On the sample complexity of learning smooth cuts on a manifold

Hariharan Narayanan
Department of Computer Science
University of Chicago
Chicago, IL 60637
hari@cs.uchicago.edu

Partha Niyogi
Department of Computer Science
University of Chicago
Chicago, IL 60637
niyogi@cs.uchicago.edu

Abstract

Modern data sets, though typically high dimensional, are often generated by processes possessing few essential degrees of freedom, as is the case with human speech. In recent years, such considerations have lead to the notion that high dimensional data may be modeled to lie on a submanifold of low intrinsic dimension. We derive bounds on the number of random samples needed before it is possible to approximately separate data into two classes using smooth decision boundaries with high probability.

1 Introduction

Over the last several years, manifold based methods have been developed and applied to a variety of problems. Much of this work is empirically and algorithmically oriented and there is a need to better understand the learning-theoretic foundations of this class of machine learning problems. Our paper is a contribution in this direction with the hope that it will better delineate the possibilities and limitations.

In the manifold setting, one is canonically interested in learning a function $f : \mathcal{M} \to \mathbb{R}$ (regression) or $f : \mathcal{M} \to \{0, 1\}$ (classification/clustering). For regression therefore, the natural objects of study are classes of real-valued functions on the manifold leading one to eventually consider functional analysis on the manifold. Thus, for example, the Laplace-Beltrami operator and its eigenfunctions have been studied with a view to function learning [1, 2].

Our interest in this paper is the setting for classification or clustering where the function is $0/1$ valued and therefore divides the manifold into two disjoint pieces $\mathcal{M}_1$ and $\mathcal{M}_2$. A natural class of such functions may be associated with smooth cuts on the manifold. We will consider *smooth cuts* where each cut corresponds to a submanifold (say $P \subset \mathcal{M}$) that divides $\mathcal{M}$ into two pieces. Since P is a submanifold of $\mathcal{M}$ and hence $\mathbb{R}^m$, one can associate to it a measure of complexity given by its condition number $1/\tau$. The condition number is defined as follows.

Definition 1 (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 $\mathcal{M}$ at different points intersect.*

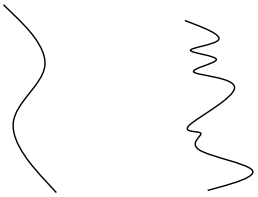

Figure 1: Curves of low and high condition number at the left and right respectively.

Given two linear subspaces V, W, let $\sphericalangle(V, W)$ be the angle between V and W, defined as

$$\sphericalangle(V, W) = \arccos\left(\sup_{v\in V}\inf_{w\in W}\frac{v\cdot w}{\|v\|\|w\|}\right). \tag{1}$$

For any manifold $\mathcal{M}$,

$$c(\mathcal{M}) = \inf_{x,y\in L}\frac{2\sin(\frac{\sphericalangle(T_x,T_y)}{2})}{\|x-y\|}, \tag{2}$$

where the infimum is taken over distinct points $x, y \in \mathcal{M}$ and T_x and T_y are the tangent spaces at x and y.

We can define the following function class (concept class in PAC terminology.)

Definition 2 *Let*

$$\mathcal{S}_\tau := \left\{S \middle| S = \overline{S} \subseteq \mathcal{M} \text{ and } c(S \cap \overline{\mathcal{M}\setminus S}) \leq \frac{1}{\tau}\right\},$$

where $\overline{S}$ is the closure of S. Let

$$\mathcal{C}_\tau := \left\{f \middle| f : \mathcal{M} \to \{0, 1\} \text{ and } f^{-1}(1) \in \mathcal{S}_\tau\right\}.$$

Thus, the concept class $\mathcal{C}_\tau$ is the collection of indicators of all closed sets in $\mathcal{M}$ whose boundaries are $1/\tau$-conditioned $d-1$ dimensional submanifolds of $\mathbb{R}^m$.

Note that when $\tau = \infty$, $\mathcal{C}_\tau$ contains the indicators of all affine half-subspaces of dimension d that are contained in $\mathcal{M}$. By letting τ vary, we obtain a structured family of cuts. We now consider the following basic question.

[Question:] Let $\mathcal{M}$ be a d-dimensional submanifold of $\mathbb{R}^m$ and let $\mathcal{C}_\tau$ be a concept class of $0/1$ valued functions corresponding to a family of smooth cuts with condition number $\frac{1}{\tau}$. Then what is the sample complexity of learning the elements of $\mathcal{C}_\tau$?

Our contributions in this paper are as follows.

1. We show that distribution-free learning of $\mathcal{C}_\tau$ is impossible in general since for some $\mathcal{M}$, it is a space of infinite VC dimension. We prove that this is the case for a natural embedding in $\mathbb{R}^m$ of the $d-$dimensional sphere of radius $\kappa > \tau$.

2. On the the other hand, it is possible to provide *distribution-specific* sample complexity bounds that hold uniformly for a large class of probability measures on $\mathcal{M}$. These are the measures for which there exists a Radon Nikodym derivative with respect to the uniform measure on $\mathcal{M}$ such that there is an upper bound $\rho_{\max}$ on the associated density function. The sample complexity is seen to depend on the intrinsic dimension d, curvature bounds τ and κ, density bound $\rho_{\max}$, but is independent of the ambient dimension m.

3. The proof technique used for obtaining these distribution specific bounds (Poissonization etc.) may be useful prove distribution specific learning in other settings.

Our sample complexity bounds depend on an upper bound $\rho_{max} \geq 1$ on the maximum density of $\mathcal{P}$ with respect to the volume measure, (normalized to be a probability measure), the curvatures and the intrinsic dimension of $\mathcal{M}$ and the class boundary P, but are independent of the ambient dimension m. We also show that the dependence on the maximum density ρ_{max} of $\mathcal{P}$ is unavoidable by proving that for any fixed τ the VC-dimension of the function class associated with cuts that are submanifolds with a condition number $\frac{1}{\tau}$ is infinite (Lemma 9) for certain compact submanifolds.

2 Preliminaries

Suppose that $\mathcal{P}$ is a probability measure supported on a d-dimensional Riemannian submanifold $\mathcal{M}$ of $\mathbb{R}^m$ having condition number $\leq \frac{1}{\kappa}$. Suppose that data samples $\{x_i\}_{i\geq 1}$ are randomly drawn from $\mathcal{P}$ in an i.i.d fashion. Let each data point x be associated with a label $f(x) \in \{0, 1\}$.

