned}$$

The theorem follows by combining Equations (8.4) and (8.5).

**Definition 8.7 (shatter coefficient)** *Given a concept $C$ and a sample $\mathbf{X}:=(x_1,\ldots,x_n)$ from some underlying space $\mathcal{X}$, define the labeling function $\mathcal{L}(C,\mathbf{X})$ as how the concept $C$ labels the sample $\mathbf{X}$. That is,*

$$\mathcal{L}(C,\mathbf{X}):=(\mathbf{1}[x_1\in C],\ldots,\mathbf{1}[x_n\in C]).$$

*Now let $\mathcal{C}$ be any data dependent concept class. Then, for $m>n$, the* shatter coefficient *is defined as*

$$S(\mathcal{C},m,n):=\sup_{\mathbf{X}_m\subset\mathcal{X}}\left|\bigcup_{\mathbf{X}_n\subset\mathbf{X}_m,C\in\mathcal{C}_{\mathbf{X}_n}}\mathcal{L}(C,\mathbf{X}_n)\right|$$

*That is, it is the worst case number of distinct dichotomies of $m$ points generated as the concepts vary over the union of data dependent concept classes of all size $n$ subsets of the $m$ points.*

**Lemma 8.8 (data dependent relative uniform convergence)** *Suppose a sample $\mathbf{X}:=(x_1,\ldots,x_n)$ of size $n$ is drawn i.i.d. from a fixed probability measure $\mu$ over a measurable space $\mathcal{X}$, with resulting empirical measure $\mu_n$. Fix a concept class $\mathcal{C}_{\mathbf{X}}$ over $\mathcal{X}$ that may depend on the sample. Then for any $\delta>0$, with probability at least $1-\delta$ we have the following: for all $C\in\mathcal{C}_{\mathbf{X}}$*

$$\mu(C)\ \leq\ \mu_n(C)+2\sqrt{\mu_n(C)\frac{\ln(S(\mathcal{C}_{\mathbf{X}},2n,n))+\ln(4/\delta)}{n}}+4\frac{\ln(4S(\mathcal{C}_{\mathbf{X}},2n,n)/\delta)}{n}.$$

*Proof.* The proof follows from closely following the arguments presented in Section 5.1 of Boucheron et al. [2005] for the specified notion for shatter coefficient (Definition 8.7). ∎

**Lemma 8.9 (shatter coefficients of unions and intersections of concept classes)** *Let $C_1$ and $C_2$ be two concept class over a measurable space $\mathcal{X}$. Define $C_\cup := C_1 \cup C_2$ and $C_\cap := \{c_1 \cap c_2 : c_1 \in C_1, c_2 \in C_2\}$. Then for $m > n$, (i) $S(C_\cup, m, n) \leq S(C_1, m, n) + S(C_2, m, n)$, (ii) $S(C_\cap, m, n) \leq S(C_1, m, n) S(C_2, m, n)$.*

**Lemma 8.10 (Sauer's lemma [1972])** *Let $\mathcal{C}$ be any concept class with VC-dimension at most $\mathcal{V} < \infty$. Then, for $m > n$*

$$S(\mathcal{C}, m, n) \leq \left(\frac{me}{\mathcal{V}}\right)^{\mathcal{V}}.$$

# Chapter 9

# Conclusion

This dissertation focused on theoretical and practical issues in designing effective learning algorithms (both unsupervised and supervised) when the given high-dimensional data conforms to some low dimensional intrinsic structure. Our results on manifold structured data show that the complexity of various learning algorithms is intimately tied with the geometric properties (such as dimension, volume and curvature) of the underlying manifold.

We then explored a holistic way to formalize and test the *intrinsic dimension hypothesis* on modern datasets. The accompanying sampling complexity results on different learning algorithms highlight good performance guarantees when datasets have low intrinsic dimension.

These results provide a theoretical justification why we can expect learning algorithms—that suffer from the traditional *curse of dimensionality*—perform well on modern high-dimensional structured datasets.

# Appendix A

# Properties of a Well-Conditioned Manifold

Throughout this section we will assume that $M$ is a compact submanifold of $\mathbb{R}^D$ of dimension $n$, and condition number $1/\tau$. The following are some properties of such a manifold that would be useful throughout the text.

