4-10-2012

Density-Sensitive Semisupervised Inference

Martin Azizyan
Carnegie Mellon University, mazizyan@andrew.cmu.edu

Aarti Singh
Carnegie Mellon University, aarti@cs.cmu.edu

Larry Wasserman
Carnegie Mellon, larry@stat.cmu.edu

DENSITY-SENSITIVE SEMISUPERVISED INFERENCE

By Martin Azizyan[‡], Aarti Singh[*,‡] and Larry Wasserman[†,‡]

Carnegie Mellon University[‡]

April 10, 2012

Semisupervised methods are techniques for using labeled data $(X_1, Y_1), \ldots, (X_n, Y_n)$ together with unlabeled data $X_{n+1}, \ldots, X_N$ to make predictions. These methods invoke some assumption that links the marginal distribution P_X of X to the regression function $f(x)$. For example, it is common to assume that f is very smooth over high density regions of P_X. Many of the methods are ad-hoc and have been shown to work in specific examples but are lacking a theoretical foundation. We provide a minimax framework for analyzing semisupervised methods. In particular, we study methods based on metrics that are sensitive to the distribution P_X. Our model includes a parameter α that controls the strength of the semisupervised assumption. We then use the data to adapt to α.

[*]Supported by Air Force grant FA9550-10-1-0382 and NSF grant IIS-1116458.

[†]Supported by NSF Grant DMS-0806009 and Air Force Grant FA95500910373.

AMS 2000 subject classifications: Primary 62G15; secondary 62G07

Keywords and phrases: nonparametric inference, semisupervised, kernel density, efficiency

CONTENTS

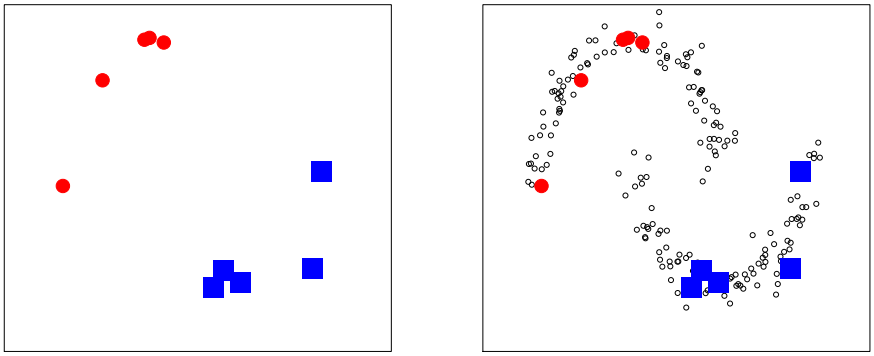

FIG 1. *The covariate $X = (X_1, X_2)$ is two-dimensional. The response Y is binary and is shown as a square or a circle. Left: The labaleled data. Right: Labeled and unlabeled data.*

1. Introduction. Suppose we have data $(X_1, Y_1), \ldots, (X_n, Y_n)$ from a distribution P, where $X_i \in \mathbb{R}^d$ and $Y_i \in \mathbb{R}$. Further, we have a second set of data $X_{n+1}, \ldots, X_N$ from the same distribution but without the $Y's$. We refer to $\mathcal{L} = \{(X_i, Y_i) : \ i = 1, \ldots, n\}$ as the *labeled data* and $\mathcal{U} = \{X_i : \ i = n+1, \ldots, N\}$ as the *unlabeled data*. There has been a major effort, mostly in the machine learning literature, to find ways to use the unlabeled data together with the labeled data to constuct good predictors of Y. These methods are known as *semisupervised methods*. It is generally assumed that the the $m = N - n$ unobserved labels $Y_{n+1}, \ldots, Y_N$ are missing completely at random and we shall assume this throughout.

To motivate semisupervised inference, consider the following example. We download a large number N of webpages X_i. We select a small subset of size n and label these with some attribute Y_i. The downloading process is cheap whereas the labeling process is expensive so typically N is huge while n is much smaller.

Figure 1 shows a toy example of how unlabeled data can help with prediction. In this case, Y is binary, $X \in \mathbb{R}^2$ and we want to find the decision boundary $\{x : \ P(Y = 1|X = x) = 1/2\}$. The left plot shows a few labeled data points from which it would be challenging to find the boundary. The right plot shows labeled and unlabeled points. The unlabeled data show that there are two clusters. If we make the seemingly reasonable assumption that $f(x) = P(Y = 1|X = x)$ is very smooth over the two clusters, then iden-

tifying the decision boundaty becomes much easier. In other words, if we assume some link between P_X and f, then we can use the unlabaled data; see Figure 2.

The assumption that the regression function $f(x) = \mathbb{E}(Y|X = x)$ is very smooth over the clusters is known as the *cluster assumption.* In the special case where the clusters are low dimensional submanifolds, the assumption is called the *manifold assumption.* These assumptions link the regression function f to the distribution P_X of X.

Many semisupervised methods are developed based on the above assumptions although this is not always made explicit. And even with such a link, it is not obvious that semisupervised methods will outperform supervised methods. Making precise how and when these assumptions actually improve inferences is surprisingly elusive and most papers do not address this issue; some exceptions are Rigollet (2007), Singh, Nowak and Zhu (2008a), Lafferty and Wasserman (2007), Nadler, Srebro and Zhou (2009), Ben-David, Lu and Pal (2008), Sinha and Belkin (2009), Belkin and Niyogi (2004) and Niyogi (2008). These authors have shown that the degree to which unlabeled data improves performance is very sensitive to the cluster and manifold assumptions. In this paper, we introduce *adaptive semisupervised inference.* We define a parameter α that controls the sensitivity of the distance metric to the density, and hence the strength of the semisupervised assumption. When $\alpha = 0$ there is no semisupervised assumption, that is, there is no link between f and P_X. When $\alpha = \infty$ there is a very strong semisupervised assumption. We use the data to estimate α and hence we adapt to the appropriate assumption linking f and P_X.

This paper makes the following contributions:

1. We formalize the link between the regression function f and the marginal distribution P_X by defining a class of functions spaces based on a metric that depends on P_X. This is called a *density sensitive metric.*
2. We show how to consistently estimate the density-sensitive metric.
3. We propose a semi-supervised kernel estimator based on the density-sensitive metric.
4. We provide some minimax bounds and show that under some conditions the semisupervised method has smaller predictive risk than any supervised method.
5. The function classes depend on a parameter α that controls how strong the semisupervised assumption is. We show that it is possible to adapt to α.
6. We provide numerical simulations to support the theory.

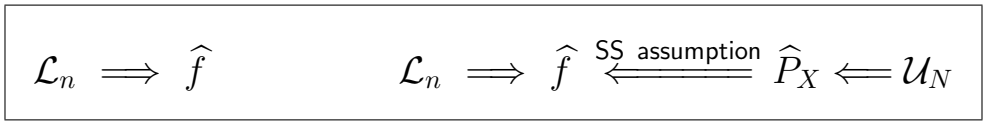

FIG 2. *Supervised learning (left) uses only the labeled data $\mathcal{L}_n$. Semisupervised learning (right) uses the unlabaled data $\mathcal{U}_N$ to estimate the marginal distribution P_X which helps estimate f if there is some link between P_X and f. This link is the semisupervised (SS) assumption.*

In addition, we should add that we focus on regression while most previous literature only deals with binary outcomes (classification).

Related Work. There are a number of papers that discuss conditions under which semisupervised methods can succeed or that discuss metrics that are useful for semisupervised methods. These include Bousquet, Chapelle and Hein (2004), Singh, Nowak and Zhu (2008a), Lafferty and Wasserman (2007), Sinha and Belkin (2009), Ben-David, Lu and Pal (2008), Nadler, Srebro and Zhou (2009), Sajama and Orlitsky (2005), Bijral, Ratliff and Srebro (2011), Belkin and Niyogi (2004), Niyogi (2008) and references therein. Papers on semisupervised inference in the statistics literature are rare; some exceptions include Culp and Michailidis (2008), Culp (2011a) and Liang, Mukherjee and West (2007). To the best of our knowledge, there are no papers that explicitly study adaptive methods that allow the data to choose the strength of the semisupervised assumption.

There is a connection between our work on the semisupervised classification method in Rigollet (2007). He divides the the covariate space $\mathcal{X}$ into clusters $C_1, \ldots, C_k$ defined by the upper level sets $\{p_X > \lambda\}$ of the density p_X of P_X. He assumes that the indicator function $I(x) = I(p(y|x) > 1/2)$ is constant over each cluster C_j. In our regression framework, we could similarly assume that

$$f(x) = \sum_{j=1}^{k} f_{\theta_j}(x) I(x \in C_j) + g(x) I(x \in C_0)$$

where $f_\theta(x)$ is a parametric regression function, g is a smooth (but nonparametric function) and $C_0 = \mathcal{X} - \bigcup_{j=1}^{k} C_j$. This yields parametric, dimension-free rates over $\mathcal{X} - C_0$. However, this creates a rather unnatural and harsh boundary at $\{x : \ p_X(x) = \lambda\}$. Our approach may be seen as a smoother version of this idea.

Outline. This paper is organized as follows. In Section 2 we give definitions and assumptions. In Section 3 we define density sensitive metrics and the function spaces defined by these metrics. In Section 4 we present results on estimating density sensitive metrics. In Section 5 we define a density sensitive semisupervised estimator and we bound its risk. In Section 6 we present some minimax results. We discuss adaptation in Section 8. We provide simulations in 9. Section 10 contains closing discussion. Additional proofs are contained in Section 11.

2. Definitions. Recall that $X_i \in \mathbb{R}^d$ and $Y_i \in \mathbb{R}$. Let

$$\mathcal{L}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\} \tag{1}$$

be an iid sample from P. Let P_X denote the X-marginal of P and let

$$\mathcal{U}_N = \{X_{n+1}, \ldots, X_N\} \tag{2}$$

be an iid sample from P_X.

Let $f(x) \equiv f_P(x) = \mathbb{E}(Y|X = x)$. An estimator of f that is a function of $\mathcal{L}_n$ is called a *supervised learner* and the set of such estimators is denoted by $\mathcal{S}_n$. An estimator that is a function of $\mathcal{L}_n \bigcup \mathcal{U}_N$ is called a *semisupervised learner* and the set of such estimators is denoted by $\mathcal{SS}_N$. Define the risk of an estimator $\widehat{f}$ by

$$R_P(\widehat{f}) = \mathbb{E}_P\left[\int (\widehat{f}(x) - f_P(x))^2 dP(x)\right]. \tag{3}$$

Of course, $\mathcal{S}_n \subset \mathcal{SS}_N$ and hence,

$$\inf_{\widehat{g} \in \mathcal{SS}_N} \sup_{P \in \mathcal{P}} R_P(\widehat{g}) \leq \inf_{\widehat{g} \in \mathcal{S}_n} \sup_{P \in \mathcal{P}} R_P(\widehat{g}).$$

We will show that, under certain conditions, semisupervised methods outperform supervised methods in the sense that the left hand side of the above equation is substantially smaller than the right hand side. More precisely, for certain classes of distributions $\mathcal{P}_n$, we show that

$$\frac{\inf_{\widehat{g} \in \mathcal{SS}_N} \sup_{P \in \mathcal{P}_n} R_P(\widehat{g})}{\inf_{\widehat{g} \in \mathcal{S}_n} \sup_{P \in \mathcal{P}_n} R_P(\widehat{g})} \to 0 \tag{4}$$

as $n \to \infty$. In this case we say that semisupervised learning is *effective*.

Remark: In order for the asymptotic analysis to reflect the behavior of finite samples, we need to let $\mathcal{P}_n$ to change with n and we need $N = N(n) \to$

∞ and $n/N(n) \to 0$ as $n \to \infty$. As an analogy, one needs to let the number of covariates in a regression problem increase with the sample size, to develop relevant asymptotics for high dimensional regression. Moreover, $\mathcal{P}_n$ must have distributions that get more concentrated as n increases. The reason is that, if n is very large and P_X is smooth, then there is no advantage to semisupervised inference. This is consistent with the finding in Ben-David, Lu and Pal (2008) who show that if P_X is smooth, then " ... knowledge of that distribution cannot improve the labeled sample complexity by more than a constant factor."'

Other Notation. If A is a set and $\delta \geq 0$ we define

$$A \oplus \delta = \bigcup_{x \in A} B(x, \delta)$$

where $B(x, \delta)$ denotes a ball of radius δ centered at x. Given a set $A \subseteq \mathbb{R}^d$, define $d_A(x_1, x_2)$ to be the length of the shortest path in A connecting x_1 and x_2.

We write $a_n = O(b_n)$ if $|a_n/b_n|$ is bounded for all large n. Similarly, $a_n = \Omega(b_n)$ if $|a_n/b_n|$ is bounded away from 0 for all large n. We write $a_n \asymp b_n$ if $a_n = O(a_n)$ and $a_n = \Omega(b_n)$. We also write $a_n \preceq b_n$ if there exists $C > 0$ such that $a_n \leq C b_n$ for all large n. Define $a_n \succeq b_n$ similarly. We use symbols of the form $c, c_1, c_2, \ldots, C, C_1, C_2, \ldots$ to denote generic positive constants whose value can change in different expressions.

To prove lower bounds, we will use Assouad's Lemma (see Lemma 24.3 in van der Vaart (1998)). Recall that the Hamming distance between two vectors v and w is $\rho(v, w) = \sum_j I(v_j \neq w_j)$.

Lemma 1 (Assouad's Lemma) *Let $\Omega = \{0, 1\}^q$ be the collection of binary vectors of length $q \geq 1$. Let $\mathcal{P}_\Omega = \{P_\omega : \omega \in \Omega\}$ be a collection of 2^q probability measures indexed by $\omega \in \Omega$. Also let*

$$\|P_\nu \wedge P_\omega\| = 1 - \sup_A |P_\nu(A) - P_\omega(A)|$$

denote the affinity between two distributions, where the supremum is over all measurable sets A. Let $\{f_\omega : \omega \in \Omega\}$ be a collection of functions. For any semi-distance d, and any $p > 0$,

(5)

$$\inf_{\widehat{f}} \max_{\omega \in \Omega} \mathbb{E}_\omega[d^p(f_\omega, \widehat{f})] \geq \frac{q}{2^{p+1}} \left(\min_{\omega, \nu : \rho(\omega, \nu) \neq 0} \frac{d^p(f_\omega, f_\nu)}{\rho(\omega, \nu)} \right) \times \left(\min_{\omega, \nu : \rho(\omega, \nu) = 1} \|P_\omega \wedge P_\nu\| \right).$$

3. Density-Sensitive Metrics. We allow the marginal distribution P_X for X to be arbitrary. We define a smoothed version of P_X as follows. Let K denote a symmetric kernel on $\mathbb{R}^d$ with compact support, let $\sigma > 0$ and define

$$p_\sigma(x) \equiv p_{X,\sigma}(x) = \int \frac{1}{\sigma^d} K\left(\frac{||x-u||}{\sigma}\right) dP_X(x). \tag{6}$$

Thus, $p_{X,\sigma}$ is the density of the convolution $P_{X,\sigma} = P_X \star \mathbb{K}_\sigma$ where $\mathbb{K}_\sigma$ is the measure with density $K_\sigma(\cdot) = \sigma^{-d}K(\cdot/\sigma)$. $P_{X,\sigma}$ always has a density even if P_X does not. This is important because, in high dimensional problems, it is not uncommon to find that P_X can be highly concentrated near a low dimensional manifold. And these are precisely the cases where semisupervised methods are often useful (Ben-David, Lu and Pal (2008)). Indeed, this was one of the original motivations for semisupervised inference. We define $P_{X,0} = P_X$. For notational simplicity, we shall sometimes drop the X and simply write p_σ instead of $p_{X,\sigma}$.

3.1. *The Exponential Metric.* Following previous work in the area, we will assume that the regression function is smooth in regions where P_X puts lots of mass. To make this precise, we define a *density sensitive metric* as follows. For any pair x_1 and x_2 let $\Gamma(x_1, x_2)$ denote the set of all continuous finite curves from x_1 to x_2 with unit speed everywhere and let $L(\gamma)$ be the length of curve γ; hence $\gamma(L(\gamma)) = x_2$. For any $\alpha \geq 0$ define the *exponential metric*

$$D(x_1, x_2) \equiv D_{P,\alpha,\sigma}(x_1, x_2) = \inf_{\gamma \in \Gamma(x_1,x_2)} \int_0^{L(\gamma)} \exp\Big[-\alpha\, p_{X,\sigma}(\gamma(t))\Big] dt. \tag{7}$$

In Section 7 we also consider a second metric, the *reciprocal metric*. Large α makes points connected by high density paths closer; see Figure 3. Note that $\alpha = 0$ corresponds to Euclidean distance. Similar definitions are used in Sajama and Orlitsky (2005), Bijral, Ratliff and Srebro (2011) and Bousquet, Chapelle and Hein (2004).

3.2. *The Regression Function.* Recall that $f(x) \equiv f_P(x) = E(Y|X = x)$ denotes the regression function. We assume that $X \in [0,1]^d \equiv \mathcal{X}$ and that $|Y| \leq M$ for some finite constant M.[1] We formalize the semisupervised smoothness assumption by defining the following scale of function spaces.

