

Manifold Regularization and Semi-supervised Learning: Some Theoretical Analyses

Partha Niyogi*

January 22, 2008

Abstract

Manifold regularization (Belkin, Niyogi, Sindhwani, 2004) is a geometrically motivated framework for machine learning within which several semi-supervised algorithms have been constructed. Here we try to provide some theoretical understanding of this approach. Our main result is to expose the natural structure of a class of problems on which manifold regularization methods are helpful. We show that for such problems, no supervised learner can learn effectively. On the other hand, a manifold based learner (that knows the manifold or “learns” it from unlabeled examples) can learn with relatively few labeled examples. Our analysis follows a minimax style with an emphasis on finite sample results (in terms of n : the number of labeled examples). These results allow us to properly interpret manifold regularization and its potential use in semi-supervised learning.

1 Introduction

A learning problem is specified by a probability distribution p on $X \times Y$ according to which labelled examples $z_i = (x_i, y_i)$ pairs are drawn and presented to a learning algorithm (estimation procedure). We are interested in an understanding of the case in which $X = \mathbb{R}^D$,

*Departments of Computer Science, Statistics, The University of Chicago

$Y \subset \mathbb{R}$ but p_X (the marginal distribution of p on X) is supported on some submanifold $\mathcal{M} \subset X$. In particular, we are interested in understanding how knowledge of this submanifold may potentially help a learning algorithm¹. To this end, we will consider two kinds of learning algorithms:

1. Algorithms that have no knowledge of the submanifold \mathcal{M} but learn from (x_i, y_i) pairs in a purely supervised way.
2. Algorithms that have perfect knowledge of the submanifold. This knowledge may be acquired by a manifold learning procedure through unlabeled examples x_i 's and having access to an essentially infinite number of them. Such a learner may be viewed as a semi-supervised learner.

Our main result is to elucidate the structure of a class of problems on which there is a difference in the performance of algorithms of Type 1 and 2.

Let \mathcal{P} be a collection of probability distributions p and thus denote a class of learning problems. For simplicity and ease of comparison with other classical results, we place some regularity conditions on \mathcal{P} . Every $p \in \mathcal{P}$ is such that its marginal p_X has support on a k -dimensional manifold $\mathcal{M} \subset X$. Different p 's may have different supports. For simplicity, we will consider the case where p_X is uniform on \mathcal{M} : this corresponds to a situation in which the marginal is the most regular.

Given such a \mathcal{P} we can naturally define the class $\mathcal{P}_{\mathcal{M}}$ to be

$$\mathcal{P}_{\mathcal{M}} = \{p \in \mathcal{P} \mid p_X \text{ is uniform on } \mathcal{M}\}$$

Clearly, we have

$$\mathcal{P} = \cup_{\mathcal{M}} \mathcal{P}_{\mathcal{M}}$$

Consider $p \in \mathcal{P}_{\mathcal{M}}$. This denotes a learning problem and the regression function m_p is defined as

$$m_p(x) = E[y|x] \text{ when } x \in \mathcal{M}$$

¹Some recent results (e.g. Bickel and Li, 2006 and Lafferty and Wasserman, 2007) have noted that in the setting in which data lives on a lower dimensional manifold, it may be possible to achieve good rates of learning using classical methods that are suitably adapted without actually knowing very much about the manifold in question. This observation, in part, motivates the question: are there cases when one needs to know the manifold in order to reap the full benefit of the manifold structure? This paper is a response to this question.

Note that $m_p(x)$ is not defined outside of \mathcal{M} . We will be interested in cases when m_p belongs to some restricted family of functions $H_{\mathcal{M}}$ (for example, a Sobolev space). Thus assuming a family $H_{\mathcal{M}}$ is equivalent to assuming a restriction on the class of conditional probability distributions $p(y|x)$ where $p \in \mathcal{P}$. For simplicity, we will assume the noiseless case where $p(y|x)$ is the probability mass function concentrated on $m_p(x)$, i.e., there is no noise in the Y space.

Since $X \setminus \mathcal{M}$ has measure zero (with respect to p_X), we can define $m_p(x)$ to be anything we want when $x \in X \setminus \mathcal{M}$. We define $m_p(x) = 0$ when $x \notin \mathcal{M}$.

For a learning problem p , the learner is presented with a collection of labeled examples $\{z_i = (x_i, y_i), i = 1, \dots, n\}$ where each z_i is drawn i.i.d. according to p . A learning algorithm A maps the collection of data $\bar{z} = (z_1, \dots, z_n)$ into a function $A(\bar{z})$. Now we can define the following minimax rate (for the class \mathcal{P}) as

$$R(n, \mathcal{P}) = \inf_A \sup_{p \in \mathcal{P}} E_{\bar{z}} \|A(\bar{z}) - m_p\|_{L^2(p_X)}$$

This is the rate achieved by a learner that has *no knowledge of the manifold* \mathcal{M} . We will contrast it with a learner that has oracle access endowing it with knowledge of the manifold.

To begin, note that since $\mathcal{P} = \cup_{\mathcal{M}} \mathcal{P}_{\mathcal{M}}$, we see that

$$R(n, \mathcal{P}) = \inf_A \sup_{\mathcal{M}} \sup_{p \in \mathcal{P}_{\mathcal{M}}} E_{\bar{z}} \|A(\bar{z}) - m_p\|_{L^2(p_X)}$$

By commuting \inf_A and $\sup_{\mathcal{M}}$, we obtain

$$Q(n, \mathcal{P}) = \sup_{\mathcal{M}} \inf_A \sup_{p \in \mathcal{P}_{\mathcal{M}}} E_{\bar{z}} \|A(\bar{z}) - m_p\|_{L^2(p_X)}$$

Note here that A is allowed knowledge of \mathcal{M} and may modify its estimate of the target function based on this knowledge. One may formally denote this dependence by writing $A(\bar{z}, \mathcal{M})$ but we let this dependence be implicit in our notation.

1.1 The Manifold Assumption for Semi-supervised Learning

So the question at hand is: for what class of problems \mathcal{P} with the structure as described above, might one expect a gap between $R(n, \mathcal{P})$

and $Q(n, \mathcal{P})$. This is a class of problems for which knowing the manifold confers an advantage to the learner.

The assumption behind the manifold based approach to semi-supervised learning is that the support of the probability distribution is on some low dimensional manifold. But this is not all. It is further assumed that harmonic analysis on this manifold makes sense, the eigenfunctions of the Laplacian on this manifold form a distinguished basis, and that the target function (regression function) has a compact representation in this basis in some sense. Thus, for a manifold \mathcal{M} , let ϕ_1, ϕ_2, \dots , be the eigenfunctions of the manifold Laplacian. Then, $m_p(x)$ may be related to this basis by

$$m_p = \sum_i \alpha_i \phi_i$$

or

$$m_p = \text{sign}\left(\sum_i \alpha_i \phi_i\right)$$

where the α_i 's have a sharp decay to zero. For the intuitions surrounding these ideas see the papers related to manifold regularization.

Thus, the space $H_{\mathcal{M}}$ is highly constrained. But knowledge of $H_{\mathcal{M}}$ (equivalently knowledge of $\mathcal{P}_{\mathcal{M}}$) only comes with knowledge of \mathcal{M} and knowledge of \mathcal{M} is obtained through unlabeled examples alone. It is possible, therefore, that a large number of unlabeled examples provide knowledge of \mathcal{M} after which a small number of labeled examples pin down the target function in $\mathcal{H}_{\mathcal{M}}$. On the other hand, with labeled examples alone, it may take a long time for the learner to acquire the target function. To illustrate this point, we will show a specific construction where we will see that $R(n) = \Omega(1)$ while $\lim_{n \rightarrow \infty} Q(n) = 0$ at a fast rate. Thus the learner with knowledge of the manifold learns easily while the learner with no knowledge cannot learn at all.

2 A Prototypical Example: Embeddings of the Circle into Euclidean Space

Let $\phi : S^1 \rightarrow X$ be an isometric embedding of the circle into Euclidean space. Now consider the family of such isometric embeddings and let this be the family of one-dimensional submanifolds that

we will deal with. Thus each $\mathcal{M} \subset X$ is of the form $\mathcal{M} = \phi(S^1)$ for some ϕ .