Definition 3 (Annealed Entropy) *Let $\mathcal{P}$ be a probability measure supported on a manifold $\mathcal{M}$. Given a class of indicator functions Λ and a set of points $Z = \{z_1, \ldots, z_\ell\} \subset \mathcal{M}$, let $N(\Lambda, Z)$ be the the number of ways of partitioning $z_1, \ldots, z_\ell$ into two sets using indicators belonging to Λ . We define $G(\Lambda, \mathcal{P}, \ell)$ to be the expected value of $N(\Lambda, Z)$. Thus*

$$G(\Lambda, \mathcal{P}, \ell) := \mathbb{E}_{Z\vdash\mathcal{P}^{\times\ell}} N(\Lambda, Z),$$

where expectation is with respect to Z and $\vdash$ signifies that Z is drawn from the Cartesian product of ℓ copies of $\mathcal{P}$. The annealed entropy of Λ with respect to ℓ samples from $\mathcal{P}$ is defined to be

$$H_{ann}(\Lambda, \mathcal{P}, \ell) := \ln G(\Lambda, \mathcal{P}, \ell).$$

Definition 4 *The risk $R(\alpha)$ of a classifier α is defined as the probability that α misclassifies a random data point x drawn from $\mathcal{P}$. Formally, $R(\alpha) := \mathbb{E}_{\mathcal{P}}[\alpha(x) \neq f(x)]$. Given a set of ℓ labeled data points $(x_1, f(x_1)), \ldots, (x_\ell, f(x_\ell))$, the empirical risk is defined to be $R_{emp}(\alpha, \ell) := \frac{\sum_{i=1}^{\ell} \mathcal{I}[\alpha(x_i)\neq f(x_i)]}{\ell}$, where $\mathcal{I}[\cdot]$ denotes the indicator of the respective event and $f(x)$ is the label of point x.*

Theorem 5 (Vapnik [6], Thm 4.2) *For any ℓ the inequality*

$$\mathbb{P}\left[\sup_{\alpha\in\Lambda} \frac{R(\alpha) - R_{emp}(\alpha, \ell)}{\sqrt{R(\alpha)}} > \epsilon\right] \quad < \quad 4e^{\left(\frac{H_{ann}(\Lambda,\mathcal{P},2\ell)}{\ell} - \frac{\epsilon^2}{4}\right)\ell}$$

holds true, where random samples are drawn from the distribution $\mathcal{P}$.

2.1 Remarks

Our setting is the natural generalization of half-space learning applied to data on a $d-$dimensional sphere. In fact, when the sphere has radius τ, $\mathcal{C}_\tau$ corresponds to half-spaces, and the VC dimension is $d + 2$. However, when $\tau < \kappa$, as we show in Lemma 9, on a $d-$dimensional sphere of radius κ, the VC dimension of $\mathcal{C}_\tau$ is infinite. Interestingly, for these spheres, if $\tau > \kappa$, $\mathcal{C}_\tau$ contains only the function that always taked value 1 and the function that always takes value 0, since there are normals of length κ from center of the sphere to any point of a submanifold embedded in the sphere. In this case, the VC dimension is 1.

If the decision surface is not thin, but there is a margin within which misclassification in not penalized, our results can be adapted to show that the VC dimension is finite.

Our results pertain to the sample complexity of classification of smooth cuts, and does not address algorithmic issues. We are not aware of a way to generate arbitrary $\frac{1}{\tau}-$conditioned cuts. One direction towards addressing algorithmic issues would be to prove bounds on the annealed entropy of the family of linear classifiers in Gaussian Hilbert space. Since the Hilbert space of Gaussians with a fixed width has infinite VC dimension, distribution independent bounds cannot be found and annealed entropy could be a useful tool. Since SVMs based on Gaussian kernels are frequently used for classification, such a result would have algorithmic implications as well.

3 Learning Smooth Class Boundaries

Following Definition 2, let $\mathcal{C}_\tau$ be the collection of indicators of all open sets in $\mathcal{M}$ whose boundaries are $1/\tau$-conditioned submanifolds of $\mathbb{R}^m$ of dimension $d - 1$.
Our main theorem is the following.

Definition 6 (Packing number) *Let $N_p(\epsilon_r)$ be the largest number N such that $\mathcal{M}$ contains N disjoint balls $B_{\mathcal{M}}(x_i, \epsilon_r)$, where $B_{\mathcal{M}}(x, \epsilon_r)$ is a geodesic ball in $\mathcal{M}$ around x of radius ϵ_r.*

Notation 1 *Without loss of generality, let ρ_{max} be greater or equal to 1. Let $\epsilon_r = \min(\frac{\tau}{4}, \frac{\kappa}{4}, 1)\epsilon/(2\rho_{max})$. For some sufficiently large universal constant C, let*

$$\ell := C\left(\frac{\ln\frac{1}{\delta} + N_p(\epsilon_r/2)d\ln(d\rho_{max}/\epsilon)}{\epsilon^2}\right).$$

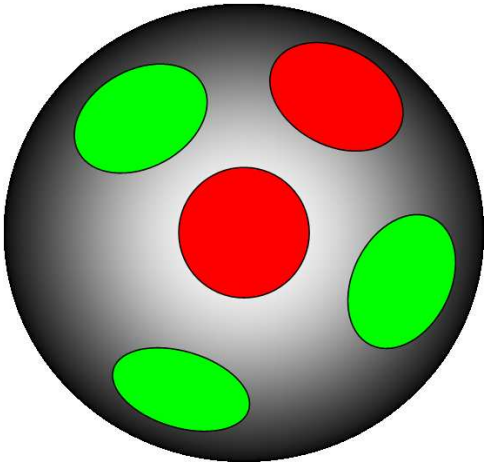

Figure 2: This illustrates the distribution from Lemma 8. The intersections of $f^{-1}(1)$ and $f^{-1}(0)$ with the support of $\mathcal{P}$ are respectively green and red.

