The Annals of Statistics
2013, Vol. 41, No. 2, 751–771
DOI: 10.1214/13-AOS1092

DENSITY-SENSITIVE SEMISUPERVISED INFERENCE

BY MARTIN AZIZYAN[1], AARTI SINGH[1] AND LARRY WASSERMAN[2]

Carnegie Mellon University

Semisupervised methods are techniques for using labeled data $(X_1, Y_1), \dots, (X_n, Y_n)$ together with unlabeled data $X_{n+1}, \dots, X_N$ to make predictions. These methods invoke some assumptions that link 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.

1. Introduction. Suppose we have data $(X_1, Y_1), \dots, (X_n, Y_n)$ 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 $X_{n+1}, \dots, X_N$ from the same distribution but without the Y's. We refer to $\mathcal{L} = \{(X_i, Y_i) : i = 1, \dots, n\}$ as the *labeled data* and $\mathcal{U} = \{X_i : i = n + 1, \dots, N\}$ 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 $m = N - n$ unobserved labels $Y_{n+1}, \dots, Y_N$ 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 identifying the decision boundary becomes

Received April 2012; revised January 2013.

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

[2] Supported by NSF Grant DMS-08-06009 and Air Force Grant FA95500910373.

MSC2010 subject classifications. Primary 62G15; secondary 62G07.

Key words and phrases. Nonparametric inference, semisupervised, kernel density, efficiency.

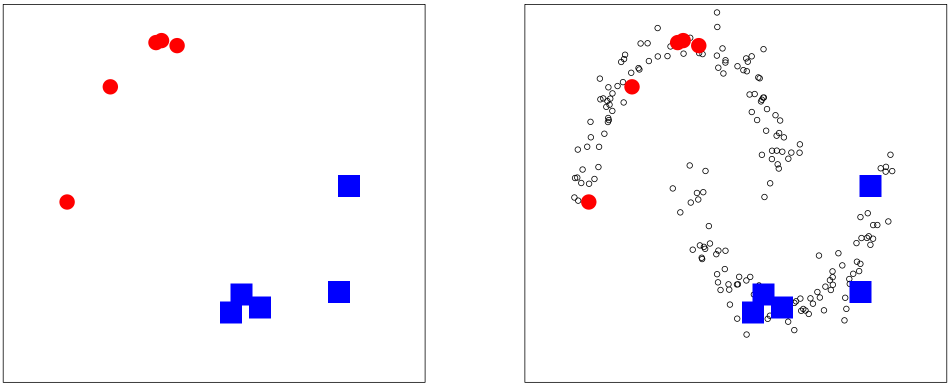

FIG. 1. *The covariate $X = (X_1, X_2)$ is two dimensional. The response Y is binary and is shown as a square or a circle. Left: the labeled data. Right: labeled and unlabeled data.*

much easier. In other words, if we assume some link between P_X and f, then we can use the unlabeled 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. 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 (2008), 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

$$\mathcal{L}_n \Longrightarrow \widehat{f} \qquad\qquad \mathcal{L}_n \Longrightarrow \widehat{f} \xLeftarrow{\text{SS assumption}} \widehat{P}_X \Longleftarrow \mathcal{U}_N$$

FIG. 2. *Supervised learning (left) uses only the labeled data $\mathcal{L}_n$. Semisupervised learning (right) uses the unlabeled 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.*

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

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 function 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.

We now give an informal statement of our main results. In Section 5 we define a nonparametric class of distributions $\mathcal{P}_n$. Let $0 < \xi < d - 3$ and assume that $m \geq n^{2/(2+\xi)}$. Let $\mathcal{S}_n$ denote the set of supervised estimators; these estimators use only the labeled data. Let $\mathcal{SS}_N$ denote the set of semisupervised estimators; these estimators use the labeled data and unlabeled data. Then:

(1) (Theorem 4.1 and Corollary 4.2.) There is a semisupervised estimator $\widehat{f}$ such that

$$\sup_{P \in \mathcal{P}_n} R_P(\widehat{f}) \leq \left(\frac{C}{n}\right)^{2/(2+\xi)}, \tag{1}$$

where $R_P(\widehat{f})$ is the risk of the estimator $\widehat{f}$ under distribution P.