Let H_{S^1} be the set of functions defined on the circle that take the value $+1$ on half the circle and -1 on the other half. Thus in local coordinates (θ denoting the coordinate of a point in S^1), we can write the class H_{S^1} as

$$H_{S^1} = \{h_\alpha : S^1 \rightarrow \mathbb{R} \mid h_\alpha(\theta) = \text{sign}(\sin(\theta + \alpha)); \alpha \in [0, 2\pi)\}$$

Now for each $\mathcal{M} = \phi(S^1)$ we can define the class $H_{\mathcal{M}}$ as

$$H_{\mathcal{M}} = \{h : \mathcal{M} \rightarrow \mathbb{R} \mid h(x) = h_\alpha(\phi^{-1}(x)) \text{ for some } h_\alpha \in H_{S^1}\} \quad (1)$$

This defines a class of regression functions (also classification functions) for our setting. Correspondingly, in our noiseless setting, we can now define $\mathcal{P}_{\mathcal{M}}$ as follows. For each, $h \in H_{\mathcal{M}}$, we can define the probability distribution $p^{(h)}$ on $X \times Y$ by letting the marginal $p_X^{(h)}$ be uniform on \mathcal{M} and the conditional $p^{(h)}(y|x)$ be a probability mass function concentrated on two points $y = +1$ and $y = -1$ such that

$$p^{(h)}(y = +1|x) = 1 \iff h(x) = +1$$

Thus

$$\mathcal{P}_{\mathcal{M}} = \{p^{(h)} \mid h \in H_{\mathcal{M}}\}$$

In our setting, we can therefore interpret the learning problem as an instantiation either of regression or of classification based on our interest.

Now that $\mathcal{P}_{\mathcal{M}}$ is defined, the set $\mathcal{P} = \cup_{\mathcal{M}} \mathcal{P}_{\mathcal{M}}$ follows naturally. A picture of the situation is shown in fig. 1.

Remark. Note that H_{S^1} and correspondingly $H_{\mathcal{M}}$ as defined above is not a Sobolev space. However, it is obtained by thresholding functions in a Sobolev space. In particular, we can write

$$H_{S^1} = \{\text{sign}(h) \mid h = \alpha\phi + \beta\psi\}$$

where $\phi(\theta) = \sin(\theta)$ and $\psi(\theta) = \cos(\theta)$ are eigenfunctions of the Laplacian Δ_{S^1} on the circle. These are the eigenfunctions corresponding to $\lambda = 1$ and define the corresponding two dimensional eigenspace. More generally one could consider a family of functions obtained by thresholding functions in a Sobolev space of any chosen order and clearly H_{S^1} is contained in any such family. Finally it is worth noting that the arguments presented below do not depend on thresholding and would work with functions that are bandlimited or in a Sobolev space just as well.

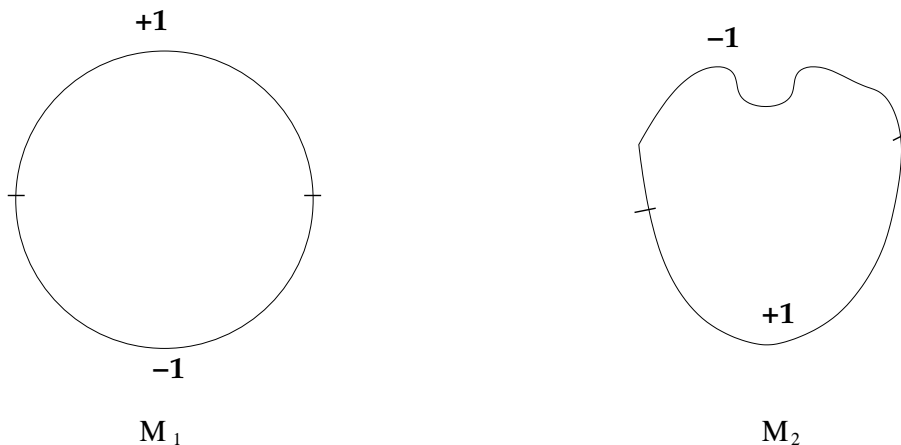


Figure 1: Shown are two embeddings M_1 and M_2 of the circle in Euclidean space (the plane in this case). The two functions, one from $M_1 \rightarrow \mathbb{R}$ and the other from $M_2 \rightarrow \mathbb{R}$ are denoted by labelings $+1, -1$ corresponding to half circles as shown.

2.1 Upper Bound on $Q(n, \mathcal{P})$

Let us begin by noting that if the manifold \mathcal{M} is known, the learner knows the class $\mathcal{P}_{\mathcal{M}}$. The learner merely needs to approximate one of the target functions in $H_{\mathcal{M}}$. It is clear that the space $H_{\mathcal{M}}$ is a family of $0 - 1$ valued functions whose VC-dimension is 2. Therefore, an algorithm that does empirical risk minimization over the class $H_{\mathcal{M}}$ will yield an upperbound of $O(\sqrt{\frac{\log(n)}{n}})$ by the usual arguments. Therefore the following theorem is obtained.

Theorem 2.1 *Following the notation of Sec. 1, let $H_{\mathcal{M}}$ be the family of functions defined by eq. 1 and \mathcal{P} be the corresponding family of learning problems. Then the learner with knowledge of the manifold converges at a fast rate given by*

$$Q(n, \mathcal{P}) \leq 2\sqrt{\frac{3 \log(n)}{n}}$$

and this rate is optimal.

Remark. If the class $H_{\mathcal{M}}$ is a parametric family of the form $\sum_{i=1}^p \alpha_i \phi_i$ where ϕ_i are the eigenfunctions of the Laplacian, one obtains the same parametric rate. Similarly, if the class $H_{\mathcal{M}}$ is a ball in a Sobolev

space of appropriate order, suitable rates on the family may be obtained by the usual arguments.

2.2 Lower Bound on $R(n, \mathcal{P})$

We now prove the following

Theorem 2.2 *Let $\mathcal{P} = \cup_{\mathcal{M}} \mathcal{P}_{\mathcal{M}}$ where each $\mathcal{M} = \phi(S^1)$ is an isometric embedding of the circle into X as shown. For each $p \in \mathcal{P}$, the marginal p_X is uniform on some \mathcal{M} and the conditional $p(y|x)$ is given by the construction in the previous section. Then*

$$R(n, \mathcal{P}) = \inf_A \sup_{p \in \mathcal{P}} E_{\bar{z}} \|A(\bar{z}) - m_p\|_{L^2(p_X)} = \Omega(1)$$

PROOF: Given n , choose a number $d = 2n$. Following Construction 1, there exist a set (denoted by $\mathcal{P}_d \subset \mathcal{P}$) of 2^d probability distributions that may be defined. Our proof uses the probabilistic method. We show that there exists a universal constant K (independent of n) such that

$$\forall A, \frac{1}{2^d} \sum_{p \in \mathcal{P}_d} E_{\bar{z}} \|A(\bar{z}) - m_p\|_{L^2(p_X)} \geq K$$

from which we conclude that

$$\forall A, \sup_{p \in \mathcal{P}_d} E_{\bar{z}} \|A(\bar{z}) - m_p\|_{L^2(p_X)} \geq K$$

Since $\mathcal{P}_d \subset \mathcal{P}$, the result follows.

To begin, consider a $p \in \mathcal{P}$. Let $\bar{z} = (z_1, \dots, z_n)$ be a set of i.i.d. examples drawn according to p . Note that this is equivalent to drawing $\bar{x} = (x_1, \dots, x_n)$ i.i.d. according to p_X and for each x_i , drawing y_i according to $p(y|x_i)$. Since the conditional $p(y|x)$ is concentrated on one point, the y_i 's are deterministically assigned. Accordingly, we can denote this dependence by writing $\bar{z} = \bar{z}_p(\bar{x})$.

