Finite sample analysis of semi-supervised learning

Technical Report ECE-08-03

Aarti Singh, Robert D. Nowak
Department of Electrical and Computer Engineering
University of Wisconsin - Madison
Madison, WI 53706
{singh@cae,nowak@engr}.wisc.edu

Xiaojin Zhu
Department of Computer Sciences
University of Wisconsin - Madison
Madison, WI 53706
jerryzhu@cs.wisc.edu

Abstract

Empirical evidence shows that in favorable situations *semi-supervised* learning (SSL) algorithms can capitalize on the abundance of *unlabeled* training data to improve the performance of a learning task, in the sense that fewer *labeled* training data are needed to achieve a target error bound. However, in other situations unlabeled data do not seem to help. Recent attempts at theoretically characterizing SSL gains only provide a partial and sometimes apparently conflicting explanations of whether, and to what extent, unlabeled data can help. In this paper, we attempt to bridge the gap between the practice and theory of semi-supervised learning. We develop a finite sample analysis that characterizes the value of unlabeled data and quantifies the performance improvement of SSL compared to supervised learning. We show that there are large classes of problems for which SSL can significantly outperform supervised learning, in finite sample regimes and sometimes also in terms of error convergence rates.

1 Introduction

Labeled training data can be expensive, time-consuming and difficult to obtain in many applications. For example, hand-written character or speech recognition and document classification require an experienced human annotator, or in some applications each label might be the outcome of a specially designed experiment. Semi-supervised learning (SSL) aims to capitalize on the abundance of unlabeled training data to improve learning performance. A thorough survey of semi-supervised learning literature is available in [1]. Empirical evidence suggests that in certain favorable situations unlabeled data can help, while in other situations it does not. As a result, there have been several recent attempts [2, 3, 4, 5, 6, 7] at developing a theoretical understanding of semi-supervised learning. It is well-accepted that unlabeled data can help only if there exists a *link* between the marginal data distribution and the target function to be learnt. Two common types of links considered are the cluster assumption [8, 4, 5] which states that the target function is locally smooth over subsets of the feature space delineated by some property of the marginal density (but may not be globally smooth), and the manifold assumption [5, 7] which assumes that the target function lies on a low-dimensional manifold. In the cluster case, knowledge of these sets reduces the problem of estimating an inhomogeneous function to a homogeneous function, and in the manifold case, knowledge of the manifold reduces a high-dimensional problem to a low-dimensional problem. Thus, knowledge of these sets which can be gleaned from unlabeled data, simplify the learning task. However, recent attempts at characterizing the amount of improvement possible under these links only provide a partial and sometimes apparently conflicting (for example, [5] vs. [7]) explanations of whether or not, and to what extent semi-supervised learning helps. In this paper, we bridge the gap between these seemingly conflicting views and develop a minimax framework based on finite sample bounds to identify

situations in which unlabeled data help to improve learning. Our results quantify both the amount of improvement possible using SSL as well as the the relative value of unlabeled data.

In this work, we focus on learning under the cluster assumption. We formalize this assumption in the next section and go on to establish that there exist nonparametric classes of distributions, denoted $\mathcal{P}_{XY}$, for which the decision sets (over which the target function is smooth) are discernable from unlabeled data. Moreover, we show that there exist *clairvoyant* supervised learners that, given perfect knowledge of the decision sets denoted by $\mathcal{D}$, can significantly outperform any generic supervised learner f_n based on the n labeled samples in these classes. That is, if $\mathcal{R}$ denotes a risk of interest, $\widehat{f}_{\mathcal{D},n}$ denotes the clairvoyant supervised learner, and $\mathbb{E}$ denotes expectation with respect to training data, then $\sup_{\mathcal{P}_{XY}} \mathbb{E}[\mathcal{R}(\widehat{f}_{\mathcal{D},n})] < \inf_{f_n} \sup_{\mathcal{P}_{XY}} \mathbb{E}[\mathcal{R}(f_n)]$. This would imply that knowledge of the decision sets simplifies the supervised learning task. Based on this, we establish that there also exist semi-supervised learners, denoted $\widehat{f}_{m,n}$, that use m unlabeled examples in addition to the n labeled examples in order to estimate the decision sets, which perform as well as $\widehat{f}_{\mathcal{D},n}$, provided that m grows appropriately relative to n. Specifically, if the error bound for $\widehat{f}_{\mathcal{D},n}$ decays polynomially (exponentially) in n, then the number of unlabeled data m needs to grow polynomially (exponentially) with the number of labeled data n. We provide general results for a broad range of learning problems using finite sample error bounds. Then we consider regression problems in detail, and examine a concrete instantiation of these general results by deriving minimax lower bounds on the performance of any supervised learner and compare that to upper bounds on the errors of $\widehat{f}_{\mathcal{D},n}$ and $\widehat{f}_{m,n}$.

In their seminal papers, Castelli and Cover [9, 10] had suggested that, in the binary classification setting, the marginal distribution can be viewed as a mixture of class conditional distributions:

$$P_X(x) = aP(x|Y=1) + (1-a)P(x|Y=0),$$

where $a = P(Y=1)$. If this mixture is identifiable, that is, learning P_X is sufficient to resolve the component distributions, then the classification problem reduces to a simple hypothesis testing problem of deciding the label $(0/1)$ for each component. For hypothesis testing problems, the error converges exponentially fast in the number of labeled examples, whereas the error convergence is typically polynomial for classification. The ideas in this paper are similar, except that we do not require identifiability of the mixture component densities, and show that it suffices to only approximately learn the decision sets over which the label is smooth. More recent attempts at theoretically characterizing SSL have been relatively pessimistic. Rigollet [4] establishes that for a fixed collection of distributions satisfying a cluster assumption, unlabeled data do not provide an improvement in convergence rate. A similar argument was made by Lafferty and Wasserman [5], based on the work of Bickel and Li [11], for the manifold case. However, in a recent paper, Niyogi [7] gives a constructive example of a class of distributions supported on a manifold whose complexity increases with the number of labeled examples, and he shows a lower bound of $\Omega(1)$ for any supervised learner (that is, the error of any supervised learner is bounded from below by a constant), whereas there exists a semi-supervised learner that can provide an error bound of $O(n^{-1/2})$, assuming infinite unlabeled data. We bridge the gap between these seemingly conflicting views. Our arguments can be understood by the simple example shown in Fig. 1, where the distribution is supported on two components separated by a margin γ and the target function is smooth over each component. Given a finite sample of data, these density sets may or may not be discernable depending on the sampling density (see Fig. 1(b), (c)). If γ is fixed (this is similar to fixing the class of cluster-based distributions in [4] or the manifold in [5, 11]), then given enough labeled data a supervised learner can achieve optimal performance (since, eventually, it operates in regime (c) of Fig. 1) and unlabeled data may not help. Thus, in this example, there is no improvement due to unlabeled data in terms of the rate of error convergence for a fixed collection of distributions. However, since the underlying true separation between the components is unknown, given a finite sample of data, there always exists a distribution for which these density sets are indiscernible (e.g., $\gamma \to 0$). This perspective is similar in spirit to the argument in [7]. We claim that meaningful characterizations of SSL performance and quantifications of the value of unlabeled data require finite sample error bounds, and that rates of convergence and asymptotic analysis may not capture the distinctions between SSL and supervised learning. Simply stated, if the component density sets are discernable from a finite sample size m of unlabeled data but not from a finite sample size $n < m$ of labeled data, then SSL can provide better performance than supervised learning. Further, we also show that there are certain plausible situations in which SSL yields rates of convergence that cannot be achieved by any supervised learner.