Theorem 7 *Let $\mathcal{M}$ be a $d-$dimensional submanifold of $\mathbb{R}^m$ whose condition number is $\leq \frac{1}{\kappa}$. Let $\mathcal{P}$ be a probability measure on $\mathcal{M}$, whose density relative to the uniform probability measure on $\mathcal{M}$ is bounded above by ρ_{max}. Then the number of random samples needed before the empirical risk and the true risk are close uniformly over $\mathcal{C}_\tau$ can be bounded above as follows. Let ℓ be defined as in Notation 1. then*

$$\mathbb{P}\left[\sup_{\alpha\in\mathcal{C}_\tau}\frac{R(\alpha)-R_{emp}(\alpha,\ell)}{\sqrt{R(\alpha)}}>\sqrt{\epsilon}\right] \quad < \quad \delta$$

Proof: The proof follows from Theorem 11 and Theorem 5. The former provides a bound on the annealed entropy of $\mathcal{C}_\tau$ with respect to samples from $\mathcal{P}$. The latter relates the sample complexity of learning an element of a class of indicators such as $\mathcal{C}_\tau$ using random samples drawn from a distribution $\mathcal{P}$, to the annealed entropy of that class. ■

Lemma 8 provides a lower bound on the sample complexity that shows that some dependence on the packing number cannot be avoided in Theorem 7. Further, Lemma 9 shows that it is impossible to learn an element of $\mathcal{C}_\tau$ in a distribution-free setting in general.

Lemma 8 *Let $\mathcal{M}$ be a $d-$dimensional sphere in $\mathbb{R}^m$. Let the $\mathcal{P}$ have a uniform density over the disjoint union of $N_p(2\tau)$ identical spherical caps*

$$S=\{B_{\mathcal{M}}(x_i,\tau)\}_{1\leq i\leq N_p(2\tau)}$$

of radius τ, whose mutual distances are all $\geq 2\tau$. Then, if $s<(1-\epsilon)N_p(2\tau)$,

$$\mathbb{P}\left[\sup_{\alpha\in\mathcal{C}_\tau}\frac{R(\alpha)-R_{emp}(\alpha,s)}{\sqrt{R(\alpha)}}>\sqrt{\epsilon}\right] \quad = \quad 1.$$

Proof: Suppose that the labels are given by $f:\mathcal{M}\to\{0,1\}$, such that $f^{-1}(1)$ is the union of some of the caps in S as depicted in Figure 1. Suppose s random samples $z_1,\ldots,z_s$ are chosen from $\mathcal{P}$. Then at least $\epsilon N_p(2\tau)$ of the caps in S do not contain any of the z_i. Let X be the union of these caps. Let $\alpha:\mathcal{M}\to\{0,1\}$ satisfy $\alpha(x)=1-f(x)$ if $x\in X$ and $\alpha(x)=f(x)$ if $x\in\mathcal{M}\setminus X$. Note that $\alpha\in\mathcal{C}_\tau$. However, $R_{emp}(\alpha,s)=0$ and $R(\alpha)\geq\epsilon$. Therefore $\frac{R(\alpha)-R_{emp}(\alpha,s)}{\sqrt{R(\alpha)}}>\sqrt{\epsilon}$, which completes the proof. ■

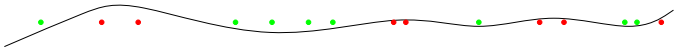

Figure 3: The $1-$dimensional submanifold P of $\mathbb{R}^2$ traced out by all points $(x,f_S(x))$.

Lemma 9 *For any $m>d\geq 2$, and $\tau>0$, there exist compact $d-$dimensional manifolds on which the VC dimension of $\mathcal{C}_\tau$ is infinite. In particular, this is true for the standard $d-$dimensional Euclidean sphere of radius κ embedded in $\mathbb{R}^m$, where $m>d\geq 2$ and $\kappa>\tau$.*

Proof: First consider the two dimensional plane $\mathbb{R}^2$. Suppose that for $i=0$ to $n-1$, $x_i=(i\tau/n,0)$. If there is no bound on the condition number, we make the following claim.

Claim 10 *For every subset $S\subseteq[n]$ there exists a boundary given by a graph $(x,f_S(x))$, $f_S:\mathbb{R}\to\mathbb{R}$ such that the following hold.*

1. *$f_S(x_i)>0$ if $i\in S$ (see Figure 2) and $f_S(x_i)<0$ if $i\in[n]\setminus S$.*
2. *f is thrice continuously differentiable.*
3. *For all $x\in\mathbb{R}$, $|f_S''(x)|<\frac{1}{\gamma}:=\frac{1}{2M\tau}$ for some large constant $M>>1$ and for all x such that $|x|\geq\tau$, $f_S(x)=0$.*

It is clear that for any S a function g_S exists that satisfy the first two conditions. We will use g_S to obtain f_S.

To see this, note that the radius of curvature at any point $(x,g_S(x))$ is given by $\frac{(1+g_S'(x)^2)^{\frac{3}{2}}}{|g_S''(x)|}$. Now, let

$$\alpha=\sup_{S\subseteq[n],x\in[-\tau,\tau]}\max(|g_S'(x)|,|g_S''(x)|).$$

Let $f_S(x)=\frac{g_S(x)}{\gamma\alpha}$. The $1-$dimensional submanifold P of $\mathbb{R}^2$ traced out by all points $(x,f_S(x))$ has curvature $\leq\frac{1}{\gamma}$ because for all $x\in[-\tau,\tau]$, for all S,

$$\frac{\alpha\gamma\left(1+\left(\frac{g_S'(x)}{\alpha\gamma}\right)^2\right)^{3/2}}{|g_S''(x)|}\geq\Omega(\gamma).$$

Let $S_\kappa^2=\{(x,y,z)\big|x^2+y^2+(z-\kappa)^2=\kappa^2\}$ be the $2-$sphere of radius $\kappa>\tau$ tangent to the (x,y) plane at the origin. Consider the stereographic projection υ_κ of $S_\kappa^2\setminus\{0,0,2\kappa\}$ onto $\mathbb{R}^2$ (embedded in $\mathbb{R}^3$), defined by

$$\upsilon_\kappa(x,y,z):=\left(\frac{2\kappa x}{2\kappa-z},\frac{2\kappa y}{2\kappa-z},0\right).$$