Now consider

$$E_{\bar{z}} \|A(\bar{z}) - m_p\|_{L^2(p_X)}$$

This is equal to

$$\int_{Z^n} dP(\bar{z}) \|A(\bar{z}) - m_p\|_{L^2(p_X)} = \int_{X^n} dp_X^n(\bar{x}) \|A(\bar{z}_p(\bar{x})) - m_p\|_{L^2(p_X)}$$

(To clarify notation, we observe that dp_X^n is the singular measure on X^n with support on \mathcal{M}^n which is the natural product measure corresponding to the distribution of n data points x_1, \dots, x_n drawn i.i.d. with each x_i distributed according to p_X .) The above in turn is lower-bounded by

$$\geq \sum_{l=0}^n \int_{\bar{x} \in S_l} dp_X^n(\bar{x}) \|A(\bar{z}_p(\bar{x})) - m_p\|_{L^2(p_X)}$$

where

$$S_l = \{\bar{x} \in X^n \mid \text{exactly } l \text{ segments contain data and links do not}\}$$

More formally,

$$S_l = \{\bar{x} \in X^n \mid \bar{x} \cap c_i \neq \phi \text{ for exactly } l \text{ segments } c_i \text{ and } \bar{x} \cap B = \phi\}$$

Now we concentrate on lowerbounding $\int_{\bar{x} \in S_l} dp_X^n(\bar{x}) \|A(\bar{z}_p(\bar{x})) - m_p\|_{L^2(p_X)}$. Using the fact that p_X is uniform, we have that $dp_X^n(\bar{x}) = cd(\bar{x})$ (where c is a normalizing constant and $d(\bar{x})$ is the Lebesgue measure or volume form on the associated product space) and therefore

$$\int_{\bar{x} \in S_l} dp_X^n(\bar{x}) \|A(\bar{z}_p(\bar{x})) - m_p\|_{L^2(p_X)} = \int_{\bar{x} \in S_l} cd(\bar{x}) \|A(\bar{z}_p(\bar{x})) - m_p\|_{L^2(p_X)}$$

Thus, we have

$$E_{\bar{z}} \|A(\bar{z}) - m_p\|_{L^2(p_X)} \geq \sum_{l=0}^n \int_{\bar{x} \in S_l} cd(\bar{x}) \|A(\bar{z}_p(\bar{x})) - m_p\|_{L^2(p_X)} \quad (2)$$

Now we see that

$$\begin{aligned} \frac{1}{2^d} \sum_{p \in \mathcal{P}_d} E_{\bar{z}} \|A(\bar{z}_p(\bar{x})) - m_p\|_{L^2(p_X)} &\geq \frac{1}{2^d} \sum_{p \in \mathcal{P}_d} \sum_{l=0}^n c \int_{\bar{x} \in S_l} d(\bar{x}) \|A - m_p\| \\ &\geq \sum_{l=0}^n c \int_{\bar{x} \in S_l} \left(\frac{1}{2^d} \sum_p \|A - m_p\| \right) d(\bar{x}) \end{aligned}$$

By lemma 2.1, we see that for each $\bar{x} \in S_l$, we have

$$\frac{1}{2^d} \sum_p \|A - m_p\| \geq (1 - \alpha - \beta) \frac{d - n}{4d}$$

from which we conclude that

$$\frac{1}{2^d} \sum_p E_{\bar{z}} \|A(\bar{z}) - m_p\|_{L^2(p_X)} \geq (1 - \alpha - \beta) \frac{d - n}{4d} \sum_{l=0}^n \int_{\bar{x} \in S_l} cd(\bar{x})$$

Now we note that

$$\sum_{l=0}^n \int_{\bar{x} \in S_l} cd(\bar{x}) = \text{Prob}(\bar{x} \cap B = \phi) \geq (1 - \beta)^n$$

Therefore,

$$\sup_p E_{\bar{z}} \|A(\bar{z}) - m_p\|_{L^2(p_X)} \geq (1 - \alpha - \beta) \frac{d - n}{4d} (1 - \beta)^n \geq (1 - \alpha - \beta) \frac{1}{8} (1 - \beta)^n \quad (3)$$

Since α and β (and for that matter, d) are in our control, we can choose them to make the righthand side of inequality 3 greater than some constant. This proves our theorem. \square

Construction 1

Consider a set of 2^d manifolds where each manifold has a structure shown in fig. 2. Each manifold has three disjoint subsets: A (loops), B (links), and C (chain) such that

$$\mathcal{M} = A \cup B \cup C$$

The chain C consists of d segments denoted by C_1, C_2, \dots, C_d such that $C = \cup C_i$. The links connect the loops to the segments as shown in fig. 2 so that one obtains a closed curve corresponding to an embedding of the circle into \mathbb{R}^D . For each choice $S \subset \{1, \dots, d\}$ one constructs a manifold (we can denote this by \mathcal{M}_S) such that the links connect C_i (for $i \in S$) to the “upper half” of the loop and they connect C_j (for $j \in \{1, \dots, d\} \setminus S$) to the “bottom half” of the loop as indicated in the figure. Thus there are 2^d manifolds altogether where each \mathcal{M}_S differs from the others in the link structure but the loops and chain are common to all, i.e.,

$$A \cup C \subset \cap_S \mathcal{M}_S$$

For manifold \mathcal{M}_S , let

$$\frac{l(A)}{l(\mathcal{M}_S)} = \int_A p_X^{(S)}(x) dx = \alpha_S$$

where $p_X^{(S)}$ is the probability density function on the manifold \mathcal{M}_S . Similarly

$$\frac{l(B)}{l(\mathcal{M}_S)} = \int_B p_X^{(S)}(x) dx = \beta_S$$

and

$$\frac{l(C)}{l(\mathcal{M}_S)} = \int_C p_X^{(S)}(x) dx = \gamma_S$$

It is easy to check that one can construct these manifolds so that

$$\beta_S \leq \beta; \gamma_S \geq \gamma$$

Thus for each manifold \mathcal{M}_S , we have the associated class of probability distributions $\mathcal{P}_{\mathcal{M}_S}$. These are used in the construction of the lower bound. Now for each such manifold \mathcal{M}_S , we pick one probability distribution $p^{(S)} \in \mathcal{P}_{\mathcal{M}_S}$ such that for every $k \in S$, we have

$$\text{For all } k \in S, p^{(S)}(y = +1|x) = 1 \text{ for all } x \in C_k$$

and for every $k \in \{1, \dots, d\} \setminus S$, we have

$$\text{For all } k \in \{1, \dots, d\} \setminus S, p^{(S)}(y = -1|x) = 1 \text{ for all } x \in C_k$$

Furthermore, the $p^{(S)}$ are all chosen so that the associated conditionals $p^{(S)}(y = +1|x)$ agree on the loops, i.e., for any $S, S' \in \{1, \dots, d\}$,

$$p^{(S)}(y = +1|x) = p^{(S')}(y = +1|x) \text{ for all } x \in A$$

This defines 2^d different probability distributions that satisfy for each p (i) the support of the marginal p_X includes $A \cup C$, (ii) the support of p_X for different p have different link structures (iii) the conditionals $p(y|x)$ disagree on the the chain. □

Lemma 2.1 *Let $\bar{x} \in S_l$ be a collection of n points such that no point belongs to the links and exactly l segments contain at least one point.*

$$\frac{1}{2^d} \sum_{p \in \mathcal{P}_d} \|A(\bar{z}_p(\bar{x}) - m_p)\|_{L^2(p_X)} \geq (1 - \alpha - \beta) \frac{d - n}{4d}$$

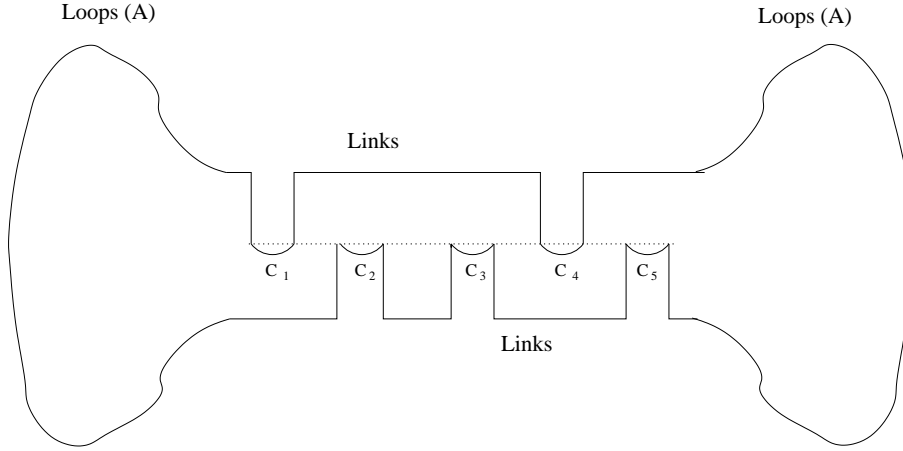


Figure 2: Figure accompanying Construction 1.