[1] The results can be extended to unbounded Y with suitable conditions on the tails of the distribution of Y.

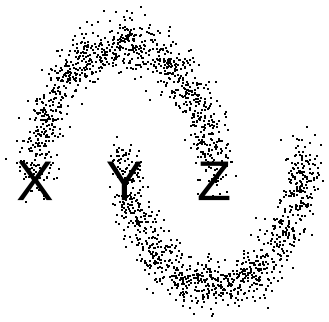

FIG 3. *With a density density metric, the points X and Z are closer than the points X and Y because there is a high density path connecting X and Z.*

.

Let $\mathcal{F} \equiv \mathcal{F}(P, \alpha, \sigma, L)$ denote the set functions $f : [0,1]^d \to \mathbb{R}$ such that, for all $x_1, x_2 \in \mathcal{X}$,

$$|f(x_1) - f(x_2)| \leq L \; D_{P,\alpha,\sigma}(x_1, x_2). \tag{8}$$

Let $\mathcal{P}(\alpha, \sigma, L)$ denote all joint distributions for (X, Y) such that $f_P \in \mathcal{F}(P, \alpha, \sigma, L)$ and such that P_X is supported on $\mathcal{X}$.

3.3. *Properties of the Function Spaces.* The variance of our estimator will depend on

$$\int \frac{dP(x)}{P(B_{P,\alpha,\sigma}(x,\epsilon))} \tag{9}$$

where $B_{P,\alpha,\sigma}(x,\epsilon) = \{z : \; D_{P,\alpha,\sigma}(x,z) \leq \epsilon\}$. Let S_P denote the support of P and let $\mathcal{N}_{P,\alpha,\sigma}(\epsilon)$ denote the covering number, the smallest number of balls of the form $B_{P,\alpha,\sigma}(x,\epsilon)$ required to cover S_P. A simple argument shows that

$$\int \frac{dP(x)}{P(B_{P,\alpha,\sigma}(x,\epsilon))} \leq \mathcal{N}_{P,\alpha,\sigma}(\epsilon/2). \tag{10}$$

In the Euclidean case $\alpha = 0$, we have $\mathcal{N}_{P,0,\sigma}(\epsilon) \leq (C/\epsilon)^d$. But when $\alpha > 0$ and P is concentrated on or near a set of dimension less than d, the $\mathcal{N}_{P,\alpha,\sigma}(\epsilon)$ can be much smaller than $(C/\epsilon)^d$. The next result gives a few examples showing that concentrated distributions have small covering numbers. We say that a set A is *regular* if there is a $C > 0$ such that, for all small $\epsilon > 0$,

$$\sup_{\substack{x,y \in A \\ ||x-y|| \leq \epsilon}} \frac{d_A(x,y)}{||x-y||} \leq C. \tag{11}$$

Recall that S_P denotes the support of P.

Lemma 2 *Suppose that S_P is regular.*

1. *For all α, σ and P, $\mathcal{N}_{P,\alpha,\sigma}(\epsilon) \preceq \epsilon^{-d}$.*
2. *Suppose that $P = \sum_{j=1}^{k} \delta_{x_j}$ where δ_x is a point mass at x. Then, for any $\alpha \geq 0$ and any $\epsilon > 0$, $\mathcal{N}_{P,\alpha,\sigma}(\epsilon) \leq k$.*
3. *Suppose that $\mathsf{dim}(S_P) = r < d$. Then, $\mathcal{N}_{P,\alpha,\sigma}(\epsilon) \preceq \epsilon^{-r}$.*
4. *Suppose that $S_P = W \oplus \gamma$ where $\mathsf{dim}(W) = r < d$. Then, for $\epsilon \geq C\gamma$, $\mathcal{N}_{P,\alpha,\sigma}(\epsilon) \preceq \left(\frac{1}{\epsilon}\right)^r$.*

Proof. (1) The first statement is follows since the covering number of S_P is no more than the covering number of $[0,1]^d$ and on $[0,1]^d$, $D_{P,\alpha,\sigma}(x,y) \leq ||x-y||$. Now $[0,1]^d$ can be covered $O(\epsilon^{-d})$ Euclidean balls.
(2) The second statement follows since $\{\{x_1\}, \ldots, \{x_k\}\}$ forms an ϵ-covering for any ϵ.
(3) We have that $D_{P,\alpha,\sigma}(x,y) \leq d_{S_P}(x,y)$. Regularity implies that, for small $d_{S_P}(x,y)$, $D_{P,\alpha,\sigma}(x,y) \leq c||x-y||$. We can thus cover S_P by $C\epsilon^{-r}$ balls of size ϵ.
(4) As in (3), cover W with $N = O(\epsilon^{-r})$ balls of D size ϵ. Denote these balls by $B_1, \ldots, B_N$. Define $C_j = \{x \in S_P : \ d_{S_P}(x, B_j) \leq \gamma\}$. The C_j form a covering of size N and each C_j has $D_{P,\alpha,\sigma}$ diameter $\max\{\epsilon, \gamma\}$. □

4. Estimating Density-Sensitive Metrics.

In this section we consider estimating the density-sensitive metrics.

4.1. *Estimating The Density.*

Let $m = N - n$ denote the number of unlabeled points and let

$$\widehat{p}_\sigma(x) = \frac{1}{m} \sum_{i=1}^{m} \frac{1}{\sigma^d} K\left(\frac{||x - X_{i+n}||}{\sigma}\right) \tag{12}$$

be the usual kernel estimator of p_σ.

To estimate the distance, we need to bound $||\widehat{p}_\sigma - p_\sigma||_\infty$ uniformly over a range of values of σ. The uniform in bandwidth result by Einmahl and Mason (2005) provides almost sure bounds of this type. For example, their result implies that, almost surely, there is an m_0 such that for all $m \geq m_0$ and all $c > 0$, and for all $(c \log m/m)^{1/d} \leq \sigma \leq 1$,

$$||\widehat{p}_\sigma - p_\sigma||_\infty \leq \frac{K(c)\sqrt{d \log(1/\sigma) \vee \log\log m}}{\sqrt{m\sigma^d}}. \tag{13}$$

However, such a bound is not uniform over a class of distributions. Instead we use the following result whose proof is in Section 11.

Theorem 4.1 *Let $X_1, \ldots, X_m \sim P$ where P has support on a compact set $\mathcal{X} \subset \mathbb{R}^d$. Let $\widehat{p}_\sigma(x) = \frac{1}{\sigma^d}\sum_i K(||x - X_i||/\sigma)$. Let $p_\sigma(x) = \mathbb{E}(\widehat{p}_\sigma(x))$. Suppose that $K(x) \leq K(0)$ for all x and that*

$$|K(y) - K(x)| \leq L||x - y||$$

for all x, y. Let $0 < a \leq A < \infty$. If $\epsilon \leq 2/3$ then

$$\sup_{P\in\mathcal{P}} P^m\left(\sup_{a\leq\sigma\leq A} ||\widehat{p}_\sigma - p_\sigma||_\infty > \epsilon\right) \leq \left(\frac{C}{a^{2d}\epsilon}\right)^{d+1} \exp\left(-\frac{3m\epsilon^2 a^d}{28K(0)}\right) \tag{14}$$

where $\mathcal{P}$ is the set of distributions such that P_X is supported on $\mathcal{X}$.

Thus, for large m, $\pi_m \preceq e^{-cm\epsilon^2 a^d}$. Now let $a_m \asymp \left(\frac{1}{m}\right)^{\frac{1}{d(1+\gamma)}}$ where γ is any small, positive number. Then, with probability at least $1 - 1/m$,

$$\sup_{a_m\leq\sigma\leq A} ||\widehat{p}_\sigma - p_\sigma||_\infty < \sqrt{\frac{C\log m}{a_m^d m}}. \tag{15}$$

4.2. *Estimating the Exponential Distance.* Define

$$\widehat{D}_{\alpha,\sigma}(x_1, x_2) = \inf_{\gamma\in\Gamma(x_1,x_2)} \int_0^{L(\gamma)} \exp\left[-\alpha\widehat{p}_\sigma(\gamma(t))\right] dt. \tag{16}$$

Lemma 3 *Suppose that $||\widehat{p}_\sigma(x) - p_\sigma(x)||_\infty \leq \epsilon_m$. For all x_1, x_2*

$$e^{-\alpha\epsilon_m} D_{P,\alpha,\sigma}(x_1, x_2) \leq \widehat{D}_{\alpha,\sigma}(x_1, x_2) \leq e^{\alpha\epsilon_m} D_{P,\alpha,\sigma}(x_1, x_2). \tag{17}$$

Proof. Follows easily from the definition of the metric. □

4.3. *A Computable Estimator.* Although the above estimator $\widehat{D}_{\alpha,\sigma}$ is consistent, it is not easily computable because it involves searching over all possible paths connecting each pair of points. In fact, even if P is known, $D_{P,\alpha,\sigma}$ is not computable. A computable estimator for the exponential distance was proposed by Sajama and Orlitsky (2005) but this estimator is

only consistent for $\alpha = 1$. Bijral, Ratliff and Srebro (2011) presented a computable estimator but it is not consistent. Here we given an algorithm that approximates $\widehat{D}_{\alpha,\sigma}$ and is consistent, uniformly over a range of values of α.

Remark: In this section we assume that P_X is known and we show how to approximate $D_{P,\alpha,\sigma}$. When P_X is unknown, simply subsitute $\widehat{p}_{X,\sigma}$ for $p_{X,\sigma}$. The bounds then get multiplied by a factor of $e^{\alpha\epsilon_m}$.

In what follows, "path" always means piecewise differentiable, continuous, finite length curve with unit speed where differentiable. Define $K^*_{\max} = \sup_u \|\nabla K(u)\|$ and $K_{\max} = \sup_u K(u) = K(0)$, and suppose that K is supported on the unit ball. Let $\sigma_{\max} \geq \sigma_{\min} > 0$ and $\alpha_{\max} \geq 0$. Let $\mathcal{X}^*$ be the convex hull of $\mathcal{X} \oplus \sigma_{\max}$. Let

$$\mathcal{C} = \{u_1, \ldots, u_J\}$$

be a Euclidean ζ-covering of $\mathcal{X}^*$. (The cover can, but need not, include the observed data.)

For any $0 \leq \alpha \leq \alpha_{\max}$ and $\sigma_{\min} \leq \sigma \leq \sigma_{\max}$, define the graph $G_{\alpha,\sigma} = (V, E, W_{\alpha,\sigma})$ where $V = \{v_1, ..., v_J\}$, $(v_i, v_j) \in E$ iff $\|u_i - u_j\| \leq \xi$, and for i, j s.t. $(v_i, v_j) \in E$ define the edge weight

$$W^{i,j}_{\alpha,\sigma} = \|u_i - u_j\| \exp\left[-\alpha p_{X,\sigma}\left(\frac{u_i + u_j}{2}\right)\right].$$

Note that each node v_j corresponds to one point u_j in the cover. Also define $\widetilde{G}_{\alpha,\sigma} = (V, E, \widetilde{W}_{\alpha,\sigma})$ where $\widetilde{W}^{i,j}_{\alpha,\sigma} = D_{P,\alpha,\sigma}(u_i, u_j)$ for i, j s.t. $(v_i, v_j) \in E$.

For any $0 \leq \alpha \leq \alpha_{\max}$ and $\sigma_{\min} \leq \sigma \leq \sigma_{\max}$, for $i, j \in \{1, ..., J\}$, define the estimated distance $\widehat{D}^*_{P,\alpha,\sigma}(u_i, u_j)$ and the intermediate distance $\widetilde{D}_{P,\alpha,\sigma}(u_i, u_j)$ to be the graph (i.e. shortest-path) distances between vertices v_i and v_j on $G_{\alpha,\sigma}$ and $\widetilde{G}_{\alpha,\sigma}$, resp. Note that the distances are only defined for points in $\{u_i\}_{i=1}^J$. Let

$$\Lambda = \exp\left[\frac{\alpha K_{\max}}{\sigma^d}\right]\frac{\alpha K^*_{\max}}{\sigma^{d+1}} \leq \exp\left[\frac{\alpha_{\max} K_{\max}}{\sigma^d_{\min}}\right]\frac{\alpha_{\max} K^*_{\max}}{\sigma^{d+1}_{\min}}.$$

Theorem 4.2 *If $\zeta \leq 7/(32\Lambda)$ then, for any $0 \leq \alpha \leq \alpha_{\max}$ and $\sigma_{\min} \leq \sigma \leq \sigma_{\max}$ and for any $i, j \in \{1, ..., J\}$,*

$$\frac{1}{8} D_{P,\alpha,\sigma}(u_i, u_j) \leq \widehat{D}_{P,\alpha,\sigma}(u_i, u_j) \leq 8 D_{P,\alpha,\sigma}(u_i, u_j). \tag{18}$$

Note that this bounds $\widehat{D}$, which is sufficient for our purposes. It is possible to modify the proof to show that, in fact, $\widehat{D}$ consistently estimates D.

Proof. Let $\xi = 4\zeta$. By Proposition 2 in Section 11 we have that for i, j s.t. $(v_i, v_j) \in E$

$$(1 - \Lambda\xi)W^{i,j}_{\alpha,\sigma} \leq \widetilde{W}^{i,j}_{\alpha,\sigma} \leq (1 + \Lambda\xi/2)W^{i,j}_{\alpha,\sigma}.$$

So for all $i, j \in \{1, ..., J\}$,

$$\frac{1}{1 + \Lambda\xi/2}\widetilde{D}_{\alpha,\sigma,P}(u_i, u_j) \leq \widehat{D}_{\alpha,\sigma,P}(u_i, u_j) \leq \frac{1}{1 - \Lambda\xi}\widetilde{D}_{\alpha,\sigma,P}(u_i, u_j).$$

Clearly for all $i, j \in \{1, ..., J\}$, $D_{P,\alpha,\sigma}(u_i, u_j) \leq \widetilde{D}_{\alpha,\sigma,P}(u_i, u_j)$. Also for i, j such that $\|u_i - u_j\| \leq \xi$ $D_{P,\alpha,\sigma}(u_i, u_j) = \widetilde{D}_{\alpha,\sigma,P}(u_i, u_j)$. Suppose i, j such that $\|u_i - u_j\| > \xi$. Let γ be the path such that $D_{P,\alpha,\sigma}(u_i, u_j) = \int_0^{L(\gamma)} \exp\left[-\alpha p_{X,\sigma}(\gamma(t))\right] dt$. Of course, $L(\gamma) > \xi$. Divide γ into a sequence of paths $\gamma_0, \gamma_1, ..., \gamma_Q$ such that $\gamma_0(0) = u_i$, $\gamma_Q(L(\gamma_Q)) = u_j$, $L(\gamma_0) \in (0, \xi - 2\zeta]$, and for $k \in \{1, ..., Q\}$ $L(\gamma_k) = \xi - 2\zeta$ and $\gamma_{k-1}(L(\gamma_{k-1})) = \gamma_k(0)$. Clearly $Q \geq 1$. For $k \in \{1, ..., Q\}$ let $r_k = \gamma_k(0)$, let $r_{Q+1} = u_j$, and let n_k such that $\|r_k - u_{n_k}\| \leq \zeta$. By Proposition 1 in Section 11,

$$\begin{aligned}
\widetilde{D}_{\alpha,\sigma,P}(u_i, u_j) &\leq \widetilde{W}^{i,n_1}_{\alpha,\sigma} + \sum_{k=1}^{Q} \widetilde{W}^{n_k,n_{k+1}}_{\alpha,\sigma} \\
&\leq D_{P,\alpha,\sigma}(u_i, r_1) + \sum_{k=1}^{Q} \left(2D_{P,\alpha,\sigma}(r_k, u_{n_k}) + D_{P,\alpha,\sigma}(r_k, r_{k+1})\right) \\
&= D_{P,\alpha,\sigma}(u_i, r_1) + \sum_{k=1}^{Q} D_{P,\alpha,\sigma}(r_k, r_{k+1}) \left(1 + 2\frac{D_{P,\alpha,\sigma}(r_k, u_{n_k})}{D_{P,\alpha,\sigma}(r_k, r_{k+1})}\right) \\
&\leq D_{P,\alpha,\sigma}(u_i, r_1) + \sum_{k=1}^{Q} D_{P,\alpha,\sigma}(r_k, r_{k+1}) \left[1 + 2\zeta\left(\Lambda + \frac{1 + \Lambda\zeta}{\xi - 2\zeta}\right)\right] \\
&\leq \left[1 + 2\zeta\left(\Lambda + \frac{1 + \Lambda\zeta}{\xi - 2\zeta}\right)\right] D_{P,\alpha,\sigma}(u_i, u_j)
\end{aligned}$$

and the result follows since $\xi = 4\zeta$ and $\zeta \leq 7/(32\Lambda)$. □

Remark: For any points x, y not in the cover $\mathcal{C}$, we can define $\widehat{D}(x, y) =$

1. Construct a Euclidean ζ-covering $\{u_i\}_{i=1}^J$ of the convex hull of $\mathcal{X} \oplus \sigma$.

2. Construct a graph with the covering points as nodes, and edges between pairs of points closer than ξ for some $\xi > 2\zeta$.