(2) (Theorem 5.1.) For supervised estimators $\mathcal{S}_n$ we have

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

(3) Combining these two results we conclude that

$$\frac{\inf_{\widehat{f} \in \mathcal{SS}_N} \sup_{P \in \mathcal{P}_n} R_P(\widehat{f})}{\inf_{\widehat{f} \in \mathcal{S}_n} \sup_{P \in \mathcal{P}_n} R_P(\widehat{f})} \leq \left(\frac{C}{n}\right)^{2(d-3-\xi)/((2+\xi)(d-1))} \to 0 \tag{3}$$

and hence, semisupervised estimation dominates supervised estimation.

REMARK. We assume, as is standard in the literature on semisupervised learning, that the margial P_X is the same for the labeled and unlabeled data. Extensions to the case where the marginal distribution changes are possible, but are beyond the scope of the paper.

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 Castelli and Cover (1995, 1996), Ratsaby and Venkatesh (1995), Bousquet, Chapelle and Hein (2004), Singh, Nowak and Zhu (2008), 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 covariate space $\mathcal{X}$ into clusters $C_1, \dots, C_k$ 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 $\{x : p_X(x) = \lambda\}$. Also, this does not yield improved rates over C_0. 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 define a density sensitive semisupervised estimator, and we bound its risk. In Section 5 we present some minimax results. We discuss adaptation in Section 6. We provide simulations in Section 7. Section 8 contains the closing discussion. Many technical details and extensions are contained in the supplemental article [Azizyan, Singh and Wasserman (2013)].

2. Definitions.

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

$$\mathcal{L}_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\} \tag{4}$$

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

$$\mathcal{U}_N = \{X_{n+1}, \dots, X_N\} \tag{5}$$

be an i.i.d. 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 $\mathcal{L}_n \cup \mathcal{U}_N$ 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

$$R_P(\widehat{f}) = \mathbb{E}_P\left[\int (\widehat{f}(x) - f_P(x))^2 \, dP(x)\right], \tag{6}$$

where $\mathbb{E}_P$ denotes the expectation over data drawn from the distribution P. Of course, $\mathcal{S}_n \subset \mathcal{SS}_N$ and hence

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

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

$$\frac{\inf_{\widehat{g} \in \mathcal{SS}_N} \sup_{P \in \mathcal{P}_n} R_P(\widehat{g})}{\inf_{\widehat{g} \in \mathcal{S}_n} \sup_{P \in \mathcal{P}_n} R_P(\widehat{g})} \to 0 \tag{7}$$

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 $N = N(n) \to \infty$ 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 $\delta \ge 0$, 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 $a_n = O(b_n)$ and $a_n = \Omega(b_n)$. We also write $a_n \preceq b_n$ if there exists $C > 0$ such that $a_n \le Cb_n$ 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.

3. Density-sensitive function spaces. We define a smoothed version of P_X as follows. (This is needed since we allow the marginal distribution P_X to be singular.) Let K denote a symmetric kernel on $\mathbb{R}^d$ with compact support, let $\sigma > 0$ and define

$$p_\sigma(x) \equiv p_{X,\sigma}(x) = \int \frac{1}{\sigma^d} K\left(\frac{\|x-u\|}{\sigma}\right) dP_X(x). \tag{8}$$

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. 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*

$$D(x_1, x_2) \equiv D_{P,\alpha,\sigma}(x_1, x_2) = \inf_{\gamma \in \Gamma(x_1,x_2)} \int_0^{L(\gamma)} \exp[-\alpha p_{X,\sigma}(\gamma(t))]\, dt. \tag{9}$$

In the supplement, 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).