PROOF: Since $\bar{x} \in S_l$, there are $d - l$ segments of the chain C such that no data is seen from them. We let $A(\bar{z}_p(\bar{x}))$ be the function hypothesized by the learner on receiving the dataset $\bar{z}_p(\bar{x})$. We begin by noting that the family \mathcal{P}_d may be naturally divided into 2^l subsets in the following way. Following the notation of Construction 1, recall that every element of \mathcal{P}_d may be identified with a set $S \subset \{1, \dots, d\}$. We denote this element by $p^{(S)}$. Now let L denote the set of indices of the segments C_i that contain data, i.e.,

$$L = \{i \mid C_i \cap \bar{x} \neq \phi\}$$

Then for every subset $D \subset L$, we have

$$\mathcal{P}_D = \{p^{(S)} \in \mathcal{P}_d \mid S \cap L = D\}$$

Thus all the elements of \mathcal{P}_D agree in their labelling of the segments containing data but disagree in their labelling of segments not containing data. Clearly there are 2^l possible choices for D and each such choice leads to a family containing 2^{d-l} probability distributions. Let us denote these 2^l families by \mathcal{P}_1 through \mathcal{P}_{2^l} .

Consider \mathcal{P}_i . By construction, for all probability distributions $p, q \in \mathcal{P}_i$, we have that $\bar{z}_p(\bar{x}) = \bar{z}_q(\bar{x})$. Let us denote this by $\bar{z}_i(\bar{x})$, i.e., $\bar{z}_i(\bar{x}) = \bar{z}_p(\bar{x})$ for all $p \in \mathcal{P}_i$.

Now $f = A(\bar{z}_i(\bar{x}))$ is the function hypothesized by the learner on receiving the dataset $\bar{z}_i(\bar{x})$. For any $p \in \mathcal{P}$ and any segment c_k , we

say that p “disagrees” with f on c_k if $|f(x) - m_p(x)| \geq 1$ on a majority of c_k , i.e.,

$$\int_A p_X(x) \geq \int_{c_k \setminus A} p_X(x)$$

where $A = \{x \in c_k \mid |f(x) - m_p(x)| \geq 1\}$. Therefore, if f and p disagree on c_k , we have

$$\int_{c_k} (f(x) - m_p(x))^2 p_X(x) \geq \frac{1}{2} \int_{c_k} p_X(x) \geq \frac{1}{2d}(1 - \alpha - \beta)$$

It is easy to check that for every choice of j unseen segments, there exists a $p \in \mathcal{P}_i$ such that p disagrees with f on each of the chosen segments. Therefore, for such a p , we have

$$\|A(\bar{z}_p(\bar{x}) - m_p)\|_{L^2(p_X)}^2 \geq \frac{1}{2} \frac{j}{d}(1 - \alpha - \beta)$$

Counting all the 2^{d-l} elements of \mathcal{P}_i based on the combinatorics of unseen segments, we see (using the fact that $\|A(\bar{z}_p(\bar{x}) - m_p)\| \geq \sqrt{\frac{1}{2} \frac{j}{d}(1 - \alpha - \beta)} \geq \frac{1}{2} \frac{j}{d}(1 - \alpha - \beta)$)

$$\sum_{p \in \mathcal{P}_i} \|A(\bar{x}_p(\bar{x}) - m_p)\| \geq \sum_{j=0}^{d-l} \binom{d-l}{j} \frac{1}{2} \frac{j}{d}(1 - \alpha - \beta) = 2^{d-l}(1 - \alpha - \beta) \frac{d-l}{4d}$$

Therefore, since $l \leq n$, we have

$$\sum_{i=1}^{2^l} \sum_{p \in \mathcal{P}_i} \|A(\bar{x}_p(\bar{x}) - m_p)\| \geq 2^d(1 - \alpha - \beta) \frac{d-n}{4d}$$

□

3 Discussion

Thus we see that knowledge of the manifold can have profound consequences for learning. The proof of the lower bound reflects the intuition that has always been at the root of manifold based methods for semi-supervised learning. Following fig. 2, if one knows the manifold, one sees that C_1 and C_4 are “close” while C_1 and C_3 are “far”. But this is only one step of the argument. We must further have the prior knowledge that the target function varies smoothly along the

manifold and so “closeness on the manifold” translates to similarity in function values (or label probabilities). However, this closeness is not obvious from the ambient distances alone. This makes the task of the learner who does not know the manifold difficult: in fact impossible in the sense described in Theorem 1.

Some further remarks are in order;

1. While we provide a detailed construction for the case of different embeddings of the circle into \mathbb{R}^N , it is clear that the argument is general and similar constructions can be made for many different classes of k -manifolds. Thus if \mathcal{M} is taken to be a k -dimensional submanifold of \mathbb{R}^N , then one could let \mathcal{M} be a family of k -dimensional submanifolds of \mathbb{R}^N and let \mathcal{P} be the naturally associated family of probability distributions that define a collection of learning problems. Our proof of Theorem 2.2 can be naturally adapted to such a setting.
2. Our example explicitly considers a class $H_{\mathcal{M}}$ that consists of a one-parameter family of functions. It is important to reiterate that many different choices of $H_{\mathcal{M}}$ would provide the same result. For one, thresholding is not necessary, and if the class $H_{\mathcal{M}}$ was simply defined as bandlimited functions, i.e., consisting of functions of the form $\sum_{i=1}^p \alpha_i \phi_i$ (where ϕ_i are the eigenfunctions of the Laplacian of \mathcal{M}), the result of theorem 2.2 holds as well. Similarly Sobolev spaces (constructed from functions $f = \sum_i \alpha_i \phi_i$ where $\alpha_i^2 \lambda_i^s < \infty$) also work with and without thresholding.
3. We have considered the simplest case where there is no noise in the Y -direction, i.e., the conditional $p(y|x)$ is concentrated at one point $m_p(x)$ for each x . Considering a more general setting with noise does not change the import of our results. The upper bound of Theorem 2.1 makes use of the fact that m_p belongs to a restricted (uniformly Glivenko-Cantelli) family $H_{\mathcal{M}}$. With a 0 – 1 loss function defined as $V(h, z) = 1_{[y \neq h(x)]}$, the rate may be as good as $O^*(\frac{1}{n})$ in the noise-free case but drops to $O^*(\frac{1}{\sqrt{n}})$ in the noisy case. The lower bound of Theorem 2.2 for the noiseless case also holds for the noisy case by immediate implication. Both upper and lower bounds are valid also for arbitrary marginal distributions p_X (not just uniform) that have support on some manifold \mathcal{M} .
4. Finally, one can consider a variety of loss functions other than

the L_2 loss function considered here. The natural 0 – 1-valued loss function (which for the special case of binary valued functions coincides with the L_2 loss) can be interpreted as the probability of error of the classifier in the classification setting.

3.1 What does it mean to know the manifold?

In the discussion so far, we have implicitly assumed that an oracle can provide perfect information about the manifold in whatever form we choose. We see that access to such an oracle can provide great power in learning from labeled examples for classes of problems that have a suitable structure. We now turn to the question of what it means to “know the manifold”? How is the manifold represented? What sort of information can the oracle provide and how is this represented? For example, one can consider the following kinds of oracles that provide different degrees of information about the manifold and in different forms.