3. Set the edge weight between connected neighbors i and j to

$$\|u_i - u_j\| \exp\left[-\alpha\, p_{X,\sigma}\left(\frac{u_i + u_j}{2}\right)\right].$$

4. Approximate the $D_{P,\alpha,\sigma}$-distance between any two points as the graph (i.e. shortest path) distance between the corresponding nearest neighbors in $\{u_i\}_{i=1}^J$.

FIG 4. *Computing Density-Sensitive Metrics*

$\widehat{D}(u_i, u_j)$ where u_i is the closest point in $\mathcal{C}$ to x and u_j is the closest point in $\mathcal{C}$ to y.

A summary of the algorithm is given in Figure 4.

To further speed up the algorithm, we have found the following heuristic to be useful. We approximate the edge weight between all pairs of points i and j by

(19)
$$W_{i,j} = \begin{cases} \|X_i - X_j\| \exp\left[-\alpha \frac{p_i + p_j}{2}\right] & \text{if } X_j \text{is a } k\text{-nearest neighbor of } X_i, \\ \|X_i - X_j\| & \text{otherwise} \end{cases}$$

where k is an integer and p_i is the k-NN density estimate at the i'th point.

5. Density-Sensitive Inference. We consider the following semisupervised learner which uses a kernel that is sensitive to the density. Let Q be a kernel and let $Q_h(x) = h^{-d} Q(x/h)$. Let

(20)
$$\widehat{f}_{h,\alpha,\sigma}(x) = \frac{\sum_{i=1}^n Y_i\, Q_h\left(\widehat{D}_{\alpha,\sigma}(x, X_i)\right)}{\sum_{i=1}^n Q_h\left(\widehat{D}_{\alpha,\sigma}(x, X_i)\right)}.$$

In the following we take, for simplicity, $Q(x) = I(||x|| \leq 1)$. Now we give an upper bound on the risk of $\widehat{f}_{h,\alpha,\sigma}$.

Theorem 5.1 *Suppose that* $|Y| \leq M$. *Define the event* $\mathcal{G}_m = \{||\widehat{p}_\sigma - p_\sigma||_\infty \leq \epsilon_m\}$ *(which depends on the unlabeled data) and suppose that* $\mathbb{P}(\mathcal{G}_m^c) \leq 1/m$. *Then, for every* $P \in \mathcal{P}(\alpha, \sigma, L)$,

$$R_P(\widehat{f}_{h,\alpha,\sigma}) \leq L^2(he^{\alpha\epsilon_m})^2 + \frac{M^2\left[4 + \frac{1}{e}\mathcal{N}(P, \alpha, \sigma, e^{-\epsilon_m\alpha}h/2)\right]}{n} + \frac{4M^2}{m}. \tag{21}$$

Proof. The risk is

$$R_P(\widehat{f}) = \mathbb{E}_{n,N}\left[(1-\mathcal{G}_m)\int(\widehat{f}_{h,\alpha,\sigma}(x) - f(x))^2 dP(x)\right] + \mathbb{E}_{n,N}\left[\mathcal{G}_m\int(\widehat{f}_{h,\alpha,\sigma}(x) - f(x))^2 dP(x)\right].$$

Since $|Y| \leq M$ and $\sup_x |\widehat{f}(x)| \leq M$,

$$\mathbb{E}_{n,N}\left[(1-\mathcal{G}_m)\int(\widehat{f}_{h,\alpha,\sigma}(x) - f(x))^2 dP(x)\right] \leq 4M^2\mathbb{P}(\mathcal{G}_m^c) \leq \frac{4M^2}{m}.$$

Now we bound the second term.

Condition on the unlabeled data. Replacing Euclidean distance with $\widehat{D}_{\alpha,\sigma}$ in the proof of Theorem 5.2 in Gyorfi et al. (2002), we have that

$$\mathbb{E}_n\left[\int(\widehat{f}_{h,\alpha,\sigma}(x) - f(x))^2 dP(x)\right] \leq LR^2 + \frac{M^2\left[4 + \frac{1}{e}\int\frac{dP(x)}{P(\widehat{B}_{\alpha,\sigma}(x,h))}\right]}{n}$$

where

$$R = \sup\Big\{D_{P,\alpha,\sigma}(x_1, x_2) : \ (x_1, x_2) \text{ such that } \widehat{D}_{\alpha,\sigma}(x_1, x_2) \leq h\Big\}$$

and $\widehat{B}_{\alpha,\sigma}(x,h) = \{z : \ \widehat{D}_{\alpha,\sigma}(x,z) \leq h\}$. On the event $\mathcal{G}_m$, we have from Lemma 3 that $e^{-\alpha\epsilon_m}D_{\alpha,\sigma}(x_1,x_2) \leq \widehat{D}_{\alpha,\sigma}(x_1,x_2) \leq e^{\alpha\epsilon_m}D_{\alpha,\sigma}(x_1,x_2)$ for all x_1, x_2. Hence, $R^2 \leq e^{2\alpha\epsilon_m}h^2$ and

$$\int\frac{dP(x)}{P(\widehat{B}_{\alpha,\sigma}(x,h))} \leq \int\frac{dP(x)}{P(B_{P,\alpha,\sigma}(x, e^{-\alpha\epsilon_m}h))}.$$

A simple covering argument (see p 76 of Gyorfi et al) shows that, for any $\delta > 0$,

$$\int\frac{dP(x)}{P(B_{P,\alpha,\sigma}(x,\delta))} \leq \mathcal{N}(P, \alpha, \sigma, \delta/2).$$

The result follows. □

Corollary 5.2 *If* $\mathcal{N}(P, \alpha, \sigma, \delta) \leq (C/\delta)^{\xi}$ *for* $\delta \geq (1/2)e^{-\alpha\epsilon_m}(ne^{2\alpha\epsilon_m})^{-\frac{1}{2+\xi}}$ *and* $N \geq 2n$ *then*

$$R_P(\widehat{f}_{\alpha,\sigma,h}) \leq e^{\alpha\epsilon_m(2\vee\xi)}\left[L^2h^2 + \frac{1}{n}\left(\frac{C}{h}\right)^{\xi}\right] + \frac{8M^2}{m}. \tag{22}$$

Hence, if $m \geq n^{2/(2+\xi)}$, *and* $h \asymp (ne^{\alpha\epsilon_m(2-\xi)})^{-\frac{1}{2+\xi}}$ *then*

$$\sup_{P\in\mathcal{P}(\alpha,\sigma,L)} R_P(\widehat{f}_{h,\alpha,\sigma}) \preceq \left(\frac{C}{n}\right)^{\frac{2}{2+\xi}}. \tag{23}$$

6. Minimax Bounds. To characterize when semisupervised methods outperform supervised methods, we show that there is a class of distributions $\mathcal{P}_n$ (which we allow to change with n) such that R_{SS} is much smaller than R_S, where

$$R_S = \inf_{\widehat{f}\in\mathcal{S}_n}\sup_{P\in\mathcal{P}_n} R_P(\widehat{f}) \quad \text{and} \quad R_{SS} = \inf_{\widehat{f}\in\mathcal{SS}_N}\sup_{P\in\mathcal{P}_n} R_P(\widehat{f})$$

To do so, it suffices to find a lower bound on R_S and an upper bound on R_{SS}. In intuitively, $\mathcal{P}_n$ should be a set distributions whose X-marginals are highly concentrated on or near lower-dimensional sets, since this is where semisupervised methods deliver improved performance. Indeed, as we mentioned earlier, for very smooth distributions P_X we do not expect semisupervised learners to offer much improvement.

6.1. *The Class* $\mathcal{P}_n$. Here we define the class $\mathcal{P}_n$. Let $N = N(n)$ and $m = m(n) = N - n$ and define

$$\epsilon_m \equiv \epsilon(m, \sigma) = \sqrt{\frac{C\log m}{m\sigma^d}}. \tag{24}$$

Let $\xi \in [0, d-3)$, $\gamma > 0$ and define

$$\mathcal{P}_n = \bigcup_{(\alpha,\sigma)\in\mathcal{A}_n\times\Sigma_n} \mathcal{Q}(\alpha, \sigma, L) \tag{25}$$

where $\mathcal{Q}(\alpha,\sigma,L) \subset \mathcal{P}(\alpha,\sigma,L)$ and $\mathcal{A}_n \times \Sigma_n \subset [0,\infty]^2$ satisfy the following conditions:

$$\textbf{(C1)} \qquad \frac{\alpha}{\sigma^d} > C\left(\frac{d-\xi-1}{(d-1)(2+\xi)}\right)\log n$$

$$\textbf{(C2)} \qquad \mathcal{Q}(\alpha,\sigma,L) = \left\{P \in \mathcal{P}(\alpha,\sigma,L):\ \mathcal{N}(P,\alpha,\sigma,\epsilon) \le \left(\frac{C}{\epsilon}\right)^{\xi} \quad \forall\, \epsilon \ge \left(\frac{1}{n}\right)^{\frac{1}{2+\xi}}\right\}$$

$$\textbf{(C3)} \qquad \alpha \le \frac{\log 2}{\epsilon(m,\sigma)}.$$

$$\textbf{(C4)} \qquad \left(\frac{1}{m}\right)^{\frac{1}{d(1+\gamma)}} \le \sigma \le \frac{1}{4C_0}\left(\frac{1}{n}\right)^{\frac{1}{d-1}}$$

where C_0 is the diameter of the support of K.

Here are some remarks about $\mathcal{P}_n$:

1. (C3) implies that $e^{\alpha\epsilon_m} \le 2$ and hence that $(1/2)D_{P,\alpha,\sigma}(x_1,x_2) \le \widehat{D}_{\alpha,\sigma}(x_1,x_2) \le 2D_{P,\alpha,\sigma}(x_1,x_2)$ with high probability.
2. (C4) implies that $m \ge (1/\sigma)^{-d(1+\gamma)}$ for each $\sigma \in \Sigma_n$. Hence, from the discussion following Theorem 4.1,
$$\sup_{P\in\mathcal{P}_n} P^m\left(\sup_{\sigma\in\Sigma_n} ||\widehat{p}_\sigma - p_\sigma||_\infty > \epsilon(m,\sigma)\right) < \frac{1}{m}$$
and thus, Theorem 5.1 and Corollary 5.2 apply.
3. The constraint in (C2) on $\mathcal{N}(\epsilon)$ holds whenever P is concentrated on or near a set of dimension less than d and α/σ^d is large. The constraint does not need to hold for arbitrarily small ϵ.
4. Some papers on semisupervised learning simply assume that $N = \infty$ since in practice N is usually very large compared to n. In that case, there is no upper bound on α and no lower bound on σ.

The class $\mathcal{P}_n$ may seem complicated. This is because showing conditions where semisupervised learning provably outperforms supervised learning is subtle. Intuitively, the class $\mathcal{P}_n$ is simply the set of high concentrated distributions with α/σ large.

6.2. *Supervised Lower Bound.*

Theorem 6.1 *There exists $C > 0$ such that*

$$R_S = \inf_{\widehat{f}\in\mathcal{S}_n}\sup_{P\in\mathcal{P}_n} R_P(\widehat{f}) \ge \left(\frac{C}{n}\right)^{\frac{2}{d-1}}. \tag{26}$$

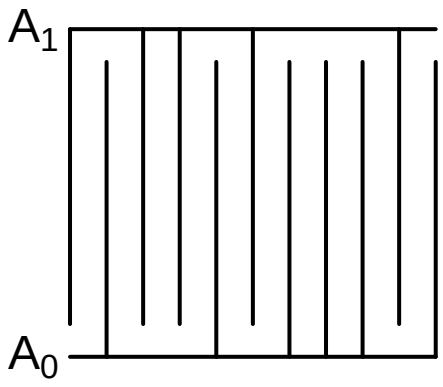

FIG 5. *The extended tendrils used in the proof of the lower bound, in the special case where $d = 2$. Each tendril has length $1 - \epsilon$ and joins up with either the top A_1 or bottom A_0 but not both.*

Proof. Let A_1 and A_0 be the top and bottom of the cube $\mathcal{X}$:

$$\begin{aligned} A_1 &= \{(x_1, \ldots, x_{d-1}, 1) : \ 0 \le x_1, \ldots, x_{d-1} \le 1\} \\ A_0 &= \{(x_1, \ldots, x_{d-1}, 0) : \ 0 \le x_1, \ldots, x_{d-1} \le 1\}. \end{aligned}$$

Fix $\epsilon = n^{-\frac{1}{d-1}}$. Let $q = (1/\epsilon)^{d-1} \asymp n$. For any integers $s = (s_1, \ldots, s_{d-1}) \in N^{d-1}$ with $0 \le s_i \le 1/\epsilon$, define the tendril

$$\{(s_1\epsilon, s_2\epsilon, \ldots, s_{d-1}\epsilon, x_d) : \ \epsilon \le x_d \le 1 - \epsilon\}.$$

There are $q = (1/\epsilon)^{d-1} \approx n$ such tendrils. Let us label the tendrils as $T_1, \ldots, T_q$. Note that the tendrils do not quite join up with A_0 or A_1.

Let

$$C = A_0 \bigcup A_1 \bigcup \left(\bigcup_{j=1}^{q} T_j \right).$$

Define a measure μ on C as follows:

$$\mu = \frac{1}{4}\mu_0 + \frac{1}{4}\mu_1 + \frac{1}{2q(1-2\epsilon)} \sum_j \nu_j$$

where μ_0 is $(d-1)$-dimensional Lebesgue measure on A_0, μ_1 is $(d-1)$-dimensional Lebesgue measure on A_1 and ν_j is one-dimensional Lebesgue measure on T_j. Thus, μ is a probability measure and $\mu(C) = 1$.

Now we define extended tendrils that are joined to the top or bottom of the cube (but not both). See Figure 5. If

$$T_j = \{(s_1\epsilon, s_2\epsilon, \ldots, s_{d-1}\epsilon, x_d) : \ \epsilon \le x_d \le 1 - \epsilon\}.$$

is a tendril, define its extensions

$$\begin{aligned} T_{j,0} &= \{(s_1\epsilon, s_2\epsilon, \ldots, s_{d-1}\epsilon, x_d) : \ 0 \le x_d \le 1-\epsilon\} \\ T_{1,j} &= \{(s_1\epsilon, s_2\epsilon, \ldots, s_{d-1}\epsilon, x_d) : \ \epsilon \le x_d \le 1\}. \end{aligned}$$

Given $\omega \in \Omega = \{0,1\}^q$ let

$$S_\omega = A_0 \bigcup A_1 \bigcup \left(\bigcup_{j=1}^{q} T_{j,\omega_j} \right)$$

and

$$P_{\omega,X} = \frac{1}{4}\mu_0 + \frac{1}{4}\mu_1 + \frac{1}{2q(1-\epsilon)} \sum_j \nu_{j,\omega_j}$$

where ν_{j,ω_j} is one-dimensional Lebesgue measure on T_{j,ω_j}. This $P_{\omega,X}$ is a probability measure supported on S_ω.

Notice that S_ω consists of two connected components, namely,

$$U_\omega^{(1)} = A_1 \bigcup \left(\bigcup_{j:\omega_j=1} T_{j,\omega_j} \right) \quad \text{and} \quad U_\omega^{(0)} = A_0 \bigcup \left(\bigcup_{j:\omega_j=0} T_{j,\omega_j} \right).$$

Let

$$f_\omega(x) = \frac{L\,\epsilon}{8}\, I(x \in U_\omega^{(1)}).$$

Finally, we define $P_\omega = P_{\omega,X} \times P_{\omega,Y|X}$ where $P_{\omega,Y|X}$ is a point mass at $f_\omega(X)$. Define $d^2(f,g) = \int (f(x) - g(x))^2 d\mu(x)$.

We complete the proof with a series of claims.

Claim 1: For each $\omega \in \Omega$, $P_\omega \in \mathcal{P}_n$.