Let B be the ball of radius 1 centered at the origin in the image of υ_κ. As $M \to \infty$, $\upsilon_\kappa^{-1}(B \cap P)$ tends uniformly to a great circle, and its tangent spaces (see (1) tend uniformly to the corresponding tangent spaces of the great circle in terms of the angle. Therefore, (by (2)) for sufficiently large M, the condition number of $\upsilon_\kappa^{-1}(P)$ is less than $\frac{1}{\tau}$, completing the proof. This argument carries over to when $S_\kappa^2 \in \mathbb{R}^m$ for $m > 3$. Now, we may extend the copy of $\mathbb{R}^2$ that we considered to $\mathbb{R}^d$ by taking the canonical embedding $\mathbb{R}^2 \to \mathbb{R}^2 \times \mathbb{R}^{d-2}$. The $1-$dimensional manifold P can similarly extended to obtain a $m-1-$dimensional submanifold $P \times \mathbb{R}^{d-2}$. We can then consider as we did in the case of $\mathbb{R}^2$, the stereographic projection that maps the $d-$sphere

$$S_\kappa^d = \{(x, y, z_1, z_2, \ldots, z_{d-1}) \big| \\ x^2 + y^2 + z_1^2 + \ldots + (z_{d-1} - \kappa)^2 = \kappa^2\}$$

onto $\mathbb{R}^d$ by the map

$$\upsilon_\kappa(x, y, z_1, \ldots, z_{d-1}) := \\ \left(\frac{2\kappa x}{2\kappa - z_{d-1}}, \frac{2\kappa y}{2\kappa - z_{d-1}}, \frac{2\kappa z_1}{2\kappa - z_{d-1}}, \ldots, \frac{2\kappa z_{d-2}}{2\kappa - z_{d-1}}, 0\right),$$

and the same argument carries through. ■

We shall nonetheless uniformly bound from above, the annealed entropy of $\mathcal{C}_\tau$ with respect to any distribution $\mathcal{P}$ on $\mathcal{M}$, whose density (with respect to the uniform probability measure) on $\mathcal{M}$ is bounded above by ρ_{max}. The number of samples that need to be taken before the empirical risk is within ϵ of the true risk, uniformly over $\mathcal{C}_\tau$ with probability $1-\delta$ is determined by the annealed entropy of $\mathcal{C}_\tau$ w.r.t $\mathcal{P}$. We have the following theorem that bounds the annealed entropy from above.

Theorem 11 *Let $\mathcal{M}$ be a $d-$dimensional submanifold of $\mathbb{R}^m$ whose condition number is $\leq \frac{1}{\kappa}$. Let $\mathcal{P}$ be a probability measure on $\mathcal{M}$, whose density relative to the uniform probability measure on $\mathcal{M}$ is bounded above by ρ_{max}. When the number n of random samples from $\mathcal{P}$ is large, the annealed entropy of $\mathcal{C}_\tau$ can be bounded from above as follows. Let $\epsilon_r = \min(\frac{\tau}{4}, frac\kappa 4, 1)\epsilon/(2\rho_{max})$. Suppose*

$$n \geq N_p(\epsilon_r/2)\frac{d\ln(2\sqrt{d}\rho_{max}^2/\epsilon)}{\epsilon^2},$$

then,

$$H_{ann}(\mathcal{C}_\tau, \mathcal{P}, \lfloor n - \sqrt{n\ln(2\pi n)}\rfloor) \quad \leq \quad 4\epsilon n + 1.$$

3.1 Overview of the Proof of Theorem 11

Our strategy is as follows.

1. Cut the manifold into small pieces $\mathcal{M}_i$ that are almost Euclidean, such that the restrictions of any cut hypersurface is almost linear.
2. Let the probability measure $\frac{\mathcal{P}|_{\mathcal{M}_i}}{\mathcal{P}(\mathcal{M}_i)}$ be denoted $\mathcal{P}_i$ for each i. Lemma 18 allows us to show, roughly, that
$$\frac{H_{ann}(\mathcal{C}_\tau, \mathcal{P}, n)}{n} \lesssim \sup_i \frac{H_{ann}(\mathcal{C}_\tau, \mathcal{P}_i, \lfloor n\mathcal{P}(\mathcal{M}_i)\rfloor)}{\lfloor n\mathcal{P}(\mathcal{M}_i)\rfloor},$$
thereby allowing us to focus on a single piece $\mathcal{M}_i$.
3. We use a projection π_i, to map $\mathcal{M}_i$ orthogonally onto the tangent space to $\mathcal{M}_i$ at a point $x_i \in \mathcal{M}_i$ and then reduce the question to a sphere inscribed in a cube $\square$ of Euclidean space.
4. We cover $\mathcal{C}_\tau\big|_\square$ by the union of classes of functions that are constant outside a thin slab (see Definition 20 and Figure 3).
5. Finally, we bound the annealed entropy of each of these classes using Lemma 21.

The rest of this chapter is devoted to a detailed treatment of the proof of Theorem 11.

3.2 Volumes of balls in a manifold

Let $\mathcal{M} \subseteq \mathbb{R}^m$ be a d-dimensional Riemannian manifold and let P be a $d-1-$dimensional submanifold of $\mathcal{M}$. Let $V_x^{\mathcal{M}}(r)$ be defined to be the volume of a ball of radius r (in the intrinsic metric) around a point $x \in \mathcal{M}$. The sectional curvature of a manifold at a point x depends on a two-dimensional plane in the tangent space at x. A formal definition of sectional curvature can be found in most textbooks of differential geometry (for example, [5]). The volumes of balls can be estimated using sectional curvatures. The Bishop-Günther inequalities tell us that if the sectional curvature $K^{\mathcal{M}}$ is upper bounded by λ, then the volume of the ball of radius r around x, $V_x^{\mathcal{M}}$ is bounded from below as follows (section 3.5, [4]).

$$V_x^{\mathcal{M}}(r) \geq \frac{2\pi^{d/2}}{\Gamma(\frac{d}{2})}\int_0^r \left(\frac{\sin(t\sqrt{\lambda})}{\sqrt{\lambda}}\right)^{d-1} dt,$$

where $\Gamma(x)$ is Euler's Γ function.

This allows us to get an explicit upper bound on the packing number $N_p(\epsilon_r/2)$, namely

$$N_p(\epsilon_r/2) \leq \frac{\text{vol}\mathcal{M}}{\frac{2\pi^{d/2}}{\Gamma(\frac{d}{2})}\int_0^{\epsilon_r/2}\left(\frac{\sin(t\sqrt{\lambda})}{\sqrt{\lambda}}\right)^{d-1} dt}.$$

3.3 Partitioning the Manifold

The next step is to partition the manifold $\mathcal{M}$ into disjoint pieces $\{\mathcal{M}_i\}$ such that each piece $\mathcal{M}_i$ is contained in the geodesic ball $B_{\mathcal{M}}(x_i, \epsilon_r)$. Such a partition can be constructed by the following natural greedy procedure.