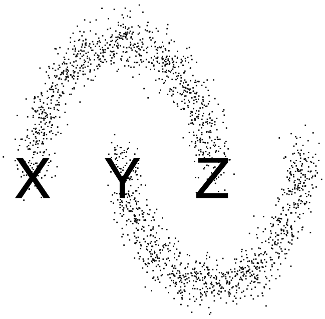

FIG. 3. *With a density metric, the points X and Z are closer than the points X and Y because there is a high density path connecting X and Z.*

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.[3] We formalize the semisupervised smoothness assumption by defining the following scale of function spaces. 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 $x_1, x_2 \in \mathcal{X}$,

$$\bigl|f(x_1) - f(x_2)\bigr| \leq L D_{P,\alpha,\sigma}(x_1, x_2). \tag{10}$$

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.* Let $B_{P,\alpha,\sigma}(x, \varepsilon) = \{z : D_{P,\alpha,\sigma}(x, z) \leq \varepsilon\}$ be a ball of size ε. Let S_P denote the support of P, and let $\mathcal{N}_{P,\alpha,\sigma}(\varepsilon)$ denote the covering number, the smallest number of balls of size ε required to cover S_P. The covering number measures the size of the function space, and the variance of any regression estimator on the space $\mathcal{F}(P, \alpha, \sigma, L)$ depends on this covering number. Here, we mention a few properties of $\mathcal{N}_{P,\alpha,\sigma}(\varepsilon)$.

In the Euclidean case $\alpha = 0$, we have $\mathcal{N}_{P,0,\sigma}(\varepsilon) \leq (C/\varepsilon)^d$. But when $\alpha > 0$ and P is concentrated on or near a set of dimension less than d, the $\mathcal{N}_{P,\alpha,\sigma}(\varepsilon)$ can be much smaller than $(C/\varepsilon)^d$. 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 $\varepsilon > 0$,

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

where $d_A(x_1, x_2)$ is the length of the shortest path in A connecting x_1 and x_2. Recall that S_P denotes the support of P.

LEMMA 1. *Suppose that S_P is regular.*

(1) *For all α, σ and P, $\mathcal{N}_{P,\alpha,\sigma}(\varepsilon) \preceq \varepsilon^{-d}$.*

(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 $\varepsilon > 0$, $\mathcal{N}_{P,\alpha,\sigma}(\varepsilon) \leq k$.*

(3) *Suppose that* $\mathsf{dim}(S_P) = r < d$. *Then, $\mathcal{N}_{P,\alpha,\sigma}(\varepsilon) \preceq \varepsilon^{-r}$.*

(4) *Suppose that $S_P = W \oplus \gamma$ where* $\mathsf{dim}(W) = r < d$. *Then, for $\varepsilon \geq C\gamma$, $\mathcal{N}_{P,\alpha,\sigma}(\varepsilon) \preceq (\frac{1}{\varepsilon})^r$.*

PROOF. (1) The first statement 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 $O(\varepsilon^{-d})$ Euclidean balls.

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

(2) The second statement follows since $\{\{x_1\}, \dots, \{x_k\}\}$ forms an ε-covering for any ε.

(3) We have that $D_{P,\alpha,\sigma}(x, y) \le d_{S_P}(x, y)$. Regularity implies that, for small $d_{S_P}(x, y)$, $D_{P,\alpha,\sigma}(x, y) \le c\|x - y\|$. We can thus cover S_P by $C\varepsilon^{-r}$ balls of size ε.

(4) As in (3), cover W with $N = O(\varepsilon^{-r})$ balls of D size ε. Denote these balls by $B_1, \dots, B_N$. Define $C_j = \{x \in S_P : d_{S_P}(x, B_j) \le \gamma\}$. The C_j form a covering of size N and each C_j has $D_{P,\alpha,\sigma}$ diameter $\max\{\varepsilon, \gamma\}$. □

4. Semisupervised kernel estimator. We consider the following semisupervised estimator 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

$$\widehat{f}_{h,\alpha,\sigma}(x) = \frac{\sum_{i=1}^n Y_i Q_h(\widehat{D}_{\alpha,\sigma}(x, X_i))}{\sum_{i=1}^n Q_h(\widehat{D}_{\alpha,\sigma}(x, X_i))}, \tag{12}$$

where

$$\widehat{D}_{\alpha,\sigma}(x_1, x_2) = \inf_{\gamma \in \Gamma(x_1,x_2)} \int_0^{L(\gamma)} \exp[-\alpha \widehat{p}_\sigma(\gamma(t))]\, dt, \tag{13}$$

$$\widehat{p}_\sigma(x) = \frac{1}{m} \sum_{i=1}^m \frac{1}{\sigma^d} K\left(\frac{\|x - X_{i+n}\|}{\sigma}\right), \tag{14}$$

and $m = N - n$ denotes the number of unlabeled points. We use a kernel estimator for the regression function because it is simple, commonly used and, as we shall see, has a fast rate of convergence in the semisupervised case.