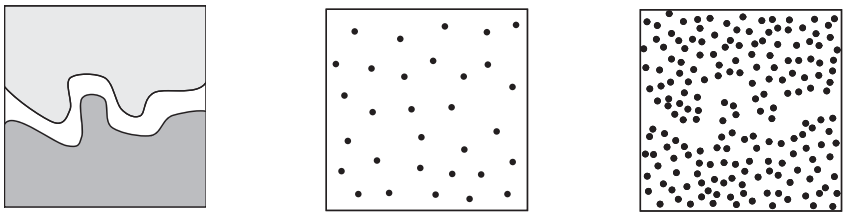

Figure 1: (a) Two separated high density sets with different labels that (b) cannot be discerned if the sample size is too small, but (c) can be estimated if sample density is high enough.

The rest of this paper is organized as follows. In the next section, we describe a mathematical model for the cluster assumption. Section 3 describes a procedure for learning the decision sets using unlabeled data. Our main result characterizing the relative performance of supervised and semi-supervised learning is presented in Section 4, and Section 5 applies the result to the regression problem. Conclusions are discussed in Section 6, and proofs are deferred to Section 7.

2 Characterization of model distributions under the cluster assumption

In this section, we describe a mathematical model for the cluster assumption. We define the collection of joint distributions $\mathcal{P}_{XY}(\gamma) = \mathcal{P}_X \times \mathcal{P}_{Y|X}$ indexed by a margin parameter γ as follows. Let X, Y be bounded random variables with marginal distribution $P_X \in \mathcal{P}_X$ and conditional label distribution $P_{Y|X} \in \mathcal{P}_{Y|X}$, supported on the domain $\mathcal{X} = [0,1]^d$.

The marginal density $p(x) = \sum_{k=1}^{K} a_k p_k(x)$ is the mixture of a finite, but unknown, number of component densities $\{p_k\}_{k=1}^{K}$, where $K < \infty$. Here the unknown mixing proportions $a_k \geq a > 0$ and $\sum_{k=1}^{K} a_k = 1$. In addition, we place the following assumptions on the mixture component densities $\{p_k\}_{k=1}^{K}$:

1. p_k is supported on a unique compact, connected set $C_k \subseteq \mathcal{X}$ with Lipschitz boundaries. Specifically, we assume the following form for the component support sets:

   $$C_k = \{x \equiv (x_1, \ldots, x_d) \in \mathcal{X} : g_k^{(1)}(x_1, \ldots, x_{d-1}) \leq x_d \leq g_k^{(2)}(x_1, \ldots, x_{d-1})\},$$

   where $g_k^{(1)}(\cdot), g_k^{(2)}(\cdot)$ are $d-1$ dimensional Lipschitz boundary functions with Lipschitz constant L. See Figure 2 for an illustrative example with $d = 2$.

   This form is a slight generalization of the boundary fragment class of sets which is used as a common tool for analysis of learning problems [12]. Boundary fragment sets capture the salient characteristics of more general decision sets since, locally, the boundaries of general sets are like fragments in a certain orientation.
2. p_k is bounded from above and below, $0 < b \leq p_k \leq B$.
3. p_k is Hölder-α_1 smooth on C_k with Hölder constant κ_1. Formally, p_k has continuous partial derivatives of up to order $[\alpha_1]$, where $[\alpha_1]$ denotes the maximal integer that is $< \alpha_1$, and $\exists \delta > 0$ such that

   $$\forall z, x \in C_k : ||z - x|| \leq \delta \Rightarrow |p_k(z) - TP_x(z, [\alpha_1])| \leq \kappa_1 \|z - x\|^{\alpha_1}$$

   where $\kappa_1, \alpha_1 > 0$, $TP_x(\cdot, [\alpha_1])$ denotes the degree $[\alpha_1]$ Taylor polynomial approximation of p_k expanded around x, and $||\cdot||$ denotes Euclidean norm.

Let the conditional label density on each component C_k be denoted by $p_k(Y|X = x)$. Thus, a labeled training point (X, Y) is obtained as follows. With probability a_k, X is drawn from p_k and Y is drawn from $p_k(Y|X = x)$. In the supervised setting, we assume access to n labeled training data $\mathcal{L} = \{X_i, Y_i\}_{i=1}^{n}$ drawn i.i.d according to $P_{XY} \in \mathcal{P}_{XY}(\gamma)$, and in the semi-supervised setting, we assume access to m additional unlabeled training data $\mathcal{U} = \{X_i\}_{i=1}^{m}$ drawn i.i.d according to $P_X \in \mathcal{P}_X$.

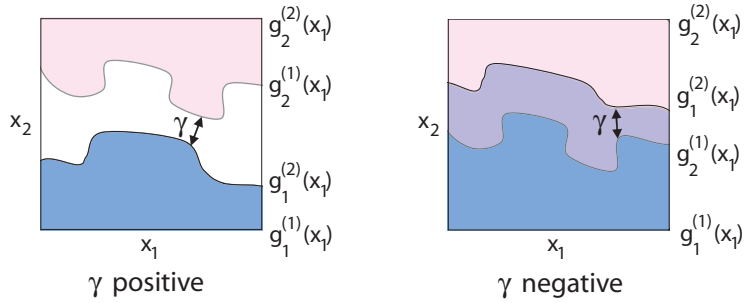

Figure 2: The margin γ measures the minimal width of a decision set, or separation between support sets of the marginal mixture component densities. The margin is positive if there is no overlap between the component support sets, and negative otherwise.

Let $\mathcal{D}$ denote the collection of all non-empty sets obtained as intersections of $\{C_k\}_{k=1}^K$ or their complements $\{C_k^c\}_{k=1}^K$, excluding the set $\cap_{k=1}^K C_k^c$ that does not lie in the support of the marginal density. Observe that $|\mathcal{D}| \leq 2^K$, and in practical situations the cardinality of $\mathcal{D}$ is much smaller as only a few of the sets are non-empty. The cluster assumption is that the target function will be smooth on each set $D \in \mathcal{D}$, hence the sets in $\mathcal{D}$ are called *decision sets*. At this point, we do not consider a specific target function; in Section 5, we will specify the smoothness assumptions on the target function in the regression setting.

The collection $\mathcal{P}_{XY}$ is indexed by a margin parameter γ, which denotes the minimum width of a decision set or separation between the component support sets C_k. The margin γ is assigned a positive sign if there is no overlap between components, otherwise it is assigned a negative sign as illustrated in Figure 2. Formally, for $j, k \in \{1, \ldots, K\}$, let

$$\begin{aligned} d_{jk} &:= \min_{p,q \in \{1,2\}} \|g_j^{(p)} - g_k^{(q)}\|_\infty \qquad\qquad j \neq k, \\ d_{kk} &:= \|g_k^{(1)} - g_k^{(2)}\|_\infty, \end{aligned}$$

where $\|\cdot\|_\infty$ denotes the sup-norm, and

$$\sigma = \begin{cases} 1 & \text{if } C_j \cap C_k = \emptyset \; \forall j \neq k, \text{ where } j, k \in \{1, \ldots, K\} \\ -1 & \text{otherwise} \end{cases}$$

Then the margin is defined as

$$\gamma = \sigma \cdot \min_{j,k \in \{1,\ldots,K\}} d_{jk}.$$

3 Learning Decision Sets

Ideally, we would like to break a given learning task into separate subproblems on each $D \in \mathcal{D}$, since the cluster assumption is that the target function is smooth on each decision set. In the section, we show that the decision sets are learnable using unlabeled data. Note that the marginal density p is smooth within each decision set $D \in \mathcal{D}$, but exhibits jumps at the decision boundaries since the component marginal mixture densities are bounded away from zero. Hence, the collection $\mathcal{D}$ can be learnt by estimating the marginal density from unlabeled data as follows:

1) *Marginal density estimation* — The procedure is based on the sup-norm kernel density estimator proposed in [13]. Consider a uniform square grid over the domain $\mathcal{X} = [0, 1]^d$ with spacing $2h_m$, where $h_m = \kappa_0 \left((\log m)^2/m\right)^{1/d}$ and $\kappa_0 > 0$ is a constant. For any point $x \in \mathcal{X}$, let $\bar{x}$ denote the closest point on the grid. Let G denote the kernel and $H_m = h_m \mathbf{I}$, then the estimator of $p(x)$ is

$$\widehat{p}(x) = \frac{1}{m h_m^d} \sum_{i=1}^m G(H_m^{-1}(X_i - \bar{x})).$$

2) *Decision set estimation* — Two points $x_1, x_2 \in \mathcal{X}$ are said to be *connected*, denoted by $x_1 \leftrightarrow x_2$, if there exists a sequence of points $x_1 = z_1, z_2, \ldots, z_{l-1}, z_l = x_2$ such that $z_2, \ldots, z_{l-1} \in \mathcal{U}$, $\|z_j - z_{j+1}\| \leq 2\sqrt{d}h_m$. That is, there exists a sequence of $2\sqrt{d}h_m$-dense unlabeled data points between x_1 and x_2. Two points $x_1, x_2 \in \mathcal{X}$ are said to be *p-connected* if in addition to being connected, the sequence is such that for all points that satisfy $\|z_i - z_j\| \leq h_m \log m$, $|\widehat{p}(z_i) - \widehat{p}(z_j)| \leq \delta_m := (\log m)^{-1/3}$. That is, there exists a sequence of $2\sqrt{d}h_m$-dense unlabeled data points between x_1 and x_2 such that the marginal density varies smoothly along the sequence. All points that are pairwise p-connected specify an empirical decision set. This decision set estimation procedure is similar in spirit to the semi-supervised learning algorithm proposed in [14]. In practice, p-connectedness only need to be evaluated for the test point X and the training points with labels, that is $\{X_i\}_{i=1}^n \in \mathcal{L}$.

The following lemma shows that if the margin is large relative to the average spacing between unlabeled data points ($m^{-1/d}$), then with high probability, two points are p-connected (lie in the same empirical decision set) if and only if they lie in the same decision set $D \in \mathcal{D}$, provided the points are not too close to the decision boundaries.

Lemma 1. *Denote the set of boundary points as*

$$\mathcal{B} := \{z : z_d = g_k^{(p)}(z_1, \ldots, z_{d-1}), k \in \{1, \ldots, K\}, p \in \{1, 2\}\}$$

and define the boundary set as

$$\mathcal{R}_{\mathcal{B}} := \{x : \inf_{z \in \mathcal{B}} \|x - z\| \leq 2\sqrt{d}h_m\}.$$

If $|\gamma| > C_o(m/(\log m)^2)^{-1/d}$, where $C_o = 6\sqrt{d}\kappa_0$, then for all $p \in \mathcal{P}_X$, all pairs of points $x_1, x_2 \in$ supp$(p) \setminus \mathcal{R}_{\mathcal{B}}$ and all $D \in \mathcal{D}$, with probability $> 1 - 1/m$,

$$x_1 \overset{p}{\leftrightarrow} x_2 \quad \text{if and only if} \quad x_1, x_2 \in D,$$

for large enough $m \geq m_0 \equiv m_0(p_{\min}, K, \kappa_1, d, \alpha_1, B, G, \kappa_0)$.[1]

The proof is given in Section 7.1.

Remark: If we are only concerned with distributions that have a positive margin, then only connectedness is needed to identify the decision sets. In fact, for the positive margin case, the decision sets correspond to connected components of the support set, and Hausdorff accurate support set estimation proposed in [15] (also see [16]) can be used to estimate the decision sets instead of identifying connecting sequences. One advantage of using Hausdorff accurate support set estimation over connecting sequences is that we can also handle densities that do not jump (are not bounded away from zero) but transition gradually to zero. However, in the negative margin case p-connectedness is needed since the supports of the mixture constituents in this case are overlapping and the decision sets are characterized by a sharp transition in the density.

4 SSL Performance and the Value of Unlabeled Data

We now state our main result that characterizes the performance of SSL relative to a clairvoyant supervised learner (with perfect knowledge of the decision sets), and follows as a corollary to the lemma stated above. Let $\mathcal{R}(f)$ denote a risk of interest for a learner f and the excess risk $\mathcal{E}(f) = \mathcal{R}(f) - \mathcal{R}^*$, where $\mathcal{R}^*$ is the infimum risk over all possible learners. The risk is given by the probability of error $P_{XY}(f(X) \neq Y)$ for classification and the mean square error $\mathbb{E}_{XY}[(f(X) - Y)^2]$ for regression.

Corollary 1. *Assume that the excess risk $\mathcal{E}$ is bounded. Suppose there exists a clairvoyant supervised learner $\widehat{f}_{\mathcal{D},n}$, with perfect knowledge of the decision sets $\mathcal{D}$, for which the following finite sample upper bound holds*

$$\sup_{\mathcal{P}_{XY}(\gamma)} \mathbb{E}[\mathcal{E}(\widehat{f}_{\mathcal{D},n})] \leq \epsilon_2(n).$$

[1] Dependence of a constant on G implies the constant depends on a norm or moment of the kernel G.

Then there exists a semi-supervised learner $\widehat{f}_{m,n}$ such that if $|\gamma| > C_o(m/(\log m)^2)^{-1/d}$, then

$$\sup_{\mathcal{P}_{XY}(\gamma)} \mathbb{E}[\mathcal{E}(\widehat{f}_{m,n})] \;\leq\; \epsilon_2(n) + O\left(\frac{1}{m} + n\left(\frac{m}{(\log m)^2}\right)^{-1/d}\right).$$

The proof is given in Section 7.2. This result captures the essence of the relative characterization of semi-supervised and supervised learning for the margin based model distributions. It suggests that if the sets $\mathcal{D}$ are discernable using unlabeled data (the margin is large enough compared to average spacing between unlabeled data points), then there exists a semi-supervised learner that can perform as well as a supervised learner with clairvoyant knowledge of the decision sets, provided $m \gg n$, so that $(n/\epsilon_2(n))^d = O(m/(\log m)^2)$ and the additonal term in the performance bound of the semi-supervised learner is small compared to $\epsilon_2(n)$. This implies that if $\epsilon_2(n)$ decays polynomially (exponentially) in n, then m needs to grow polynomially (exponentially) in n.

Further, suppose that the following finite sample lower bound holds for any supervised learner based on n labeled data:

$$\inf_{f_n} \sup_{\mathcal{P}_{XY}(\gamma)} \mathbb{E}[\mathcal{E}(f_n)] \geq \epsilon_1(n).$$

If $\epsilon_2(n) < \epsilon_1(n)$, then there exists a clairvoyant supervised learner with perfect knowledge of the decision sets that outperforms any supervised learner that does not have this knowledge. Hence, Corollary 1 implies that SSL can provide better performance than any supervised learner provided (i) $m \gg n$ so that $(n/\epsilon_2(n))^d = O(m/(\log m)^2)$, and (ii) knowledge of the decision sets simplifies the supervised learning task, so that $\epsilon_2(n) < \epsilon_1(n)$. In the next section, we provide a concrete application of this result in the regression setting. As a simple example in the binary classification setting, if $p(x)$ is supported on two disjoint sets and if $P(Y = 1|X = x)$ is strictly greater than $1/2$ on one set and strictly less than $1/2$ on the other (that is, the label is constant on each set), then perfect knowledge of the decision sets reduces the problem to a hypothesis testing problem for which $\epsilon_2(n) = O(e^{-Cn})$, for some constant $C > 0$. However, if γ is small relative to the average spacing $n^{-1/d}$ between labeled data points, then $\epsilon_1(n) = c\, n^{-1/d}$ where $c > 0$ is a constant. This is because in this case the decision set boundaries can only be localized to an accuracy of $n^{-1/d}$, the average spacing between labeled data points. Since the boundaries are Lipschitz, the expected volume that is incorrectly assigned to any decision set is greater than $c\, n^{-1/d}$, where $c > 0$ is a constant. This implies that the overall expected excess risk is greater than $c\, n^{-1/d}$. A formal proof for the lower bound can be derived along the lines of the minimax lower bound proof for regression in the next section. Thus, an exponential improvement is possible using semi-supervised learning provided the number of unlabeled data examples m grows exponentially in n, the number of labeled data examples. In other words, to obtain the same performance bound as a supervised learner with n labeled examples, a semi-supervised learner only needs $n' \equiv \log n$ labeled examples in the binary classification setting, and the number m of unlabeled examples needed is exponential in n', that is, polynomial in n.