- Choose $N_p(\epsilon_r/2)$ disjoint balls $B_{\mathcal{M}}(x_i, \epsilon_r/2)$, $1 \leq i \leq N_p(\epsilon_r/2)$ where $N_p(\epsilon_r/2)$ is the packing number as in Definition 6.
- Let $\mathcal{M}_1 := B_{\mathcal{M}}(x_1, \epsilon_r)$.
- Iteratively, for each $i \geq 2$, let $\mathcal{M}_i := B_{\mathcal{M}}(x_i, \epsilon_r) \setminus \{\cup_{k=1}^{i-1}\mathcal{M}_k\}$.

3.4 Constructing charts by projecting onto Euclidean Balls

In this section, we show how the question can be reduced to Euclidean space using a family of charts. The strategy is the following. Let ϵ_r be as defined in Notation 1. Choose a set of points $X = \{x_1, \ldots, x_N\}$ belonging to $\mathcal{M}$ such that

the union of geodesic balls in $\mathcal{M}$ (measured in the intrinsic Riemannian metric) of radius ϵ_r centered at these points in $\mathcal{M}$ covers all of $\mathcal{M}$.

$$\bigcup_{i\in[N]} B_{\mathcal{M}}(x_i,\epsilon_r) = \mathcal{M}.$$

Definition 12 *For each $i \in [N_p(\epsilon_r/2)]$, let the $d-$dimensional affine subspace of $\mathbb{R}^m$ tangent to $\mathcal{M}$ at x_i be denoted $\mathbb{A}_i$, and let the d-dimensional ball of radius ϵ_r contained in $\mathbb{A}_i$, centered at x_i be $B_{\mathbb{A}_i}(x_i,\epsilon_r)$. Let the orthogonal projection from $\mathbb{R}^m$ onto $\mathbb{A}_i$ be denoted π_i.*

Lemma 13 *The image of $B_{\mathcal{M}}(x_i,\epsilon_r)$ under the projection π_i is contained in the corresponding ball $B_{\mathcal{M}}(x_i,\epsilon_r)$ in $\mathbb{A}_i$.*

$$\pi_i(B_{\mathcal{M}}(x_i,\epsilon_r)) \subseteq B_{\mathbb{A}_i}(x_i,\epsilon_r).$$

Proof: This follows from the fact that the length of a geodesic segment on $B_{\mathcal{M}}(x_i,\epsilon_r)$ is greater or equal to the length of its image under a projection. ■

Let P be a smooth $1/\tau$-conditioned boundary (i. e. $c(P) \leq \frac{1}{\tau}$) separating $\mathcal{M}$ into two parts. and $c(\mathcal{M}) \leq \frac{1}{\kappa}$.

Lemma 14 *Let $\epsilon_r \leq \min(1,\tau/4,\kappa/4)$. Let $\pi_i(B_{\mathcal{M}}(x_i,\epsilon_r)\cap P)$ be the image of P restricted to $B_{\mathcal{M}}(x_i,\epsilon_r)$ under the projection π_i. Then, the condition number of $\pi_i(B_{\mathcal{M}}(x_i,\epsilon_r)\cap P)$ is bounded above by $\frac{2}{\tau}$.*

Proof:

Let $T_{\pi_i(x)}$ and $T_{\pi_i(y)}$ be the spaces tangent to $\pi_i(B_{\mathcal{M}}(x_i,\epsilon_r)\cap P)$ at $\pi_i(x)$ and $\pi_i(y)$ respectively. Then, for any $x,y \in B_{\mathcal{M}}(x_i,\epsilon_r)\cap P$, because the kernel of π_i is nearly orthogonal to $T_{\pi_i(x)}$ and $T_{\pi_i(y)}$,

$$\sphericalangle(T_{\pi_i(x)},T_{\pi_i(y)}) \leq \sqrt{2}\sphericalangle(T_x,T_y). \tag{3}$$

$B_{\mathcal{M}}(x_i,\epsilon_r)\cap P$ is contained in a neighborhood of the affine space tangent to $B_{\mathcal{M}}(x_i,\epsilon_r)\cap P$ at x_i, which is orthogonal to the kernel of π_i. After some calculation, this can be used to show that for all $x,y \in B_{\mathcal{M}}(x_i,\epsilon_r)\cap P$,

$$\frac{1}{\sqrt{2}} \leq \frac{\|\pi_i(x)-\pi_i(y)\|}{\|x-y\|} \leq 1. \tag{4}$$

The lemma follows from (2). ■

3.5 Proof of Theorem 11

We shall organize this proof into several Lemmas, which will be proved immediately after their respective statements. The following Lemma allows us to work with a random rather than deterministic number of samples. The purpose of allowing the number of samples to be a Poisson random variable is that we are able make the set of numbers of samples $\{\nu_i\}$ from different $\mathcal{M}_i$, a collection of independent random variables.

Lemma 15 (Poissonization) *Let ν be a Poisson random variable with mean λ, where $\lambda > 0$. Then, for any $\epsilon > 0$ the expected value of the annealed entropy of a class of indicators with respect to ν random samples from a distribution $\mathcal{P}$ is asymptotically greater or equal to the annealed entropy of $\lfloor(1-\epsilon)\lambda\rfloor$ random samples from the distribution $\mathcal{P}$. More precisely, for any $\epsilon > 0$, $\ln \mathbb{E}_\nu G(\Lambda,\mathcal{P},\nu) \geq \ln G(\Lambda,\mathcal{P},\lfloor\lambda(1-\epsilon)\rfloor) - \exp\left(-\epsilon^2\lambda + \frac{\ln(2\pi\lambda)}{2}\right)$.*

Proof:

$$\begin{aligned}\ln \mathbb{E}_\nu G(\Lambda,\mathcal{P},\nu) &= \ln \sum_{n\in\mathbb{N}} \mathbb{P}[\nu=n] H_{ann}(\Lambda,\mathcal{P},n) \\ &\geq \ln \sum_{n\geq\lfloor\lambda(1-\epsilon)\rfloor} \mathbb{P}[\nu=n] G(\Lambda,\mathcal{P},n).\end{aligned}$$

$G(\Lambda,\mathcal{P},n)$ is monotonically increasing as a function of n. Therefore the above expression can be lower bounded by $\ln \mathbb{P}[\nu \geq \lfloor\lambda(1-\epsilon)\rfloor] G(\Lambda,\mathcal{P},\nu) \geq H_{ann}(\Lambda,\mathcal{P},\lfloor\lambda(1-\epsilon)\rfloor) - \exp\left(-\epsilon^2\lambda + \frac{\ln(2\pi\lambda)}{2}\right)$. ■

Definition 16 *For each $i \in [N_p(\epsilon_r/2)]$, let $\mathcal{P}_i$ be the restriction of $\mathcal{P}$ to $\mathcal{M}_i$. Let $|\mathcal{P}_i|$ denote the total measure of $\mathcal{P}_i$. Let λ_i denote $\lambda|\mathcal{P}_i|$. Let $\{\nu_i\}$ be a collection of independent Poisson random variables such that for each $i \in [N_p(\epsilon_r/2)]$, the mean of ν_i is λ_i.*

The following Lemma allows us to focus our attention to small pieces $\mathcal{M}_i$ which are almost Euclidean.