The estimator $\widehat{D}_{\alpha,\sigma}(x_1, x_2)$ is discussed in detail in the supplement where we study its properties and we give an algorithm for computing it.

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

THEOREM 4.1. *Suppose that $|Y| \le M$. Define the event $\mathcal{G}_m = \{\|\widehat{p}_\sigma - p_\sigma\|_\infty \le \varepsilon_m\}$ (which depends on the unlabeled data) and suppose that $\mathbb{P}(\mathcal{G}_m^c) \le 1/m$. Then, for every $P \in \mathcal{P}(\alpha, \sigma, L)$,*

$$R_P(\widehat{f}_{h,\alpha,\sigma}) \le L^2(he^{\alpha\varepsilon_m})^2 + \frac{M^2(2 + 1/e)\mathcal{N}(P, \alpha, \sigma, e^{-\varepsilon_m\alpha}h/2)}{n} + \frac{4M^2}{m}. \tag{15}$$

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

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

Now we bound the second term.

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

$$\begin{aligned}&\mathbb{E}_n\biggl[\int\bigl(\widehat{f}_{h,\alpha,\sigma}(x)-f(x)\bigr)^2\,dP(x)\biggr]\\&\qquad\leq L^2R^2+\frac{M^2(2+1/e)\int dP(x)/P(\widehat{B}_{\alpha,\sigma}(x,h))}{n},\end{aligned}$$

where

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

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 2 in the supplement that $e^{-\alpha\varepsilon_m}D_{\alpha,\sigma}(x_1,x_2)\leq\widehat{D}_{\alpha,\sigma}(x_1,x_2)\leq e^{\alpha\varepsilon_m}D_{\alpha,\sigma}(x_1,x_2)$ for all x_1,x_2. Hence, $R^2\leq e^{2\alpha\varepsilon_m}h^2$ and

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

A simple covering argument [see page 76 of Györfi et al. (2002)] 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 4.2. *If $\mathcal{N}(P,\alpha,\sigma,\delta)\leq(C/\delta)^\xi$ for $\delta\geq(1/2)e^{-\alpha\varepsilon_m}(n\times e^{2\alpha\varepsilon_m})^{-1/(2+\xi)}$ and $N\geq 2n$, then*

$$R_P(\widehat{f}_{\alpha,\sigma,h})\leq e^{\alpha\varepsilon_m(2\vee\xi)}\biggl[L^2h^2+\frac{1}{n}\biggl(\frac{C}{h}\biggr)^\xi\biggr]+\frac{4M^2}{m}. \tag{16}$$

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

$$\sup_{P\in\mathcal{P}(\alpha,\sigma,L)}R_P(\widehat{f}_{h,\alpha,\sigma})\preceq\biggl(\frac{C}{n}\biggr)^{2/(2+\xi)}. \tag{17}$$

5. 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}. 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 semisuspervised 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.

5.1. *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

$$\varepsilon_m \equiv \varepsilon(m,\sigma) = \sqrt{\frac{C\log m}{m\sigma^d}}. \tag{18}$$

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

$$\mathcal{P}_n = \bigcup_{(\alpha,\sigma)\in\mathcal{A}_n\times\Sigma_n} \mathcal{Q}(\alpha,\sigma,L), \tag{19}$$

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:

(C1) $\mathcal{Q}(\alpha,\sigma,L)$

$$= \left\{P \in \mathcal{P}(\alpha,\sigma,L) : \mathcal{N}(P,\alpha,\sigma,\varepsilon) \le \left(\frac{C}{\varepsilon}\right)^{\xi}\ \forall\varepsilon \ge \left(\frac{1}{n}\right)^{1/(2+\xi)}\right\};$$

(C2) $\alpha \le \dfrac{\log 2}{\varepsilon(m,\sigma)}$;

(C3) $\left(\dfrac{1}{m}\right)^{1/(d(1+\gamma))} \le \sigma \le \dfrac{1}{4C_0}\left(\dfrac{1}{n}\right)^{1/(d-1)}$,

where C_0 is the diameter of the support of K.

