Sample complexity of testing the manifold hypothesis

Hariharan Narayanan
Laboratory for Information and Decision Systems
EECS, MIT
Cambridge, MA 02139
har@mit.edu

Sanjoy Mitter
Laboratory for Information and Decision Systems
EECS, MIT
Cambridge, MA 02139
mitter@mit.edu

Abstract

The hypothesis that high dimensional data tends to lie in the vicinity of a low dimensional manifold is the basis of a collection of methodologies termed Manifold Learning. In this paper, we study statistical aspects of the question of fitting a manifold with a nearly optimal least squared error. Given upper bounds on the dimension, volume, and curvature, we show that Empirical Risk Minimization can produce a nearly optimal manifold using a number of random samples that is *independent* of the ambient dimension of the space in which data lie. We obtain an upper bound on the required number of samples that depends polynomially on the curvature, exponentially on the intrinsic dimension, and linearly on the intrinsic volume. For constant error, we prove a matching minimax lower bound on the sample complexity that shows that this dependence on intrinsic dimension, volume and curvature is unavoidable. Whether the known lower bound of $O(\frac{k}{\epsilon^2} + \frac{\log\frac{1}{\delta}}{\epsilon^2})$ for the sample complexity of Empirical Risk minimization on $k-$means applied to data in a unit ball of arbitrary dimension is tight, has been an open question since 1997 [3]. Here ϵ is the desired bound on the error and δ is a bound on the probability of failure. We improve the best currently known upper bound [16] of $O(\frac{k^2}{\epsilon^2} + \frac{\log\frac{1}{\delta}}{\epsilon^2})$ to $O\left(\frac{k}{\epsilon^2}\left(\min\left(k, \frac{\log^4\frac{k}{\epsilon}}{\epsilon^2}\right)\right) + \frac{\log\frac{1}{\delta}}{\epsilon^2}\right)$. Based on these results, we devise a simple algorithm for $k-$means and another that uses a family of convex programs to fit a piecewise linear curve of a specified length to high dimensional data, where the sample complexity is independent of the ambient dimension.

1 Introduction

We are increasingly confronted with very high dimensional data in speech signals, images, gene-expression data, and other sources. Manifold Learning can be loosely defined to be a collection of methodologies that are motivated by the belief that this hypothesis (henceforth called the manifold hypothesis) is true. It includes a number of extensively used algorithms such as Locally Linear Embedding [19], ISOMAP [21], Laplacian Eigenmaps [4] and Hessian Eigenmaps [10]. The sample complexity of classification is known to be independent of the ambient dimension [17] under the manifold hypothesis, (assuming the decision boundary is a manifold as well,) thus obviating the curse of dimensionality. A recent empirical study [6] of a large number of 3×3 images, represented as points in $\mathbb{R}^9$ revealed that they approximately lie on a two dimensional manifold known as the Klein bottle. On the other hand, knowledge that the manifold hypothesis is false with regard to

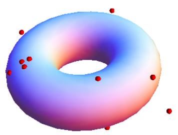

Figure 1: Fitting a torus to data.

certain data would give us reason to exercise caution in applying algorithms from manifold learning and would provide an incentive for further study.

It is thus of considerable interest to know whether given data lie in the vicinity of a low dimensional manifold. Our primary technical results are the following.

1. We obtain uniform bounds relating the empirical squared loss and the true squared loss over a class $\mathcal{F}$ consisting of manifolds whose dimensions, volumes and curvatures are bounded in Theorems 1 and 2. These bounds imply upper bounds on the sample complexity of Empirical Risk Minimization (ERM) that are *independent* of the ambient dimension, exponential in the intrinsic dimension, polynomial in the curvature and almost linear in the volume.
2. We obtain a minimax lower bound on the sample complexity of *any* rule for learning a manifold from $\mathcal{F}$ in Theorem 6 showing that for a fixed error, the the dependence of the sample complexity on intrinsic dimension, curvature and volume must be at least exponential, polynomial, and linear, respectively.
3. We improve the best currently known upper bound [16] on the sample complexity of Empirical Risk minimization on $k-$means applied to data in a unit ball of arbitrary dimension from $O(\frac{k^2}{\epsilon^2} + \frac{\log \frac{1}{\delta}}{\epsilon^2})$ to $O\left(\frac{k}{\epsilon^2}\left(\min\left(k, \frac{\log^4 \frac{k}{\epsilon}}{\epsilon^2}\right)\right) + \frac{\log \frac{1}{\delta}}{\epsilon^2}\right)$. Whether the known lower bound of $O(\frac{k}{\epsilon^2} + \frac{\log \frac{1}{\delta}}{\epsilon^2})$ is tight, has been an open question since 1997 [3]. Here ϵ is the desired bound on the error and δ is a bound on the probability of failure.

One technical contribution of this paper is the use of dimensionality reduction via random projections in the proof of Theorem 5 to bound the Fat-Shattering dimension of a function class, elements of which roughly correspond to the squared distance to a low dimensional manifold. The application of the probabilistic method involves a projection onto a low dimensional random subspace. This is then followed by arguments of a combinatorial nature involving the VC dimension of halfspaces, and the Sauer-Shelah Lemma applied with respect to the low dimensional subspace. While random projections have frequently been used in machine learning algorithms, for example in [2, 7], to our knowledge, they have not been used as a tool to bound the complexity of a function class. We illustrate the algorithmic utility of our uniform bound by devising an algorithm for $k-$means and a convex programming algorithm for fitting a piecewise linear curve of bounded length. For a fixed error threshold and length, the dependence on the ambient dimension is linear, which is optimal since this is the complexity of reading the input.

2 Connections and Related work

In the context of curves, [12] proposed "Principal Curves", where it was suggested that a natural curve that may be fit to a probability distribution is one where every point on the curve is the center of mass of all those points to which it is the nearest point. A different definition of a principal curve was proposed by [14], where they attempted to find piecewise linear curves of bounded length which minimize the expected squared distance to a random point from a distribution. This paper studies the decay of the error rate as the number of samples tends to infinity, but does not analyze the dependence of the error rate on the ambient dimension and the bound on the length. We address this in a more general setup in Theorem 4, and obtain sample complexity bounds that are independent of the ambient dimension, and depend linearly on the bound on the length. There is a significant amount of recent research aimed at understanding topological aspects of data, such its homology [22, 18]. It has been an open question since 1997 [3], whether the known lower bound of $O(\frac{k}{\epsilon^2} + \frac{\log\frac{1}{\delta}}{\epsilon^2})$ for the sample complexity of Empirical Risk minimization on $k-$means applied to data in a unit ball of arbitrary dimension is tight. Here ϵ is the desired bound on the error and δ is a bound on the probability of failure. The best currently known upper bound is $O(\frac{k^2}{\epsilon^2} + \frac{\log\frac{1}{\delta}}{\epsilon^2})$ and is based on Rademacher complexities. We improve this bound to $O\left(\frac{k}{\epsilon^2}\left(\min\left(k, \frac{\log^4\frac{k}{\epsilon}}{\epsilon^2}\right)\right) + \frac{\log\frac{1}{\delta}}{\epsilon^2}\right)$, using an argument that bounds the Fat-Shattering dimension of the appropriate function class using random projections and the Sauer-Shelah Lemma. Generalizations of principal curves to parameterized principal manifolds in certain regularized settings have been studied in [20]. There, the sample complexity was related to the decay of eigenvalues of a Mercer kernel associated with the regularizer. When the manifold to be fit is a set of k points ($k-$means), we obtain a bound on the sample complexity s that is independent of m and depends at most linearly on k, which also leads to an approximation algorithm with additive error, based on sub-sampling. If one allows a multiplicative error of 4 in addition to an additive error of ϵ, a statement of this nature has been proven by Ben-David (Theorem 7, [5]).