Lemma 17 (Factorization) *The quantity $\ln \mathbb{E}_\nu G(\mathcal{C}_\tau,\mathcal{P},\nu)$ is less or equal to the sum over i of the corresponding quantities $\mathcal{C}_\tau$ with respect to ν_i random samples from $\mathcal{P}_i$. i. e.*

$$\ln \mathbb{E}_\nu G(\mathcal{C}_\tau,\mathcal{P},\nu) \leq \sum_{i\in N_p(\epsilon_r/2)} \ln \mathbb{E}_{\nu_i} G(\mathcal{C}_\tau,\mathcal{P}_i,\nu_i).$$

Proof:

$$G(\mathcal{C}_\tau,\mathcal{P},\ell) := \ln \mathbb{E}_{X\vdash\mathcal{P}^{\times\ell}} N(\mathcal{C}_\tau,X),$$

where expectation is with respect to X and $\vdash$ signifies that X is drawn from the Cartesian product of ℓ copies of $\mathcal{P}$. The number of ways of splitting $X = \{x_1,\ldots,x_k,\ldots,x_\ell\}$ using elements of $\mathcal{C}_\tau$, $N(\mathcal{C}_\tau,X)$ satisfies a sub-multiplicative property, namely

$$N(\mathcal{C}_\tau,\{x_1,\ldots,x_\ell\}) \leq$$
$$N(\mathcal{C}_\tau,\{x_1,\ldots,x_k\})N(\mathcal{C}_\tau,\{x_{k+1},\ldots,x_\ell\}).$$

This can be iterated to generate inequalities where the right side involves a partition with any integer number of parts. Note that $\mathcal{P}$ is a mixture of the $\mathcal{P}_i$, and can be expressed as

$$\mathcal{P} = \sum_i \frac{\lambda_i}{\lambda}\mathcal{P}_i.$$

A draw from $\mathcal{P}$ of a Poisson number of samples can be decomposed as the union of independently chosen sets of samples. The i^{th} set is a draw of size ν_i from $\mathcal{P}_i$, ν_i being a Poisson random variable having mean λ_i. These facts imply that
$\ln \mathbb{E}_\nu G(\mathcal{C}_\tau,\mathcal{P},\nu) \leq \sum_{i\in N_p(\epsilon_r/2)} \ln \mathbb{E}_{\nu_i} G(\mathcal{C}_\tau,\mathcal{P}_i,\nu_i)$.

■

Lemma 17 can be used together with an upper bound on annealed entropy based on the number of samples to obtain

Lemma 18 (Localization) *For any $\epsilon' > 0$*

$$\frac{\ln \mathbb{E}_\nu G(\mathcal{C}_\tau,\mathcal{P},\nu)}{\lambda} \leq \sup_{i\ s.t\ |\mathcal{P}_i|\geq\frac{\epsilon'}{N_p(\epsilon_r/2)}} \frac{\ln \mathbb{E}_{\nu_i} G(\mathcal{C}_\tau,\mathcal{P}_i,\nu_i)}{\lambda_i} + \epsilon'.$$

Proof: Lemma 18 allows us to reduce the question to a single $\mathcal{M}_i$ in the following way.

$$\frac{\ln \mathbb{E}_\nu G(\mathcal{C}_\tau, \mathcal{P}, \nu)}{\lambda} \leq \sum_{i \in N_p(\epsilon_r/2)} \frac{\lambda_i}{\lambda} \frac{\ln \mathbb{E}_{\nu_i} G(\mathcal{C}_\tau, \mathcal{P}_i, \nu_i)}{\lambda_i}$$

Allowing all summations to be over i s.t $|\mathcal{P}_i| \geq \frac{\epsilon'}{N_p(\epsilon_r/2)}$, the right side can be split into

$$\sum_i \frac{\lambda_i}{\lambda} \frac{\ln \mathbb{E}_{\nu_i} G(\mathcal{C}_\tau, \mathcal{P}_i, \nu_i)}{\lambda_i} + \sum_i \ln \mathbb{E}_{\nu_i} G(\mathcal{C}_\tau, \mathcal{P}_i, \nu_i).$$

$G(\mathcal{C}_\tau, \mathcal{P}_i, \nu_i)$ must be less or equal to the expression obtained in the case of complete shattering, which is 2^{ν_i}. Therefore the second term in the above expression can be bounded above as follows,

$$\begin{aligned} \sum_i \ln \mathbb{E}_{\nu_i} G(\mathcal{C}_\tau, \mathcal{P}_i, \nu_i) &\leq \sum_i \ln \mathbb{E}_{\nu_i} 2^{\nu_i} \\ &= \sum_i \lambda_i \\ &\leq \epsilon'. \end{aligned}$$

Therefore,

$$\begin{aligned} \frac{\ln \mathbb{E}_\nu G(\mathcal{C}_\tau, \mathcal{P}, \nu)}{\lambda} &\leq \sum_i \frac{\lambda_i}{\lambda} \frac{\ln \mathbb{E}_{\nu_i} G(\mathcal{C}_\tau, \mathcal{P}_i, \nu_i)}{\lambda_i} + \epsilon' \\ &\leq \sup_i \frac{\ln \mathbb{E}_{\nu_i} G(\mathcal{C}_\tau, \mathcal{P}_i, \nu_i)}{\lambda_i} + \epsilon'. \end{aligned}$$