1. One could know the manifold \mathcal{M} as a set through some kind of set-membership oracle (similar in spirit to the kinds of membership oracles used in the literature on volume estimation in convex bodies). Since the submanifold is of lower dimension than the Euclidean space \mathbb{R}^N in which it is embedded, the set \mathcal{M} has measure 0. To deal with this complication, a membership oracle that makes sense is of the following sort: given a point x and a number $r > 0$, the oracle tells us whether x is in a tubular neighborhood of radius r around the manifold.
2. One could know a system of coordinate charts on the manifold. This would correspond to having maps of the form $\psi_i : U_i \rightarrow \mathbb{R}^D$ where $U_i \subset \mathbb{R}^k$ is an open set.
3. One could know in some explicit form the harmonic functions on the manifold. In particular, one might know the Laplacian $\Delta_{\mathcal{M}}$, the Heat Kernel $H_t(p,q)$ on the manifold. Moreover, one might know the eigenvalues λ_i and eigenfunctions ϕ_i of the Laplace operator on the manifold.
4. One could know the manifold upto some geometric or topological invariants. For example, one might know just the dimension of the manifold. Alternatively, one might know the homology, the homeomorphism or diffeomorphism type, etc. of the manifold.

5. One could have metric information on the manifold. One might know the metric tensor at points on the manifold, one might know the geodesic distances between points on the manifold, or the Gromov-Hausdorff distance, or one might know the heat kernel from which various derived distances (such as diffusion distance) are obtained.

Depending upon the kind of oracle access we have, the task of the learner might vary from simple to impossible. For example, in the problem described in Sec. 2 of this paper, the natural algorithm that realizes the upper bound of Theorem 2.1 performs empirical risk minimization over the class $H_{\mathcal{M}}$. To do this it needs, of course, to be able to represent $H_{\mathcal{M}}$ in a computationally efficient manner. In order to do this, it needs to know the eigenfunctions (in the specific example, only the first two, but in general some arbitrary number depending on the choice of $H_{\mathcal{M}}$) of the Laplacian on the \mathcal{M} . This is immediately accessible from oracle 3. It can be computed from oracle 2 but this computation may be nontrivial in general. There is a closed formula for the value of the $\Delta f(p)$ for every point $p \in \mathcal{M}$ but the eigenfunctions are a global quantity and need not have a straightforward representation. For certain special cases where the manifold has a structure that may be exploited, the eigenfunctions have been worked out. This includes the eigenfunctions on the sphere (the spherical harmonics) and covers therefore the specific example we provide in this paper. But this is far from typical. In fact, the eigenfunctions are typically computed numerically and this may itself involve a triangulation of the manifold which may not be at hand. With access to oracles that provide partial knowledge (oracles 4 and 5) the eigenfunctions cannot be computed at all and so the learning algorithm is unable to exploit the oracle to achieve the bound of Theorem 2.1.

Oracles of the form in 4 and 5 provide partial information about the manifold and while this is not enough for our specific problem, this information may still be very useful for learning in other contexts. In fact many manifold based algorithms including classical algorithms that may be adapted to a manifold setting sometimes need to begin with knowledge of the dimension of the manifold. Access to such an oracle then becomes significant in such a setting.

3.2 Simulating the oracle: manifold learning

The significance of manifold based methods for semi-supervised learning is this: it may be possible to know the manifold by collecting points on it, i.e., by collecting *unlabeled* data alone. This naturally leads us to a crucial piece of our overall task. Which of the above oracles can be simulated by unlabeled data? This question is of independent algorithmic interest in its own right even outside the context of our current discussion. One may ask: given a collection of points \bar{x} randomly sampled from a submanifold \mathcal{M} of Euclidean space, what geometric and topological properties of this manifold may be estimated from the point cloud?

There has been some concentrated activity in this direction over the last several years. Considerable progress has been made in estimating the dimension, various distance metrics, the Laplace operator (and related operators), the homology, and so on. Much of this has happened in the larger context of manifold learning and a significant class of algorithms construct an adjacency graph from the data and then use the graph and graph-theoretic operators as a proxy for the underlying manifold and corresponding manifold-theoretic operators. A unifying concept in this line of work is the relationship between the Laplace-Beltrami operator on the manifold, the heat kernel, harmonic functions, and the graph Laplacian. Let us recall the following fundamental fact from Belkin and Niyogi (2005,2007) that has some significance for the problem in this paper.

Let \mathcal{M} be a compact, Riemannian submanifold (without boundary) of \mathbb{R}^N and let $\Delta_{\mathcal{M}}$ be the Laplace operator (on functions) on this manifold. Let $\bar{x} = \{x_1, \dots, x_m\}$ be a collection of m points sampled in i.i.d. fashion according to the uniform probability distribution on \mathcal{M} . Then one may define the point cloud Laplace operator L_m^t as follows:

$$L_m^t f(x) = \frac{1}{t} \frac{1}{(4\pi t)^{d/2}} \frac{1}{m} \sum_{i=1}^m (f(x) - f(x_i)) e^{-\frac{\|x-x_i\|^2}{4t}}$$

The point cloud Laplacian is a random operator that is the natural extension of the graph Laplacian operator to the whole space. For any thrice differentiable function $f : \mathcal{M} \rightarrow \mathbb{R}$, we have

Theorem 3.1

$$\lim_{t \rightarrow 0, m \rightarrow \infty} L_m^t f(x) = \Delta_{\mathcal{M}} f(x)$$

Some remarks are in order:

1. Given $\bar{x} \in \mathcal{M}$ as above, consider the graph with vertices (let V be the vertex set) identified with the points in \bar{x} and adjacency matrix $W_{ij} = \frac{1}{mt} \frac{1}{(4\pi t)^{d/2}} e^{-\frac{\|x-x_i\|^2}{4t}}$. Given $f : \mathcal{M} \rightarrow \mathbb{R}$, the restriction $f_V : \bar{x} \rightarrow \mathbb{R}$ is a function defined on the vertices of this graph. Correspondingly, the graph Laplacian $L = (D - W)$ acts on f_V and it is easy to check that

$$(L_m^t f)|_{x_i} = (L f_V)|_{x_i}$$

In other words, the point cloud Laplacian and graph Laplacian agree on the data. However, the point cloud Laplacian is defined everywhere while the graph Laplacian is only defined on the data.

2. The quantity t (similar to a bandwidth) needs to go to zero at a suitable rate ($tm^{d+2} \rightarrow \infty$) so there exists a sequence t_m such that the point cloud Laplacian converges to the manifold Laplacian as $m \rightarrow \infty$.
3. It is possible to show (see Belkin and Niyogi, Coifman and Lafon, Hein et al, Gine and Koltchinski) that this basic convergence is true for arbitrary probability distributions (not just the uniform distribution as stated in the above theorem) in which case the point cloud Laplacian converges to an operator of the Laplace type that may be related to the weighted Laplacian (Grigoryan).
4. While the above convergence is pointwise, it also holds uniformly over classes of functions with suitable conditions on their derivatives (Belkin and Niyogi, Gine and Koltchinski).
5. Finally, and most crucially (see Belkin and Niyogi, 2006), if $\lambda_m^{(i)}$ and $\phi_m^{(i)}$ are the i th (in increasing order) eigenvalue and corresponding eigenfunction respectively of the operator $L_m^{t_m}$, then with probability 1, as m goes to infinity,

$$\lim_{m \rightarrow \infty} |\lambda_i - \lambda_m^{(i)}| = 0 \quad (4)$$

and

$$\lim_{m \rightarrow \infty} \|\phi_m^{(i)} - \phi_i\|_{L_2(\mathcal{M})} = 0 \quad (5)$$

In other words, the eigenvalues and eigenfunctions of the point cloud Laplacian converge to those of the manifold Laplacian as the number of data points m go to infinity.