5 Density-adaptive Regression

Let Y denote a continuous and bounded random variable. Under squared error loss, the optimal decision rule $f^*(x) = \mathbb{E}[Y|X = x]$, and the excess risk $\mathcal{E}(f) = \mathbb{E}[(f(X) - f^*(X))^2]$. Recall that $p_k(Y|X = x)$ is the conditional density on the k-th component and let $\mathbb{E}_k$ denote expectation with respect to the corresponding conditional distribution. The regression function on each component is $f_k(x) = \mathbb{E}_k[Y|X = x]$ and we assume that for $k = 1, \ldots, K$

1. f_k is uniformly bounded, $|f_k| \leq M$.
2. f_k is Hölder-α_2 smooth on C_k with Hölder constant κ_2.

This implies that the overall regression function $f^*(x)$, given as

$$f^*(x) = \sum_{k=1}^{K} \frac{a_k p_k(x)}{\sum_{j=1}^{K} a_j p_j(x)} f_k(x),$$

Margin range γ	SSL upper bound $\epsilon_2(n)$	SL lower bound $\epsilon_1(n)$	SSL helps
$\gamma \geq \gamma_0$	$n^{-2\alpha/(2\alpha+d)}$	$n^{-2\alpha/(2\alpha+d)}$	No
$\gamma \geq c_o n^{-1/d}$	$n^{-2\alpha/(2\alpha+d)}$	$n^{-2\alpha/(2\alpha+d)}$	No
$c_o n^{-1/d} > \gamma \geq C_o(\frac{m}{(\log m)^2})^{-1/d}$	$n^{-2\alpha/(2\alpha+d)}$	$n^{-1/d}$	Yes
$C_o(\frac{m}{(\log m)^2})^{-1/d} > \gamma \geq -C_o(\frac{m}{(\log m)^2})^{-1/d}$	$n^{-1/d}$	$n^{-1/d}$	No
$-C_o(\frac{m}{(\log m)^2})^{-1/d} > \gamma$	$n^{-2\alpha/(2\alpha+d)}$	$n^{-1/d}$	Yes
$-\gamma_0 > \gamma$	$n^{-2\alpha/(2\alpha+d)}$	$n^{-1/d}$	Yes

Table 1: Comparison of finite sample lower bounds on the mean square error for supervised learning, with finite sample upper bounds on the mean square error for semi-supervised learning, for the margin based model distributions. These bounds hold for $m \gg n^{2d}$ and $d \geq 2\alpha/(2\alpha-1)$, and suppress constants and log factors.

is piecewise Hölder-α smooth, where $\alpha = \min(\alpha_1, \alpha_2)$. That is, f^* is Hölder-α smooth on each $D \in \mathcal{D}$, except possibly at the decision boundaries. Since a Hölder-α smooth function can be locally well-approximated by a Taylor polynomial, we propose the following semi-supervised learner that performs local polynomial fits within each empirical decision set, that is, using labeled training data that are p-connected as per the definition in Section 3. While a spatially uniform estimator suffices to estimate a Hölder-α smooth function, we use the following spatially adaptive estimator proposed in Section 4.1 of [17] which is shown to yield minimax optimal performance for piecewise-smooth functions. This ensures that when the decision sets are indiscernible using unlabeled data, the semi-supervised learner still achieves an error bound that is, up to logarithmic factors, no worse than the minimax lower bound for supervised learners.

$$\widehat{f}_{m,n,x}(\cdot) = \arg\min_{f' \in \Gamma} \sum_{i=1}^{n} (Y_i - f'(X_i))^2 \mathbf{1}_{x \overset{p}{\leftrightarrow} X_i} + \text{pen}(f')$$

and

$$\widehat{f}_{m,n}(x) \equiv \widehat{f}_{m,n,x}(x).$$

Here Γ denotes a collection of piecewise polynomials with quantized coefficients of degree $[\alpha]$ (the maximal integer $< \alpha$), defined over recursive dyadic partitions of the domain $\mathcal{X} = [0,1]^d$ with cells of sidelength between $2^{-\lceil \log(n/\log n)/(2\alpha+d) \rceil}$ and $2^{-\lceil \log(n/\log n)/d \rceil}$ (see [17] for details). The penalty term pen(f') is proportional to $\log(\sum_{i=1}^n \mathbf{1}_{x \overset{p}{\leftrightarrow} X_i}) \cdot \#f'$, where $\sum_{i=1}^n \mathbf{1}_{x \overset{p}{\leftrightarrow} X_i}$ simply denotes the number of labeled training data that are p-connected to x, that is are in the same empirical decision set as x, and $\#f'$ denotes the number of cells in the recursive dyadic partition on which f' is defined. It is shown in [17] that, under the Hölder-α assumption, this estimator obeys a finite sample error bound of $n^{-2\alpha/(2\alpha+d)}$, ignoring a logarithmic factor. Also, it is shown that for piecewise Hölder-α smooth functions, this estimator yields a finite sample error bound of $\max(n^{-2\alpha/(2\alpha+d)}, n^{-1/d})$, ignoring a logarithmic factor.

Using these results from [17] and Corollary 1, in Section 7.3.1, we derive finite sample upper bounds on the mean square excess risk of the semi-supervised learner (SSL) described above. Also, we derive finite sample minimax lower bounds on the performance of any supervised learner (SL) based on n labeled examples in Section 7.3.2. Our results are summarized in Table 1, for model distributions characterized by various values of the margin parameter γ. In the table, we suppress constants and log factors in the error bounds, and assume that $m \gg n^{2d}$ so that the performance bound on the semi-supervised learner given in Corollary 1 essentially scales as $\epsilon_2(n)$. The constants c_o and C_o characterizing the margin only depend on the fixed parameters of the class $\mathcal{P}_{XY}(\gamma)$. Also, γ_0 denotes a constant, and thus the cases $\gamma \geq \gamma_0$ and $-\gamma_0 > \gamma$ correspond to considering a fixed collection of distributions (whose complexity does not change with the amount of data).