Proof: Since the definition of the $\{P_\omega\}$ does not depend on (C1), (C3) or (C4), we may simply choose α and σ to satisfy these three constraints. We must then verify (C2). If x and y are in the same connected component then $|f_\omega(x) - f_\omega(y)| = 0$. Now let x and y be in different components, i.e. $x \in U_\omega^{(1)}, y \in U_\omega^{(0)}$. Let us choose x and y as close as possible in Euclidean distance; hence $||x - y|| = \epsilon$. Let γ be any path connecting x to y. Since x and y lie on different components, there exists a subset γ_0 of γ of length at least ϵ on which P_ω puts zero mass. By assumption (C4), $\sigma \le \epsilon/(4C_0)$ and hence $P_{X,\sigma}$ puts zero mass on the portion of γ_0 that is at least $C_0\sigma$

away from the support of P_ω. This has length at least $\epsilon - 2C_0\sigma \geq \epsilon/2$. Since $p_{X,\sigma}(x) = 0$ on a portion of γ_0,

$$D_{P,\alpha,\sigma}(x,y) \geq \frac{\epsilon}{2} = \frac{||x-y||}{2}.$$

Hence, $||x-y|| \leq 2D_{P,\alpha,\sigma}(x,y)$. Then

$$\frac{|f_\omega(x) - f_\omega(y)|}{D_{P,\alpha,\sigma}(x,y)} \leq \frac{2|f_\omega(x) - f_\omega(y)|}{||x-y||}$$

and the latter is maximized by finding two points x and y as close together with nonzero numerator. In this case, $||x-y|| = \epsilon$ and $|f_\omega(x) - f_\omega(y)| = L\epsilon/8$. Hence, $|f_\omega(x) - f_\omega(y)| \leq LD_{P,\alpha,\sigma}(x,y)$ as required. Now we show that each $P = P_\omega$ satisfies

$$\mathcal{N}(P,\alpha,\sigma,\epsilon) \leq \left(\frac{C}{\epsilon}\right)^\xi$$

for all $\epsilon \geq n^{-\frac{1}{2+\xi}}$. Cover the top A_1 and bottom A_0 of the cubes with Euclidean spheres of radius δ. There are $O((1/\delta)^{d-1})$ such spheres. The $D_{P,\alpha,\sigma}$ radius of each sphere is at most $\delta e^{-\alpha K(0)/\sigma^d}$. Thus, these form an ϵ covering as long as $\delta e^{-\alpha K(0)/\sigma^d} \leq \epsilon$. Thus the covering number of the top and bottom is at most $2(1/\delta)^{d-1} \leq 2(1/(e^{\alpha K(0)/\sigma^d}\epsilon))^{d-1}$. Now cover the tendris with one-dimensional segments of length δ. The $D_{P,\alpha,\sigma}$ radius of each segment is at most $\delta e^{-\alpha/\sigma^d}$. Thus, these form an ϵ covering as long as $\delta e^{-\alpha K(0)/\sigma^d} \leq \epsilon$. Thus the covering number of the tendrils is at most $q/\delta = n/\delta \leq n/(\epsilon e^{\alpha K(0)/\sigma^2})$. Thus we can cover the support with

$$N(\epsilon) \leq 2\left(\frac{1}{e^{\alpha/\sigma^d}\epsilon}\right)^{d-1} + \frac{n}{\epsilon e^{\alpha/\sigma^2}}$$

balls of size ϵ. (C2) then implies that $N(\epsilon) \leq (1/\epsilon)^\xi$ for $\epsilon \geq n^{-\frac{1}{2+\xi}}$ as required.

Claim 2: For any ω, and any $g \geq 0$, $\int g(x)dP_\omega(x) \geq \frac{1}{2}\int g(x)d\mu(x)$.

Proof: We have

$$\begin{aligned}\int_{S_\omega} g dP_\omega &\geq \int_C g dP_\omega = \frac{1}{4}\int_{A_0} g d\mu_0 + \frac{1}{4}\int_{A_1} g d\mu_1 + \frac{\sum_j \int_{T_j} g d\nu_{j,\omega}}{2q(1-\epsilon)} \\ &= \frac{1}{4}\int_{A_0} g d\mu_0 + \frac{1}{4}\int_{A_1} g d\mu_1 + \frac{\frac{1-2\epsilon}{1-\epsilon}\sum_j \int_{T_j} g d\nu_j}{2q(1-2\epsilon)} \times \frac{\frac{1}{2}+q(1-2\epsilon)}{\frac{1}{2}+q(1-\epsilon)} \\ &\geq \frac{1}{2}\left(\frac{1}{4}\int_{A_0} g d\mu_0 + \frac{1}{4}\int_{A_1} g d\mu_1 + \frac{\sum_j \int_{T_j} g d\nu_j}{2q(1-2\epsilon)}\right) = \frac{1}{2}\int g\, d\mu.\end{aligned}$$

Claim 3: For any $\omega, \nu \in \Omega$,

$$d^2(f_\omega, f_\nu) = \frac{\rho(\omega,\nu)L^2\epsilon^2(1-2\epsilon)}{2q(1-2\epsilon)}.$$

Proof: This follows from direct calculation.

Claim 4: If $\rho(\omega,\nu) = 1$ then $||P_\omega^n \wedge P_\nu^n|| \geq 1/(16e)$.

Proof: Suppose that $\rho(\omega,\nu) = 1$. P_ω and P_ν are the same everywhere except $T_{j,0} \cup T_{j,1}$, where j is the index where ω and ν differ (assume $\omega_j = 0$ and $\nu_j = 1$). Define $A = T_{j,0} \times \{0\}$ and $B = T_{j,1} \times \{L\epsilon\}$. Note that $A \cap B = \emptyset$. So,

$$P_\omega(T_{j,0} \cup T_{j,1}) = P_\omega(A) = P_\nu(T_{j,0} \cup T_{j,1}) = P_\nu(B) = \frac{1-\epsilon}{2q(1-\epsilon)}$$

and

$$\begin{aligned}\mathsf{TV}(P_\omega, P_\nu) &= |P_\omega(A) - P_\nu(A)| = |P_\omega(B) - P_\nu(B)| \\ &= \frac{1-\epsilon}{2q(1-\epsilon)} = \frac{1}{2q} = \frac{\epsilon^{d-1}}{2}.\end{aligned}$$

Thus,

$$\|P_\omega^n \wedge P_\nu^n\| \geq \frac{1}{8}\left(1 - \mathsf{TV}(P_\omega, P_\nu)\right)^{2n} \geq \frac{1}{8}\left(1 - \epsilon^{d-1}/2\right)^{2n}.$$

Since $\epsilon = n^{-\frac{1}{d-1}}$, this implies that

$$\|P_\omega^n \wedge P_\nu^n\| \geq \frac{1}{8}\left(1 - \frac{1}{2n}\right)^{2n} \geq \frac{1}{16e}$$

for all large n.

Completion of the proof. Recall that $\epsilon = n^{-\frac{1}{d-1}}$. Combining Assouad's Lemma with the above claims, we have

$$\begin{aligned} R_S &= \inf_{\widehat{f}\in\mathcal{S}_n}\sup_{P\in\mathcal{P}_{n,\xi}} R_P(\widehat{f}) \geq \inf_{\widehat{f}\in\mathcal{S}_n}\sup_{P\in\mathcal{P}_{\Omega}} R_P(\widehat{f}) \geq \frac{1}{2}\inf_{\widehat{f}}\max_{\omega\in\Omega}\mathbb{E}_\omega[d^2(f_\omega,\widehat{f})] \\ &\geq \frac{q}{16}\times\frac{(L/8)^2\epsilon^2(1-2\epsilon)}{2q(1-2\epsilon)}\times\frac{1}{16e} = C\frac{q\epsilon^2(1-2\epsilon)}{2q(1-2\epsilon)} \\ &\geq C\epsilon^2 = Cn^{-\frac{2}{d-1}} \end{aligned}$$

□

6.3. *Semisupervised Upper Bound.* Now we state the upper bound for this class.

Theorem 6.2 *Let* $h = (ne^{2(2-\xi)})^{-\frac{1}{2+\xi}}$. *Then*

$$\sup_{P\in\mathcal{P}_n} R(\widehat{f}_{h,\alpha,\sigma}) \leq \left(\frac{C}{n}\right)^{\frac{2}{2+\xi}}. \tag{27}$$

Proof. This follows from (C2), (C3) and Corollary 5.2. □

6.4. *Comparison of Lower and Upper Bound.* Combining the last two theorems we have:

Corollary 6.3 *Under the conditions of the previous theorem, and assuming that* $d > \xi + 3$,

$$\frac{R_{SS}}{R_S} \preceq \left(\frac{1}{n}\right)^{\frac{2(d-3-\xi)}{(2+\xi)(d-1)}} \to 0 \tag{28}$$

as $n \to \infty$.

This establishes the effectiveness of semi-supervised inference in the minimax sense.

7. The Reciprocal Distance. In this section we consider a second densitive-sensitive metric, called the reciprocal distance. This distance is more difficult to implement but it provides a more dramatic distinction between supervised and semisupervised methods.

Define the reciprocal distance

$$D(x_1, x_2) \equiv D_{P,\alpha}(x_1, x_2) = \inf_{\gamma \in \Gamma(x_1, x_2)} \int_0^{L(\gamma)} \left[\frac{1}{p_X(\gamma(t))} \right]^\alpha dt. \tag{29}$$

Let $\mathcal{N}(P, \alpha, \epsilon)$ denote the covering number of S_P under this distance. Let $\mathcal{F} \equiv \mathcal{F}(P, \alpha, L)$ denote the set functions $f : [0,1]^d \to \mathbb{R}$ such that, for all $x_1, x_2 \in \mathcal{X}$,

$$|f(x_1) - f(x_2)| \leq L \; D_{P,\alpha}(x_1, x_2). \tag{30}$$

Let $\mathcal{P}(\alpha, L)$ denote all joint distributions for (X, Y) such that $f_P \in \mathcal{F}(P, \alpha, L)$ and such that P_X is supported on $\mathcal{X}$. The rest of the section shows that there is a class $\mathcal{P}_n$ where semisupervised inference provably outperforms supervised inference under the reciprocal distance.

7.1. *The Class $\mathcal{P}_n$.* The *condition number* $\tau(S)$ of a set S with boundary ∂S is the largest real number $\tau > 0$ such that, if $d(x, \partial S) \leq \tau$ then x has a unique projection onto the boundary of S. Here, $d(x, \partial S) = \inf_{z \in \partial S} ||x - z||$. When τ is large, S cannot be too thin, the boundaries of S cannot be too curved and S cannot get too close to being self-intersecting. If S consists of more than one connected component, then τ large also means that the connected components cannot be too close to each other.

Let $\epsilon_m = c_1 (\log m)^{-1/2}$ and $\delta_m = 2c_2 \sqrt{d} \left((\log^2 m)/m \right)^{\frac{1}{d}}$. Let $\mathcal{W}(K, \lambda, \tau_n)$ denote all distributions P such that $\tau(S_P) \geq \tau_n$, the number of connected components of S_P is at most K and

$$1 < \lambda \leq \inf_{x \in S_P} p(x) \leq \sup_{x \in S_P} p(x) \leq \Lambda < \infty.$$

As before, let $N = N(n)$ and $m = m(n) = N - n$. Also, let $\eta > 0$. Define

$$\mathcal{P}_n = \left[\bigcup_{\alpha \in \mathcal{A}_n} \mathcal{P}(\alpha, L) \right] \bigcap \mathcal{W}(K, \lambda, \tau_n) \tag{31}$$

where $\mathcal{A}_n \subset [0, \infty]$ satisfies the following conditions:

(**C1**) $\tau_n \succeq n^{-(d-1)}$.

(**C2**) $\epsilon_m \leq \min\Big\{\lambda/2, \lambda[2^{1/\alpha} - 1]\Big\}$.

(**C3**) $\left(\frac{\lambda}{2}\right)^{\alpha} > \frac{2n}{\tau_n^{d-1}}$.

(**C4**) $\frac{\delta_m}{\tau_n} \leq \frac{1}{n}$.

(**C5**) $m \succeq n^{d^2}$.

(**C6**) p is Holder(η) smooth over its support.

7.2. *Supervised Lower Bound.*

Theorem 7.1 *Assume $d \geq 2$. Then, there exists $C > 0$ such that*

$$\inf_{\widehat{f} \in \mathcal{S}_n} \sup_{P \in \mathcal{P}_n} \mathbb{E}_n \int (\widehat{f}(x) - f(x))^2 dP(x) \geq C.$$

First we provide some intuition regarding the proof strategy. We construct a set of joint distributions over X and Y that depends on n, and apply Assouad's Lemma. Intuitively, we need to take advantage of the decreasing condition number τ_n. This is because if τ_n were to be kept fixed, as n increases the semi-supervised assumption would reduce to familiar Euclidean smoothness.

We construct the distributions as follows. We split the unit cube in $\mathbb{R}^d$ into two rectangle sets with a small gap in between, and let the marginal density p be uniform over these sets. Then we add a series of "bumps" between the two rectangles, as shown schematically in Figure 6. Over one of the sets we set $f \equiv M$, and over the other we set $f \equiv -M$. The number of bumps increases with n, implying that the condition number must decrease. The sets are designed specifically so that the condition number can be lower bounded easily as a function of n. In essence, as n increases these boundaries become space-filling, so that there is a region where the regression function could be M or $-M$, and it is not possible to tell which with only labeled data.

Proof. Step 1: Constructing the hypercube. Let $l = \lfloor c_0 n^{1/(d-1)} \rfloor$ with $c_0 > 1$ a constant, $q = l^{d-1}$, $\Omega = \{0,1\}^q$ and $\epsilon = \frac{1}{l+2}$. For $i \in \{1, ..., l\}$, let $a_i = \frac{i+0.5}{l+2}$. For $\vec{i} \in \{1, ..., l\}^{d-1}$, let $v_{\vec{i}} = (a_{\vec{i}_1}, ..., a_{\vec{i}_{d-1}})$. Define $g : \mathbb{R}^{d-1} \to$

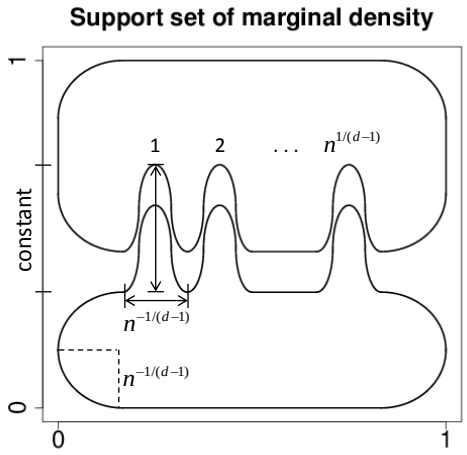

FIG 6. *A two-dimensional cross-section of the support of a marginal density p used in the proof of Theorem 7.1.*

$\mathbb{R}$ as $g(\widetilde{x}) =$

$$g(\widetilde{x}) = \begin{cases} r + \sqrt{\left(\frac{1}{2} - r\right)^2 - \|\widetilde{x}\|_2^2} & \text{for} \quad \|\widetilde{x}\|_2 < \frac{1}{2} - r \\ r - \sqrt{r^2 - \left(\frac{1}{2} - \|\widetilde{x}\|_2\right)^2} & \text{for} \quad \frac{1}{2} - r \le \|\widetilde{x}\|_2 < \frac{1}{2} \\ 0 & \text{o.w.} \end{cases} \tag{32}$$

for $\widetilde{x} \in \mathbb{R}^{d-1}$, where $r \in (0, 1/4)$ will be specified later (see Equation 33). Let

$$B = \{(\widetilde{x}, x_d) \in [-0.5, 0.5]^{d-1} \times [0, 1] : x_d \le g(\widetilde{x})\}.$$

For $\vec{i} \in \{1, ..., l\}^{d-1}$, let

$$\underline{B}_{\vec{i}} = \{(\widetilde{x}, x_d) \in \mathbb{R}^{d-1} \times \mathbb{R} : ((\widetilde{x} - v_{\vec{i}})/\epsilon, x_d - 1/8) \in B\}$$

and

$$\overline{B}_{\vec{i}} = \{(\widetilde{x}, x_d) \in \mathbb{R}^{d-1} \times \mathbb{R} : ((\widetilde{x} - v_{\vec{i}})/\epsilon, x_d - (1/8 + r)) \in B\}.$$

Let

$$\underline{S} = \{x \in \mathbb{R}^d : \exists x' = (\widetilde{x}', x_d') \in [\epsilon, 1 - \epsilon]^{d-1} \times [\epsilon, \frac{1}{8} - \epsilon] \text{ s.t. } \|x - x'\|_2 \le \epsilon\}$$

and

$$\overline{S} = \{x \in \mathbb{R}^d : \exists x' = (\widetilde{x}', x_d') \in [\epsilon, 1-\epsilon]^{d-1} \times [\frac{1}{8} + r + \epsilon, 1-\epsilon] \text{ s.t. } \|x-x'\|_2 \le \epsilon\}.$$

For any $\Gamma \subseteq \{1,...,l\}^{d-1}$, let $\underline{S}_\Gamma = \underline{S} \cup \left(\bigcup_{\vec{i} \in \Gamma} \underline{B}_{\vec{i}} \right)$ and $\overline{S}_\Gamma = \overline{S} \backslash \left(\bigcup_{\vec{i} \in \Gamma} \overline{B}_{\vec{i}} \right)$. Let $\vec{\Gamma}$ be an arbitrary ordering of $\{1,...,l\}^{d-1}$. Given $\omega \in \Omega$, let $\Gamma(\omega) = \{\vec{\Gamma}_i : \omega_i = 1\}$, and let $\underline{S}^\omega = \underline{S}_{\Gamma(\omega)}$, $\overline{S}^\omega = \overline{S}_{\Gamma(\omega)}$, and $S^\omega = \underline{S}^\omega \cup \overline{S}^\omega$.

Let $p^\omega(x) = \frac{I_{S^\omega}(x)}{\mathrm{Leb}(S^\omega)}$, $f^\omega(x) = MI_{\underline{S}^\omega}(x) - MI_{\overline{S}^\omega}(x)$, and let $P^\omega_{Y|X}$ be a point mass at $f^\omega(x)$. Finally, let P^ω denote the measure on $\mathbb{R}^{d+1}$ defined by the X marginal $p^\omega(x)$ and the conditional distributionand $P^\omega_{Y|X}$.