3.3 What is manifold regularization?

Against this backdrop, one might now consider manifold regularization to get some better understanding of when and why it might be expected to provide good semi-supervised learning. First off, it is worthwhile to clarify what is meant by manifold regularization. The term “manifold regularization” was introduced in Belkin, Niyogi, and Sindhwani (2004, 2005, 2006) to describe a class of algorithms in which geometrically motivated regularization penalties were used. One unifying framework adopts a setting of Tikhonov regularization over a Reproducing Kernel Hilbert Space of functions to yield algorithms that arise as special cases of the following:

$$\hat{f} = \arg \min_{f \in H_K} \frac{1}{n} \sum_{i=1}^n V(f(x_i), y_i) + \gamma_A \|f\|_K^2 + \gamma_I \|f\|_I^2 \quad (6)$$

Here $K : X \times X \rightarrow \mathbb{R}$ is a p.d. kernel that defines a suitable RKHS (H_K) of functions that are ambiently defined. The ambient RKHS norm $\|f\|_K$ and an “intrinsic norm” $\|f\|_I$ are traded-off against each other. Intuitively the intrinsic norm $\|f\|_I$ penalizes functions by considering only $f_{\mathcal{M}}$ the restriction of f to \mathcal{M} and essentially considering various smoothness functionals. Since the eigenfunctions of the Laplacian provide a basis for L^2 functions intrinsically defined on \mathcal{M} , one might express $f_{\mathcal{M}} = \sum_i \alpha_i \phi_i$ in this basis and consider constraints on the coefficients.

Remarks:

1. Various choices of $\|f\|_I^2$ include (i) *iterated* Laplacian given by $\int_{\mathcal{M}} f(\Delta^i f) = \sum_j \alpha_j^2 \lambda_j^i$ (ii) *heat kernel* given by $\sum_j e^{t\lambda_j} \alpha_j^2$ (iii) *band limiting* given by $\|f\|_I^2 = \sum_i \mu_i \alpha_i^2$ where $\mu_i = \infty$ for all $i > p$.
2. The loss function V can vary from squared loss to hinge loss to the 0 – 1 loss for classification giving rise to different kinds of algorithmic procedures.
3. While eq. 6 is regularization in the Tikhonov form, one could consider other kinds of model selection principles that are in

the spirit of manifold regularization. For example, the method of Belkin and Niyogi (2003) is a version of the method of sieves that may be interpreted as manifold regularization with bandlimited functions where one allows the bandwidth to grow as more and more data becomes available.

The point of these remarks is that manifold regularization admits a variety of different algorithms that incorporate a geometrically motivated complexity penalty. Let us now consider one such canonical algorithm for the class of problems considered in Sec. 2 of this paper.

3.3.1 A manifold regularization algorithm for semi-supervised learning

Let $\bar{z} = (z_1, \dots, z_n)$ be a set of n i.i.d. labeled examples drawn according to p and $\bar{x} = (x_1, \dots, x_m)$ be a set of m i.i.d. unlabeled examples drawn according to p_X . Then a semi-supervised learner's estimate may be denoted by $A(\bar{z}, \bar{x})$. Let us consider the following kind of manifold regularization based semi-supervised learner.

1. Construct the point cloud Laplacian operator $L_m^{t_m}$ from the *unlabeled* data \bar{x} .
2. Solve for the eigenfunctions of $L_m^{t_m}$ and take the first two (orthogonal to the constant function). Let these be ϕ_m and ψ_m respectively.
3. Perform empirical risk minimization with the empirical eigenfunctions by minimizing

$$\hat{f}_m = \arg \min_{f=\alpha\phi_m+\beta\psi_m} \frac{1}{n} \sum_{i=1}^n V(f(x_i), y_i)$$

subject to $\alpha_i^2 + \beta_i^2 = 1$. Here $V(f(x), y) = \frac{1}{4}|y - \text{sign}(f(x))|^2$ is the 0 – 1 loss. This is equivalent to Ivanov regularization with an intrinsic norm that forces candidate hypothesis functions to be bandlimited.

Note that if the empirical risk minimization was performed with the true eigenfunctions (ϕ and ψ respectively), then the resulting algorithm achieves the rate of Theorem 2.1. Since for large m , the empirical eigenfunctions are close to the true ones by the result in Belkin and Niyogi (2006), we may expect the above algorithm to perform

well. Thus we may compare the two manifold regularization algorithms (an empirical one with unlabeled data and an oracle one that knows the manifold):

$$A(\bar{z}, \bar{x}) = \text{sign}(\hat{f}_m) = \text{sign}(\hat{\alpha}_m \phi_m + \hat{\beta}_m \psi_m)$$

and

$$A_{\text{oracle}}(\bar{z}, \mathcal{M}) = \text{sign}(\hat{f}) = \text{sign}(\hat{\alpha} \phi + \hat{\beta} \psi)$$

We can now state the following:

Theorem 3.2 *For any $\epsilon > 0$, we have*

$$\sup_{p \in \mathcal{P}_{\mathcal{M}}} E_{\bar{z}} \|m_p - A\|_{L^2(p_X)}^2 \leq \frac{4}{2\pi} (\arcsin(\epsilon)) + \frac{1}{\epsilon^2} (\|\phi - \phi_m\|^2 + \|\psi - \psi_m\|^2) + 3\sqrt{\frac{3 \log(n)}{n}}$$

PROOF: Consider $p \in \mathcal{P}$ and let $m_p = \text{sign}(\alpha_p \phi + \beta_p \psi)$. Let $g_m = \alpha_p \phi_m + \beta_p \psi_m$. Now, first note that by the fact of empirical risk minimization, we have

$$\frac{1}{n} \sum_{z \in \bar{z}} V(A(x), y) \leq \frac{1}{n} \sum_{z \in \bar{z}} V(\text{sign}(g_m(x)), y)$$

Second, note that the set of functions $\mathcal{F} = \{\text{sign}(f) \mid f = \alpha \phi_m + \beta \psi_m\}$ has VC dimension equal to 2. Therefore the empirical risk converges to the true risk uniformly over this class so that with probability $> 1 - \delta$, we have

$$\begin{aligned} E_z[V(A(x), y)] - \sqrt{\frac{2 \log(n) + \log(1/\delta)}{n}} &\leq \frac{1}{n} \sum_{z \in \bar{z}} V(A(x), y) \\ &\leq \frac{1}{n} \sum_{z \in \bar{z}} V(\text{sign}(g_m(x)), y) \leq E_z[V(\text{sign}(g_m(x)), y)] + \sqrt{\frac{2 \log(n) + \log(1/\delta)}{n}} \end{aligned}$$

Using the fact that $V(h(x), y) = \frac{1}{4}(y - h)^2$, we have in general for any h

$$E_z[V(h(x), y)] = \frac{1}{4} E_z(y - m_p)^2 + \frac{1}{4} \|m_p - h\|_{L^2(p_X)}^2$$

from which we obtain with probability $> 1 - \delta$ (over choices of labeled training sets \bar{z} ,

$$\|m_p - A\|^2 \leq \|m_p - \text{sign}(g_m)\|^2 + 2\sqrt{\frac{2 \log(n) + \log(1/\delta)}{n}}$$

Setting $\delta = \frac{1}{n}$ and noting that $\|m_p - A\|^2 \leq 1$, we have after some straightforward manipulations,

$$E_{\bar{z}} \|m_p - A\|^2 \leq \|m_p - \text{sign}(g_m)\|^2 + 3\sqrt{\frac{3 \log(n)}{n}}$$

Using lemma 3.1, we get for any $\epsilon > 0$,

$$\sup_{p \in \mathcal{P}_{\mathcal{M}}} E_{\bar{z}} \|m_p - A\|^2 \leq \frac{4}{2\pi} (\arcsin(\epsilon)) + \frac{1}{\epsilon^2} (\|\phi - \phi_m\|^2 + \|\psi - \psi_m\|^2) + 3\sqrt{\frac{3 \log(n)}{n}}$$

□

Lemma 3.1 *Let f, g be any two functions. Then for any $\epsilon > 0$,*

$$\|\text{sign}(f) - \text{sign}(g)\|_{L^2(p_X)}^2 \leq \mu(X_\epsilon) + \frac{1}{\epsilon^2} \|f - g\|_{L^2(p_X)}^2$$

where $X_{\epsilon, f} = \{x \mid |f(x)| > \epsilon\}$ and μ is the measure corresponding to the marginal distribution p_X .