3 Upper bounds on the sample complexity of Empirical Risk Minimization

In the remainder of the paper, C will always denote a universal constant which may differ across the paper. For any submanifold $\mathcal{M}$ contained in, and probability distribution $\mathcal{P}$ supported on the unit ball B in $\mathbb{R}^m$, let $\mathcal{L}(\mathcal{M}, \mathcal{P}) := \int \mathbf{d}(\mathcal{M}, x)^2 d\mathcal{P}(x)$. Given a set of i.i.d points $x = \{x_1, \dots, x_s\}$ from $\mathcal{P}$, a tolerance ϵ and a class of manifolds $\mathcal{F}$, Empirical Risk Minimization (ERM) outputs a manifold in $\mathcal{M}_{erm}(x) \in \mathcal{F}$ such that $\sum_{i=1}^{s} \mathbf{d}(x_i, \mathcal{M}_{erm})^2 \leq \epsilon/2 + \inf_{\mathcal{N}\in\mathcal{F}} \mathbf{d}(x_i, \mathcal{N})^2$. Given error parameters ϵ, δ, and a rule $\mathcal{A}$ that outputs a manifold in $\mathcal{F}$ when provided with a set of samples, we define the sample complexity $s = s(\epsilon, \delta, \mathcal{A})$ to be the least number such that for any probability distribution $\mathcal{P}$ in the unit ball, if the result of $\mathcal{A}$ applied to a set of at least s i.i.d random samples from $\mathcal{P}$ is $\mathcal{N}$, then $\mathbb{P}\left[\mathcal{L}(\mathcal{N}, \mathcal{P}) < \inf_{\mathcal{M}\in\mathcal{F}} \mathcal{L}(\mathcal{M}, \mathcal{P}) + \epsilon\right] > 1 - \delta$.

3.1 Bounded intrinsic curvature

Let $\mathcal{M}$ be a Riemannian manifold and let $p \in \mathcal{M}$. Let ζ be a geodesic starting at p.

Definition 1. *The first point on ζ where ζ ceases to minimize distance is called the cut point of p along $\mathcal{M}$. The cut locus of p is the set of cut points of $\mathcal{M}$. The injectivity radius is the minimum taken over all points of the distance between the point and its cut locus. $\mathcal{M}$ is complete if it is complete as a metric space.*

Let $\mathcal{G}_i = \mathcal{G}_i(d, V, \lambda, \iota)$ be the family of all isometrically embedded, complete Riemannian submanifolds of B having dimension less or equal to d, induced $d-$dimensional volume less or equal to V, sectional curvature less or equal to λ and injectivity radius greater or equal to ι. Let $U_{int}(\frac{1}{\epsilon}, d, V, \lambda, \iota) := V\left(C\left(\frac{d}{\min(\epsilon, \iota, \lambda^{-1/2})}\right)\right)^d$, which for brevity, we denote U_{int}.

Theorem 1. *Let ϵ and δ be error parameters. If*

$$s \geq C\left(\min\left(\left(\frac{1}{\epsilon^2}\right)\log^4\left(\frac{U_{int}}{\epsilon}\right), U_{int}\right)\frac{U_{int}}{\epsilon^2} + \frac{1}{\epsilon^2}\log\frac{1}{\delta}\right),$$

and $x = \{x_1, \dots, x_s\}$ is a set of i.i.d points from $\mathcal{P}$ then,

$$\mathbb{P}\left[\mathcal{L}(\mathcal{M}_{erm}(x), \mathcal{P}) - \inf_{\mathcal{M}\in\mathcal{G}_i} \mathcal{L}(\mathcal{M}, \mathcal{P}) < \epsilon\right] > 1 - \delta.$$

The proof of this theorem is deferred to Section 4.

3.2 Bounded extrinsic curvature

We will consider submanifolds of B that have the property that around each of them, for any radius $r < \tau$, the boundary of the set of all points within a distance r is smooth. This class of submanifolds has appeared in the context of manifold learning [18, 17]. The condition number is defined as follows.

Definition 2 (Condition Number). *Let $\mathcal{M}$ be a smooth $d-$dimensional submanifold of $\mathbb{R}^m$. We define the condition number $c(\mathcal{M})$ to be $\frac{1}{\tau}$, where τ is the largest number to have the property that for any $r < \tau$ no two normals of length r that are incident on $\mathcal{M}$ have a common point unless it is on $\mathcal{M}$.*

Let $\mathcal{G}_e = \mathcal{G}_e(d, V, \tau)$ be the family of Riemannian submanifolds of B having dimension $\leq d$, volume $\leq V$ and condition number $\leq \frac{1}{\tau}$. Let ϵ and δ be error parameters. Let $U_{ext}(\frac{1}{\epsilon}, d, \tau) := V\left(C\left(\frac{d}{\min(\epsilon,\tau)}\right)\right)^d$, which for brevity, we denote by U_{ext}.

Theorem 2. *If*

$$s \geq C\left(\min\left(\left(\frac{1}{\epsilon^2}\right)\log^4\left(\frac{U_{ext}}{\epsilon}\right), U_{ext}\right)\frac{U_{ext}}{\epsilon^2} + \frac{1}{\epsilon^2}\log\frac{1}{\delta}\right),$$

and $x = \{x_1, \dots, x_s\}$ is a set of i.i.d points from $\mathcal{P}$ then,

$$\mathbb{P}\left[\mathcal{L}(\mathcal{M}_{erm}(x), \mathcal{P}) - \inf_{\mathcal{M}} \mathcal{L}(\mathcal{M}, \mathcal{P}) < \epsilon\right] > 1 - \delta. \tag{1}$$

4 Relating bounded curvature to covering number

In this subsection, we note that that bounds on the dimension, volume, sectional curvature and injectivity radius suffice to ensure that they can be covered by relatively few $\epsilon-$balls. Let $V_p^{\mathcal{M}}$ be the volume of a ball of radius $\mathcal{M}$ centered around a point p. See ([11], page 51) for a proof of the following theorem.

Theorem 3 (Bishop-Günther Inequality). *Let $\mathcal{M}$ be a complete Riemannian manifold and assume that r is not greater than the injectivity radius ι. Let $K^{\mathcal{M}}$ denote the sectional curvature of $\mathcal{M}$ and let $\lambda > 0$ be a constant. Then, $K^{\mathcal{M}} \leq \lambda$ implies $V_p^{\mathcal{M}}(r) \geq \frac{2\pi^{\frac{n}{2}}}{\Gamma(\frac{n}{2})}\int_0^r \left(\frac{\sin(t\sqrt{\lambda})}{\sqrt{\lambda}}\right)^{n-1} dt$.*

Thus, if $\epsilon < \min(\iota, \frac{\pi\lambda^{-\frac{1}{2}}}{2})$, then, $V_p^{\mathcal{M}}(\epsilon) > \left(\frac{\epsilon}{Cd}\right)^d$.