Step 2: Lower Bound. Note that $\mathrm{Leb}(\underline{B}_{\vec{i}}) = \mathrm{Leb}(\overline{B}_{\vec{i}})$, and so for any $\omega, \omega', \mathrm{Leb}(S^\omega) = \mathrm{Leb}(S^{\omega'}) = \mathrm{Leb}(\underline{S}) + \mathrm{Leb}(\overline{S})$. Let $\lambda = 1/(\mathrm{Leb}(\underline{S}) + \mathrm{Leb}(\overline{S}))$, i.e. $\lambda = 1/\mathrm{Leb}(S^\omega)$ for any ω. Let $\omega, \omega' \in \Omega$ such that $\rho(\omega, \omega') = 1$ (where ρ denotes the Hamming distance), and without loss of generality assume $\omega_i = 0$ and $\omega'_i = 1$. Also denote $\vec{i} = \vec{\Gamma}_i$. Then the L_1 distance between P^ω and $P^{\omega'}$ is

$$\begin{aligned}
d_1(P^\omega, P^{\omega'}) &= \int_{\mathbb{R}^d} \int_{\mathbb{R}} |p^\omega(x) dP^\omega_{Y|X} - p^{\omega'}(x) dP^{\omega'}_{Y|X}| dx \\
&= \int_{\underline{S}^\omega \cup \overline{S}^{\omega'}} \int_{\mathbb{R}} \lambda \left| dP^\omega_{Y|X} - dP^{\omega'}_{Y|X} \right| dx + \int_{\overline{B}_{\vec{i}} \backslash \underline{B}_{\vec{i}}} \int_{\mathbb{R}} \lambda P^\omega_{Y|X} dx + \int_{\underline{B}_{\vec{i}} \backslash \overline{B}_{\vec{i}}} \int_{\mathbb{R}} \lambda P^{\omega'}_{Y|X} dx \\
&+ \int_{\underline{B}_{\vec{i}} \cap \overline{B}_{\vec{i}}} \int_{\mathbb{R}} \lambda \left| dP^\omega_{Y|X} - dP^{\omega'}_{Y|X} \right| dx \\
&= 0 + \lambda \,\mathrm{Leb}(\overline{B}_{\vec{i}} \backslash \underline{B}_{\vec{i}}) + \lambda \,\mathrm{Leb}(\underline{B}_{\vec{i}} \backslash \overline{B}_{\vec{i}}) + 2\lambda \,\mathrm{Leb}(\underline{B}_{\vec{i}} \cap \overline{B}_{\vec{i}}) \\
&= \lambda(\mathrm{Leb}(\underline{B}_{\vec{i}}) + \mathrm{Leb}(\overline{B}_{\vec{i}})) = 2\lambda \epsilon^{d-1} \,\mathrm{Leb}(B)
\end{aligned}$$

where in the first step we have used the fact that $x \notin S^\omega \cup S^{\omega'} \Rightarrow p^\omega(x) = p^{\omega'}(x) = 0$, and divided $S^\omega \cup S^{\omega'}$ into four non-intersecting components. Then we can bound the affinity of the product measures P^ω_n and $P^{\omega'}_n$ for $\rho(\omega, \omega') = 1$ as

$$\|P^\omega_n \wedge P^{\omega'}_n\| \geq \frac{1}{8}(1 - d_1(P^\omega, P^{\omega'})/2)^{2n} = \frac{1}{8}(1 - \lambda \epsilon^{d-1} \,\mathrm{Leb}(B))^{2n}.$$

For any $\omega \neq \omega'$, we have, for arbitrary $\vec{j} \in \{1, ..., l\}^{d-1}$,

$$\begin{aligned} d^2(f^\omega, f^{\omega'}) &= \sum \int\limits_{\underline{B}_{\vec{i}} \Delta \overline{B}_{\vec{i}}} M^2 dx + \int\limits_{\underline{B}_{\vec{i}} \cap \overline{B}_{\vec{i}}} 4M^2 dx \\ &= \rho(\omega, \omega')(M^2 \operatorname{Leb}(\underline{B}_{\vec{j}} \Delta \overline{B}_{\vec{j}}) + 4M^2 \operatorname{Leb}(\underline{B}_{\vec{j}} \cap \overline{B}_{\vec{j}})) \\ &= 2\rho(\omega, \omega') M^2 (\operatorname{Leb}(\underline{B}_{\vec{j}}) + \operatorname{Leb}(\underline{B}_{\vec{j}} \cap \overline{B}_{\vec{j}})) \\ &= 2\rho(\omega, \omega') M^2 \epsilon^{d-1} (\operatorname{Leb}(B) + \operatorname{Leb}(B_r)) \end{aligned}$$

where the sum is only over indices where ω and ω' differ and $B_r = \{x \in B : x - (0, ..., 0, r) \in B\}$. Then, by Assouad's lemma,

$$\inf_{\widehat{\omega}} \max_{\omega \in \Omega} \mathbb{E}_\omega [d^2(f^\omega, f^{\widehat{\omega}})] \geq \frac{M^2 (l\epsilon)^{d-1}}{32} (\operatorname{Leb}(B) + \operatorname{Leb}(B_r))(1 - \lambda \epsilon^{d-1} \operatorname{Leb}(B))^n.$$

Also we have

$$\begin{aligned} &\frac{1}{\lambda} \int (f^\omega(x) - f^{\omega'}(x))^2 p^\omega(x) dx = \int\limits_{S^\omega} (f^\omega(x) - f^{\omega'}(x))^2 dx \\ &= \int (f^\omega(x) - f^{\omega'}(x))^2 dx - \int\limits_{S^{\omega'} \backslash S^\omega} (f^\omega(x) - f^{\omega'}(x))^2 dx \\ &= d^2(f^\omega, f^{\omega'}) - M^2 \operatorname{Leb}(S^{\omega'} \backslash S^\omega) \geq d^2(f^\omega, f^{\omega'}) - M^2 q \epsilon^{d-1} \operatorname{Leb}(B \backslash B_r) \\ &= d^2(f^\omega, f^{\omega'}) - M^2 \left(\frac{l}{l+2} \right)^{d-1} (\operatorname{Leb}(B) - \operatorname{Leb}(B_r)). \end{aligned}$$

Since $\lambda > 1$,

$$\begin{aligned} &\inf_{\widehat{f}} \sup_{P \in \mathcal{P}_n} \mathbb{E}_n \int (\widehat{f}(x) - f(x))^2 dP(x) \quad \geq \quad \inf_{\widehat{\omega}} \max_{\omega \in \Omega} \mathbb{E}_\omega \int (f^{\widehat{\omega}}(x) - f^\omega(x))^2 p^\omega(x) dx \\ &\geq \frac{M^2 (l\epsilon)^{d-1}}{32} (\operatorname{Leb}(B) + \operatorname{Leb}(B_r))(1 - \lambda \epsilon^{d-1} \operatorname{Leb}(B))^n - M^2 (l\epsilon)^{d-1} (\operatorname{Leb}(B) - \operatorname{Leb}(B_r)). \end{aligned}$$

As soon as $n \geq 2^d$, $l \geq 2$ and $(\epsilon l)^{d-1} \geq \frac{1}{2^{d-1}}$. Clearly $\operatorname{Leb}(B) \leq \frac{1}{2}$. Let $c_0 \geq 3$. Then $\epsilon \leq 1/8$ and $\lambda \leq (1 - 2\epsilon)^{-(d-1)} (1 - 4\epsilon - r)^{-1} \leq 2^{d+1}$, so

$$(1 - \lambda \epsilon^{d-1} \operatorname{Leb}(B))^n \geq \left(1 - \frac{2^d}{c_0^{d-1} n} \right)^{2n} \to \left(e^{-2^d / c_0^{d-1}} \right)^2.$$

So if we let $c_0 > (2^d / \log(5/4))^{1/(d-1)}$, then $e^{-2^d / c_0^{d-1}} > 4/5$ and for sufficiently large n we will have $(1 - \lambda \epsilon^{d-1} \operatorname{Leb}(B))^{2n} \geq 8/25$. Hence,

$$\inf_{\widehat{f}} \sup_{P \in \mathcal{P}_n} \mathbb{E}_n \int (\widehat{f}(x) - f(x))^2 dP(x) \geq \frac{M^2}{50 \cdot 2^{d-1}} \left(\operatorname{Leb}(B_r) - 50 \operatorname{Leb}(B \backslash B_r) \right).$$

Since

$$\mathrm{Leb}(B_r) = \frac{1}{2}\left(\frac{1}{2}-r\right)^d \frac{\pi^{d/2}}{\Gamma(d/2+1)} \quad \text{and} \quad \mathrm{Leb}(B\backslash B_r) \le \frac{r\pi^{(d-1)/2}}{2^{d-1}\Gamma((d-1)/2+1)},$$

then

$$\begin{aligned}\mathrm{Leb}(B_r) - 50\,\mathrm{Leb}(B\backslash B_r) &\ge \frac{1}{2}\left(\frac{1}{2}-r\right)^d \frac{\pi^{d/2}}{\Gamma(d/2+1)} - \frac{50r\pi^{(d-1)/2}}{2^{d-1}\Gamma((d-1)/2+1)} \\ &\ge \frac{\pi^{d/2}}{2^d\Gamma\left(\frac{d+1}{2}\right)(d+1)}\left[(1-2r)^d - \frac{100(d+1)r}{\sqrt{\pi}}\right].\end{aligned}$$

Now let r be such that

$$(1-2r)^d - \frac{100(d+1)r}{\sqrt{\pi}} = \frac{1}{2}. \tag{33}$$

Then

$$\inf_{\widehat{f}} \sup_{P\in\mathcal{P}_n} \mathbb{E}_n \int (\widehat{f}(x) - f(x))^2 dP(x) \ge \frac{M^2\pi^{d/2}}{50\cdot 4^d\Gamma\left(\frac{d+1}{2}\right)(d+1)}.$$

Step 3: Verifying condition number: For any ω,

$$\tau(S^\omega) = \min\left\{\tau(\underline{S}^\omega), \tau(\overline{S}^\omega), \frac{1}{2}\inf_{u\in\underline{S}^\omega}\inf_{v\in\overline{S}^\omega}\|u-v\|_2\right\}.$$

Due to the shape of the function g, for arbitrary $\vec{i}\in\{1,...,l\}^{d-1}$ we have

$$\tau(\underline{S}^\omega) \ge \min\left\{\tau(\partial\underline{S}), \tau(\partial\underline{B}_{\vec{i}}\backslash\partial\underline{S})\right\}$$

By definition of $\underline{S}$ it is easy to see that $\tau(\partial\underline{S}) = \epsilon$. Also

$$\begin{aligned}\tau(\partial\underline{B}_{\vec{i}}\backslash\partial\underline{S}) &= \tau(\{(\widetilde{x}, x_d)\in[-\epsilon/2,\epsilon/2]^{d-1}\times[0,1] : x_d = g(\widetilde{x}/\epsilon)\}) \\ &\ge \epsilon\tau(\{(\widetilde{x}, x_d)\in[-1/2,1/2]^{d-1}\times[0,1] : x_d = g(\widetilde{x})\}) = \epsilon r.\end{aligned}$$

Since $r<1$, we have $\tau(\underline{S}^\omega)\ge r\epsilon$, and similarly $\tau(\overline{S}^\omega)\ge r\epsilon$. Now,

$$\begin{aligned}\frac{1}{2}\inf_{u\in\underline{S}^\omega}\inf_{v\in\overline{S}^\omega}\|u-v\|_2 &\ge \frac{\epsilon}{2}\inf_{u,v\in[-0.5,0.5]^{d-1}}\|(u,g(u))-(v,g(v)+r)\|_2 \\ &= \frac{\epsilon}{2}\left(\sqrt{\frac{1}{4}+r^2}-\frac{1}{2}\right)\end{aligned}$$

which is smaller than ϵr, so for n sufficiently large,

$$\tau(S^\omega) \geq \frac{\epsilon}{2}\left(\sqrt{\frac{1}{4}+r^2}-\frac{1}{2}\right) \geq \frac{1}{2(c_0 n^{1/(d-1)}+2)}\left(\sqrt{\frac{1}{4}+r^2}-\frac{1}{2}\right)$$
$$\geq n^{-\frac{1}{d-1}}\frac{1}{2(c_0+1)}\left(\sqrt{\frac{1}{4}+r^2}-\frac{1}{2}\right)$$

which completes the proof.

Step 4: Each P^ω is in $\mathcal{P}_n$. This follows by construction. □

7.3. *Semisupervised Upper Bound.* Define $\widehat{p}_m$ to be the kernel density estimator based on the unlabeled data with bandwidth $(\log m/m)^{1/(4+d)}$. Let $\widehat{S} = \{x : \widehat{p}_m(x) > 0\}$. Recall that $\epsilon_m = c_1(\log m)^{-1/2}$ and $\delta_m = 2c_2\sqrt{d}\left((\log^2 m)/m\right)^{\frac{1}{d}}$. Define

$$\widehat{D}_\alpha(x_1,x_2) = \inf_{\gamma\in\widehat{\Gamma}(x_1,x_2)} \int_0^{L(\gamma)} \frac{1}{\widehat{p}_m(\gamma(t))^\alpha}dt$$

where

$$\widehat{\Gamma}(x_1,x_2) = \Big\{\gamma\in\Gamma(x_1,x_2) : \forall t\in[0,L(\gamma)],\ \gamma(t)\in\widehat{S}\backslash\widehat{\mathcal{R}}_{\partial S}\Big\},$$

$$\widehat{\mathcal{R}}_{\partial S} = \left\{x : \inf_{z\in\partial\widehat{S}}\|x-z\|_2 < 2\delta_m\right\}.$$

We define $\widehat{D}_\alpha(x_1,x_2)=\infty$ if $\widehat{\Gamma}(x_1,x_2)=\emptyset$. Let

$$\widehat{f}_{h,\alpha}(x) = \frac{\sum_{i=1}^n Y_i\, Q_h(\widehat{D}_\alpha(x,X_i))}{\sum_{i=1}^n Q_h(\widehat{D}_\alpha(x,X_i))}.$$

Theorem 7.2 *Let $h=\sqrt{1/n}$. Then,*

$$\sup_{P\in\mathcal{P}_n} R_P(\widehat{f}_{h,\alpha}) \preceq \frac{1}{n}.$$

Proof. Let $S = S_P$. Let $S_m = \left\{ x \in S : \inf_{z \in \partial S} \|x - z\|_2 \geq 3\delta_m \right\}$. Then,

$$\int (\widehat{f}_{h,\alpha}(x) - f(x))^2 dP(x) = \int_{S_m} (\widehat{f}_{h,\alpha}(x) - f(x))^2 dP(x) + \int_{S - S_m} (\widehat{f}_{h,\alpha}(x) - f(x))^2 dP(x).$$

Now

$$\int_{S - S_m} (\widehat{f}_{\alpha}(x) - f(x))^2 dP(x) \leq 2M^2 P(S \backslash S_m) \leq 2\Lambda M^2 \, \mathrm{Leb}(S \backslash S_m).$$

Since the radius of curvature of ∂S is at least τ_n, and $\tau_n > 3\delta_m$, we have by Proposition 6,

$$\begin{aligned} \mathrm{Leb}(S \backslash S_m) &\leq \mathrm{Vol}(\partial S) \frac{(\tau_n + 3\delta_m)^d - \tau_n^d}{\tau_n^{d-1}} \leq c_3 \left[\left(1 + \frac{3\delta_m}{\tau_n} \right)^d - 1 \right] \\ &\leq c_3 \sum_{i=1}^{d} \binom{d}{i} \frac{3\delta_m}{\tau_n} \leq 3c_3 2^d \frac{\delta_m}{\tau_n} \leq \left(\frac{1}{n} \right) \end{aligned}$$

from (C4) where Vol denotes the $d-1$-dimensional volume on ∂S.

Arguing as in Theorem 5.1, we have, for each $P \in \mathcal{P}_n$, that

$$\mathbb{E} \int_{S_m} (\widehat{f}_{h,\alpha}(x) - f(x))^2 dP(x) \leq L^2 \left(R(h/2) \right)^2 + \frac{M^2 \left[4 + \frac{1}{e} \widehat{\mathcal{N}} \left(P, \alpha, \frac{h}{2} \right) \right]}{n} + \frac{4M^2}{m}$$

where $R = \sup\{ D_{P,\alpha}(x_1, x_2) / \widehat{D}_\alpha(x_1, x_2) : x_1, x_2 \in S_m \}$ and $\widehat{\mathcal{N}}(P, \alpha, \epsilon)$ denotes the covering number of S_m under $\widehat{D}_\alpha$. In Proposition 5 in Section 11.3 we show that $D_{P,\alpha}(x_1, x_2) \leq [(\lambda + \epsilon_m)/\lambda]^\alpha \widehat{D}_\alpha(x_1, x_2)$ for $x_1, x_2 \in S_m$. By (C2) this implies that $R \leq 2$. We also show in Proposition 5 that $\widehat{D}_\alpha(x_1, x_2) \leq d_{S_m}(x_1, x_2)/(\lambda - \epsilon_m)^\alpha$ for $x_1, x_2 \in S_m$. Here, $d_{S_m}(x_1, x_2)$ is the length (in Euclidean distance) of the shortest path in S_m connecting x_1 and x_2. Thus

$$\widehat{\mathcal{N}}(P, \alpha, h/2) \leq \mathcal{N}_m \left(\frac{h}{2} (\lambda - \epsilon_m)^\alpha \right)$$

where $\mathcal{N}_m$ denotes the covering number under d_{S_m}. (C3) implies that $h(\lambda - \epsilon_m)^\alpha > \tau^{-(d-1)}$. By Proposition 7, each connected component of S_P may be covered by one set and hence $\mathcal{N}_m \left(\frac{h}{2} (\lambda - \epsilon_m)^\alpha \right) \leq K$. We thus have that

$$\mathbb{E} \int_{S_m} (\widehat{f}_{h,\alpha}(x) - f(x))^2 dP(x) \leq L^2 (2h)^2 + \frac{M^2 \left[4 + \frac{1}{e} K \right]}{n} + \frac{4M^2}{m}$$

and the result follows since $h = n^{-1/2}$ and $m \geq n$. □

7.4. *Comparison of Lower and Upper Bound.* Finally we have:

Corollary 7.3 *Under the conditions of the previous theorem,*

$$\frac{R_{SS}}{R_S} \preceq \left(\frac{1}{n}\right) \to 0 \tag{34}$$

as $n \to \infty$.