■

As mentioned earlier, Lemma 18 allows us to reduce the proof to a question concerning a single piece $\mathcal{M}_i$. This is more convenient because $\mathcal{M}_i$ can be projected onto a single Euclidean ball in the way described in Section 3.4 without incurring significant distortion. By Lemmas 13 and 14, the question can be transferred to one about the annealed entropy of the induced function class $\mathcal{C}_\tau \circ \pi_i^{-1}$ on chart $B_{\mathbb{A}_i}(x_i, \epsilon_r)$ with respect to ν_i random samples from the projected probability distribution $\pi_i(\nu_i)$. $\mathcal{C}_\tau \circ \pi_i^{-1}$ is contained in $\mathcal{C}_{\tau/2}(\mathbb{A}_i)$ which is the analogue of $\mathcal{C}_{\tau/2}$ on $\mathbb{A}_i$. For simplicity, henceworth we shall abbreviate $\mathcal{C}_{\tau/2}(\mathbb{A}_i)$ as $\mathcal{C}_{\tau/2}$. Then,

$$\begin{aligned} \frac{\ln \mathbb{E}_{\nu_i} G(\mathcal{C}_\tau, \mathcal{P}_i, \nu_i)}{\lambda_i} &= \frac{\ln \mathbb{E}_{\nu_i} G(\mathcal{C}_\tau \circ \pi_i^{-1}, \pi_i(\mathcal{P}_i), \nu_i)}{\lambda_i} \\ &\leq \frac{\ln \mathbb{E}_{\nu_i} G(\mathcal{C}_{\tau/2}, \pi_i(\mathcal{P}_i), \nu_i)}{\lambda_i}. \end{aligned}$$

We inscribe $B_{\mathbb{A}_i}(x_i, \epsilon_r)$ in a cube of side $2\epsilon_r$ for convenience, and proceed to find the desired upper bound on $G(\mathcal{C}_{\tau/2}, \pi_i(\mathcal{P}_i), \nu_i)$. We shall indicate how to achieve this using covers. For convenience, let this cube be dilated until we have the cube of side 2. The measure $\pi_i(\mathcal{P}_i)$ assigns to it must be scaled to a probability measure that we call $\mathcal{P}_\circ$, which is actually supported on the inscribed ball. We shall normalize all quantities appropriately when the calculations are over. The $\tau_\square$ that we shall work with below is a rescaled version of the original, $\tau_\square = \tau/\epsilon_r$. Let B_∞^d be the cube of side 2 centered at the origin and ι_∞^d be its indicator. Let B_2^d be the unit ball inscribed in B_∞^d.

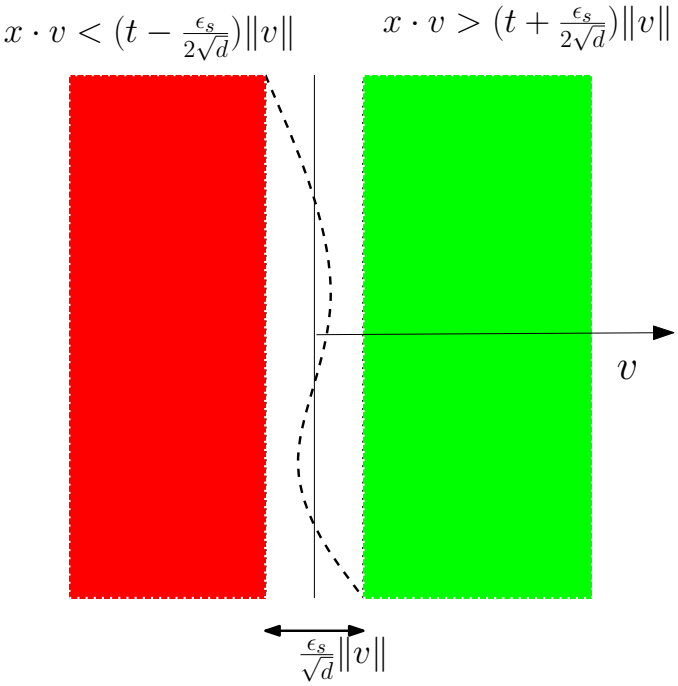

Figure 4: Each class of the form $\tilde{\mathcal{C}}_{\epsilon_s}^{(v,t)}$ contains a subset of the set of indicators of the form $\mathcal{I}_c \cdot \iota_\infty^d$

Definition 19 *Let $\tilde{C_{\tau_\square}}$ be defined to be the set of all indicators of the form $\iota_\infty^d \cdot \iota$, where ι is the indicator of some set in $\mathcal{C}_{\tau_\square}$.*

In other words, $\tilde{\mathcal{C}}_{\tau_\square}$ is the collection of all functions that are indicators of sets that can be expressed as the intersection of the unit cube and an element of $\mathcal{C}_{\tau_\square}$.

$$\tilde{C}_{\tau_\square} = \{f \,|\, \exists c \in \mathcal{C}_{\tau_\square}, \text{ for which } f = \mathcal{I}_c \cdot \iota_\infty^d\}, \tag{5}$$

where $\mathcal{I}_c$ is the indicator of c.

Definition 20 *For every $v \in \mathbb{R}^d$ where $\|v\| = 1$, $t \in \mathbb{R}$ and $\epsilon > 0$ and $\epsilon_s = \epsilon^2/\rho_{max}$. Let $\tilde{\mathcal{C}}_{\epsilon_s}^{(v,t)}$ be a class of indicator functions consisting of all those measurable indicators ι that satisfy the following.*

1. $x \cdot v < (t - \frac{\epsilon_s}{2\sqrt{d}})\|v\|$ *or* $x \notin B_\infty^d \Rightarrow \iota(x) = 0$ *and*
2. $x \cdot v > (t + \frac{\epsilon_s}{2\sqrt{d}})\|v\|$ *and* $x \in B_\infty^d \Rightarrow \iota(x) = 1$.

The VC dimension of the above class is clearly infinite since any samples lying within the slab of thickness $\epsilon_s/\sqrt{d}$ get shattered. However if a distribution is sufficiently uniform, most samples would lie outside the slab and so the annealed entropy can be bounded from above. We shall construct a finite set W of tuples (v, t) such that the union of the corresponding classes $\tilde{C}_{\epsilon_s}^{(v,t)}$ contains $\tilde{\mathcal{C}}_{\tau_\square}$. Let tv take values in an $\frac{\tau_\square}{2}$-grid contained in B_∞^d, i. e. $tv \in \frac{\epsilon_s}{2\sqrt{d}}\mathbb{Z}^d \cap B_\infty^d$. It is then the case (see Figure 3) that any indicator in $\tilde{\mathcal{C}}_{\tau_\square}$ agrees over B_2^d with a member in some class $\tilde{C}_{\epsilon_s}^{(v,t)}$, if $\epsilon_s \geq \frac{2}{\tau_\square}$, i. e.

$$\tilde{\mathcal{C}}_{\tau_\square} \subseteq \bigcup_{tv \in \frac{\epsilon_s}{2\sqrt{d}}\mathbb{Z}^d \cap B_\infty^d} \tilde{C}_{\epsilon_s}^{(v,t)}.$$

A bound on the volume of the band where $(t-\frac{\epsilon_s}{2\sqrt{d}})\|v\| < x \cdot v < (t+\frac{\epsilon_s}{2\sqrt{d}})\|v\|$ in B_2^d follows from the fact that the maximum volume hyperplane section is a bisecting hyperplane, whose volume is $< 2\sqrt{d}\,\mathrm{vol}(B_2^d)$.