Consider the case when the dimension is large or the target function is smooth enough so that $d \geq 2\alpha/(2\alpha-1)$. If $d < 2\alpha/(2\alpha-1)$, then the supervised learning error incurred by averaging across decision sets (which behaves like $n^{-1/d}$) is smaller than error incurred in estimating the target function away from the boundaries (which behaves like $n^{-2\alpha/(2\alpha+d)}$). Thus, when $d < 2\alpha/(2\alpha -$

$1)$, learning the decision sets does not simplify the supervised learning task, and there appears to be no benefit to using a semi-supervised learner. So focusing on the case when $d \geq 2\alpha/(2\alpha - 1)$, the results of Table 1 state that if the margin γ is large relative to the average spacing between labeled data points $n^{-1/d}$, then a supervised learner can discern the decision sets accurately and SSL provides no gain. When $\gamma \geq \gamma_0$, we consider a fixed collection of distributions, and this argument is similar in spirit to the argument made by Lafferty and Wasserman [5]. However, if $\gamma > 0$ is small relative to $n^{-1/d}$, but large with respect to the spacing between unlabeled data points $m^{-1/d}$, then the proposed semi-supervised learner provides improved error bounds compared to *any* supervised learner. This is similar in spirit to the argument made by Niyogi [7] that the true underlying distribution can be more complex than can be discerned using labeled data. If $|\gamma|$ is smaller than $m^{-1/d}$, the decision sets are not discernable even with unlabeled data and SSL provides no gain. However, notice that the performance of the semi-supervised learner is no worse than the minimax lower bound for supervised learners since we chose an estimator that is also optimal for piecewise smooth functions (recall that the overall target function is piecewise smooth). In the $\gamma < 0$ case, when the component support sets can overlap, if the magnitude of the margin $|\gamma|$ larger than $m^{-1/d}$, then the semi-supervised learner can discern the decision sets and achieves smaller error bounds ($n^{-2\alpha/(2\alpha+d)}$), whereas these sets cannot be as accurately discerned by any supervised learner. For the overlap case ($\gamma < 0$), the supervised learners are always limited by the error incurred due to not resolving the decision sets $(n^{-1/d})$. In particular, for the fixed collection of distributions with $\gamma < -\gamma_0$, a faster rate of error convergence is attained by SSL compared to SL, provided $m \gg n^{2d}$.

6 Concluding Remarks

In this paper, we develop a framework for evaluating the performance gains possible with semi-supervised learning under a cluster assumption using finite sample error bounds. The theoretical characterization we present explains why in certain situations unlabeled data can help to improve learning, while in other situations they may not. We demonstrate that there exist general situations under which semi-supervised learning can be significantly superior to supervised learning, in terms of achieving smaller finite sample error bounds than any supervised learner, and sometimes in terms of a better rate of error convergence. Moreover, our results also provide a quantification of the relative value of unlabeled to labeled data.

While we focus on the cluster assumption in this paper, we conjecture that similar techniques can be applied to quantify the performance of semi-supervised learning under the manifold assumption as well. In the manifold case, the curvature and how close the manifold can get to itself or another manifold will play the role that the margin plays under the cluster assumption. In particular, we believe that the use of minimax lower bounding techniques is essential because many of the interesting distinctions between supervised and semi-supervised learning occur only in finite sample regimes, and rates of convergence and asymptotic analysis may not capture the complete picture.

In this work, we also show that though semi-supervised learning simplifies the learning task when the link relating the marginal and conditional distributions holds, it is possible to ensure that the performance of the semi-supervised learning does not deteriorate when the link is not discernable using unlabeled data or does not hold. For example, when the margin is small relative to the spacing between unlabeled data, the decision sets cannot be identified using unlabeled data, however by employing a more sophisticated tool (a learner that has optimal performance for piecewise smooth functions) similar to what a supervised learning algorithm would use, we ensured that the SSL performance is no worse than what a supervised learner would achieve. In this sense, the semi-supervised learner we propose is somewhat agnostic. However, if the number of decision sets can grow with n, the semi-supervised learning algorithm can perform worse because it would break the problem into a large collection of subproblems. Thus, it is of interest to develop an agnostic procedure that can identify such situations.

7 Proofs

Since the component densities are bounded from below and above, define $p_{\min} := b \min_k a_k \leq p(x) \leq B =: p_{\max}$.

7.1 Proof of Lemma 1

We present the proof in two steps - first, we establish some results about the proposed kernel density estimator, and then using the density estimation results, we establish that the decision sets $\mathcal{D}$ can be learnt based only on unlabeled data.

1) Density estimation:

Theorem 1. ***[Sup-norm density estimation of non-boundary points]*** *Consider the kernel density estimator proposed in Section 3 $\widehat{p}(x) = \frac{1}{mh_m^d}\sum_{i=1}^m G(H_m^{-1}(X_i - \bar{x}))$, where $H_m = h_m\mathbf{I}$, $h_m = \kappa_0((\log m)^2/m)^{1/d}$, $\kappa_0 > 0$ is a constant, and $\bar{x}$ denotes the point closest to x on a uniform grid over the domain $\mathcal{X} = [0,1]^d$ with spacing $2h_m$. Let the kernel G satisfy*

$$\textit{supp}(G) = [-1,1]^d, G \in (0, G_{\max}] \textit{ and } \int_{[-1,1]^d} u^j G(u)du = \begin{cases} 1 & j = 0 \\ 0 & 1 \le j \le [\alpha_1] \end{cases},$$

where supp$(\cdot)$ denotes the support of a function, then for all $p \in \mathcal{P}_X$, with probability at least $1 - 1/m$,

$$\sup_{x \in \textit{supp}(p)\setminus\mathcal{R}_\mathcal{B}} |p(x) - \widehat{p}(x)| \le c_3\left(h_m^{\min(1,\alpha_1)} + \sqrt{\frac{\log m}{mh_m^d}}\right) =: \epsilon_m,$$

for $m \ge m_1 \equiv m_1(G,B)$, where $c_3 \equiv c_3(K, \kappa_1, d, \alpha_1, B, G) > 0$ is a constant. Notice that ϵ_m decreases with increasing m.

Proof. Consider any $p \in \mathcal{P}_X$. Since $\widehat{p}(x) = \widehat{p}(\bar{x})$,

$$\sup_{x \in \text{supp}(p)\setminus\mathcal{R}_\mathcal{B}} |p(x) - \widehat{p}(x)| \le \sup_{x \in \text{supp}(p)\setminus\mathcal{R}_\mathcal{B}} |p(x) - p(\bar{x})| + \sup_{\bar{x}: x \in \text{supp}(p)\setminus\mathcal{R}_\mathcal{B}} |p(\bar{x}) - \widehat{p}(\bar{x})| \quad (1)$$

To bound the first term of (1), observe that since $x \in \text{supp}(p) \setminus \mathcal{R}_\mathcal{B}$ and $\|x - \bar{x}\| \le \sqrt{d}h_m$, by definition of $\mathcal{R}_\mathcal{B}$ if $x \in C_k$ then $\bar{x} \in C_k$ and vice versa. Thus, for all $x \in \text{supp}(p) \setminus \mathcal{R}_\mathcal{B}$,

$$\begin{aligned} |p(x) - p(\bar{x})| = \left|\sum_{k=1}^K a_k p_k(x) - a_k p_k(\bar{x})\right| &\le \sum_{k=1}^K a_k \left|p_k(x) - p_k(\bar{x})\right| \\ &= \sum_{k: x,\bar{x} \in C_k} a_k |p_k(x) - p_k(\bar{x})| \\ &\le \sum_{k: x,\bar{x} \in C_k} a_k \left(\kappa_1(\sqrt{d}h_m)^{\alpha_1} + \left|\sum_{j=1}^{[\alpha_1]} \frac{p_k^{(j)}(x)}{j!}(\bar{x} - x)^j\right|\right) \\ &\le c_1\, h_m^{\min(1,\alpha_1)}, \end{aligned}$$

where $c_1 \equiv c_1(K, \kappa_1, d, \alpha_1, B) > 0$ is a constant. The last step follows since if p_k is Hölder-α_1 smooth, then all its derivatives up to $[\alpha_1]$ are bounded and $\|x - \bar{x}\| \le \sqrt{d}h_m$.

To bound the second term, notice that for all $\bar{x} : x \in \text{supp}(p) \setminus \mathcal{R}_\mathcal{B}$,

$$|p(\bar{x}) - \widehat{p}(\bar{x})| = |p(\bar{x}) - \mathbb{E}[\widehat{p}(\bar{x})]| + |\mathbb{E}[\widehat{p}(\bar{x})] - \widehat{p}(\bar{x})|$$

We now bound the two terms in the last expression.