Proof of Theorem 1. As a consequence of Theorem 3, we obtain an upper bound of $V\left(\frac{Cd}{\epsilon}\right)^d$ on the number of disjoint sets of the form $\mathcal{M} \cap B_{\epsilon/32}(p)$ that can be packed in $\mathcal{M}$. If $\{\mathcal{M} \cap B_{\epsilon/32}(p_1), \dots, \mathcal{M} \cap B_{\epsilon/32}(p_k)\}$ is a maximal family of disjoint sets of the form $\mathcal{M} \cap B_{\epsilon/32}(p)$, then there is no point $p \in \mathcal{M}$ such that $\min_i \|p - p_i\| > \epsilon/16$. Therefore, $\mathcal{M}$ is contained in the union of balls, $\bigcup_i B_{\epsilon/16}(p_i)$. Therefore, we may apply Theorem 4 with $U\left(\frac{1}{\epsilon}\right) \leq V\left(\frac{Cd}{\min(\epsilon, \lambda^{-\frac{1}{2}}, \iota)}\right)^d$. □

The proof of Theorem 2 is along the lines of that of Theorem 1, so it has been delegated to the appendix.

5 Class of manifolds with a bounded covering number

In this section, we show that uniform bounds relating the empirical squares loss and the expected squared loss can be obtained for a class of manifolds whose covering number at a different scale ϵ has a specified upper bound. Let $U : \mathbb{R}^+ \to \mathbb{Z}^+$ be any integer valued function. Let $\mathcal{G}$ be any family of subsets of B such that for all $r > 0$ every element $\mathcal{M} \in \mathcal{G}$ can be covered using open Euclidean balls of radius r centered around $U(\frac{1}{r})$ points; let this set be $\Lambda_{\mathcal{M}}(r)$. Note that if the subsets consist of $k-$tuples of points, $U(1/r)$ can be taken to be the constant function equal to k and we recover the $k-$means question. A priori, it is unclear if

$$\sup_{\mathcal{M} \in \mathcal{G}}\left|\frac{\sum_{i=1}^{s} \mathbf{d}(x_i, \mathcal{M})^2}{s} - \mathbb{E}_{\mathcal{P}}\mathbf{d}(x, \mathcal{M})^2\right|, \tag{2}$$

is a random variable, since the supremum of a set of random variables is not always a random variable (although if the set is countable this is true). However (2) is equal to

$$\lim_{n\to\infty} \sup_{\mathcal{M}\in\mathcal{G}} \left| \frac{\sum_{i=1}^{s} \mathbf{d}(x_i, \Lambda_{\mathcal{M}}(1/n))^2}{s} - \mathbb{E}_{\mathcal{P}}\mathbf{d}(x, \Lambda_{\mathcal{M}}(1/n))^2 \right|, \tag{3}$$

and for each n, the supremum in the limits is over a set parameterized by $U(n)$ points, which without loss of generality we may take to be countable (due to the density and countability of rational points). Thus, for a fixed n, the quantity in the limits is a random variable. Since the limit as $n \to \infty$ of a sequence of bounded random variables is a random variable as well, (2) is a random variable too.

Theorem 4. *Let ϵ and δ be error parameters. If*

$$s \geq C\left(\frac{U(16/\epsilon)}{\epsilon^2} \min\left(U(16/\epsilon), \left(\frac{1}{\epsilon^2}\right) \log^4\left(\frac{U(16/\epsilon)}{\epsilon}\right)\right) + \frac{1}{\epsilon^2}\log\frac{1}{\delta}\right),$$

Then,

$$\mathbb{P}\left[\sup_{\mathcal{M}\in\mathcal{G}} \left| \frac{\sum_{i=1}^{s} \mathbf{d}(x_i, \mathcal{M})^2}{s} - \mathbb{E}_{\mathcal{P}}\mathbf{d}(x, \mathcal{M})^2 \right| < \frac{\epsilon}{2}\right] > 1 - \delta. \tag{4}$$

Proof. For every $g \in \mathcal{G}$, let $\mathbf{c}(g, \epsilon) = \{c_1, \dots, c_k\}$ be a set of $k := U(16/\epsilon)$ points in $g \subseteq B$, such that g is covered by the union of balls of radius $\epsilon/16$ centered at these points. Thus, for any point $x \in B$,

$$\begin{aligned} \mathbf{d}^2(x, g) &\leq \left(\frac{\epsilon}{16} + \mathbf{d}(x, \mathbf{c}(g, \epsilon))\right)^2 & (5)\\ &\leq \frac{\epsilon^2}{256} + \frac{\epsilon \min_i \|x - c_i\|}{8} + \mathbf{d}(x, \mathbf{c}(g, \epsilon))^2. & (6) \end{aligned}$$

Since $\min_i \|x - c_i\|$ is less or equal to 2, the last expression is less than $\frac{\epsilon}{2} + \mathbf{d}(x, \mathbf{c}(g, \epsilon))^2$. Our proof uses the "kernel trick" in conjunction with Theorem 5. Let $\Phi : (x_1, \dots, x_m)^T \mapsto 2^{-1/2}(x_1, \dots, x_m, 1)^T$ map a point $x \in \mathbb{R}^m$ to one in $\mathbb{R}^{m+1}$. For each i, let $c_i := (c_{i1}, \dots, c_{im})^T$, and $\tilde{c}_i := 2^{-1/2}(-c_{i1}, \dots, -c_{im}, \frac{\|c_i\|^2}{2})^T$. The factor of $2^{-1/2}$ is necessitated by the fact that we wish the image of a point in the unit ball to also belong to the unit ball. Given a collection of points $\mathbf{c} := \{c_1, \dots, c_k\}$ and a point $x \in B$, let $f_{\mathbf{c}}(x) := \mathbf{d}(x, \mathbf{c}(g, \epsilon))^2$. Then,

$$f_{\mathbf{c}}(x) = \|x\|^2 + 4\min(\Phi(x)\cdot\tilde{c}_1, \dots, \Phi(x)\cdot\tilde{c}_k).$$

For any set of s samples $x_1, \dots, x_s$,

$$\begin{aligned} \sup_{f_{\mathbf{c}}\in\mathcal{G}} \left| \frac{\sum_{i=1}^{s} f_{\mathbf{c}}(x_i)}{s} - \mathbb{E}_{\mathcal{P}} f_{\mathbf{c}}(x) \right| &\leq \left| \frac{\sum_{i=1}^{s} \|x_i\|^2}{s} - \mathbb{E}_{\mathcal{P}}\|x\|^2 \right| & (7)\\ &+ 4\sup_{f_{\mathbf{c}}\in\mathcal{G}} \left| \frac{\sum_{i=1}^{s} \min_i \Phi(x_i)\cdot\tilde{c}_i}{s} - \mathbb{E}_{\mathcal{P}} \min_i \Phi(x)\cdot\tilde{c}_i \right|. & (8) \end{aligned}$$

By Hoeffding's inequality,

$$\mathbb{P}\left[\left| \frac{\sum_{i=1}^{s} \|x_i\|^2}{s} - \mathbb{E}_{\mathcal{P}}\|x\|^2 \right| > \frac{\epsilon}{4}\right] < 2e^{-(\frac{1}{8})s\epsilon^2}, \tag{9}$$

which is less than $\frac{\delta}{2}$.

By Theorem 5, $\mathbb{P}\left[\sup_{f_{\mathbf{c}}\in\mathcal{G}} \left| \frac{\sum_{i=1}^{s} \min_i \Phi(x_i)\cdot\tilde{c}_i}{s} - \mathbb{E}_{\mathcal{P}} \min_i \Phi(x)\cdot\tilde{c}_i \right| > \frac{\epsilon}{16}\right] < \frac{\delta}{2}$.