This allows us to bound the annealed entropy of a single class $\tilde{C}_{\epsilon_s}^{(v,t)}$ in the following lemma, where ρ_{max} is the same maximum density with respect to the uniform density on B_2^d. (Re-scaling was unnecessary because that was with respect to the Lebesgue measure normalized to be a probability measure).

Lemma 21 *The logarithm of the expected growth function of a class $\tilde{C}_{\epsilon_s}^{(v,t)}$ with respect to $\nu_\circ$ random samples from $\mathcal{P}_\circ$, is $< 2\epsilon_s\rho_{max}\lambda_\circ$, where $\nu_\circ$ is a Poisson random variable of mean $\lambda_\circ$; i. e.*

$$\ln \mathbb{E}_{\nu_\circ} G(\mathcal{C}_{\tau_\square}, \mathcal{P}_\circ, \nu_\circ) < 2\epsilon_s\rho_{max}\lambda_\circ.$$

Proof: A bound on the volume of the band where $(t - \frac{\epsilon_s}{2\sqrt{d}})\|v\| < x \cdot v < (t + \frac{\epsilon_s}{2\sqrt{d}})\|v\|$ in B_2^d follows from the fact that the maximum volume hyperplane section is a bisecting hyperplane, whose $d-1$-dimensional volume is $< 2\sqrt{d}\,\mathrm{vol}(B_2^d)$. Therefore, the number of samples that fall in this band is a Poisson random variable whose mean is less than $2\epsilon_s\rho_{max}\lambda_\circ$. This implies the Lemma. ■

Therefore the expected annealed entropy of

$$\bigcup_{tv \in \frac{\epsilon_s}{2\sqrt{d}}\mathbb{Z}^d \cap B_\infty^d} \tilde{C}_{\epsilon_s}^{(v,t)}$$

with respect to $\nu_\circ$ random samples from $\mathcal{P}_\circ$ is bounded above by $2\epsilon_s\rho_{max}\lambda_\circ + \ln|\frac{\epsilon}{2\sqrt{d}}\mathbb{Z}^d \cap B_\infty^d|$. Putting these observations together,

$$\begin{aligned}\ln \mathbb{E}_\nu G\left(\mathcal{C}_\tau, \mathcal{P}, \nu\right)/\lambda &\leq \frac{\ln \mathbb{E}_{\nu_\circ} G(\mathcal{C}_{\tau_\square}, \mathcal{P}_\circ, \nu_\circ)}{\lambda_\circ} + \epsilon \\ &\leq 2\epsilon_s\rho_{max} + \frac{d\ln(2\sqrt{d}/\epsilon_s)}{\lambda_\circ} + \epsilon\end{aligned}$$

We know that $\lambda_\circ N_p(\epsilon_r/2) \geq \epsilon\lambda$. Then,

$$2\epsilon_s\rho_{max} + \frac{d\ln(2\sqrt{d}/\epsilon_s)}{\lambda_\circ} + \epsilon \leq$$

$$2\epsilon + N_p(\epsilon_r/2)\frac{d\ln(2\sqrt{d}\rho_{max}/\epsilon_s)}{\epsilon\lambda} + \epsilon,$$

which is

$$\leq 2\epsilon + N_p(\epsilon_r/2)\frac{d\ln(2\sqrt{d}\rho_{max}^2/\epsilon)}{\epsilon\lambda} + \epsilon.$$

Therefore, if $\lambda \geq N_p(\epsilon_r/2)\frac{d\ln(2\sqrt{d}\rho_{max}^2/\epsilon)}{\epsilon^2}$, then,

$$\ln \mathbb{E}_\nu G\left(\mathcal{C}_\tau, \mathcal{P}, \nu\right)/\lambda \leq 4\epsilon.$$

Together with Lemma 15, this shows that for any $\epsilon_1 > 0$, if

$$\lambda \geq N_p(\epsilon_r/2)\frac{d\ln(2\sqrt{d}\rho_{max}^2/\epsilon)}{\epsilon^2},$$

then

$$\begin{aligned}H_{ann}(\Lambda, \mathcal{P}, \lfloor\lambda(1-\epsilon_1)\rfloor) &\leq \ln \mathbb{E}_\nu G(\Lambda, \mathcal{P}, \nu) \\ &+ \exp\left(-\epsilon_1^2\lambda + \frac{\ln(2\pi\lambda)}{2}\right) \\ &\leq 4\epsilon\lambda + \exp\left(-\epsilon_1^2\lambda + \frac{\ln(2\pi\lambda)}{2}\right).\end{aligned}$$

Setting ϵ_1 to $\sqrt{\frac{\ln(2\pi\lambda)}{\lambda}}$, $\exp\left(-\epsilon_1^2\lambda + \frac{\ln(2\pi\lambda)}{2}\right)$ is less than 1. Therefore,

$$H_{ann}(\Lambda, \mathcal{P}, \lfloor\lambda - \sqrt{\lambda\ln(2\pi\lambda)}\rfloor) \leq 4\epsilon\lambda + 1.$$

This completes the proof of Theorem 11.

4 Acknowledgements

We are grateful to the anonymous referees for pointing out an oversight in Lemma 9.

References

[1] M. Belkin, P. Niyogi, *Laplacian Eigenmaps for Dimensionality Reduction and Data Representation*, Neural Computation, June 2003.

[2] R.R. Coifman, S. Lafon, *Diffusion maps*, Applied and Computational Harmonic Analysis: Special issue on Diffusion Maps and Wavelets, Vol 21, July 2006, pp 5-30.

[3] R.M. Dudley, S.R. Kulkarni, T.J. Richardson, O. Zeitouni, *A Metric Entropy Bound is Not Sufficient for Learnability*, IEEE Transactions on Information Theory, Vol. 40, No. 3, pp. 883-885, May, 1994.

[4] A. Gray, *Tubes*, Birkhauser Verlag, 2004

[5] S. Kobayashi and K. Nomizu, *Foundations of differential geometry*, Interscience Tracts in Pure and Applied Math., No. 15, 1963

[6] V. Vapnik, *Statistical Learning Theory*, Wiley 1998.