1. For all $\bar{x} : x \in \text{supp}(p) \setminus \mathcal{R}_\mathcal{B}$, consider

$$|p(\bar{x}) - \mathbb{E}[\widehat{p}(\bar{x})]| = \left|p(\bar{x}) - \frac{1}{h_m^d}\int_{\bar{x}-h_m}^{\bar{x}+h_m} G(H_m^{-1}(y - \bar{x}))p(y)dy\right|$$

Notice that given the conditions on the kernel,

$$\begin{aligned} p(\bar{x}) &= \int_{[-1,1]^d} \sum_{j=0}^{[\alpha_1]} \frac{p^{(j)}(\bar{x})}{j!}(h_m u)^j G(u)du \\ &= \frac{1}{h_m^d}\int_{\bar{x}-h_m}^{\bar{x}+h_m} G(H_m^{-1}(y - \bar{x})) \sum_{j=0}^{[\alpha_1]} \frac{p^{(j)}(\bar{x})}{j!}(y - \bar{x})^j dy \end{aligned}$$

Therefore, we get

$$
\begin{aligned}
&|p(\bar{x}) - \mathbb{E}[\widehat{p}(\bar{x})]| \\
&= \left| \frac{1}{h_m^d} \int_{\bar{x}-h_m}^{\bar{x}+h_m} G(H_m^{-1}(y-\bar{x})) \left(\sum_{j=0}^{[\alpha_1]} \frac{p^{(j)}(\bar{x})}{j!}(y-\bar{x})^j dy - p(y) \right) dy \right| \\
&= \left| \frac{1}{h_m^d} \int_{\bar{x}-h_m}^{\bar{x}+h_m} G(H_m^{-1}(y-\bar{x})) \sum_{k=1}^{K} a_k \left(\sum_{j=0}^{[\alpha_1]} \frac{p_k^{(j)}(\bar{x})}{j!}(y-\bar{x})^j dy - p_k(y) \right) dy \right| \\
&\leq \left| \frac{1}{h_m^d} \int_{\bar{x}-h_m}^{\bar{x}+h_m} G(H_m^{-1}(y-\bar{x})) \kappa_1 \|y-\bar{x}\|^{\alpha_1} dy \right| \\
&\leq \kappa_1 \left(\int_{[-1,1]^d} \|u\|^{\alpha_1} G(u) du \right) h_m^{\alpha_1} = c_2 \, h_m^{\alpha_1},
\end{aligned}
$$

where $c_2 \equiv c_2(\kappa_1, G, \alpha_1) > 0$ is a constant.

2. Now consider

$$
\begin{aligned}
P\left(\sup_{\bar{x}: x \in \text{supp}(p) \setminus \mathcal{R}_{\mathcal{B}}} |\mathbb{E}[\widehat{p}(\bar{x})] - \widehat{p}(\bar{x})| > \epsilon \right) &\leq \sum_{\bar{x}} P\left(|\mathbb{E}[\widehat{p}(\bar{x})] - \widehat{p}(\bar{x})| > \epsilon \right) \\
&= \sum_{\bar{x}} P\left(\left| \sum_{i=1}^{m} \mathbb{E}[Z_i] - Z_i \right| > m h_m^d \epsilon \right) \\
&\leq \sum_{\bar{x}} P\left(\sum_{i=1}^{m} |\mathbb{E}[Z_i] - Z_i| > m h_m^d \epsilon \right)
\end{aligned}
$$

where $Z_i = G(H_m^{-1}(X_i - \bar{x}))$. Now observe that $|\mathbb{E}[Z_i] - Z_i| \leq G_{\max}$ and

$$
\begin{aligned}
\text{var}(Z_i) \leq E[Z_i^2] &= \int_{\bar{x}-h_m}^{\bar{x}+h_m} G^2(H_m^{-1}(y-\bar{x})) p(y) dy \\
&= h_m^d \int_{[-1,1]^d} G^2(u) p(\bar{x} + H_m u) du \\
&= h_m^d \int_{[-1,1]^d} G^2(u)(p(\bar{x}) + o(1)) du \\
&\leq 2\|G\|_2^2 p(\bar{x}) h_m^d \leq 2\|G\|_2^2 B h_m^d
\end{aligned}
$$

Thus, using Bernstein's inequality, we get:

$$
P\left(\sum_{i=1}^{m} |\mathbb{E}[Z_i] - Z_i| > m h_m^d \epsilon \right) \leq \exp\left\{ -\frac{(m h_m^d \epsilon)^2/2}{2\|G\|_2^2 B m h_m^d + G_{\max} m h_m^d \epsilon / 3} \right\}
$$

Setting $\epsilon = 4\|G\|_2 \sqrt{B} \sqrt{\frac{\log m}{m h_m^d}}$, and observing that $G_{\max}\epsilon/3 \leq 2\|G\|_2^2 B$ for large enough $m \geq m_1 \equiv m_1(G, B)$, we get:

$$
\begin{aligned}
&P\left(\sup_{\bar{x}: x \in \text{supp}(p) \setminus \mathcal{R}_{\mathcal{B}}} |\mathbb{E}[\widehat{p}(\bar{x})] - \widehat{p}(\bar{x})| > 4\|G\|_2 \sqrt{B} \sqrt{\frac{\log m}{m h_m^d}} \right) \\
&\qquad \leq \sum_{\bar{x}} \exp\left\{ -\frac{16\|G\|_2^2 B m h_m^d \log m / 2}{4\|G\|_2^2 B m h_m^d} \right\} \\
&\qquad \leq h_m^{-d} \exp\{-2 \log m\} \\
&\qquad \leq m \cdot \frac{1}{m^2} = \frac{1}{m}
\end{aligned}
$$

Therefore we get, with probability at least $1 - 1/m$, for $m \geq m_1(G, B)$ we have the following bound on the second term

$$\sup_{\bar{x}: x \in \text{supp}(p) \setminus \mathcal{R}_\mathcal{B}} |p(\bar{x}) - \widehat{p}(\bar{x})| \leq c_2 h_m^{\alpha_1} + 4\|G\|_2 \sqrt{B} \sqrt{\frac{\log m}{m h_m^d}}.$$

And putting the bounds on the two terms together: For all $p \in \mathcal{P}_X$, with probability at least $1 - 1/m$, for $m \geq m_1(G, B)$

$$\sup_{x \in \text{supp}(p) \setminus \mathcal{R}_\mathcal{B}} |p(x) - \widehat{p}(x)| \leq c_3 \left(h_m^{\min(1,\alpha_1)} + \sqrt{\frac{\log m}{m h_m^d}} \right),$$

where $c_3 \equiv c_3(K, \kappa_1, d, \alpha_1, B, G) > 0$ is a constant. □

Remark: This bound can be tightened to $O(h_m^{\alpha_1} + \sqrt{\log m / m h_m^d})$ by also estimating the density derivatives at the grid points and defining $p(x)$ as the Taylor polynomial approximation expanded around the closest grid point $\bar{x}$, see [13]. Also, the arguments of the proof hold if $h_m = \kappa_0 (\log m / m)^{-1/(d+2\alpha_1)}$. Hence, we recover the minimax rate of $O((m/\log m)^{-\alpha_1/(d+2\alpha_1)})$ for sup-norm density estimation of a Hölder-α_1 smooth density. However, we want to characterize the largest collection of distributions (smallest margin) that a semi-supervised learner can handle, and thus we seek the smallest h_m (which determines the smallest margin that can be handled) for which the bound ϵ_m decreases with increasing m.