Therefore, $\mathbb{P}\left[\sup_{f_{\mathbf{c}}\in\mathcal{G}} \left| \frac{\sum_{i=1}^{s} f_{\mathbf{c}}(x_i)}{s} - \mathbb{E}_{\mathcal{P}} f_{\mathbf{c}}(x) \right| \leq \frac{\epsilon}{2}\right] \geq 1 - \delta.$

□

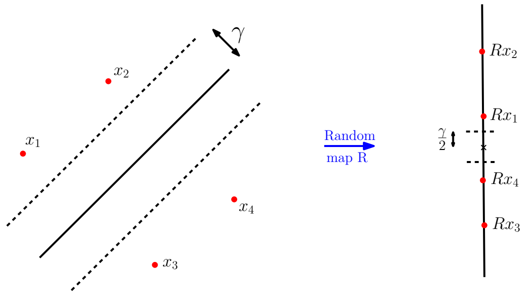

Figure 2: Random projections are likely to preserve linear separations.

6 Bounding the Fat-Shattering dimension using random projections

The core of the uniform bounds in Theorems 1 and 2 is the following uniform bound on the minimum of k linear functions on a ball in $\mathbb{R}^m$.

Theorem 5. *Let $\mathcal{F}$ be the set of all functions f from $B := \{x \in \mathbb{R}^m : \|x\| \leq 1\}$ to $\mathbb{R}$, such that for some k vectors $v_1, \dots, v_k \in B$,*

$$f(x) := \min_i (v_i \cdot x).$$

Independent of m, if

$$s \geq C\left(\frac{k}{\epsilon^2}\min\left(\frac{1}{\epsilon^2}\log^4\left(\frac{k}{\epsilon}\right), k\right) + \frac{1}{\epsilon^2}\log\frac{1}{\delta}\right),$$

then

$$\mathbb{P}\left[\sup_{F\in\mathcal{F}}\left|\frac{\sum_{i=1}^s F(x_i)}{s} - \mathbb{E}_{\mathcal{P}}F(x)\right| < \epsilon\right] > 1-\delta. \tag{10}$$

It has been open since 1997 [3], whether the known lower bound of $C\left(\frac{k}{\epsilon^2} + \frac{1}{\epsilon^2}\log\frac{1}{\delta}\right)$ on the sample complexity s is tight. Theorem 5 in [16], uses Rademacher complexities to obtain an upper bound of

$$C\left(\frac{k^2}{\epsilon^2} + \frac{1}{\epsilon^2}\log\frac{1}{\delta}\right). \tag{11}$$

(The scenarios in [3, 16] are that of $k-$means, but the argument in Theorem 4 reduces $k-$means to our setting.) Theorem 5 improves this to

$$C\left(\frac{k}{\epsilon^2}\min\left(\frac{1}{\epsilon^2}\log^4\left(\frac{k}{\epsilon}\right), k\right) + \frac{1}{\epsilon^2}\log\frac{1}{\delta}\right) \tag{12}$$

by putting together (11) with a bound of

$$C\left(\frac{k}{\epsilon^4}\log^4\left(\frac{k}{\epsilon}\right) + \frac{1}{\epsilon^2}\log\frac{1}{\delta}\right) \tag{13}$$

obtained using the Fat-Shattering dimension (Definition 3 in the appendix). Due to constraints on space, the details of the proof of Theorem 5 appear in the appendix, but the essential ideas are summarized here.

Let $u := \text{fat}_{\mathcal{F}}(\frac{\epsilon}{24})$ and $x_1, \dots, x_u$ be a set of vectors that is $\gamma-$shattered by $\mathcal{F}$. We would like to use VC theory to bound u, but doing so directly leads to a linear dependence on the ambient dimension m. In order to circumvent this difficulty, for $g := C\frac{\log(u+k)}{\epsilon^2}$, we consider a $g-$dimensional random linear subspace and the image under an appropriately scaled orthogonal projection R of the points $x_1, \dots, x_u$ onto it. We show that the expected value of the $\frac{\gamma}{2}-$shatter coefficient of $\{Rx_1, \dots, Rx_u\}$ is at least 2^{u-1} using the Johnson-Lindenstrauss Lemma [13] and the fact that $\{x_1, \dots, x_u\}$ is $\gamma-$shattered. Using Vapnik-Chervonenkis theory and the Sauer-Shelah Lemma, we then show that $\frac{\gamma}{2}-$shatter coefficient cannot be more than $u^{k(g+2)}$. This implies that $2^{u-1} \leq u^{k(g+2)}$, allowing us to conclude that $\text{fat}_{\mathcal{F}}(\frac{\epsilon}{24}) \leq \frac{Ck}{\epsilon^2}\log^2\left(\frac{k}{\epsilon}\right)$. By a well-known theorem of [1] (Theorem 8 in the appendix), a bound of $\frac{Ck}{\epsilon^2}\log^2\left(\frac{k}{\epsilon}\right)$ on $\text{fat}_{\mathcal{F}}(\frac{\epsilon}{24})$ implies the bound in (13) on the sample complexity, which implies Theorem 5.

7 Minimax lower bounds on the sample complexity

Let K be a universal constant whose value will be fixed throughout this section. In this section, we will state lower bounds on the number of samples needed for the minimax decision rule for learning from high dimensional data, with high probability, a manifold with a squared loss that is within ϵ of the optimal. We will construct a carefully chosen prior on the space of probability distributions and use an argument that can either be viewed as an application of the probabilistic method or of the fact that the Minimax risk is at least the risk of a Bayes optimal manifold computed with respect to this prior. Let U be a $K^{2d}k-$dimensional vector space containing the origin, spanned by the basis $\{e_1, \dots, e_{K^{2d}k}\}$ and S be the surface of the ball of radius 1 in $\mathbb{R}^m$. We assume that m be greater or equal to $K^{2d}k+d$. Let W be the $d-$dimensional vector space spanned by $\{e_{K^{2d}k+1}, \dots, e_{K^{2d}k+d}\}$. Let $S_1, \dots, S_{K^{2d}k}$ denote spheres, such that for each i, $S_i := S \cap (\sqrt{1-\tau^2}e_i + W)$, where $x+W$ is the translation of W by x. Note that each S_i has radius τ. Let $\ell = \binom{K^{2d}k}{K^dk}$ and $\{\mathcal{M}_1, \dots, \mathcal{M}_\ell\}$ consist of all K^dk-element subsets of $\{S_1, \dots, S_{K^{2d}k}\}$. Let ω_d be the volume of the unit ball in $\mathbb{R}^d$. The following theorem shows that no algorithm can produce a nearly optimal manifold with high probability unless it uses a number of samples that depends linearly on volume, exponentially on intrinsic dimension and polynomially on the curvature.