Further, if $f = \alpha\phi + \beta\psi$ ($\alpha^2 + \beta^2 = 1$) and $g = \alpha\phi_m + \beta\psi_m$ where ϕ, ψ are eigenfunctions of the Laplacian on \mathcal{M} while ϕ_m, ψ_m are eigenfunctions of point cloud Laplacian as defined in the previous developments. Then for any $\epsilon > 0$

$$\|\text{sign}(f) - \text{sign}(g)\|_{L^2(p_X)}^2 \leq \frac{4}{2\pi} (\arcsin(\epsilon)) + \frac{1}{\epsilon^2} (\|\phi - \phi_m\|^2 + \|\psi - \psi_m\|^2)$$

PROOF: We see that

$$\begin{aligned} \|\text{sign}(f) - \text{sign}(g)\|_{L^2(p_X)}^2 &= \int_{X_{\epsilon, f}} |\text{sign}(f(x)) - \text{sign}(g(x))|^2 + \int_{\mathcal{M} \setminus X_{\epsilon, f}} |\text{sign}(f(x)) - \text{sign}(g(x))|^2 \\ &\leq 4\mu(X_{\epsilon, f}) + \int_{\mathcal{M} \setminus X_{\epsilon, f}} |\text{sign}(f(x)) - \text{sign}(g(x))|^2 \end{aligned}$$

Note that if $x \in \mathcal{M} \setminus X_{\epsilon, f}$, we have that

$$|\text{sign}(f(x)) - \text{sign}(g(x))| \leq \frac{2}{\epsilon} |f(x) - g(x)|$$

Therefore,

$$\int_{\mathcal{M} \setminus X_{\epsilon, f}} |\text{sign}(f(x)) - \text{sign}(g(x))|^2 \leq \frac{4}{\epsilon^2} \int_{\mathcal{M} \setminus X_{\epsilon, f}} |f(x) - g(x)|^2 \leq \frac{4}{\epsilon^2} \|f - g\|_{L^2(p_X)}^2$$

This proves the first part. The second part follows by a straightforward calculation on the circle. □

Remarks:

1. While we have stated the above theorem for our running example of embeddings of the circle into \mathbb{R}^D , it is clear that the results can be generalized to cover arbitrary k -manifolds, more general classes of functions $H_{\mathcal{M}}$, noise, and loss functions V . Many of these extensions are already implicit in the proof and associated technical discussions.
2. A corollary of the above theorem is relevant for the ($m = \infty$) case that has been covered in Castelli and Cover. We will discuss this in the next section. The corollary is

Corollary 3.1 *Let $\mathcal{P} = \cup_{\mathcal{M}} \mathcal{P}_{\mathcal{M}}$ be a collection of learning problems with the structure described in Section 2, i.e., each $p \in \mathcal{P}$ is such that the marginal p_X has support on a submanifold \mathcal{M} of \mathbb{R}^D which corresponds to a particular isometric embedding of the circle into Euclidean space. For each such p , the regression function $m_p = E[p(y|x)]$ belongs to a class of functions $H_{\mathcal{M}}$ which consists of thresholding bandlimited functions on \mathcal{M} . Then no supervised learning algorithm exists (Theorem 2) but the semi-supervised manifold regularization algorithm described above (with infinite amount of unlabeled data) converges at a fast rate as a function of labelled data. In other words,*

$$\sup_{\mathcal{M}} \lim_{m \rightarrow \infty} \sup_{\mathcal{P}_{\mathcal{M}}} \|m_p - A(\bar{z}, \bar{x})\|_{L^2(p_X)}^2 = 3 \sqrt{\frac{3 \log(n)}{n}}$$

3. It is natural to ask what it would take to move the limit $m \rightarrow \infty$ outside. In order to do this, one will need to put additional constraints on the class of possible manifolds \mathcal{M} that we are allowed to consider. But putting such constraints we can construct classes of learning problems where for any realistic number of labeled examples n , there is a gap between the performance of a supervised learner and the manifold based semi-supervised learner. An example of such a theorem is:

Theorem 3.3 *Fix any number N . Then there exists a class of learning problems \mathcal{P}_N such that for all $n < N$*

$$R(n, \mathcal{P}_N) = \inf_A \sup_{p \in \mathcal{P}_N} E_{\bar{z}} \|m_p - A(\bar{z})\| \geq 1/100$$

while

$$Q(n, \mathcal{P}_N) = \lim_{m \rightarrow \infty} \sup_{p \in \mathcal{P}_N} \|m_p - A_{manreg}(\bar{z}, \bar{x})\|^2 \leq \sqrt{\frac{3 \log(n)}{n}}$$

However, asymptotically, $R(n)$ and $Q(n)$ have the same rate for $n \gg N$. Since N can be arbitrarily chosen to be astronomically large, this asymptotic rate is of little consequence in practical learning situations. This suggests the limitations of asymptotic analysis without a careful consideration of the finite sample situation.

3.4 The structure of semi-supervised learning

It is worthwhile to reflect on why the manifold regularization algorithm is able to display improved performance in semi-supervised learning. The manifold assumption is a device that allows us to link the marginal p_X with the conditional $p(y|x)$. Through unlabeled data \bar{x} , we can learn the manifold \mathcal{M} thereby greatly reducing the class of possible conditionals $p(y|x)$ that we need to consider. More generally, semi-supervised learning will be feasible only if such a link is made. To clarify the structure of problems on which semi-supervised learning is likely to be meaningful, let us define a map $\pi : p \rightarrow p_X$ that takes any probability distribution p on $X \times Y$ and maps it to the marginal p_X .

Given any collection of learning problems \mathcal{P} , we have

$$\pi : \mathcal{P} \rightarrow \mathcal{P}_X$$

where $\mathcal{P}_X = \{p_X | p \in \mathcal{P}\}$. Consider the case in which the structure of \mathcal{P} is such that for any $q \in \mathcal{P}_X$, the family of conditionals $\pi^{-1}(q) = \{p \in \mathcal{P} | p_X = q\}$ is “small”. For a situation like this, knowing the marginal tells us a lot about the conditional and therefore unlabeled data can be useful.

3.4.1 Castelli and Cover interpreted

Let us consider the structure of the class of learning problems considered by Castelli and Cover (1996). They consider a two-class problem with the following structure. The class of learning problems \mathcal{P} is such that for each $p \in \mathcal{P}$, the marginal $q = p_X$ can be uniquely expressed as

$$q = \mu f + (1 - \mu)g$$

where $0 \leq \mu \leq 1$ and f, g belong to some class \mathcal{G} of possible probability distributions. In other words, the marginal is always a mixture

(identifiable) of two distributions. Furthermore, the class \mathcal{P} of possible probability distributions is such that there are precisely two probability distributions $p_1, p_2 \in \mathcal{P}$ such that their marginals are equal to q . In other words,

$$\pi^{-1}(q) = \{p_1, p_2\}$$

where $p_1(y = 1|x) = \frac{\mu f(x)}{q(x)}$ and $p_2(y = 1|x) = \frac{(1-\mu)g(x)}{q(x)}$.

In this setting, unlabeled data allows the learner to estimate the marginal q . Once the marginal is obtained, the class of possible conditionals is reduced to *exactly two functions*. Castelli and Cover (1996) show that the risk now converges to the Bayes' risk exponentially as a function of labeled data (i.e., the analog of an upper bound on $Q(n, \mathcal{P})$ is approximately e^{-n}). The reason semi-supervised learning is successful in this setting is that the marginal q tells us a great deal about the class of possible conditionals. It seems that a precise lower bound on purely supervised learning (the analog of $R(n, \mathcal{P})$) has never been clearly stated in that setting.

3.4.2 Spectral Geometry and Semi-supervised learning

In its most general form, manifold regularization encompasses a class of geometrically motivated approaches to learning. Spectral geometry provides the unifying point of view and the spectral analysis of a suitable geometrically motivated operator yields a “distinguished basis”. Since (i) only unlabeled examples are needed for the spectral analysis and the learning of this basis, and (ii) the target function is assumed to be compactly representable in this basis, the idea has the possibility to succeed in semi-supervised learning. Indeed, the previous theorems clarify the theoretical basis of this approach. This, together with the empirical success of algorithms based on these intuitions suggest there is some merit in this point of view.