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

(1) (C2) implies that $e^{\alpha\varepsilon_m} \le 2$ and hence, (C3) and Theorem 1.3 in the supplement $(1/2)D_{P,\alpha,\sigma}(x_1,x_2) \le \widehat{D}_{\alpha,\sigma}(x_1,x_2) \le 2D_{P,\alpha,\sigma}(x_1,x_2)$ with probability at least $1 - 1/m$.

(2) The constraint in (C1) on $\mathcal{N}(\varepsilon)$ 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.

(3) 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.

5.2. *Supervised lower bound.*

THEOREM 5.1. *Suppose that* $m \geq n^{d(1+\gamma)/(d-1)}$. *There exists* $C > 0$ *such that*

$$R_S = \inf_{\widehat{f} \in \mathcal{S}_n} \sup_{P \in \mathcal{P}_n} R_P(\widehat{f}) \geq \left(\frac{C}{n}\right)^{2/(d-1)}. \tag{20}$$

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

$$A_1 = \{(x_1, \dots, x_{d-1}, 1) : 0 \leq x_1, \dots, x_{d-1} \leq 1\},$$
$$A_0 = \{(x_1, \dots, x_{d-1}, 0) : 0 \leq x_1, \dots, x_{d-1} \leq 1\}.$$

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

$$\{(s_1\varepsilon, s_2\varepsilon, \dots, s_{d-1}\varepsilon, x_d) : \varepsilon \leq x_d \leq 1 - \varepsilon\}.$$

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

Let

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

Define a measure μ on C as follows:

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

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 4. If

$$T_j = \{(s_1\varepsilon, s_2\varepsilon, \dots, s_{d-1}\varepsilon, x_d) : \varepsilon \leq x_d \leq 1 - \varepsilon\}$$

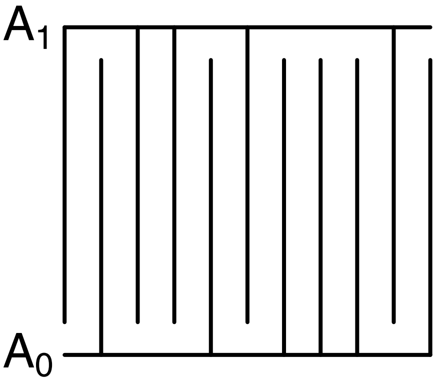

FIG. 4. *The extended tendrils used in the proof of the lower bound, in the special case where $d = 2$. Each tendril has length $1 - \varepsilon$ and joins up with either the top A_1 or bottom A_0 but not both.*

is a tendril, define its extensions

$$T_{j,0} = \{(s_1\varepsilon, s_2\varepsilon, \dots, s_{d-1}\varepsilon, x_d) : 0 \le x_d \le 1 - \varepsilon\},$$
$$T_{1,j} = \{(s_1\varepsilon, s_2\varepsilon, \dots, s_{d-1}\varepsilon, x_d) : \varepsilon \le x_d \le 1\}.$$

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

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

and

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

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,

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

Let

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

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: Let

$$\sigma = \left(\frac{1}{m}\right)^{1/(d(1+\gamma))}$$

and let

$$\frac{3}{2+\xi}\frac{\log m}{m^{1/(1+\gamma)}} \le \alpha \le \sqrt{\frac{m^{\gamma/(1+\gamma)}}{\log m}}. \tag{21}$$

It follows that (C2) and (C3) hold. We must verify (C1). 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, that is, $x \in U_\omega^{(1)}$, $y \in U_\omega^{(0)}$. Let us choose x and y as close as possible in Euclidean distance; hence $\|x - y\| = \varepsilon$. 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 (C3), $\sigma \le \varepsilon/(4C_0)$ 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 $\varepsilon - 2C_0\sigma \ge \varepsilon/2$. Since $p_{X,\sigma}(x) = 0$ on a portion of γ_0,