**Lemma A.1 (relating closeby tangent vectors – implicit in the proof of Proposition 6.2 Niyogi et al. [2008])** *Pick any two (path-connected) points $p, q \in M$. Let $u \in T_pM$ be a unit length tangent vector and $v \in T_qM$ be its parallel transport along the (shortest) geodesic path to $q$. Then*[1]*, i) $u \cdot v \geq 1 - D_G(p,q)/\tau$, ii) $\|u - v\| \leq \sqrt{2D_G(p,q)/\tau}$.*

**Lemma A.2 (relating geodesic distances to ambient distances – Proposition 6.3 of Niyogi et al. [2008])** *If $p, q \in M$ such that $\|p - q\| \leq \tau/2$, then $D_G(p,q) \leq \tau(1 - \sqrt{1 - 2\|p-q\|/\tau}) \leq 2\|p - q\|$.*

**Lemma A.3 (projection of a section of a manifold onto the tangent space)** *Pick any $p \in M$ and define $M_{p,r} := \{q \in M : \|q - p\| \leq r\}$. Let $f$*

[1]Technically, it is not possible to directly compare two vectors that reside in different tangent spaces. However, since we only deal with manifolds that are immersed in some ambient space, we can treat the tangent spaces as $n$-dimensional affine subspaces. We can thus parallel translate the vectors to the origin of the ambient space, and do the necessary comparison (such as take the dot product, etc.). We will make a similar abuse of notation for any calculation that uses vectors from different affine subspaces to mean to first translate the vectors and then perform the necessary calculation.

*denote the orthogonal linear projection of $M_{p,r}$ onto the tangent space $T_pM$. Then, for any $r \leq \tau/2$*

*(i) the map $f : M_{p,r} \to T_pM$ is one-to-one. (see Lemma 5.4 of Niyogi et al. [2008])*

*(ii) for any $x, y \in M_{p,r}$, $\|f(x) - f(y)\|^2 \geq (1 - (r/\tau)^2) \cdot \|x - y\|^2$. (implicit in the proof of Lemma 5.3 of Niyogi et al. [2008])*

**Lemma A.4 (coverings of a section of a manifold)** *Pick any $p \in M$ and define $M_{p,r} := \{q \in M : \|q - p\| \leq r\}$. If $r \leq \tau/2$, then there exists $C \subset M_{p,r}$ of size at most $9^n$ with the property: for any $p' \in M_{p,r}$, exists $c \in C$ such that $\|p' - c\| \leq r/2$.*

*Proof.* The proof closely follows the arguments presented in the proof of Theorem 22 of Dasgupta and Freund [2008].

For $r \leq \tau/2$, note that $M_{p,r} \subset \mathbb{R}^D$ is (path-)connected. Let $f$ denote the projection of $M_{p,r}$ onto $T_pM \cong \mathbb{R}^n$. Quickly note that $f$ is one-to-one (see Lemma A.3(i)). Then, $f(M_{p,r}) \subset \mathbb{R}^n$ is contained in an $n$-dimensional ball of radius $r$. By standard volume arguments, $f(M_{p,r})$ can be covered by at most $9^n$ balls of radius $r/4$. WLOG we can assume that the centers of these covering balls are in $f(M_{p,r})$. Now, noting that the inverse image of each of these covering balls (in $\mathbb{R}^n$) is contained in a $D$-dimensional ball of radius $r/2$ (see Lemma A.3(ii)) finishes the proof. ∎

**Lemma A.5 (relating closeby manifold points to tangent vectors)** *Pick any point $p \in M$ and let $q \in M$ (distinct from $p$) be such that $D_G(p, q) \leq \tau$. Let $v \in T_pM$ be the projection of the vector $q - p$ onto $T_pM$. Then, i) $\left| \frac{v}{\|v\|} \cdot \frac{q-p}{\|q-p\|} \right| \geq 1 - (D_G(p,q)/2\tau)^2$, ii) $\left\| \frac{v}{\|v\|} - \frac{q-p}{\|q-p\|} \right\| \leq D_G(p,q)/\tau\sqrt{2}$.*