Theorem 6. *Let $\mathcal{F}$ be equal to either $\mathcal{G}_e(d, V, \tau)$ or $\mathcal{G}_i(d, V, \frac{1}{\tau^2}, \pi\tau)$. Let $k = \lfloor \frac{V}{d\omega_d(K^{\frac{5}{4}}\tau)^d} \rfloor$. Let $\mathcal{A}$ be an arbitrary algorithm that takes as input a set of data points $x = \{x_1, \dots, x_k\}$ and outputs a manifold $\mathcal{M}_\mathcal{A}(x)$ in $\mathcal{F}$. If $\epsilon + 2\delta < \frac{1}{3}\left(\frac{1}{2\sqrt{2}} - \tau\right)^2$ then,*

$$\inf_{\mathcal{P}} \mathbb{P}\left[\mathcal{L}(\mathcal{M}_\mathcal{A}(x), \mathcal{P}) - \inf_{\mathcal{M}\in\mathcal{F}} \mathcal{L}(\mathcal{M}, \mathcal{P}) < \epsilon\right] < 1 - \delta,$$

where $\mathcal{P}$ ranges over all distributions supported on B and $x_1, \dots, x_k$ are i.i.d draws from $\mathcal{P}$.

Proof. Observe from Lemma 7 and Theorem 3 that $\mathcal{F}$ is a class of a manifolds such that each manifold in $\mathcal{F}$ is contained in the union of $K^{\frac{3d}{2}}k$ $m-$dimensional balls of radius τ, and $\{\mathcal{M}_1, \dots, \mathcal{M}_\ell\} \subseteq \mathcal{F}$. (The reason why we have $K^{\frac{3d}{2}}$ rather than $K^{\frac{5d}{4}}$ as in the statement of the theorem is that the parameters of $\mathcal{G}_i(d, V, \tau)$ are intrinsic, and to transfer to the extrinsic setting of the last sentence, one needs some leeway.) Let $\mathcal{P}_1, \dots, \mathcal{P}_\ell$ be probability distributions that are uniform on $\{\mathcal{M}_1, \dots, \mathcal{M}_\ell\}$ with respect to the induced Riemannian measure. Suppose A is an algorithm that takes as input a set of data points $x = \{x_1, \dots, x_t\}$ and outputs a manifold $\mathcal{M}_\mathcal{A}(x)$. Let r be chosen uniformly at random from $\{1, \dots, \ell\}$. Then,

$$\begin{aligned}\inf_{\mathcal{P}} \mathbb{P}\left[\left|\mathcal{L}(\mathcal{M}_\mathcal{A}(x), \mathcal{P}) - \inf_{\mathcal{M}\in\mathcal{F}} \mathcal{L}(\mathcal{M}, \mathcal{P})\right| < \epsilon\right] &\leq \mathbb{E}_{\mathcal{P}_r}\mathbb{P}_x\left[\left|\mathcal{L}(\mathcal{M}_\mathcal{A}(x), \mathcal{P}_r) - \inf_{\mathcal{M}\in\mathcal{F}} \mathcal{L}(\mathcal{M}, \mathcal{P}_r)\right| < \epsilon\right] \\ &= \mathbb{E}_x\mathbb{P}_{\mathcal{P}_r}\left[\left|\mathcal{L}(\mathcal{M}_\mathcal{A}(x), \mathcal{P}_r) - \inf_{\mathcal{M}\in\mathcal{F}} \mathcal{L}(\mathcal{M}, \mathcal{P}_r)\right| < \epsilon \middle| x\right] \\ &= \mathbb{E}_x\mathbb{P}_{\mathcal{P}_r}\left[\mathcal{L}(\mathcal{M}_\mathcal{A}(x), \mathcal{P}_r) < \epsilon \middle| x\right].\end{aligned}$$

We first prove a lower bound on $\inf_x \mathbb{E}_r\left[\mathcal{L}(\mathcal{M}_\mathcal{A}(x), \mathcal{P}_r)|x\right]$.

We see that

$$\mathbb{E}_r\left[\mathcal{L}(\mathcal{M}_\mathcal{A}(x), \mathcal{P}_r)\middle|x\right] = \mathbb{E}_{r,x_{k+1}}\left[\mathbf{d}(\mathcal{M}_\mathcal{A}(x), x_{k+1})^2\middle|x\right]. \tag{14}$$

Conditioned on x, the probability of the event (say E_{dif}) that x_{k+1} does not belong to the same sphere as one of the $x_1, \dots, x_k$ is at least $\frac{1}{2}$.

Conditioned on E_{dif} and $x_1, \dots, x_k$, the probability that x_{k+1} lies on a given sphere S_j is equal to 0 if one of $x_1, \dots, x_k$ lies on S_j and $\frac{1}{K^2k-k'}$ otherwise, where $k' \leq k$ is the number of spheres in $\{S_i\}$ which contain at least one point among $x_1, \dots, x_k$.

By construction, $A(x_1, \dots, x_k)$ can be covered by $K^{\frac{3d}{2}}k$ balls of radius τ; let their centers be $y_1, \dots, y_{K^{\frac{3d}{2}}k}$.

However, it is easy to check that for any dimension m, the cardinality of the set $\mathcal{S}_y$ of all S_i that have a nonempty intersection with the balls of radius $\frac{1}{2\sqrt{2}}$ centered around $y_1, \dots, y_{K^{\frac{3d}{2}}k}$, is at most $K^{\frac{3d}{2}}k$. Therefore, $\mathbb{P}\left[\mathbf{d}(\mathcal{M}_{\mathcal{A}}(x), x_{k+1}) \geq \frac{1}{2\sqrt{2}} - \tau \big| x\right]$ is at least

$$\begin{aligned}\mathbb{P}\left[\mathbf{d}(\{y_1, \dots, y_{K^{\frac{3d}{2}}k}\}, x_{k+1}) \geq \frac{1}{2\sqrt{2}}\big|x\right] &\geq \mathbb{P}\left[E_{dif}\right]\mathbb{P}\left[x_{k+1} \not\in \mathcal{S}_y | E_{dif}\right] \\ &\geq \frac{1}{2}\frac{K^{2d}k - k' - K^{\frac{3d}{2}}k}{K^{2d}k - k'} \\ &\geq \tfrac{1}{3}.\end{aligned}$$

Therefore, $\mathbb{E}_{r, x_{k+1}}\left[\mathbf{d}(\mathcal{M}_{\mathcal{A}}(x), x_{k+1})^2 \big| x\right] \geq \frac{1}{3}\left(\frac{1}{2\sqrt{2}} - \tau\right)^2$. Finally, we observe that it is not possible for $\mathbb{E}_x \mathbb{P}_{\mathcal{P}_r}\left[\mathcal{L}(\mathcal{M}_{\mathcal{A}}(x), \mathcal{P}_r) < \epsilon \big| x\right]$ to be more than $1 - \delta$ if $\inf_x \mathbb{P}_{\mathcal{P}_r}\left[\mathcal{L}(\mathcal{M}_{\mathcal{A}}(x), \mathcal{P}_r) \big| x\right] > \epsilon + 2\delta$, because $\mathcal{L}(\mathcal{M}_{\mathcal{A}}(x), \mathcal{P}_r)$ is bounded above by 2. □

8 Algorithmic implications

8.1 $k-$means

Applying Theorem 4 to the case when $\mathcal{P}$ is a distribution supported equally on n specific points (that are part of an input) in a unit ball of $\mathbb{R}^m$, we see that in order to obtain an additive ϵ approximation for the $k-$means problem with probability $1 - \delta$, it suffices to sample

$$s \geq C\left(\frac{k}{\epsilon^2}\min\left(\frac{\log^4\left(\frac{k}{\epsilon}\right)}{\epsilon^2}, k\right) + \frac{1}{\epsilon^2}\log\frac{1}{\delta}\right)$$

points uniformly at random (which would have a cost of $O(s \log n)$ if the cost of one random bit is $O(1)$) and exhaustively solve $k-$means on the resulting subset. Supposing that a dot product between two vectors x_i, x_j can be computed using $\tilde{m}$ operations, the total cost of sampling and then exhaustively solving $k-$means on the sample is $O(\tilde{m}sk^s \log n)$. In contrast, if one asks for a multiplicative $(1+\epsilon)$ approximation, the best running time known depends linearly on n [15]. If $\mathcal{P}$ is an unknown probability distribution, the above algorithm improves upon the best results in a natural statistical framework for clustering [5].