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

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

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

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

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

for all $\varepsilon \ge n^{-1/(2+\xi)}$. 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 $\delta e^{-\alpha K(0)/\sigma^d} \le \varepsilon$. Thus the covering number of the top and bottom is at most $2(1/\delta)^{d-1} \le 2(1/(e^{\alpha K(0)/\sigma^d}\varepsilon))^{d-1}$. 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 $\delta e^{-\alpha K(0)/\sigma^d} \le \varepsilon$. Thus the covering number of the tendrils is at most $q/\delta = n/\delta \le n/(\varepsilon e^{\alpha K(0)/\sigma^d})$. Thus we can cover the support with

$$N(\varepsilon) \le 2\left(\frac{1}{e^{\alpha K(0)/\sigma^d}\varepsilon}\right)^{d-1} + \frac{n}{\varepsilon e^{\alpha K(0)/\sigma^d}}$$

balls of size ε. It follows from (21) that $N(\varepsilon) \le (1/\varepsilon)^\xi$ for $\varepsilon \ge n^{-1/(2+\xi)}$ as required.

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

Proof: We have

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

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

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

Proof: This follows from direct calculation.

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

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 $B = T_{j,1} \times \{L\varepsilon\}$. Note that $A \cap B = \varnothing$. So,

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

and

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

Thus,

$$\|P_\omega^n \wedge P_\nu^n\| \ge \tfrac{1}{8}(1 - \mathsf{TV}(P_\omega, P_\nu))^{2n} \ge \tfrac{1}{8}(1 - \varepsilon^{d-1}/2)^{2n}.$$

Since $\varepsilon = n^{-1/(d-1)}$, 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 $\varepsilon = n^{-1/(d-1)}$. Combining Assouad's lemma (see Lemma 3 in the supplement) with the above claims, we have

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

□

5.3. *Semisupervised upper bound.* Now we state the upper bound for this class.

THEOREM 5.2. *Let $h = (ne^{2(2-\xi)})^{-1/(2+\xi)}$. Then*

$$\sup_{P \in \mathcal{P}_n} R(\widehat{f}_{h,\alpha,\sigma}) \leq \left(\frac{C}{n}\right)^{2/(2+\xi)}. \tag{22}$$

PROOF. This follows from (C2), (C3) and Corollary 4.2. □

5.4. *Comparison of lower and upper bound.* Combining the last two theorems we have:

COROLLARY 5.3. *Under the conditions of the previous theorem, and assuming that $d > \xi + 3$,*

$$\frac{R_{SS}}{R_S} \preceq \left(\frac{1}{n}\right)^{2(d-3-\xi)/((2+\xi)(d-1))} \to 0 \tag{23}$$

as $n \to \infty$.

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

6. 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 $Y_i = f(X_i) + \varepsilon_i$ with $\varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$. Also, we assume that $|f(x)|, |\widehat{f}_\theta^T(x)| \le M$, where $M > 0$ is a constant.[4]

THEOREM 6.1. *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 $0 < \delta < 1$,

$$\mathbb{E}[\mathcal{E}(\widehat{f}_\theta)] \le \frac{1}{1-a}\left[\min_{\theta \in \Theta} \mathbb{E}[\mathcal{E}(\widehat{f}_\theta)] + \frac{\log(J)/\delta)}{nt}\right] + 4\delta M^2, \tag{24}$$

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:

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

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

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

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

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

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

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

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

$$\mathcal{E}(f) - \widehat{\mathcal{E}}(f) \le \frac{\log(1/\delta)}{nt} + \frac{t2(M^2+\sigma^2)\mathcal{E}(f)}{(1-c)}.$$

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 $> 1-\delta$,

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

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 $> 1-\delta$,

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

Now,

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

Taking expectation with respect to validation dataset,

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

Now taking expectation with respect to training dataset,

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

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

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

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 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.

7. 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 5 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 approximation method described in the supplement [see equation (10)] with the number of nearest neighbors used set to $k = 20$.

Figure 6 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.