Corollary 2. ***[Empirical density of unlabeled data]*** *Under the conditions of Theorem 1, for all $p \in \mathcal{P}_X$ and $m \geq m_3 \equiv m_3(p_{\min}, K, \kappa_1, d, \alpha_1, B, G, \kappa_0)$, with probability at least $1 - 1/m$, for all $x \in$ supp$(p) \setminus \mathcal{R}_\mathcal{B}$, there exists an unlabeled data point $X_i \in \mathcal{U}$ such that $\|X_i - x\| \leq \sqrt{d} h_m$.*

Proof. From Theorem 1, for all $x \in \text{supp}(p) \setminus \mathcal{R}_\mathcal{B}$, for $m \geq m_1(G, B)$

$$\widehat{p}(x) \geq p(x) - \epsilon_m \geq p_{\min} - \epsilon_m > 0$$

The last step follows for large enough $m \geq m_2 \equiv m_2(p_{\min}, K, \kappa_1, d, \alpha_1, B, G, \kappa_0)$ since ϵ_m is decreasing with m. This implies that $\sum_{i=1}^m G(H_m^{-1}(X_i - x)) > 0$ for $m \geq m_3 = \max(m_1, m_2)$, and therefore there exists an unlabeled data point within $\sqrt{d} h_m$ of x. □

2) Decision set estimation - Using the density estimation results, we now show that if $|\gamma| > 6\sqrt{d} h_m$, then for all $p \in \mathcal{P}_X$, all pairs of points $x_1, x_2 \in \text{supp}(p) \setminus \mathcal{R}_\mathcal{B}$ and all $D \in \mathcal{D}$, for $m \geq m_0 \equiv m_0(p_{\min}, K, \kappa_1, d, \alpha_1, B, G, \kappa_0)$ with probability $> 1 - 1/m$, we have $x_1 \overset{p}{\leftrightarrow} x_2$ if and only if $x_1, x_2 \in D$. We establish this in two steps:

1. $x_1 \in D, x_2 \notin D \Rightarrow x_1 \overset{p}{\not\leftrightarrow} x_2$:
   Since x_1 and x_2 belong to different decision sets and $x_1, x_2 \in \text{supp}(p) \setminus \mathcal{R}_\mathcal{B}$, all sequences connecting x_1 and x_2 through unlabeled data points pass through a region where either (i) the density is zero, or (ii) the density is positive. In case (i), there cannot exist a sequence connecting x_1 and x_2 through unlabeled data points such that for any two consecutive points z_j, z_{j+1} along the sequence $\|z_j - z_{j+1}\| \leq 2\sqrt{d} h_m$ since the region of zero density is at least $|\gamma| > 6\sqrt{d} h_m$ wide. Therefore, $x_1 \not\leftrightarrow x_2$, and hence $x_1 \overset{p}{\not\leftrightarrow} x_2$. In case (ii), since x_1 and x_2 belong to different decision sets, the marginal density $p(x)$ jumps by at least $p_{\min}$ one or more times along all sequences connecting x_1 and x_2. Suppose the first jump (in the sequence) occurs where decision set D ends and another decision set $D' \neq D$ begins. Then since D, D' are at least $|\gamma| > 6\sqrt{d} h_m$ wide, by Corollary 2 with probability $> 1 - 1/m$ for $m \geq m_3$, for all sequences connecting x_1 and x_2 through unlabeled data points, there exist points z, z' in the sequence that lie in $D \setminus \mathcal{R}_\mathcal{B}$, $D' \setminus \mathcal{R}_\mathcal{B}$, respectively, and $\|z - z'\| \leq h_m \log m$. We will show that $|p(z) - p(z')| \geq p_{\min} - O((h_m \log m)^{\min(1,\alpha_1)})$ which using Theorem 1 implies that $|\widehat{p}(z) - \widehat{p}(z')| \geq p_{\min} - O((h_m \log m)^{\min(1,\alpha_1)}) - 2\epsilon_m > \delta_m$ for m large enough. Hence $x_1 \overset{p}{\not\leftrightarrow} x_2$.

To see these claims, observe that since D' and D are adjacent decision sets, if $D = \cap_{k=1}^{K} d_k$ where $d_k \in \{C_k, C_k^c\}$ and $D' = \cap_{k=1}^{K} d'_k$, then $\exists k_0$ such that $d_k = d'_k$ for all $k \neq k_0$. Thus, $\{k : z \in C_k \text{ or } z' \in C_k\} = k_0$. Since $\|z - z'\| \leq h_m \log m$, we get:

$$\begin{aligned}
|p(z) - p(z')| &= \left| \sum_{k=1}^{K} a_k p_k(z) - \sum_{k=1}^{K} a_k p_k(z') \right| \\
&= \left| \sum_{k: z \in C_k \text{ or } z' \in C_k} a_k (p_k(z) - p_k(z')) + \sum_{k: z, z' \in C_k} a_k (p_k(z) - p_k(z')) \right| \\
&\geq |a_{k_0}(p_{k_0}(z) - p_{k_0}(z'))| - \left| \sum_{k: z, z' \in C_k} a_k (p_k(z) - p_k(z')) \right| \\
&\geq ab - \left| \sum_{k: z, z' \in C_k} a_k (p_k(z) - p_k(z')) \right| \\
&\geq p_{\min} - c_4 (h_m \log m)^{\min(1, \alpha_1)},
\end{aligned}$$

where $c_4 > 0$ is a constant. The fourth step follows since $d_{k_0} \neq d'_{k_0}$ and hence either $p_{k_0}(z)$ is zero or $p_{k_0}(z')$ is zero, and since p_{k_0} is bounded from below by b and $a_k \geq a$. To see the last step, recall that the component densities p_k are Hölder-α_1 smooth and $\|z' - z\| \leq h_m \log m$. Thus, we have:

$$\begin{aligned}
\left| \sum_{k: z, z' \in C_k} a_k (p_k(z) - p_k(z')) \right| &\leq \sum_{k: z, z' \in C_k} a_k |p_k(z) - p_k(z')| \\
&\leq \sum_{k: z, z' \in C_k} a_k \left(\kappa_1 (h_m \log m)^{\alpha_1} + \left| \sum_{j=0}^{[\alpha_1]} \frac{p_k^{(j)}(z)}{j!} (z' - z)^j \right| \right) \\
&\leq c_4 (h_m \log m)^{\min(1, \alpha_1)},
\end{aligned}$$

where $c_4 \equiv c_4(K, \kappa_1, \alpha_1, B) > 0$ is a constant. Here the last step follows since if p_k is Hölder-α_1 smooth, then all its derivatives up to $[\alpha_1]$ are bounded.

Now since $z, z' \in \text{supp}(p) \setminus \mathcal{R}_{\mathcal{B}}$, using Theorem 1, we get with probability $> 1 - 1/m$, for $m \geq \max(m_1, m_3)$

$$\begin{aligned}
|\widehat{p}(z) - \widehat{p}(z')| &= |\widehat{p}(z) - p(z) + p(z) - p(z') + p(z') - \widehat{p}(z')| \\
&\geq |p(z) - p(z')| - |\widehat{p}(z) - p(z)| - |p(z') - \widehat{p}(z')| \\
&\geq p_{\min} - c_4 (h_m \log m)^{\min(1, \alpha_1)} - 2\epsilon_m \\
&> \frac{1}{(\log m)^{1/3}} = \delta_m.
\end{aligned}$$

The last step holds for large enough $m \geq m_4 \equiv m_4(p_{\min}, K, \kappa_1, d, \alpha_1, B, G, \kappa_0)$. Thus, for case (ii) we have shown that,for $m \geq \max(m_1, m_3, m_4)$ with probability $> 1 - 1/m$, for all sequences connecting x_1 and x_2 through $2\sqrt{d}h_m$-dense unlabeled data points, there exists points z, z' in the sequence such that $\|z - z'\| \leq h_m \log m$ but $|\widehat{p}(z) - \widehat{p}(z')| > \delta_m$. Thus,

$$x_1 \in D, x_2 \notin D \Rightarrow x_1 \overset{p}{\not\leftrightarrow} x_2.$$