*Proof.* If vectors $v$ and $q - p$ are in the same direction, we are done. Otherwise, consider the plane spanned by vectors $v$ and $q - p$. Then since $M$ has condition number $1/\tau$, we know that the point $q$ cannot lie within any $\tau$-ball tangent to $M$ at $p$ (see Figure A.1). Consider such a $\tau$-ball (with center $c$) whose center is closest

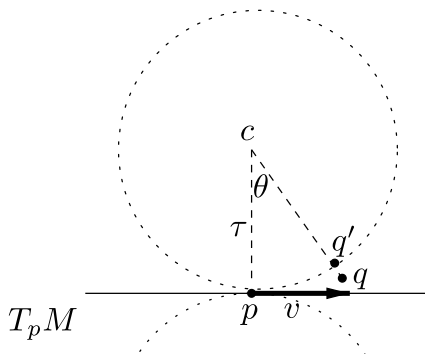



**Figure A.1**: Plane spanned by vectors $q-p$ and $v \in T_pM$ (where $v$ is the projection of $q-p$ onto $T_pM$), with $\tau$-balls tangent to $p$. Note that $q'$ is the point on the ball such that $\angle pcq = \angle pcq' = \theta$.

to $q$ and let $q'$ be the point on the surface of the ball which subtends the same angle ($\angle pcq'$) as the angle formed by $q$ ($\angle pcq$). Let this angle be called $\theta$. Then using cosine rule, we have $\cos\theta = 1 - \|q'-p\|^2/2\tau^2$.

Define $\alpha$ as the angle subtended by vectors $v$ and $q-p$, and $\alpha'$ the angle subtended by vectors $v$ and $q'-p$. WLOG we can assume that the angles $\alpha$ and $\alpha'$ are less than $\pi$. Then, $\cos\alpha \geq \cos\alpha' = \cos\theta/2$. Using the trig identity $\cos\theta = 2\cos^2\left(\frac{\theta}{2}\right) - 1$, and noting $\|q-p\|^2 \geq \|q'-p\|^2$, we have

$$\left|\frac{v}{\|v\|}\cdot\frac{q-p}{\|q-p\|}\right| = \cos\alpha \geq \cos\frac{\theta}{2} \geq \sqrt{1-\|q-p\|^2/4\tau^2} \geq 1-(D_G(p,q)/2\tau)^2.$$

Now, by applying the cosine rule, we have $\left\|\frac{v}{\|v\|} - \frac{q-p}{\|q-p\|}\right\|^2 = 2(1-\cos\alpha)$. The lemma follows. ∎

**Lemma A.6 (approximating tangent space by closeby samples)** *Let $0 < \delta \leq 1$. Pick any point $p_0 \in M$ and let $p_1, \ldots, p_n \in M$ be $n$ points distinct from $p_0$ such that (for all $1 \leq i \leq n$)*

*(i)* $D_G(p_0, p_i) \leq \tau\delta/\sqrt{n}$,

*(ii)* $\left|\frac{p_i-p_0}{\|p_i-p_0\|}\cdot\frac{p_j-p_0}{\|p_j-p_0\|}\right| \leq 1/2n$ *(for $i \neq j$).*

*Let $\hat{T}$ be the $n$ dimensional subspace spanned by vectors $\{p_i-p_0\}_{i\in[n]}$. For any unit vector $\hat{u} \in \hat{T}$, let $u$ be the projection of $\hat{u}$ onto $T_{p_0}M$. Then, $\left|\hat{u}\cdot\frac{u}{\|u\|}\right| \geq 1-\delta$.*

*Proof.* Define the vectors $\hat{v}_i := \frac{p_i - p_0}{\|p_i - p_0\|}$ (for $1 \le i \le n$). Observe that $\{\hat{v}_i\}_{i\in[n]}$ forms a basis of $\hat{T}$. For $1 \le i \le n$, define $v_i$ as the projection of vector $\hat{v}_i$ onto $T_{p_0}M$. Also note that by applying Lemma A.5, we have that for all $1 \le i \le n$, $\|\hat{v}_i - v_i\|^2 \le \delta^2/2n$.