8. Adaptive Semisupervised Inference. We have established a bound on the risk of the density-sensitive semisupervised kernel estimator. The bound is achieved by using an estimate $\widehat{D}_{\alpha,\sigma}$ of the density-sensitive distance. However, this requires knowing the density-sensitive parameter α, along with other parameters. It is critical to choose α (and h) appropriately, otherwise we might incur a large error if the semisupervised assumption does not hold or holds with a different density sensitivity value α. We consider two methods for choosing the parameters.

The following result shows that we can adapt to the correct degree of semisupervisedness if cross-validation is used to select the appropriate α, σ and h. This implies that the estimator gracefully degrades to a supervised learner if the semisupervised assumption (sensitivity of regression function to marginal density) does not hold ($\alpha = 0$).

For any f, define the risk $R(f) = \mathbb{E}[(f(X) - Y)^2]$ and the excess risk $\mathcal{E}(f) = R(f) - R(f^*) = \mathbb{E}[(f(X) - f^*(X))^2]$ where f^* is the true regression function. Let $\mathcal{H}$ be a finite set of bandwidths, let $\mathcal{A}$ be a finite set of values for α and let Σ be a finite set of values for σ. Let $\theta = (h, \alpha, \sigma)$, $\Theta = \mathcal{H} \times \mathcal{A} \times \Sigma$ and $J = |\Theta|$.

Divide the data into training data T and validation data V. For notational simplicity, let both sets have size n. Let $\mathcal{F} = \{\widehat{f}_\theta^T\}_{\theta \in \Theta}$ denote the semisupervised kernel estimators trained on data T using $\theta \in \Theta$. For each $\widehat{f}_\theta^T \in \mathcal{F}$ let

$$\widehat{R}^V(\widehat{f}_\theta^T) = \frac{1}{n}\sum_{i=1}^{n}(\widehat{f}_\theta^T(X_i) - Y_i)^2$$

where the sum is over V. Let $Y_i = f(X_i) + \epsilon_i$ with $\epsilon_i \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)$. Also, we assume that $|f(x)|, |\widehat{f}_\theta^T(x)| \leq M$, where $M > 0$ is a constant.[2]

[2] Note that the estimator can always be truncated if necessary.

Theorem 8.1 *Let $\mathcal{F} = \{\widehat{f}_\theta^T\}_{\theta\in\Theta}$ denote the semisupervised kernel estimators trained on data T using $\theta \in \Theta$. Use validation data V to pick*

$$\widehat{\theta} = \arg\min_{\theta\in\Theta} \widehat{R}^V(\widehat{f}_\theta^T)$$

and define the corresponding estimator $\widehat{f} = \widehat{f}_{\widehat{\theta}}$. Then, for every $0 < \delta < 1$,

$$\mathbb{E}[\mathcal{E}(\widehat{f}_{\widehat{\theta}})] \leq \frac{1}{1-a}\left[\min_{\theta\in\Theta}\mathbb{E}[\mathcal{E}(\widehat{f}_\theta)] + \frac{\log(J)/\delta)}{nt}\right] + 4\delta M^2 \tag{35}$$

where $0 < a < 1$ and $0 < t < 15/(38(M^2+\sigma^2))$ are constants. $\mathbb{E}$ denotes expectation over everything that is random.

Proof. First, we derive a general concentration of $\widehat{\mathcal{E}}(f)$ around $\mathcal{E}(f)$ where $\widehat{\mathcal{E}}(f) = \widehat{R}(f) - \widehat{R}(f^*) = -\frac{1}{n}\sum_{i=1}^n U_i$, and $U_i = -(Y_i - f(X_i))^2 + (Y_i - f^*(X_i))^2$.

If the variables U_i satisfy the following moment condition:

$$\mathbb{E}[|U_i - \mathbb{E}[U_i]|^k] \leq \frac{\mathsf{Var}(U_i)}{2}k!r^{k-2}$$

for some $r > 0$, then the Craig-Bernstein (CB) inequality (Craig 1933) states that with probability $> 1-\delta$,

$$\frac{1}{n}\sum_{i=1}^n (U_i - \mathbb{E}[U_i]) \leq \frac{\log(1/\delta)}{nt} + \frac{t\ \mathsf{Var}(U_i)}{2(1-c)}$$

for $0 \leq tr \leq c < 1$. The moment conditions are satisfied by bounded random variables as well as Gaussian random variables (see e.g. Haupt and Nowak (2006)).

To apply this inequality, we first show that $\mathsf{Var}(U_i) \leq 4(M^2+\sigma^2)\mathcal{E}(f)$ since $Y_i = f(X_i) + \epsilon_i$ with $\epsilon_i \overset{i.i.d}{\sim} \mathcal{N}(0,\sigma^2)$. Also, we assume that $|f(x)|$, $|\widehat{f}(x)| \leq M$, where $M > 0$ is a constant.

$$\begin{aligned}
\mathsf{Var}(U_i) &\leq \mathbb{E}[U_i^2] = \mathbb{E}[(-(Y_i - f(X_i))^2 + (Y_i - f^*(X_i))^2)^2] \\
&= \mathbb{E}[(-(f^*(X_i) + \epsilon_i - f(X_i))^2 + (\epsilon_i)^2)^2] \\
&= \mathbb{E}[(-(f^*(X_i) - f(X_i))^2 - 2\epsilon_i(f^*(X_i) - f(X_i)))^2] \\
&\leq 4M^2\mathcal{E}(f) + 4\sigma^2\mathcal{E}(f) = 4(M^2+\sigma^2)\mathcal{E}(f)
\end{aligned}$$

Therefore using CB inequality we get, with probability $> 1-\delta$,

$$\mathcal{E}(f) - \widehat{\mathcal{E}}(f) \leq \frac{\log(1/\delta)}{nt} + \frac{t\ 2(M^2+\sigma^2)\mathcal{E}(f)}{(1-c)}$$

Now set $c = tr = 8t(M^2 + \sigma^2)/15$ and let $t < 15/(38(M^2 + \sigma^2))$. With this choice, $c < 1$ and define

$$a = \frac{t2(M^2 + \sigma^2)}{(1-c)} < 1.$$

Then, using a and rearranging terms, with probability $> 1 - \delta$,

$$(1-a)\mathcal{E}(f) - \widehat{\mathcal{E}}(f) \leq \frac{\log(1/\delta)}{nt}$$

where $t < 15/(38(M^2 + \sigma^2))$.

Then, using the previous concentration result, and taking union bound over all $f \in \mathcal{F}$, we have with probability $> 1 - \delta$,

$$\mathcal{E}(f) \leq \frac{1}{1-a}\left[\widehat{\mathcal{E}}^V(f) + \frac{\log(J/\delta)}{nt}\right].$$

Now,

$$\begin{aligned}
\mathcal{E}(\widehat{f}_{\widehat{\theta}}) &= R(\widehat{f}_{\widehat{\theta}}) - R(f^*) \\
&\leq \frac{1}{1-a}\left[\widehat{R}^V(\widehat{f}_{\widehat{\theta}}) - \widehat{R}^V(f^*) + \frac{\log(J/\delta)}{nt}\right] \\
&\leq \frac{1}{1-a}\left[\widehat{R}^V(f) - \widehat{R}^V(f^*) + \frac{\log(J/\delta)}{nt}\right]
\end{aligned}$$

Taking expectation with respect to validation dataset,

$$\mathbb{E}_V[\mathcal{E}(\widehat{f}_{\widehat{\theta}})] \leq \frac{1}{1-a}\left[R(f) - R(f^*) + \frac{\log(J/\delta)}{nt}\right] + 4\delta M^2.$$

Now taking expectation with respect to training dataset,

$$\mathbb{E}_{TV}[\mathcal{E}(\widehat{f}_{\widehat{\theta}})] \leq \frac{1}{1-a}\left[\mathbb{E}_T[R(f) - R(f^*)] + \frac{\log(J/\delta)}{nt}\right] + 4\delta M^2.$$

Since this holds for all $f \in \mathcal{F}$, we get:

$$\mathbb{E}_{TV}[\mathcal{E}(\widehat{f}_{\widehat{\theta}})] \leq \frac{1}{1-a}\left[\min_{f\in\mathcal{F}} \mathbb{E}_T[\mathcal{E}(f)] + \frac{\log(J/\delta)}{nt}\right] + 4\delta M^2.$$

The result follows. □

In practice, both Θ may be taken to be of size n^a for some $a > 0$. Then we can approximate the optimal h, σ and α with sufficient accuracy to achieve the optimal rate. Setting $\delta = 1/(4M^2 n)$, we then see that the penalty for

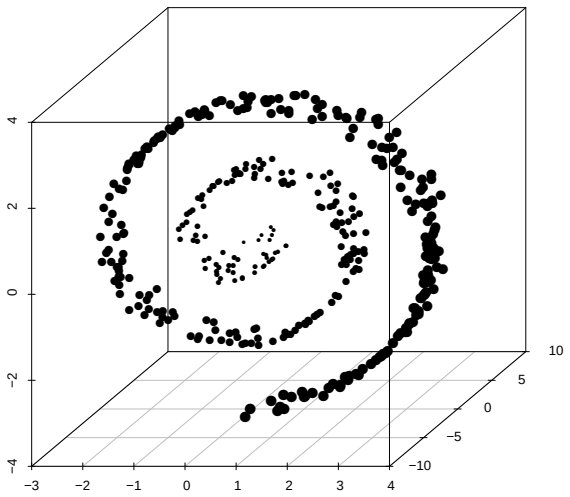

FIG 7. *The swiss roll data set. Point size represents regression function.*

adaptation is $\frac{\log(J/\delta)}{nt} + \delta M = O(\log n/n)$ and hence introduces only a logarithmic term.

Remark: Cross-validation is not the only way to adapt. For example, the adaptive method in Kpotufe (2011) can also be used here.

9. Simulation Results. In this section we describe the results of a series of numerical experiments on a simulated data set to demonstrate the effect of using the exponential version of the density sensitive metric for small, labeled sample sizes. For the marginal distribution of X, we used a slightly modified version of the swiss roll distribution used in Culp (2011b). Figure 7 shows a sample from this distribution, where the point size represents the response Y. We repeatedly sampled $N = 400$ points from this distribution, and computed the mean squared error of the kernel regression estimator using a set of values for α and for labeled sample size ranging from $n = 5$ to $n = 320$. We used the approxmation (19) with $k = 20$.

Figure 8 shows the average results after 300 repetitions of this procedure with error bars indicating a 95% confidence interval. As expected, we observe that for small labeled sample sizes increasing α can decrease the error. But as the labeled sample size increases, using the density sensitive metric becomes decreasingly beneficial, and can even hurt.

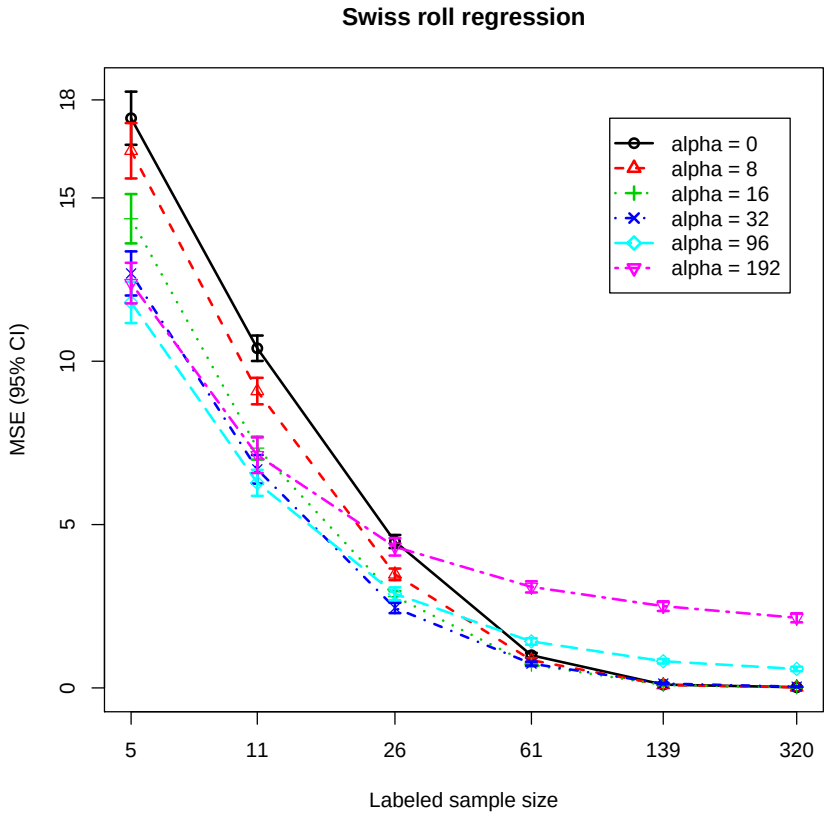

FIG 8. *MSE of kernel regression on the swiss roll data set for a range of labeled sample sizes using different values of* α.

10. Discussion. Semisupervised methods are very powerful but, like all methods, they only work under certain conditions. We have shown that, under certain conditions, semisupervised methods provably outperform supervised methods. In particular, the advantage of semisupervised methods is mainly when the distribution P_X of X is concentrated near a low dimensional set rather than when P_X is smooth.

We introduced a family of estimators indexed by a parameter α. This parameter controls the strength of the semi-supervised assumption. The behavior of the semi-supervised method depends critically on α. Finally, we showed that cross-validation can be used to automatically adapt to α so that α does not need to be known. Hence, our method takes advantage of the unlabeled data when the semi-supervised assumption holds, but does not add extra bias when the assumption fails. Our simulations confirm that our proposed estimator adapts well to alpha and has good risk when the semi-supervised smoothness holds and when it fails.

The analysis in this paper can be extended in several ways. First, it is possible to use other density sensitive metrics such as the diffusion distance (Lee and Wasserman, 2008). Second, we defined a method to estimate the density sensitive metric that works under broader conditions than the two existing methods due to Sajama and Orlitsky (2005) and Bijral, Ratliff and Srebro (2011). We suspect that faster methods can be developed. Finally, other estimators besides kernel estimators can be used. We will report on these extensions elsewhere.

11. Additional Proofs.

11.1. *Proof of Theorem 4.1.* To prove Theorem 4.1 We use the approach in Yukich (1985). (See also Gine and Guillou (2002) and Prakasa-Rao (1983).) If $\ell \leq u$, define the bracket $[\ell, u] = \{h : \ell \leq h \leq u\}$. A collection $(\ell_1, u_1), \ldots, (\ell_N, u_N)$ is a ϵ bracketing of a class of functions $\mathcal{F}$ if $\mathcal{F} \subset \bigcup_{j=1}^{N} [\ell_j, u_j]$ and $\int |u_j - \ell_j|^p dP \leq \epsilon^p$ for $j = 1, \ldots, N$. The bracketing number $N_{[]}(\epsilon, \mathcal{F}, L_p(P))$ is the size of the smallest ϵ bracketing.

Theorem 11.1 *Let $X_1, \ldots, X_n \sim P$. Define $P(f) = \int f(z) dP(z)$ and $P_n(f) = \frac{1}{n}\sum_{i=1}^{n} f(X_i)$. Let $A = \sup_f \int |f| dP$ and $B = \sup_f ||f||_\infty$. Then*

$$P^n\left(\sup_{f\in\mathcal{F}} |P_n(f) - P(f)| > \epsilon\right) \leq 2N_{[]}(\epsilon/8, \mathcal{F}, L_1(P)) \exp\left(-\frac{3n\epsilon^2}{4B[6A+\epsilon]}\right) + 2N_{[]}(\epsilon/8, \mathcal{F}, L_1(P)) \exp\left(-\frac{3n\epsilon}{64B}\right).$$

Hence, if $\epsilon \le 2A/3$,

(36)
$$P^n\left(\sup_{f\in\mathcal{F}}|P_n(f)-P(f)|>\epsilon\right)\le 4N_{[]}(\epsilon/8,\mathcal{F},L_1(P))\exp\left(-\frac{3n\epsilon^2}{4B[6A+\epsilon]}\right).$$

Proof. (This proof follows Yukich (1985).) For notational simplicity in the proof, let us write, $N(\epsilon)\equiv N_{[]}(\epsilon,\mathcal{F},L_1(P))$. Define $z_n(f)=\int f(dP_n-dP)$. Let $[\ell_1,u_1],\dots,[\ell_N,u_N]$ be a minimal $\epsilon/8$ bracketing. We may assume that for each j, $||u_j||\le B$ and $||\ell_j||\le B$. (Otherwise, we simply truncate the brackets.) For each j, choose some $f_j\in[\ell_j,u_j]$.