8. 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

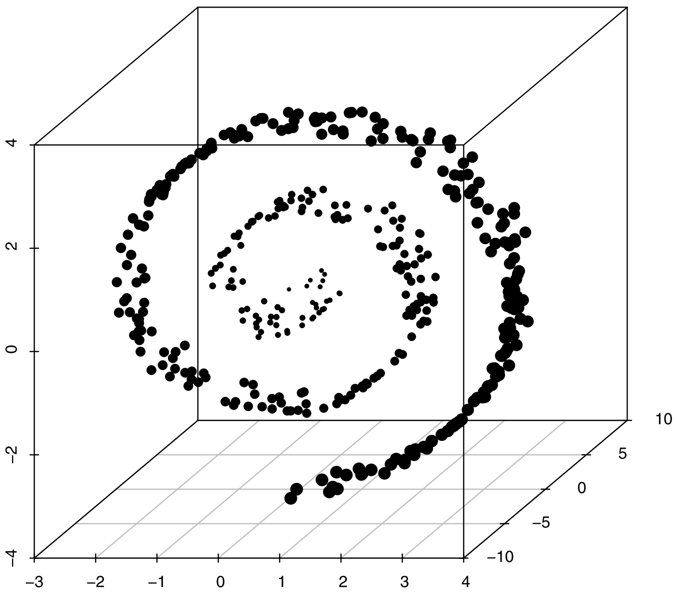

FIG. 5. *The swiss roll data set. Point size represents regression function.*

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 has good risk when the semi-supervised smoothness holds.

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.

SUPPLEMENTARY MATERIAL

Supplement to "Density-sensitive semisupervised inference" (DOI: 10.1214/13-AOS1092SUPP; .pdf). Contains technical details, proofs and extensions.

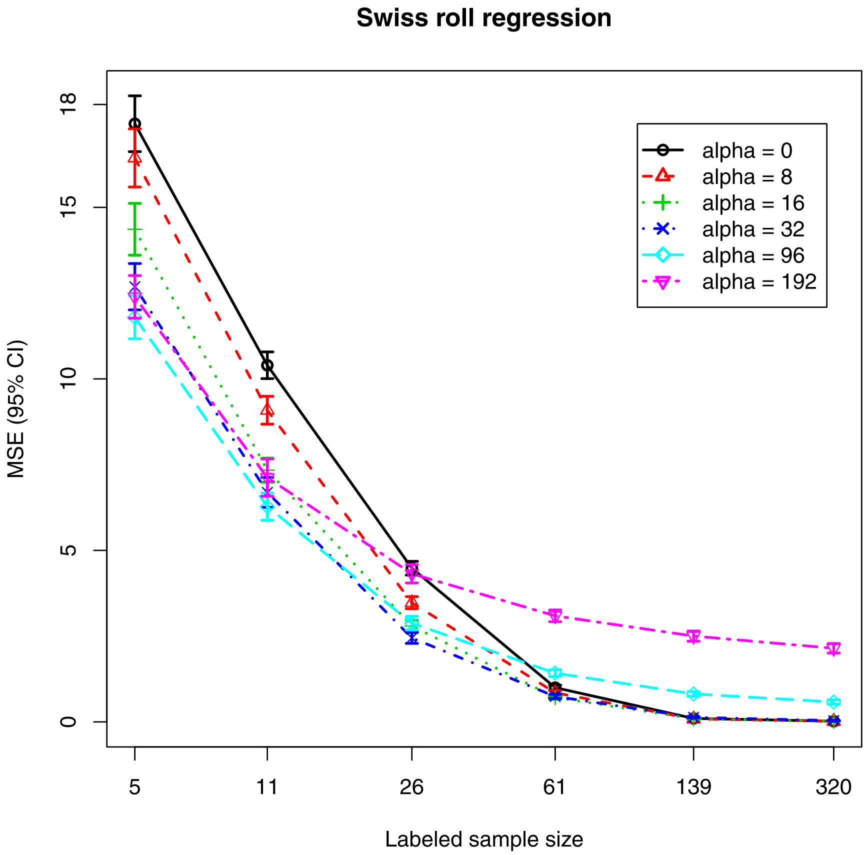

FIG. 6. *MSE of kernel regression on the swiss roll data set for a range of labeled sample sizes using different values of* α.