Let $V = [\hat{v}_1, \ldots, \hat{v}_n]$ be the $D \times n$ matrix. We represent the unit vector $\hat{u}$ as $V\alpha = \sum_i \alpha_i \hat{v}_i$. Also, since $u$ is the projection of $\hat{u}$, we have $u = \sum_i \alpha_i v_i$. Then, $\|\alpha\|^2 \le 2$. To see this, we first identify $\hat{T}$ with $\mathbb{R}^n$ via an isometry $S$ (a linear map that preserves the lengths and angles of all vectors in $\hat{T}$). Note that $S$ can be represented as an $n \times D$ matrix, and since $V$ forms a basis for $\hat{T}$, $SV$ is an $n \times n$ invertible matrix. Then, since $S\hat{u} = SV\alpha$, we have $\alpha = (SV)^{-1}S\hat{u}$. Thus, (recall $\|S\hat{u}\| = 1$)

$$\begin{aligned}
\|\alpha\|^2 &\le \max_{x\in S^{n-1}} \|(SV)^{-1}x\|^2 = \lambda_{\max}((SV)^{-\mathsf{T}}(SV)^{-1}) \\
&= \lambda_{\max}((SV)^{-1}(SV)^{-\mathsf{T}}) = \lambda_{\max}((V^{\mathsf{T}}V)^{-1}) = 1/\lambda_{\min}(V^{\mathsf{T}}V) \\
&\le 1/1 - ((n-1)/2n) \le 2,
\end{aligned}$$

where i) $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ denote the largest and smallest eigenvalues of a square symmetric matrix $A$ respectively, and ii) the second inequality is by noting that $V^{\mathsf{T}}V$ is an $n \times n$ matrix with 1's on the diagonal and at most $1/2n$ on the off-diagonal elements, and applying the Gershgorin circle theorem.

Now we can bound the quantity of interest. Note that

$$\begin{aligned}
\left|\hat{u} \cdot \frac{u}{\|u\|}\right| &\ge |\hat{u}^{\mathsf{T}}(\hat{u} - (\hat{u} - u))| \ge 1 - \|\hat{u} - u\| = 1 - \Big\|\sum_i \alpha_i(\hat{v}_i - v_i)\Big\| \\
&\ge 1 - \sum_i |\alpha_i| \|\hat{v}_i - v_i\| \ge 1 - (\delta/\sqrt{2n}) \sum_i |\alpha_i| \ge 1 - \delta,
\end{aligned}$$

where the last inequality is by noting $\|\alpha\|_1 \le \sqrt{2n}$. ∎

**Lemma A.7 (manifold covers – see Section 2.4 of Clarkson [2008])** *Let $M \subset \mathbb{R}^N$ be a compact $n$-dimensional manifold with $\mathrm{VOL}(M) \le V$ and $\mathrm{COND}(M) \le 1/\tau$. Pick any $0 < \epsilon \le \tau/2$. There exists an $\epsilon$-covering of $M$ of size at most $2^{c_0 n}(V/\epsilon^n)$, where $c_0$ is an absolute constant. That is, there exists $C \subset M$ such that $|C| \le 2^{c_0 n}(V/\epsilon^n)$ with the property: for all $p \in M$, $\exists\ q \in C$ such that $\|p - q\| \le \epsilon$.*

**Lemma A.8 (manifold volumes – see Lemma 5.3 Niyogi et al. [2008])** *Let $M \subset \mathbb{R}^N$ be a compact $n$-dimensional manifold with $\mathrm{COND}(M) \leq 1/\tau$. Pick any $p \in M$ and let $A_\epsilon := M \cap B(p, \epsilon)$, where $B(p, \epsilon)$ is a Euclidean ball in $\mathbb{R}^N$ centered at $p$ of radius $\epsilon$. If $A_\epsilon$ does not contain any boundary points of $M$, then $\mathrm{VOL}(A_\epsilon) \geq (\cos(\arcsin(\epsilon/2\tau)))^n \mathrm{VOL}(B^n_\epsilon)$, where $B^n_\epsilon$ is a Euclidean ball in $\mathbb{R}^n$ of radius $\epsilon$. In particular, noting that $\mathrm{VOL}(B^n_\epsilon) \geq \epsilon^{c_0 n}$ for some absolute constant $c_0$, if $\epsilon \leq \tau$, we have $\mathrm{VOL}(A_\epsilon) \geq \epsilon^{c_0 n}$.*

# Bibliography


S. Ali and M. Shah. Human action recognition in videos using kinematic features and multiple instance learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 32(2):288–303, 2008.