Consider any $f\in\mathcal{F}$ and let $[\ell_j,u_j]$ denote a bracket containing f. Then

$$|z_n(f)|\le|z_n(f_j)|+|z_n(f-f_j)|.$$

Furthermore,

$$\begin{aligned}|z_n(f-f_j)| &= \left|\int(f-f_j)(dP_n-dP)\right|\le\int|f-f_j|\,(dP_n+dP)\le\int|u_j-\ell_j|\,(dP_n+dP)\\ &= \int|u_j-\ell_j|\,(dP_n-dP)+2\int|u_j-\ell_j|dP\\ &= \int|u_j-\ell_j|\,(dP_n-dP)+2\left(\frac{\epsilon}{8}\right)=z_n(|u_j-\ell_j|)+\frac{\epsilon}{4}.\end{aligned}$$

Hence,

$$|z_n(f)|\le|z_n(f_j)|+\left[[z_n(|u_j-\ell_j|)+\frac{\epsilon}{4}\right].$$

Thus,

$$\begin{aligned}P^n(\sup_{f\in\mathcal{F}}|z_n(f)|>\epsilon) &\le P^n(\max_j|z_n(f_j)|>\epsilon/2)+P^n(\max_j|z_n(|u_j-\ell_j|)|+\epsilon/4>\epsilon/2)\\ &\le P^n(\max_j|z_n(f_j)|>\epsilon/2)+P^n(\max_j|z_n(|u_j-\ell_j|)|>\epsilon/4).\end{aligned}$$

Now

$$\mathsf{Var}(f_j)\le\int f_j^2\,dP=\int|f_j|\,|f_j|dP\le||f_j||_\infty\int|f_j|dP\le AB.$$

Hence, by Bernstein's inequality,

$$P^n\Big(\max_j|z_n(f_j)|>\epsilon/2\Big)\le 2\sum_{j=1}^N\exp\left(-\frac{1}{2}\frac{n(\epsilon/2)^2}{AB+B\epsilon/6}\right)\le 2N(\epsilon/8)\exp\left(-\frac{3}{4B}\frac{n\epsilon^2}{6A+\epsilon}\right).$$

Similarly,

$$\begin{aligned}\mathsf{Var}(|u_j-\ell_j|) &\leq \int (u_j-\ell_j)^2 dP \leq \int |u_j-\ell_j|\,|u_j-\ell_j| dP \\ &\leq ||u_j-\ell_j||_\infty \int |u_j-\ell_j| dP \leq 2B\frac{\epsilon}{8} = \frac{B\epsilon}{4}.\end{aligned}$$

Also, $||u_j-\ell_j||_\infty \leq 2B$. Hence, by Bernstein's inequality,

$$\begin{aligned}P^n\Big(\max_j z_n(|u_j-\ell_j|) > \epsilon/4\Big) &\leq 2\sum_{j=1}^N \exp\left(-\frac{1}{2}\frac{n(\epsilon/4)^2}{2B\frac{\epsilon}{4}+2B(\epsilon/4)/3}\right) \\ &\leq 2N(\epsilon/8)\exp\left(-\frac{3n\epsilon}{64B}\right).\end{aligned}$$

□

The following result is from Example 19.7 from van der Vaart (1998).

Lemma 4 *Let $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$ where Θ is a bounded subset of $\mathbb{R}^d$. Suppose there exists a function m such that, for every θ_1, θ_2,*

$$|f_{\theta_1}(x) - f_{\theta_2}(x)| \leq m(x)\,||\theta_1-\theta_2||.$$

Then,

$$N_{[]}(\epsilon, \mathcal{F}, L_q(P)) \leq \left(\frac{4\sqrt{d}\ \mathsf{diam}(\Theta)\int |m(x)|^q dP(x)}{\epsilon}\right)^d.$$

Proof. Let

$$\delta = \frac{\epsilon}{4\sqrt{d}\int |m(x)|^q dP(x)}.$$

We can cover Θ with (at most) $N = (\mathsf{diam}(\Theta)/\delta)^d$ cubes $C_1, \ldots, C_N$ of size δ. Let $c_1, \ldots, c_N$ denote the centers of the cubes. Note that $C_j \subset B(c_j, \sqrt{d}\delta)$ where $B(x,r)$ denotes a ball of radius r centered at x. Hence, $\bigcup_j B(c_j, \sqrt{d}\delta)$ covers Θ. Let θ_j be the projection of c_j onto Θ. Then $\bigcup_j B(\theta_j, 2\delta\sqrt{d})$ covers Θ. In summary, for every $\theta \in \Theta$ there is a $\theta_j \in \{\theta_1, \ldots, \theta_N\}$ such that

$$||\theta - \theta_j|| \leq 2\delta\sqrt{d} \leq \frac{\epsilon}{2\int |m(x)|^q dP(x)}.$$

Define $\ell_j = f_{\theta_j} - \epsilon m(x)/2\int m$ and $u_j = f_{\theta_j} + \epsilon m(x)/2\int m$. We claim that the brackets $[\ell_1, u_1], \ldots, [\ell_N, u_N]$ cover $\mathcal{F}$. To see this, choose any $f_\theta \in \mathcal{F}$. Let θ_j be the closest element $\{\theta_1, \ldots, \theta_N\}$ to θ. Then

$$\begin{aligned}f_\theta(x) &= f_{\theta_j}(x) + f_\theta(x) - f_{\theta_j}(x) \leq f_{\theta_j}(x) + |f_\theta(x) - f_{\theta_j}(x)| \\ &\leq f_{\theta_j}(x) + m(x)||\theta-\theta_j|| \leq f_{\theta_j}(x) + \frac{m(x)\epsilon}{2\int |m(x)|^q dP(x)} = u_j(x).\end{aligned}$$

By a similar argument, $f_\theta(x) \geq \ell_j(x)$. Also, $\int (u_j - \ell_j)^q dP \leq \epsilon^q$. Finally, note that the number of brackets is

$$N = (\mathsf{diam}(\Theta)/\delta)^d = \left(\frac{4\sqrt{d}\ \mathsf{diam}(\Theta) \int |m(x)|^q dP(x)}{\epsilon} \right)^d .$$

□

Now we prove Theorem 4.1.

Proof. Let $\theta = (x, \sigma)$, $\Theta = \mathcal{X} \times [a, A]$, $f_\theta(u) = \sigma^{-d} K(||x - u||/\sigma)$ and $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$. We apply Theorem 11.1 with $A = 1$ and $B = K(0)/a^d$. We need to bound $N_{[]}(\epsilon, \mathcal{F}, L_1(P)$. Let $\theta = (x, \sigma^d)$ and $\nu = (y, \tau^d)$. Some algebra shows that

$$|f_\theta(u) - f_\nu(u)| \leq \frac{C}{a^{2d}} ||\theta - \nu||.$$

Apply Lemma 4 to get

$$N_{[]}(\epsilon, \mathcal{F}, L_1(P)) \leq \left(\frac{C}{a^{2d}\epsilon} \right)^{d+1} .$$

Hence, Theorem 11.1 yields,

$$\begin{aligned} P^n\Big(\sup_x |\widehat{p}_\sigma(x) - p_\sigma(x)| > \epsilon \Big) \quad &\leq \quad 2 \left(\frac{C}{a^{2d}\epsilon} \right)^{d+1} \\ &\quad \times \left[\exp\left(-\frac{3n\epsilon^2 a^d}{4K(0)(6+\epsilon)} \right) + \exp\left(-\frac{3n\epsilon a^d}{64K(0)} \right) \right] . \end{aligned}$$

□

Note that the proofs of the last two results did not depend on P. Hence, the results hold uniformly over P.

11.2. *Propositions for Section 4.3.*

Proposition 1 *Given $x, y, z \in \mathbb{R}^d$, let γ_1 be the path from x to y such that $D_{P,\alpha,\sigma}(x, y) = \int_0^{L(\gamma_1)} \exp\left[-\alpha p_{X,\sigma}(\gamma_1(t))\right] dt$. Then*

$$D_{P,\alpha,\sigma}(y, z) \leq D_{P,\alpha,\sigma}(x, y) \frac{\|y - z\|}{L(\gamma_1)} \left(1 + \exp\left[\frac{\alpha K_{\max}}{\sigma^d} \right] \frac{\alpha K^*_{\max}}{\sigma^{d+1}} (L(\gamma_1) + \|y - z\|) \right) .$$

Proof. Let γ_2 be the straight line from y to z (i.e. $L(\gamma_2) = \|y-z\|$). Then

$$
\begin{aligned}
D_{P,\alpha,\sigma}(y,z) &= \inf_{\gamma\in\Gamma(y,z)} \int_0^{L(\gamma)} \exp\left[-\alpha p_{X,\sigma}(\gamma(t))\right] dt \\
&\leq \int_0^{L(\gamma_2)} \exp\left[-\alpha p_{X,\sigma}(\gamma_2(t))\right] dt \leq L(\gamma_2) \sup_{t\in[0,L(\gamma_2)]} \exp\left[-\alpha p_{X,\sigma}(\gamma_2(t))\right] \\
&\leq L(\gamma_2)\left(\frac{D_{P,\alpha,\sigma}(x,y)}{L(\gamma_1)} + \sup_{u_0}\left\|\nabla_u \exp\left[-\alpha p_{X,\sigma}(u)\right]\big|_{u=u_0}\right\| \sup_{\substack{t_1\in[0,L(\gamma_1)]\\ t_2\in[0,L(\gamma_2)]}} \|\gamma_1(t_1)-\gamma_2(t_2)\|\right) \\
&\leq L(\gamma_2)\left(\frac{D_{P,\alpha,\sigma}(x,y)}{L(\gamma_1)} + \alpha \sup_{u_0}\left[\exp\left[-\alpha p_{X,\sigma}(u)\right]\|\nabla_u p_{X,\sigma}(u)\|\right]_{u=u_0}(L(\gamma_1)+L(\gamma_2))\right) \\
&\leq L(\gamma_2)\left(\frac{D_{P,\alpha,\sigma}(x,y)}{L(\gamma_1)} + \alpha \sup_{u_0}\left[\|\nabla_u p_{X,\sigma}(u)\|\right]_{u=u_0}(L(\gamma_1)+L(\gamma_2))\right) \\
&\leq L(\gamma_2)\left(\frac{D_{P,\alpha,\sigma}(x,y)}{L(\gamma_1)} + \frac{\alpha}{\sigma^{d+1}} K^*_{\max}(L(\gamma_1)+L(\gamma_2))\right) \\
&\leq D_{P,\alpha,\sigma}(x,y)\frac{L(\gamma_2)}{L(\gamma_1)}\left(1+\exp\left[\frac{\alpha K_{\max}}{\sigma^d}\right]\frac{\alpha K^*_{\max}}{\sigma^{d+1}}(L(\gamma_1)+L(\gamma_2))\right)
\end{aligned}
$$

where the second to last inequality is due to the dominated convergence theorem (exchanging differentiation and expectation with respect to P), and Jensen's inequality (the L_2 norm is a convex function). □

Proposition 2 *Given $x, y \in \mathbb{R}^d$,*

$$
\|x-y\|\exp\left[-\alpha p_{X,\sigma}\left(\frac{x+y}{2}\right)\right]\left(1-\exp\left[\frac{\alpha K_{\max}}{\sigma^d}\right]\frac{\alpha K^*_{\max}\|x-y\|}{\sigma^{d+1}}\right) \leq D_{P,\alpha,\sigma}(x,y)
$$

and

$$
D_{P,\alpha,\sigma}(x,y) \leq \|x-y\|\exp\left[-\alpha p_{X,\sigma}\left(\frac{x+y}{2}\right)\right]\left(1+\exp\left[\frac{\alpha K_{\max}}{\sigma^d}\right]\frac{\alpha K^*_{\max}\|x-y\|}{2\sigma^{d+1}}\right)
$$

Proof. Let γ be the straight line from x to y.

$$\begin{aligned}
D_{P,\alpha,\sigma}(x,y) &\leq \int_0^{L(\gamma)} \exp\left[-\alpha p_{X,\sigma}\left(\gamma(t)\right)\right] dt \\
&\leq L(\gamma) \sup_{t\in[0,L(\gamma)]} \exp\left[-\alpha p_{X,\sigma}\left(\gamma(t)\right)\right] \\
&\leq \|x-y\|\left(\exp\left[-\alpha p_{X,\sigma}\left(\frac{x+y}{2}\right)\right] + \frac{\|x-y\|}{2}\sup_{u_0}\left\|\nabla_u \exp\left[-\alpha p_{X,\sigma}(u)\right]\big|_{u=u_0}\right\|\right) \\
&\leq \|x-y\|\left(\exp\left[-\alpha p_{X,\sigma}\left(\frac{x+y}{2}\right)\right] + \frac{\alpha K^*_{\max}\|x-y\|}{2\sigma^{d+1}}\right) \\
&\leq \|x-y\|\exp\left[-\alpha p_{X,\sigma}\left(\frac{x+y}{2}\right)\right]\left(1+\exp\left[\frac{\alpha K_{\max}}{\sigma^d}\right]\frac{\alpha K^*_{\max}\|x-y\|}{2\sigma^{d+1}}\right).
\end{aligned}$$

Now the ball $B\left(\frac{x+y}{2}, \|x-y\|\right)$ contains the balls $B_1 = B\left(x, \|x-y\|/2\right)$ and $B_2 = B\left(y, \|x-y\|/2\right)$. The integral over any path γ connecting x and y is at least as large as the integral over $\gamma \cap B\left(\frac{x+y}{2}, \|x-y\|\right)$. Hence,

$$\begin{aligned}
D_{P,\alpha,\sigma}(x,y) &\geq \|x-y\|\left(\inf_{u\in B\left(\frac{x+y}{2},\|x-y\|\right)} \exp\left[-\alpha p_{X,\sigma}\left(u\right)\right]\right) \\
&\geq \|x-y\|\left(\exp\left[-\alpha p_{X,\sigma}\left(\frac{x+y}{2}\right)\right] - \frac{\alpha K^*_{\max}\|x-y\|}{\sigma^{d+1}}\right) \\
&\geq \|x-y\|\exp\left[-\alpha p_{X,\sigma}\left(\frac{x+y}{2}\right)\right]\left(1-\exp\left[\frac{\alpha K_{\max}}{\sigma^d}\right]\frac{\alpha K^*_{\max}\|x-y\|}{\sigma^{d+1}}\right).
\end{aligned}$$

□

11.3. *Proofs For Section 7.3.*

Proposition 3 *If $m \geq m_0$, where $m_0 \equiv m_0(\lambda,\Lambda)$ is a constant, then for all marginal densities p of distributions in $\mathcal{P}_n$, we have with probability $> 1-1/m$,*

$$\sup_{x\in S\setminus\mathcal{R}_{\partial S}} |p(x) - \widehat{p}_m(x)| \leq \epsilon_m \text{ and } \partial\widehat{S} \subset \mathcal{R}_{\partial S}$$

where $\epsilon_m = c_1(\log m)^{-1/2}$ for constant $c_1 \equiv c_1(K, C_2, d, \eta, \Lambda)$, $\widehat{S} = \{x : \widehat{p}_m(x) > 0\}$, and $\mathcal{R}_{\partial S} = \left\{x : \inf_{z\in\partial S}\|x-z\|_2 < \delta_m\right\}$ where $\delta_m = 2c_2\sqrt{d}\left(\frac{\log^2 m}{m}\right)^{\frac{1}{d}}$ for some constant $c_2 > 0$.