REFERENCES

AZIZYAN, M., SINGH, A. and WASSERMAN, L. (2013). Supplement to "Density-sensitive semisupervised inference." DOI:10.1214/13-AOS1092SUPP.

BELKIN, M. and NIYOGI, P. (2004). Semi-supervised learning on Riemannian manifolds. *Machine Learning* **56** 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 21*st Annual Conference on Learning Theory* (*COLT*). Available at http://www.informatik.uni-trier.de/~ley/db/conf/colt/colt2008.html.

BIJRAL, A., RATLIFF, N. and SREBRO, N. (2011). Semi-supervised learning with density based distances. In 27*th Conference on Uncertainty in Artificial Intelligence*. Available at http://auai.org/uai2011/accepted.html.

BOUSQUET, O., CHAPELLE, O. and HEIN, M. (2004). Measure based regularization. In *Advances in Neural Information Processing Systems* **16**. MIT Press, Cambridge, MA.

CASTELLI, V. and COVER, T. M. (1995). On the exponential value of labeled samples. *Pattern Recognition Letters* **16** 105–111.

CASTELLI, V. and COVER, T. M. (1996). The relative value of labeled and unlabeled samples in pattern recognition with an unknown mixing parameter. *IEEE Trans. Inform. Theory* **42** 2102–2117. MR1447517

CRAIG, C. C. (1933). On the Tchebychef inequality of Bernstein. *Ann. Math. Statist.* **4** 94–102.

CULP, M. (2011a). On propagated scoring for semisupervised additive models. *J. Amer. Statist. Assoc.* **106** 248–259. MR2816718

CULP, M. (2011b). spa: Semi-supervised semi-parametric graph-based estimation in R. *Journal of Statistical Software* **40** 1–29.

CULP, M. and MICHAILIDIS, G. (2008). An iterative algorithm for extending learners to a semi-supervised setting. *J. Comput. Graph. Statist.* **17** 545–571. MR2451341

GYÖRFI, L., KOHLER, M., KRZYŻAK, A. and WALK, H. (2002). *A Distribution-Free Theory of Nonparametric Regression*. Springer, New York. MR1920390

HAUPT, J. and NOWAK, R. (2006). Signal reconstruction from noisy random projections. *IEEE Trans. Inform. Theory* **52** 4036–4048. MR2298532

KPOTUFE, S. (2011). *k*-NN regression adapts to local intrinsic dimension. In *Advances in Neural Information Processing Systems* **24** 729–737. MIT Press, Cambridge, MA.

LAFFERTY, J. and WASSERMAN, L. (2007). Statistical analysis of semi-supervised regression. In *Advances in Neural Information Processing Systems* **20** 801–808. MIT Press, Cambridge, MA.

LEE, A. B. and WASSERMAN, L. (2008). Spectral connectivity analysis. Preprint. Available at arXiv:0811.0121.

LIANG, F., MUKHERJEE, S. and WEST, M. (2007). The use of unlabeled data in predictive modeling. *Statist. Sci.* **22** 189–205. MR2408958

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. MIT Press, Cambridge, MA.

NIYOGI, P. (2008). Manifold regularization and semi-supervised learning: Some theoretical analyses. Technical Report TR-2008-01, Computer Science Dept., Univ. Chicago. Available at http://people.cs.uchicago.edu/~niyogi/papersps/ssminimax2.pdf.

RATSABY, J. and VENKATESH, S. S. (1995). Learning from a mixture of labeled and unlabeled examples with parametric side information. In *Proceedings of the Eighth Annual Conference on Computational Learning Theory* 412–417. ACM, New York.

RIGOLLET, P. (2007). Generalized error bounds in semi-supervised classification under the cluster assumption. *J. Mach. Learn. Res.* **8** 1369–1392. MR2332435

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. ACM, New York.

SINGH, A., NOWAK, R. D. and ZHU, X. (2008). Unlabeled data: Now it helps, now it doesn't. Technical report, ECE Dept., Univ. Wisconsin–Madison. Available at http://www.cs.cmu.edu/~aarti/pubs/SSL_TR.pdf.

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. MIT Press, Cambridge, MA.

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