N. Amenta and M. Bern. Surface reconstruction by Voronoi filtering. *Symposium on Computational Geometry (SoCG)*, pages 39–48, 1998.

S. Andrews, T. Hofmann, and I. Tsochantaridis. Multiple instance learning with generalized support vector machines. *Association for the Advancement of Artificial Intelligence (AAAI)*, pages 943–944, 2002.

M. Anthony and P.L. Bartlett. *Neural network learning: Theoretical foundations*. Cambridge University Press, 1999.

P. Auer, P. M. Long, and A. Srinivasan. Approximating hyper-rectangles: Learning and pseudo-random sets. *ACM Symposium on Theory of Computing (STOC)*, pages 314–323, 1997.

B. Babenko, N. Verma, P. Dollár, and S. Belongie. Multiple instance learning with manifold bags. *International Conference on Machine Learning (ICML)*, pages 81–88, 2011.

R. Baraniuk and M. Wakin. Random projections of smooth manifolds. *Foundations of Computational Mathematics (FoCM)*, 9(1):51–77, 2009.

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

P. Bickel and B. Li. Local polynomial regression on unknown manifolds. Technical report, UC Berkeley, 2006.

A. Blum and A. Kalai. A note on learning from multiple-instance examples. *Machine Learning*, 30(1):23–29, 1998.

S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification: a survey of some recent advances. *ESAIM: Probability and Statistics*, 9:323–375, 2005.

P. Buehler, A. Zisserman, and M. Everingham. Learning sign language by watching TV (using weakly aligned subtitles). *Computer Vision and Pattern Recognition (CVPR)*, pages 2961–2968, 2009.

K. Clarkson. Nearest-neighbor searching and metric space dimensions. *Nearest-Neighbor Methods for Learning and Vision: Theory and Practice*, pages 15–59, 2006.

K. Clarkson. Tighter bounds for random projections of manifolds. *Symposium on Computational Geometry (SoCG)*, 24:39–48, 2008.

D. Cohn, L. Atlas, and R. Ladner. Improving generalization with active learning. *Machine Learning*, pages 201–221, 1994.

C. Cutler. A review of the theory and estimation of fractal dimension. *Nonlinear Time Series and Chaos, Vol. I: Dimension Estimation and Models*, 1993.

N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. *Computer Vision and Pattern Recognition (CVPR)*, 1:886–893, 2005.

S. Dasgupta and Y. Freund. Random projection trees and low dimensional manifolds. *ACM Symposium on Theory of Computing (STOC)*, pages 537–546, 2008.

S. Dasgupta and A. Gupta. An elementary proof of the Johnson-Lindenstrauss lemma. Technical Report 99-006, UC Berkeley, March 1999.

S. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing (TASSP)*, 88(4):357–366, 1980.

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

T. Dey, J. Giesen, S. Goswami, and W. Zhao. Shape dimension and approximation from samples. *Symposium on Discrete Algorithms (SODA)*, pages 772–780, 2002.

T.G. Dietterich, R.H. Lathrop, and T. Lozano-Perez. Solving the multiple-instance problem with axis parallel rectangles. *Artificial Intelligence*, 89(1-2):31–71, 1997.

M. do Carmo. *Riemannian Geometry*. Birkhäuser, 1992.

P. Dollár, Z. Tu, P. Perona, and S. Belongie. Integral channel features. *British Machine Vision Conference (BMVC)*, 2009.

D.P.W. Ellis. PLP and RASTA (and MFCC, and inversion) in Matlab. http://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat/, 2005.

P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 32(9):1627–1645, 2009.

Y. Freund, S. Dasgupta, M. Kabra, and N. Verma. Learning the structure of manifolds using random projections. *Neural Information Processing Systems (NIPS)*, pages 473–480, 2007.

J.S. Garofolo et al. TIMIT Acoustic-Phonetic Continuous Speech Corpus. http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC93S1, 1993.