8.2 Fitting piecewise linear curves

In this subsection, we illustrate the algorithmic utility of the uniform bound in Theorem 4 by obtaining an algorithm for fitting a curve of length no more than L, to data drawn from an unknown probability distribution $\mathcal{P}$ supported in B, whose sample complexity is independent of the ambient dimension. This curve, with probability $1 - \delta$, achieves a mean squared error of less than ϵ more than the optimum. The proof of its correctness and analysis of its run-time have been deferred to the appendix. The algorithm is as follows:

1. Let $k := \lceil \frac{L}{\epsilon} \rceil$ and $s \geq C\left(\frac{k}{\epsilon^2}\min\left(\frac{\log^4\left(\frac{k}{\epsilon}\right)}{\epsilon^2}, k\right) + \frac{1}{\epsilon^2}\log\frac{1}{\delta}\right)$. Sample points $x_1, \dots, x_s$ i.i.d from $\mathcal{P}$ for $s =$, and set $J := \text{span}(\{x_i\}_{i=1}^s)$.
2. For every permutation σ of $[s]$, minimize the convex objective function $\sum_{i=1}^n d(x_{\sigma(i)}, y_i)^2$ over the convex set of all $s-$tuples of points $(y_1, \dots, y_s)$ in J, such that $\sum_{i=1}^{s-1} \|y_{i+1} - y_i\| \leq L$.
3. If the minimum over all $(y_1, \dots, y_s)$ (and σ) is achieved for $(z_1, \dots, z_s)$, output the curve obtained by joining z_i to z_{i+1} for each i by a straight line segment.

References

[1] Noga Alon, Shai Ben-David, Nicolò Cesa-Bianchi, and David Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *J. ACM*, 44(4):615–631, 1997.

[2] Rosa I. Arriaga and Santosh Vempala. An algorithmic theory of learning: Robust concepts and random projection. In *FOCS*, pages 616–623, 1999.

[3] Peter Bartlett. The minimax distortion redundancy in empirical quantizer design. *IEEE Transactions on Information Theory*, 44:1802–1813, 1997.

[4] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.*, 15(6):1373–1396, 2003.

[5] Shai Ben-David. A framework for statistical clustering with constant time approximation algorithms for k-median and k-means clustering. *Mach. Learn.*, 66(2-3):243–257, 2007.

[6] Gunnar Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46:255–308, January 2009.

[7] Sanjoy Dasgupta. Learning mixtures of gaussians. In *FOCS*, pages 634–644, 1999.

[8] Sanjoy Dasgupta and Anupam Gupta. An elementary proof of a theorem of johnson and lindenstrauss. *Random Struct. Algorithms*, 22(1):60–65, 2003.

[9] Luc Devroye, Lászlo Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*. Springer, 1996.

[10] David L. Donoho and Carrie Grimes. Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proceedings of the National Academy of Sciences*, 100(10):5591–5596, May 2003.

[11] A. Gray. *Tubes*. Addison-Wesley, 1990.

[12] Trevor J. Hastie and Werner Stuetzle. Principal curves. *Journal of the American Statistical Association*, 84:502–516, 1989.

[13] William Johnson and Joram Lindenstrauss. Extensions of lipschitz mappings into a hilbert space. *Contemporary Mathematics*, 26:419–441, 1984.

[14] Balázs Kégl, Adam Krzyzak, Tamás Linder, and Kenneth Zeger. Learning and design of principal curves. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:281–297, 2000.

[15] Amit Kumar, Yogish Sabharwal, and Sandeep Sen. A simple linear time $(1+\epsilon)-$approximation algorithm for k-means clustering in any dimensions. In *FOCS*, pages 454–462, 2004.

[16] Andreas Maurer and Massimiliano Pontil. Generalization bounds for k-dimensional coding schemes in hilbert spaces. In *ALT*, pages 79–91, 2008.

[17] H. Narayanan and P. Niyogi. On the sample complexity of learning smooth cuts on a manifold. In *Proc. of the 22nd Annual Conference on Learning Theory (COLT)*, June 2009.

[18] Partha Niyogi, Stephen Smale, and Shmuel Weinberger. Finding the homology of submanifolds with high confidence from random samples. *Discrete & Computational Geometry*, 39(1-3):419–441, 2008.

[19] Sam T. Roweis and Lawrence K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *SCIENCE*, 290:2323–2326, 2000.

[20] Alexander J. Smola, Sebastian Mika, Bernhard Schölkopf, and Robert C. Williamson. Regularized principal manifolds. *J. Mach. Learn. Res.*, 1:179–209, 2001.

[21] J. B. Tenenbaum, V. Silva, and J. C. Langford. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 290(5500):2319–2323, 2000.

[22] Afra Zomorodian and Gunnar Carlsson. Computing persistent homology. *Discrete & Computational Geometry*, 33(2):249–274, 2005.

A Proofs

Let $\mathcal{G}$ be the class of submanifolds of B whose condition number is $\geq \frac{1}{\tau}$ and volume is less or equal to V. The following was proved in (Lemma 5.3, [18]).

Lemma 7. *Let $\mathcal{M}$ be a $d-$dimensional submanifold of $\mathbb{R}^m$, whose condition number is $\geq \frac{1}{\tau}$. For $p \in \mathcal{M}$, let $B_\epsilon(p)$ be the ball in $\mathbb{R}^m$ of radius ϵ centered at p, and let ω_d be the volume of the $d-$dimensional unit ball. Suppose $\theta := \arcsin(\frac{\epsilon}{2\tau})$ for $\epsilon \leq 2\tau$, then, the volume of $\mathcal{M} \cap B_\epsilon(p)$ is greater or equal to $\epsilon^d(\cos^d \theta)\omega_d$.*

Proof of Theorem 2. As a consequence of Lemma 7, we obtain an upper bound of

$$\frac{2^{5d}V}{\epsilon^d(\cos^d(\arcsin\frac{\epsilon}{64\tau}))\omega_d}$$

on the number of disjoint sets of the form $\mathcal{M} \cap B_{\epsilon/32}(p)$ that can be packed in $\mathcal{M}$. If $\{\mathcal{M} \cap B_{\epsilon/32}(p_1), \dots, \mathcal{M} \cap B_{\epsilon/32}(p_k)\}$ is a maximal family of disjoint sets of the form $\mathcal{M} \cap B_{\epsilon/32}(p)$, then there is no point $p \in \mathcal{M}$ such that $\min_i \|p - p_i\| > \epsilon/16$. Therefore, $\mathcal{M}$ is contained in the union of balls,

$$\bigcup_i B_{\epsilon/16}(p_i).$$

Therefore, we may apply Theorem 4 with

$$\begin{aligned} U\left(\frac{1}{\min(\epsilon,\tau)}\right) &\leq \frac{2^{5d}V}{\epsilon^d \cos^d \arcsin\min\left(\frac{\epsilon}{64\tau}, \frac{1}{64}\right)\omega_d} && (15)\\ &\leq V\left(C\left(\frac{d}{\min(\epsilon,\tau)}\right)\right)^d. && (16)\end{aligned}$$

□

In order to prove a bound of the form

$$\mathbb{P}\left[\sup_{F\in\mathcal{F}}\left|\frac{\sum_{i=1}^s F(x_i)}{s} - \mathbb{E}_{\mathcal{P}}F(x)\right| < \epsilon\right] > 1-\delta, \tag{17}$$

it suffices to bound a measure of the complexity of $\mathcal{F}$ known as the Fat-Shattering dimension of the function class $\mathcal{F}$. The metric entropy of $\mathcal{F}$ can be bounded using the Fat-Shattering dimension, leading to uniform bounds.