Proof. Follows from Theorem 1 in Singh, Nowak and Zhu (2008b) by noting that since the density estimate will be 0 a.s. outside the boundary region, and we have $p \geq \lambda$ on S, for sufficiently large m (i.e. small ϵ_m), we must have $S \backslash \mathcal{R}_{\partial S} \subseteq \widehat{S} \subseteq S \cup \mathcal{R}_{\partial S}$. □

Proposition 4 *Assume* $\sup_{x \in S \backslash \mathcal{R}_{\partial S}} |\widehat{p}_m(x) - p(x)| \leq \epsilon_m$ *and* $\partial \widehat{S} \subset \mathcal{R}_{\partial S}$. *Let*

$$\widetilde{D}_\alpha(x_1, x_2) = \inf_{\gamma \in \widehat{\Gamma}(x_1, x_2)} \int_0^{L(\gamma)} \frac{1}{p(\gamma(t))^\alpha} dt$$

and $\Psi = \{(x_1, x_2) : x_1, x_2 \in \widehat{S} \backslash \widehat{\mathcal{R}}_{\partial S},\ \widehat{\Gamma}(x_1, x_2) \neq \emptyset\}$. *Then for any* $(x_1, x_2) \in \Psi$,

$$\left(\frac{\lambda}{\lambda + \epsilon_m}\right)^\alpha \widetilde{D}_\alpha(x_1, x_2) \leq \widehat{D}_\alpha(x_1, x_2) \leq \left(\frac{\lambda}{(\lambda - \epsilon_m)_+}\right)^\alpha \widetilde{D}_\alpha(x_1, x_2).$$

Proof. Note that by the triangle inequality, $\mathcal{R}_{\partial S} \subseteq \widehat{\mathcal{R}}_{\partial S}$, so $\widehat{S} \backslash \widehat{\mathcal{R}}_{\partial S} \subseteq S \backslash \mathcal{R}_{\partial S}$ since $\tau_n > 2\delta_m$ for m large enough. We see that if $(x_1, x_2) \in \Psi$, then x and y must be in the same connected component of $\widehat{S} \backslash \widehat{\mathcal{R}}_{\partial S}$, and, furthermore, all points along any path in $\widehat{\Gamma}(x_1, x_2)$ must also be in the same connected component. For $(x_1, x_2) \in \Psi$,

$$\begin{aligned}
\widehat{D}_\alpha(x_1, x_2) &= \inf_{\gamma \in \widehat{\Gamma}(x_1, x_2)} \int_0^{L(\gamma)} \frac{1}{p(\gamma(t))^\alpha} \frac{p(\gamma(t))^\alpha}{\widehat{p}_m(\gamma(t))^\alpha} dt \\
&\leq \inf_{\gamma \in \widehat{\Gamma}(x_1, x_2)} \left[\int_0^{L(\gamma)} \frac{1}{p(\gamma(t))^\alpha} dt\right] \left[\sup_{t \in [0, L(\gamma)]} \left(\frac{p(\gamma(t))}{\widehat{p}_m(\gamma(t))}\right)^\alpha\right] \\
&\leq \sup_{z \in S \backslash \mathcal{R}_{\partial S}} \left(\frac{p(z)}{\widehat{p}_m(z)}\right)^\alpha \widetilde{D}_\alpha(x_1, x_2)
\end{aligned}$$

and

$$\sup_{z \in S \backslash \mathcal{R}_{\partial S}} \left(\frac{p(z)}{\widehat{p}_m(z)}\right)^\alpha \leq \sup_{z \in S \backslash \mathcal{R}_{\partial S}} \left(\frac{p(z)}{(p(z) - \epsilon_m)_+}\right)^\alpha \leq \left(\frac{\lambda}{(\lambda - \epsilon_m)_+}\right)^\alpha.$$

So

$$\widehat{D}_\alpha(x_1, x_2) \leq \left(\frac{\lambda}{(\lambda - \epsilon_m)_+}\right)^\alpha \widetilde{D}_\alpha(x_1, x_2).$$

Similarly,

$$\widehat{D}_\alpha(x_1,x_2) \geq \inf_{z\in S\setminus\mathcal{R}_{\partial S}} \left(\frac{p(z)}{p(z)+\epsilon_m}\right)^\alpha \widetilde{D}_\alpha(x_1,x_2) \geq \left(\frac{\lambda}{\lambda+\epsilon_m}\right)^\alpha \widetilde{D}_\alpha(x_1,x_2).$$

□

Proposition 5 *With the notation of Proposition 4, for all* x_1, x_2,

$$D_\alpha(x_1,x_2) \leq \widetilde{D}_\alpha(x_1,x_2).$$

Assume $\sup_{x\in S\setminus\mathcal{R}_{\partial S}} |\widehat{p}_m(x)-p(x)| \leq \epsilon_m$ *and* $\partial\widehat{S} \subset \mathcal{R}_{\partial S}$. *Then for any* $(x_1,x_2) \in \Psi$, $\widetilde{D}_\alpha(x_1,x_2) \leq \frac{d_{S\setminus\widehat{\mathcal{R}}_{\partial S}}(x_1,x_2)}{\lambda^\alpha}$ *and*

$$\left(\frac{\lambda}{\lambda+\epsilon_m}\right)^\alpha D_\alpha(x_1,x_2) \leq \widehat{D}_\alpha(x_1,x_2) \leq \frac{d_{S_m}(x_1,x_2)}{(\lambda-\epsilon_m)_+^\alpha}$$

where we recall that $S_m = \left\{x \in S : \inf_{z\in\partial S} \|x-z\|_2 \geq 3\delta_m\right\}$. *Thus*

$$D_\alpha(x_1,x_2) \leq \left(\frac{\lambda+\epsilon_m}{\lambda}\right)^\alpha \widehat{D}_\alpha(x_1,x_2).$$

Proof. Since for any x_1 and x_2, $\widehat{\Gamma}(x_1,x_2) \subseteq \Gamma(x_1,x_2)$, clearly $D_\alpha(x_1,x_2) \leq \widetilde{D}_\alpha(x_1,x_2)$. If $\partial\widehat{S} \subset \mathcal{R}_{\partial S}$,

$$\begin{aligned}\widetilde{D}_\alpha(x_1,x_2) &= \inf_{\gamma\in\widehat{\Gamma}(x_1,x_2)} \int_0^{L(\gamma)} \frac{1}{p(\gamma(t))^\alpha}dt \leq \left[\sup_{z\in\widehat{S}\setminus\widehat{\mathcal{R}}_{\partial S}} \frac{1}{p(z)^\alpha}\right]\left[\inf_{\gamma\in\widehat{\Gamma}(x_1,x_2)} \int_0^{L(\gamma)} dt\right] \\ &\leq \frac{1}{\lambda^\alpha} d_{\widehat{S}\setminus\widehat{\mathcal{R}}_{\partial S}}(x_1,x_2) \leq \frac{1}{\lambda^\alpha} d_{S_m}(x_1,x_2)\end{aligned}$$

since, by the triangle inequality, $S_m \subseteq \widehat{S}\setminus\widehat{\mathcal{R}}_{\partial S}$. Applying Proposition 4, the result follows. □

Proposition 6 *Let* $\mathcal{X}$ *be a compact subset of* $\mathbb{R}^d$, *and* $T > 0$. *Then for any* $\tau \in (0,T)$, *for all sets* $S \subseteq \mathcal{X}$ *with condition number at least* τ, $\mathrm{Vol}(\partial S) \leq c_3/\tau$ *for some* c_3 *independent of* τ, *where* Vol *is the* $d-1$*-dimensional volume.*

Proof. Let $\{z_i\}_{i=1}^N$ be a minimal Euclidean $\tau/2$-covering of ∂S, and $B_i = \{x : \|x-z_i\|_2 \leq \tau/2\}$. Let T_i be the tangent plane to ∂S at z_i. Then using the argument made in the proof of Lemma 4 in Genovese et al. (2010),

$$\mathrm{Vol}(B_i \cap \partial S) \leq C_1 \,\mathrm{Vol}(B_i \cap T_i) \frac{1}{\sqrt{1-(\tau/2)^2/\tau^2}} \leq C_2 \tau^{d-1}$$

for some constants C_1 and C_2 independent of τ. Since $\mathcal{X}$ is compact,

$$\mathcal{N}(\partial S, \|\cdot\|_2, \tau/2) \le C\left(\frac{1}{\tau}\right)^d$$

for some constant C depending only on $\mathcal{X}$ and T, where $\mathcal{N}$ denotes the covering number (note that even though ∂S is a $d-1$ dimensional set, we can't claim $\mathcal{N}(\partial S, \|\cdot\|_2, \tau) = O(\tau^{-(d-1)})$, since ∂S can become space-filling as $\tau \to 0$). So

$$\mathrm{Vol}(\partial S) \le \sum_{i=1}^{N} \mathrm{Vol}(B_i \cap \partial S) \le C_2 \tau^{d-1} \mathcal{N}(\partial S, \|\cdot\|_2, \tau/2) \le C_2 C \tau^{-1}$$

and the result follows with $c_3 = C_2 C$. □

Proposition 7 *Let $\mathcal{X}$ be a compact subset of $\mathbb{R}^d$, and $T > 0$. Then for any $\tau \in (0, T)$, for all compact, connected sets $S \subseteq \mathcal{X}$ with condition number at least τ, $\sup_{u,v \in S} d_S(u,v) \le c_4 \tau^{1-d}$ for some c_4 independent of τ.*

Proof. First consider the quantity $\sup_{u,v \in \partial S} d_S(u,v)$. Since $\partial S \subseteq S$, clearly

$$\sup_{u,v \in \partial S} d_S(u,v) \le \sup_{u,v \in \partial S} d_{\partial S}(u,v).$$

Since ∂S is closed, there must exist $u^*, v^* \in \partial S$ such that

$$\sup_{u,v \in \partial S} d_{\partial S}(u,v) = d_{\partial S}(u^*, v^*).$$

Let $\{z_i\}_{i=1}^N$ be a minimal τ-covering of ∂S in the $d_{\partial S}$ metric. Let $\{\widetilde{z}_i\}_{i=1}^{\widetilde{N}} \subseteq \{z_i\}_{i=1}^N$ such that $d_{\partial S}(u^*, \widetilde{z}_1) \le \tau$, $d_{\partial S}(v^*, \widetilde{z}_{\widetilde{N}}) \le \tau$, and for any $1 \le i \le \widetilde{N}-1$, $d_{\partial S}(\widetilde{z}_i, \widetilde{z}_{i+1}) \le 2\tau$. Then

$$d_{\partial S}(u^*, v^*) \le d_{\partial S}(u^*, \widetilde{z}_1) + d_{\partial S}(v^*, \widetilde{z}_{\widetilde{N}}) + \sum_{i=1}^{\widetilde{N}-1} d_{\partial S}(\widetilde{z}_i, \widetilde{z}_{i+1}) \le 2\tau \widetilde{N}.$$

So, $d_{\partial S}(u^*, v^*) \le 2\tau \mathcal{N}(\partial S, d_{\partial S}, \tau)$. By Proposition 6.3 in Niyogi, Smale and Weinberger (2008) (or see Lemma 3 in Genovese et al. (2010)), if $x, y \in \partial S$ such that $\|x - y\|_2 = a \le \tau/2$, then $d_{\partial S}(x,y) \le \tau - \tau\sqrt{1 - (2a)/\tau}$. In particular, if $\|x - y\|_2 \le \tau/2$, then $d_{\partial S}(x,y) \le \tau$. So any Euclidean $\tau/2$-covering of ∂S is also a τ-covering in the $d_{\partial S}$ metric. Then we have

$$\begin{aligned} \sup_{u,v \in \partial S} d_S(u,v) &\le d_{\partial S}(u^*, v^*) \le 2\tau \mathcal{N}(\partial S, d_{\partial S}, \tau) \le 2\tau \mathcal{N}(\partial S, \|\cdot\|_2, \tau/2) \\ &\le C\tau \left(\frac{1}{\tau}\right)^d = C\tau^{1-d} \end{aligned}$$

for some constant C depending only on $\mathcal{X}$ and T (note that, as in the proof of 6, even though ∂S is a $d-1$ dimensional set, we can't claim $\mathcal{N}(\partial S, \|\cdot\|_2, \tau) = O(\tau^{-(d-1)})$, since ∂S can become space-filling as $\tau \to 0$).

Now let $u^\dagger, v^\dagger \in S$ such that $\sup_{u,v \in S} d_S(u,v) = d_S(u^\dagger, v^\dagger)$ which must exist since S is compact. Let $u^\ddagger, v^\ddagger \in \partial S$ be the (not necessarily unique) projections of $u^\dagger$ and $v^\dagger$ onto ∂S. Clearly the line segment connecting $u^\dagger$ and $u^\ddagger$ is fully contained in S, and the same applies to $v^\dagger$ and $v^\ddagger$. So,

$$\begin{aligned} d_S(u^\dagger, v^\dagger) &\leq d_S(u^\dagger, u^\ddagger) + d_S(u^\ddagger, v^\ddagger) + d_S(v^\ddagger, v^\dagger) \\ &\leq \|u^\dagger - u^\ddagger\|_2 + \|v^\dagger - v^\ddagger\|_2 + d_S(u^*, v^*) \\ &\leq 2\,\mathrm{diam}(\mathcal{X}) + C\tau^{1-d} \end{aligned}$$

and setting $c_4 = 2T^{d-1}\,\mathrm{diam}(\mathcal{X}) + C$, the result follows. □

References.

Belkin, M. and Niyogi, P. (2004). Semi-supervised learning on Riemannian manifolds. *Machine Learning* **56(1-3)** 209–239.

Ben-David, S., Lu, T. and Pal, D. (2008). Does Unlabeled Data Provably Help? Worst-case Analysis of the Sample Complexity of Semi-Supervised Learning. In *21st Annual Conference on Learning Theory (COLT)*.

Bijral, A., Ratliff, N. and Srebro, N. (2011). Semi-supervised Learning with Density Based Distances. In *27th Conference on Uncertainty in Artificial Intelligence*.

Bousquet, O., Chapelle, O. and Hein, M. (2004). Measure based regularization. In *Advances in Neural Information Processing Systems*.

Culp, M. (2011a). On Propagated Scoring for Semisupervised Additive Models. *Journal of the American Statistical Association* **106** 248-259.

Culp, M. (2011b). spa: Semi-Supervised Semi-Parametric Graph-Based Estimation in R. *Journal of Statistical Software* **40**.

Culp, M. and Michailidis, G. (2008). An Iterative Algorithm for Extending Learners to a Semi-Supervised Setting. *Journal of Computational and Graphical Statistics* **17** 545-571.

Einmahl, U. and Mason, D. M. (2005). Uniform in bandwidth consistency of kernel-type function estimators. *The Annals of Statistics* **33** 1380-1403.

Genovese, C., Perone-Pacifico, M., Verdinelli, I. and Wasserman, L. (2010). Minimax Manifold Estimation. *Arxiv preprint arXiv:1007.0549*.

Gine, E. and Guillou, A. (2002). Rates of strong uniform consistency for multivariate kernel density estimators. *Annales de l'Institute Henri Poincare (B) Probability and Staistics* **38** 907-921.

Gyorfi, L., Kohler, M., Krzyzak, A. and Walk, H. (2002). *A distribution-free theory of nonparametric regression*. Springer Verlag.

Haupt, J. and Nowak, R. (2006). Signal reconstruction from noisy random projections. *IEEE Trans. Info. Th.* **52** 4036–4048.

Kpotufe, S. (2011). k-NN Regression Adapts to Local Intrinsic Dimension. *NIPS*.

LAFFERTY, J. and WASSERMAN, L. (2007). Statistical Analysis of Semi-Supervised Regression. In *Advances in Neural Information Processing Systems 20* 801–808.

LEE, A. B. and WASSERMAN, L. (2008). Spectral Connectivity Analysis. *Arxiv preprint arXiv:0811.0121*.

LIANG, F., MUKHERJEE, S. and WEST, M. (2007). The Use of Unlabeled Data in Predictive Modeling. *Statistical Science* **22** 189-205.

NADLER, B., SREBRO, N. and ZHOU, X. (2009). Statistical analysis of semi-supervised learning: The limit of infinite unlabelled data. In *Advances in Neural Information Processing Systems 22* 1330–1338.

NIYOGI, P. (2008). Manifold Regularization and Semi-supervised Learning: Some Theoretical Analyses Technical Report No. TR-2008-01, Computer Science Department, University of Chicago. URL http://people.cs.uchicago.edu/∼niyogi/papersps/ ssminimax2.pdf.

NIYOGI, P., SMALE, S. and WEINBERGER, S. (2008). Finding the Homology of Submanifolds with High Confidence from Random Samples. *Discrete & Computational Geometry* **39** 419–441.

PRAKASA-RAO, B. L. S. (1983). *Nonparametric Functional Estimation*. Academic Press.

RIGOLLET, P. (2007). Generalization error bounds in semi-supervised classification under the cluster assumption. *Journal of Machine Learning Research* **8** 1369–1392.

SAJAMA, and ORLITSKY, A. (2005). Estimating and computing density based distance metrics. In *Proceedings of the 22nd international conference on Machine learning. ICML 2005* 760–767.

SINGH, A., NOWAK, R. D. and ZHU, X. (2008a). Unlabeled data: Now it helps, now it doesn't Technical Report, University of Wisconsin - Madison, ECE Department. URL http://www.cae.wisc.edu/∼singh/SSL_TR.pdf.

SINGH, A., NOWAK, R. and ZHU, X. (2008b). Unlabeled data: Now it helps, now it doesn't. In *Neural Information Processing Systems (NIPS)*.

SINHA, K. and BELKIN, M. (2009). Semi-supervised Learning using Sparse Eigenfunction Bases. In *Advances in Neural Information Processing Systems 22* (Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams and A. Culotta, eds.) 1687–1695.

VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge.

YUKICH, J. E. (1985). Laws of Large Numbers for Classes of Functions. *Journal of Multivariate Analysis* **17** 245-260.

DEPARTMENT OF STATISTICS
AND MACHINE LEARNING DEPARTMENT
CARNEGIE MELLON UNIVERSITY
PITTSBURGH, PENNSYLVANIA 15213
USA
E-MAIL: mazizyan@cs.cmu.edu
aarti@cs.cmu.edu
larry@stat.cmu.edu