J. Giesen and U. Wagner. Shape dimension and intrinsic metric from samples of manifolds with high co-dimension. *Symposium on Computational Geometry (SoCG)*, pages 329–337, 2003.

L. Györfi, M. Kohler, A. Krzyzak, and H. Walk. *A distribution-free theory of nonparametric regression*. Springer, 2002.

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

W. Johnson and J. Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. *Conference in Modern Analysis and Probability*, 26:189–206, 1984.

M.W. Kadous. *Temporal classification: Extending the classification paradigm to multivariate time series*. PhD thesis, School of Computer Science and Engineering, University of New South Wales, 2002.

S. Kpotufe and S. Dasgupta. A tree-based regressor that adapts to intrinsic dimension. *Journal of Computer and System Sciences (JCSS)*, 78(5):1496–1515, 2012.

N. Kuiper. On $C^1$-isometric embeddings, I, II. *Indagationes Mathematicae*, 17: 545–556, 683–689, 1955.

P.M. Long and L. Tan. PAC learning axis-aligned rectangles with respect to product distributions from multiple-instance examples. *Machine Learning*, 30(1): 7–21, 1998.

M.I. Mandel and D.P.W. Ellis. Multiple-instance learning for music information retrieval. *International Society for Music Information Retrieval (ISMIR)*, pages 577–582, 2008.

O. Maron and T. Lozano-Perez. A framework for multiple-instance learning. *Neural Information Processing Systems (NIPS)*, pages 570–576, 1998.

C. McDiarmid. On the method of bounded differences. In *Surveys in Combinatorics*, pages 148–188. Cambridge University Press, 1989.

J. Milnor. *Topology from the differential viewpoint*. University of Virginia Press, 1972.

J. Nash. $C^1$ isometric imbeddings. *Annals of Mathematics*, 60(3):383–396, 1954.

P. Niyogi, S. Smale, and S. Weinberger. Finding the homology of submanifolds with high confidence from random samples. *Discrete Computational Geometry*, 39(1):419–441, 2008.

S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

S. Sabato and N. Tishby. Homogeneous multi-instance learning with arbitrary dependence. *Conference on Computational Learning Theory (COLT)*, pages 93–104, 2009.

S. Sabato, N. Srebro, and N. Tishby. Reducing label complexity by learning from bags. *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 9:685–692, 2010.

N. Sauer. On the density of families of sets. *Journal of Combinatorial Theory*, 13: 41–45, 1972.

L.K. Saul, M.G. Rahim, and J.B. Allen. A statistical model for robust integration of narrowband cues in speech. *Computer Speech and Language*, 15(2):175–194, 2001.

C. Scott and R.D. Nowark. Minimax-optimal classification with dyadic decision trees. *IEEE Transactions on Information Theory*, 52:1225–1232, 2006.

M. Stikic and B. Schiele. Activity recognition from sparsely labeled data using multi-instance learning. *Location and Context Awareness*, 5561:156–173, 2009.

J. Stoker. *Differential Geometry*. Wiley-Interscience, 1969.

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

V.N. Vapnik and A.Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Its Applications*, 16(2):264–280, 1971.

N. Verma. A note on random projections for preserving paths on a manifold. Technical Report CS2011-0971, UC San Diego, 2011.

N. Verma. Distance preserving embeddings for general $n$-dimensional manifolds. *Conference on Computational Learning Theory (COLT)*, 23:32.1–32.28, 2012.

N. Verma, S. Kpotufe, and S. Dasgupta. Which spatial partition trees are adaptive to intrinsic dimension? *Uncertainty in Artificial Intelligence (UAI)*, pages 565–574, 2009.

R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing, Theory and Applications*. 2010.

P. Viola, J.C. Platt, and C. Zhang. Multiple instance boosting for object detection. *Neural Information Processing Systems (NIPS)*, 18:1417–1426, 2005.

H. Whitney. Differentiable manifolds. *Annals of Mathematics*, 37:645–680, 1936.

Q. Zhang and S.A. Goldman. EM-DD: An improved multiple-instance learning technique. *Neural Information Processing Systems (NIPS)*, 14:1073–1080, 2002.