Definition 3. *Let $\mathcal{F}$ be a set of real valued functions. We say that a set of points $x_1, \dots, x_k$ is $\gamma-$shattered by $\mathcal{F}$ if there is a vector of real numbers $t = (t_1, \dots, t_k)$ such that for all binary vectors $\mathbf{b} = (b_1, \dots, b_k)$ and each $i \in [s] = \{1, \dots, s\}$, there is a function $f_{\mathbf{b},t}$ satisfying,*

$$f_{\mathbf{b},t}(x_i) = \begin{cases} > t_i + \gamma, & \textit{if } b_i = 1;\\ < t_i - \gamma, & \textit{otherwise.}\end{cases} \tag{18}$$

More generally, the supremum taken over $(t_1, \dots, t_s)$ of the number of binary vectors $\mathbf{b}$ for which there is a function $f_{\mathbf{b},t}$, is called the $\gamma-$shatter coefficient. For each $\gamma > 0$, the Fat-Shattering dimension $\mathrm{fat}_{\mathcal{F}}(\gamma)$ of the set $\mathcal{F}$ is defined to be the size of the largest $\gamma-$shattered set if this is finite; otherwise it is declared to be infinity.

Theorem 8. *For any permissible class of functions $\mathcal{F}$ from B into $[-1,1]$, an arbitrary probability distribution $\mathcal{P}$ supported on B $\epsilon, \delta > 0$, if*

$$s \geq C\left(\frac{\mathrm{fat}_{\mathcal{F}}(\frac{\epsilon}{24})}{\epsilon^2}\log^2\frac{\mathrm{fat}_{\mathcal{F}}(\frac{\epsilon}{24})}{\epsilon} + \frac{1}{\epsilon^2}\log\frac{1}{\delta}\right)$$

and samples $x_1, \dots, x_s$ are independently drawn from $\mathcal{P}$, then

$$\mathbb{P}\left[\sup_{F\in\mathcal{F}}\left|\frac{\sum_{i=1}^s F(x_i)}{s} - \mathbb{E}_{\mathcal{P}}F(x)\right| < \epsilon\right] > 1-\delta. \tag{19}$$

Any countable family of functions is known to be permissible. We will also need to use the notion of VC dimension, and some of its properties. These appear below.

Definition 4. *Let Λ be a collection of measurable subsets of $\mathbb{R}^m$. For $x_1, \dots, x_k \in \mathbb{R}^m$, Let the number of different sets in $\{\{x_1, \dots, x_k\} \cap H; H \in \Lambda\}$ be denoted the shatter coefficient $N_\Lambda(x_1, \dots, x_k)$. The VC dimension VC_Λ of Λ is the largest integer k such that there exist $x_1, \dots x_k$ such that $N_\Lambda(x_1, \dots, x_k) = 2^k$.*

The following result regarding the VC dimension of halfspaces is well known (Corollary 13.1, [9]).

Theorem 9. *Let Λ be the class of halfspaces in $\mathbb{R}^g$. Then $VC_\Lambda = g + 1$.*

We state the Sauer-Shelah Lemma below.

Lemma 10 (Theorem 13.2, [9])**.** *For any $x_1, \dots, x_k \in \mathbb{R}^g$, $N_\Lambda(x_1, \dots, x_k) \leq \sum_{i=0}^{VC_\Lambda} \binom{k}{i}$.*

For $VC_\Lambda > 2$, $\sum_{i=0}^{VC_\Lambda} \binom{k}{i} \leq k^{VC_\Lambda}$.

Proof of Theorem 5. It follows from Theorem 5 in [16] that if

$$s \geq C \left(\frac{k^2}{\epsilon^2} + \frac{1}{\epsilon^2} \log \frac{1}{\delta} \right),$$

then

$$\mathbb{P} \left[\sup_{F \in \mathcal{F}} \left| \frac{\sum_{i=1}^s F(x_i)}{s} - \mathbb{E}_\mathcal{P} F(x) \right| < \epsilon \right] > 1 - \delta. \tag{20}$$

It therefore suffices by Theorem 8, to show that $\operatorname{fat}_\mathcal{F}(\gamma) \leq \frac{Ck}{\gamma^2} \log^2 \frac{Ck}{\gamma^2}$.

Let $x_1, \dots, x_u$ be u points such that $\forall \mathcal{A} \subseteq X := \{x_1, \dots, x_u\}$, and $\exists f_\mathcal{A} \in \mathcal{F}$ such that $f_\mathcal{A}(x) = \min_i v_i^\mathcal{A} \cdot x$ and $V_\mathcal{A} := \{v_1^\mathcal{A}, \dots, v_u^\mathcal{A}\}$ such that for some $\mathbf{t} = (t_1, \dots, t_u)$,

$$\forall x_r \in \mathcal{A}, \min_i (v_i^\mathcal{A} \cdot x_r^\mathcal{A}) < t_r - \gamma \tag{21}$$

and

$$\forall x_r \notin \mathcal{A}, \min_i (v_i^\mathcal{A} \cdot x_r) > t_r + \gamma. \tag{22}$$

We will obtain an upper bound on u. Let $g := C\left(\gamma^{-2} \log(u + k)\right)$. Consider a particular $\mathcal{A} \in X$ and $f_\mathcal{A}(x) := \min_i v_i^\mathcal{A} \cdot x$ that satisfies (21) and (22). In what follows, by a random $g-$dimensional subspace, we mean the image of a fixed subspace of dimension g under a random orthogonal transformation of $\mathbb{R}^m$ chosen from the normalized Haar measure. We will need the following version of the Johnson-Lindenstrauss Lemma [13, 8].

Lemma 11. *Let $y_1, \dots, y_k$ be points in the unit ball in $\mathbb{R}^m$. Let $\sqrt{\frac{g}{m}} R$ be the orthogonal projection onto a random $g-$dimensional subspace where $g = C \frac{\log \ell}{\gamma^2}$ for some $\gamma > 0$. Then,*

$$\mathbb{P} \left[\sup_{i,j \in \{1, \dots, g\}} \left| (Ry_i) \cdot (Ry_j) - y_i \cdot y_j \right| > \frac{\gamma}{2} \right] < \frac{1}{2}. \tag{23}$$

Setting $\ell := u + k$ and $\{y_1, \dots, y_\ell\}$ to $\{x_1, \dots, x_u, v_1^\mathcal{A}, \dots, v_k^\mathcal{A}\}$ in the above lemma, we make the following observation.