In general, let q be a probability density function on $X = \mathbb{R}^D$. The support of q may be a submanifold of X (with possibly many connected components). Alternatively, it may lie close to a submanifold, it may be all of X , or it may be a subset of X . As long as q is far from uniform, i.e., it has a “shape”, one may consider the following “weighted Laplacian” (see Grigoryan) defined as

$$\Delta_q f(x) = \frac{1}{q(x)} \operatorname{div}(q \operatorname{grad} f)$$

where the gradient (grad) and divergence (div) are with respect to the support of q (which may simply be all of X).

The heat kernel associated with this weighted Laplacian (essentially the Fokker-Planck operator) is given by $e^{-t\Delta_q}$. Laplacian eigenmaps and Diffusion maps are thus defined in this more general setting.

If ϕ_1, ϕ_2, \dots represent an eigenbasis for this operator, then, one may consider the regression function m_q to belong to the family (parameterized by $\mathbf{s} = (s_1, s_2, \dots)$ where each $s_i \in \mathbb{R} \cup \{\infty\}$).

$$H_q^{\mathbf{s}} = \{h : X \rightarrow \mathbb{R} \text{ such that } h = \sum_i \alpha_i \phi_i \text{ and } \sum_i \alpha_i^2 s_i < \infty\}$$

Some natural choices of \mathbf{s} are (i) $\forall i > p, s_i = \infty$: this gives us bandlimited functions (ii) $s_i = \lambda_i^t$ where λ_i is the i th eigenvalue of Δ_q : this gives us spaces of Sobolev type (iii) $\forall i \in A, s_i = 1$, else $s_i = \infty$ where A is a finite set: this gives us functions that are sparse in that basis.

The class of learning problems $\mathcal{P}^{(\mathbf{s})}$ may then be factored as

$$\mathcal{P}^{(\mathbf{s})} = \cup_q \mathcal{P}_q^{(\mathbf{s})}$$

where

$$\mathcal{P}_q^{(\mathbf{s})} = \{p \mid p_X = q \text{ and } m_p \in H_q^{\mathbf{s}}\}$$

The logic of the geometric approach to semi-supervised learning is as follows:

1. Unlabeled data allow us to approximate q , the eigenvalues and eigenfunctions of Δ_q , and therefore the space $H^{\mathbf{s}}$.
2. If \mathbf{s} is such that $\pi^{-1}(q)$ is "small" for every q , then a small number of labeled examples suffice to learn the regression function m_q .

In problems that have this general structure, we expect manifold regularization and related algorithms (that use the graph Laplacian or a suitable spectral approximation) to work well. Precise theorems showing the correctness of these algorithms for a variety of choices of \mathbf{s} remains part of future work. The theorems in this paper establish results for some choices of \mathbf{s} and are a step in a broader understanding of this question.

4 Conclusions

We have considered a minimax style framework within which we have investigated the potential role of manifold learning in learning from labeled and unlabeled examples. We demonstrated the natural structure of a class of problems on which knowing the manifold makes a big difference. On such problems, we see that manifold regularization is provably better than any supervised learning algorithm.

Our proof clarifies a potential source of confusion in the literature on manifold learning. We see that if data lives on an underlying manifold but this manifold is *unknown* and belongs to a class of possible smooth manifolds, it is possible that supervised learning (classification and regression problems) may be ineffective, even impossible. In contrast, if the manifold is fixed though unknown, it may be possible to (e.g. Bickel and Li, 2006) learn effectively by a classical method suitably modified. In between these two cases lie various situations that need to be properly explored for a greater understanding of the potential benefits and limitations of manifold methods and the need for manifold learning.

Our analysis allows us to see the role of manifold regularization in semi-supervised learning in a clear way. It provides a framework for analysis and a justification for several algorithms that use manifold and graph-based methods for semi-supervised learning that have seen empirical success.

References

- [1] M. Belkin, P. Niyogi. (2003). *Laplacian Eigenmaps for Dimensionality Reduction and Data Representation*, Neural Computation, Vol. 15, No. 6, 1373-1396.
- [2] M. Belkin, P. Niyogi, *Towards a Theoretical Foundation for Laplacian-Based Manifold Methods*, COLT 2005.
- [3] M. Belkin, P. Niyogi, *Towards a Theoretical Foundation for Laplacian-Based Manifold Methods*, Journal of Computer and System Sciences, 2007.
- [4] M. Belkin, P. Niyogi, V. Sindhwani, *Manifold Regularization: A Geometric Framework for Learning from Examples*, Technical Report, Department of Computer Science, University of Chicago. 2004. Later appears in JMLR, 2006.

- [5] M. Belkin, P. Niyogi, V. Sindhvani, *On Manifold Regularization*, AI Stats 2005.
- [6] P. Bickel and Bo Li, *Local Polynomial Regression on Unknown Manifolds*, COMPLEX DATASETS AND INVERSE PROBLEMS: TOMOGRAPHY, NETWORKS AND BEYOND, IMS Lecture Notes-Monograph Series, Vol 54, 177-186, 2007
- [7] R. Coifman and S. Lafon, *Diffusion Maps*, Applied and Computational Harmonic Analysis, Vol. 21, July 2006.
- [8] R.R.Coifman, S. Lafon, A. Lee, M. Maggioni, B. Nadler, F. Warner and S. Zucker, *Geometric diffusions as a tool for harmonic analysis and structure definition of data* (Parts I and II), Proc. of Nat. Acad. Sci., (102):7426-7438, May 2005.
- [9] V. Castelli and T.M. Cover, *The Relative Value of Labeled and Unlabeled Samples in Pattern Recognition with an Unknown Mixing Parameter*, IEEE Trans. Information Theory, 42(6):2102-2117, Nov. 1996.
- [10] E. Gine, V. Koltchinski, *Empirical Graph Laplacian Approximation of Laplace-Beltrami Operators: Large Sample Results*, preprint. (also, 4th International Conference on High Dimensional Probability.)
- [11] M. Hein, J.-Y. Audibert, U. von Luxburg, *From Graphs to Manifolds – Weak and Strong Pointwise Consistency of Graph Laplacians*, COLT 2005.
- [12] J. Lafferty and L. Wasserman, *Statistical Analysis of Semi-supervised Learning*, NIPS, Dec. 2007.
- [13] F. Memoli, G. Sapiro, *Comparing Point Clouds*, IMA Technical Report, 2004.
- [14] A. Grigoryan, *Heat Kernels on Weighted Manifolds and Applications*, Cont. Math, 398, 2006.
- [15] P. Niyogi, S. Smale, S. Weinberger, *Finding the Homology of Submanifolds with High Confidence from Random Samples*, Univ. of Chicago Technical Report TR-2004-08, 2004. appears in Discrete and Computational Geometry, 2006.
- [16] S. Rosenberg, *The Laplacian on a Riemannian Manifold*, Cambridge University Press, 1997.
- [17] J. Shi, J. Malik, *Normalized Cuts and Image Segmentation*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 22, 8, 2000.

- [18] O.G. Smolyanov, H.v. Weizscker, and O. Wittich. *Brownian motion on a manifold as limit of stepwise conditioned standard Brownian motions*. In *Stochastic processes, Physics and Geometry: New Interplays. II: A Volume in Honour of S. Albeverio*, volume 29 of *Can. Math. Soc. Conf. Proc.*, pages 589-602. Am. Math. Soc., 2000.
- [19] O.G. Smolyanov, H. v Weizsacker, and O. Wittich. *Chernoff's Theorem and discrete time approximations of Brownian motion on manifolds*. Preprint. 2004.
- [20] D. Zhou, O. Bousquet, T.N. Lal, J. Weston and B. Schoelkopf, *Learning with Local and Global Consistency*, NIPS 2003.
- [21] X. Zhu, J. Lafferty and Z. Ghahramani, *Semi-supervised learning using Gaussian fields and harmonic functions*, ICML 2003.
- [22] A. Zomorodian, G. Carlsson, *Computing persistent homology*, 20th ACM Symposium on Computational Geometry, 2004.