Observation 1. *For each $\mathcal{A}$, with probability at least $\frac{1}{2}$ the following statement is true.*

$$\forall x_r \in \mathcal{A}, \exists\, i \in [\ell] \quad Rv_i^\mathcal{A} \cdot Rx_r < t_r - \frac{\gamma}{2} \tag{24}$$

and

$$\forall x_r \notin \mathcal{A}, \forall i \in [\ell] \quad Rv_i^\mathcal{A} \cdot Rx_r > t_r + \frac{\gamma}{2}. \tag{25}$$

For a random $R \in \Re$, denote by RX, the set $\{Rx_1, \dots, Rx_u\}$ and set $J := \mathrm{span}(RX)$ to be the corresponding $u-$dimensional random subspace. Let $t^J : J \to \mathbb{R}$ be the function such that for each $j \in [u]$, $t^J(Rx_j) := t_j$, and for all other $z \in J$, $t^J(z) = 0$.

Let $\mathcal{F}_{J,k}$ be the concept class consisting of all subsets of J of the form

$$\{z : \min_i \begin{pmatrix} w_i \\ 1 \end{pmatrix} \cdot \begin{pmatrix} z \\ -t^J(z) \end{pmatrix} \leq 0\},$$

where $w_1, \dots w_k$ are arbitrary vectors in J.

By Observation 1, for a random $R \in \Re$, the expected number of sets of the form $RX \cap \imath,\ \imath \in \mathcal{F}_{J,k}$ is greater or equal to 2^{u-1}. Therefore, we observe the following.

Observation 2. *There exists an $R \in \Re$ such that the number of sets of the form $RX \cap \imath,\ \imath \in \mathcal{F}_{J,k}$ is greater or equal to 2^{u-1}. Fix such an R and set $J := \mathrm{span}(RX)$.*

Classical VC theory (Theorem 9) tells us that the VC dimension of Halfspaces in the span of all vectors of the form

$$\begin{pmatrix} z \\ -t^J(z) \end{pmatrix}$$

is less or equal to $g+2$. Therefore, by the Sauer-Shelah Lemma (Lemma 10), the number $W(u, \mathcal{F}_{J,1})$ of distinct sets $\{y_1, \dots, y_u\} \cap \jmath,\ \jmath \in \mathcal{F}_{J,1}$ is not more than $\sum_{i=0}^{g+2} \binom{u}{i} \leq u^{g+2}$.

Every set of the form $\{y_1, \dots, y_u\} \cap \imath,\ \imath \in \mathcal{F}_{J,k}$ can be expressed as a union of sets of the form $\{y_1, \dots, y_u\} \cap \jmath,\ \jmath \in \mathcal{F}_{J,1}$, in which the total number of sets participating is k. Therefore, the number $W(u, \mathcal{F}_{J,k})$ of distinct sets $\{y_1, \dots, y_u\} \cap \imath,\ \imath \in \mathcal{F}_{J,1}$ is less or equal to $W(u, \mathcal{F}_{J,1})^k$, which is in turn less or equal to $u^{(g+2)k}$. In the light of Observation 2, we see that

$$u^{(g+2)k} \geq 2^{u-1}. \tag{26}$$

Therefore $u - 1 \leq \ell(g+2)\log u$. Assuming without loss of generality that $u \geq k$, and substituting $C\left(\gamma^{-2}\log(u+k)\right)$ for g, we see that

$$u \leq C\left(\ell\gamma^{-2}\log^2 u\right), \tag{27}$$

and hence $\frac{u}{\log^2 u} \leq C\left(\frac{k}{\gamma^2}\right)$, implying that $u \leq C\left(\left(\frac{k}{\gamma^2}\right)\log^2\left(\frac{k}{\gamma}\right)\right)$.

Thus, the Fat-Shattering dimension $\mathrm{fat}_{\mathcal{F}}(\gamma)$ is $C\left(\left(\frac{k}{\gamma^2}\right)\log^2\left(\frac{k}{\gamma}\right)\right)$. Substituting this back into Theorem 8 completes the proof. □

B Fitting Curves

In this section, we analyze the correctness and run-time of the algorithm for curve fitting described in Section 8.2, which we denote CurvFit. Let $\mathcal{G}_L$ be the collection of all rectifiable curves of length $\leq L$ in the $m-$dimensional unit ball B. Since any curve of length $\leq L$ can be covered using less or equal to $2\lceil\frac{L}{\epsilon}\rceil$ balls of radius $\epsilon/2$, Theorem 4 implies the following.

Theorem 12. *Let $k := \lceil\frac{L}{\epsilon}\rceil$. If*

$$s \geq C\left(\frac{k}{\epsilon^2}\min\left(\frac{\log^4\left(\frac{k}{\epsilon}\right)}{\epsilon^2}, k\right) + \frac{1}{\epsilon^2}\log\frac{1}{\delta}\right),$$

and $\mathcal{M} \in \mathcal{G}_L$ achieves within $\epsilon/2$ of the infimum of $s^{-1}\sum_{i=1}^s \mathbf{d}(x_i, \mathcal{N})^2$ taken over $\mathcal{N} \in \mathcal{G}_L$, then the probability is less than δ, that there exists $\mathcal{M}_{\mathcal{A}} \in \mathcal{G}_L$ that is a better least-squares fit to the extent that

$$\mathbb{E}_{\mathcal{P}}\,\mathbf{d}(x, \mathcal{M}_{\mathcal{A}})^2 < \mathbb{E}_{\mathcal{P}}\,\mathbf{d}(x, \mathcal{M})^2 - \epsilon.$$

B.1 Correctness

Let Γ by a curve of length less or equal to L, which minimizes the empirical squared distance to the points $x_1, \dots, x_s$. There are points $z_1, \dots, z_s$ on Γ such that this empirical squared distance is $\sum_{i=1}^s \|z_i - x_i\|^2$, where $\sum_i \|z_i - z_{i+1}\| \leq L$. The piecewise linear path joining z_i to z_{i+1} has an empirical squared error and length that are less or equal to the respective values for Γ. This piecewise linear path would be among those considered by the above algorithm, which proves that the algorithm finds a curve that achieves the minimum empirical squared error.

B.2 Approximation algorithm for curve fitting

Applying Theorem 12 to a setting where $\mathcal{P}$ is a distribution supported equally on n fixed points (that are part of an input) in a unit ball of $\mathbb{R}^m$, we obtain the following result. In order to obtain an additive ϵ approximation of the best possible mean squared loss with probability $1 - \delta$, it suffices to sample

$$s \geq C \left(\frac{k}{\epsilon^2} \min \left(\frac{\log^4 \left(\frac{k}{\epsilon}\right)}{\epsilon^2}, k \right) + \frac{1}{\epsilon^2} \log \frac{1}{\delta} \right)$$

points uniformly at random (which would have a cost of $O(s \log n)$ if the cost of one random bit is $O(1)$), where $k := \lceil \frac{L}{\epsilon} \rceil$ and find the curve whose empirical loss with respect to these points is the least using $\mathsf{CurvFit}$.

B.3 Run-time

Each of the $s!$ convex optimizations in $\mathsf{CurvFit}$ can be done using $s^{O(1)}$ operations using convex optimization in the span of the s points. Supposing that a dot product between two vectors x_i, x_j can be computed using $\tilde{m}$ operations, membership in the convex set can be tested using $O(s^{O(1)}\tilde{m})$ arithmetic operations, and so can evaluating the objective function. Thus, the total number of arithmetic operations involved is $s^{O(s)}\tilde{m} \ln n$. Note that $\tilde{m}$ can be much smaller than m if data is sparse, or if the curve is being fit in the implicit space arising from a kernel.