2. $x_1, x_2 \in D \Rightarrow x_1 \overset{p}{\leftrightarrow} x_2$:
Since D has width at least $|\gamma| > 6\sqrt{d}h_m$, there exists a set of width $> 2\sqrt{d}h_m$ contained in $D \setminus \mathcal{R}_{\mathcal{B}}$, and Corollary 2 implies that for $m \geq m_3$, with probability $> 1 - 1/m$, there exist sequence(s) contained in $D \setminus \mathcal{R}_{\mathcal{B}}$ connecting x_1 and x_2 through $2\sqrt{d}h_m$-dense unlabeled data points. Since the sequence is contained in $D \setminus \mathcal{R}_{\mathcal{B}}$, and the density on D is Hölder-α_1

smooth, we have for all points z, z' in the sequence such that $\|z - z'\| \leq h_m \log m$,

$$\begin{aligned}
|\widehat{p}(z) - \widehat{p}(z')| &= |\widehat{p}(z) - p(z) + p(z) - p(z') + p(z') - \widehat{p}(z')| \\
&\leq |\widehat{p}(z) - p(z)| + |p(z) - p(z')| + |p(z') - \widehat{p}(z')| \\
&\leq 2\epsilon_m + |p(z) - p(z')| \\
&\leq 2\epsilon_m + c_5(h_m \log m)^{\min(1,\alpha_1)} \\
&\leq \frac{1}{(\log m)^{1/3}} = \delta_m,
\end{aligned}$$

where $c_5 > 0$ is a constant, and the last step holds for large enough $m \geq m_5 \equiv m_5(K, \kappa_1, d, \alpha_1, B, G, \kappa_0)$. The third step follow since $z, z' \in \text{supp}(p) \setminus \mathcal{R}_\mathcal{B}$, and invoking Theorem 1. To see the fourth step, since $z, z' \in D$, if $z \in C_k$ then $z' \in C_k$ and vice versa. Thus,

$$\begin{aligned}
|p(z) - p(z')| &= \left| \sum_{k: z, z' \in C_k} a_k (p_k(z) - p_k(z')) \right| \leq \sum_{k: z, z' \in C_k} a_k |p_k(z) - p_k(z')| \\
&\leq \sum_{k: z, z' \in C_k} a_k \left(\kappa_1 (h_m \log m)^{\alpha_1} + \left| \sum_{j=0}^{[\alpha_1]} \frac{p_k^{(j)}(z)}{j!} (z' - z)^j \right| \right) \\
&\leq c_5 (h_m \log m)^{\min(1,\alpha_1)},
\end{aligned}$$

where $c_5 \equiv c_5(\kappa_1, K, B, \alpha_1) > 0$ is a constant. Here the third step follows since $\|z' - z\| \leq h_m \log m$, and p_k is Hölder-α_1 on C_k. The last step follows since if p_k is Hölder-α_1 smooth, then all its derivatives up to $[\alpha_1]$ are bounded. Thus, we have shown that

$$x_1, x_2 \in D \Rightarrow x_1 \overset{p}{\leftrightarrow} x_2.$$

Thus, the result of the Lemma holds for $m \geq m_0 = \max(m_1, m_3, m_4, m_5)$, where $m_0 \equiv m_0(p_{\min}, K, \kappa_1, d, \alpha_1, B, G, \kappa_0)$ is a constant.

■

7.2 Proof of Corollary 1

Let Ω_1 denote the event under which Lemma 1 holds. Then $P(\Omega_1^c) \leq 1/m$, where Ω^c denotes the complement of Ω. Let Ω_2 denote the event that the test point X and training data $X_1, \ldots, X_n \in \mathcal{L}$ don't lie in $\mathcal{R}_\mathcal{B}$. Then

$$P(\Omega_2^c) \leq (n+1) P(\mathcal{R}_\mathcal{B}) \leq (n+1) p_{\max} \text{vol}(\mathcal{R}_\mathcal{B}) = O(n h_m).$$

The last step can be explained as follows. Since the decision boundaries are Lipschitz and K is finite, the length of the decision boundaries is a finite constant, and hence $\text{vol}(\mathcal{R}_\mathcal{B}) = O(h_m)$.

Now observe that $\widehat{f}_{\mathcal{D},n}$ essentially uses the clairvoyant knowledge of the decision sets $\mathcal{D}$ to discern which labeled points $X_1, \ldots, X_n$ are in the same decision set as X. Conditioning on Ω_1, Ω_2, Lemma 1 implies that $X, X_i \in D$ if and only if $X \overset{p}{\leftrightarrow} X_i$ for all $i = 1, \ldots, n$. Thus, we can define a semi-supervised learner $\widehat{f}_{m,n}$ to be the same as $\widehat{f}_{\mathcal{D},n}$ except that instead of using clairvoyant knowledge of whether $X, X_i \in D$, $\widehat{f}_{m,n}$ is based on whether $X \overset{p}{\leftrightarrow} X_i$. It follows that $\sup_{\mathcal{P}_{XY}(\gamma)} \mathbb{E}[\mathcal{E}(\widehat{f}_{m,n}) | \Omega_1, \Omega_2] = \sup_{\mathcal{P}_{XY}(\gamma)} \mathbb{E}[\mathcal{E}(\widehat{f}_{\mathcal{D},n})]$, and since the excess risk is bounded,

$$\begin{aligned}
\sup_{\mathcal{P}_{XY}(\gamma)} \mathbb{E}[\mathcal{E}(\widehat{f}_{m,n})] &= \sup_{\mathcal{P}_{XY}(\gamma)} \mathbb{E}[\mathcal{E}(\widehat{f}_{m,n}) | \Omega_1, \Omega_2] P(\Omega_1, \Omega_2) + \mathbb{E}[\mathcal{E}(\widehat{f}_{m,n}) | \Omega_1^c \cup \Omega_2^c] P(\Omega_1^c \cup \Omega_2^c) \\
&\leq \sup_{\mathcal{P}_{XY}(\gamma)} \mathbb{E}[\mathcal{E}(\widehat{f}_{\mathcal{D},n})] + O\left(\frac{1}{m} + n h_m \right) \\
&\leq \epsilon_2(n) + O\left(\frac{1}{m} + n \left(\frac{m}{(\log m)^2} \right)^{-1/d} \right).
\end{aligned}$$

■

7.3 Density Adaptive Regression Results

7.3.1 Semi-Supervised Learning Upper Bound

If the margin $|\gamma| > C_o(m/(\log m)^2)^{-1/d}$, where $C_o = 6\sqrt{d}\kappa_0$ and $m \gg n^{2d}$, we show that the semi-supervised learner proposed in Section 5 achieves a finite sample error bound of $O\left((n/\log n)^{-2\alpha/(d+2\alpha)}\right)$. Observe that the clarivoyant counterpart of $\widehat{f}_{m,n}(x)$ is given as

$$\widehat{f}_{\mathcal{D},n}(x) = \widehat{f}_{\mathcal{D},n,x}(x),$$

where

$$\widehat{f}_{\mathcal{D},n,,x}(\cdot) = \arg\min_{f' \in \Gamma} \sum_{i=1}^{n} (Y_i - f'(X_i))^2 \mathbf{1}_{x,X_i \in D} + \text{pen}(f').$$

Observe that $\widehat{f}_{\mathcal{D},n}$ is a standard supervised learner that performs piecewise polynomial fit on each decision set $D \in \mathcal{D}$, where the regression function is Hölder-α smooth. Let $n_D = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{X_i \in D}$ denote the number of labeled training examples that fall in a decision set $D \in \mathcal{D}$. Since the regression function on each decision set is Hölder-α smooth, it follows (for example, along the lines of Theorem 8 in [17]) that

$$\mathbb{E}[(f^*(X) - \widehat{f}_{\mathcal{D},n}(X))^2 \mathbf{1}_{X \in D} | n_D] \leq C \left(\frac{n_D}{\log n_D}\right)^{-\frac{2\alpha}{d+2\alpha}}.$$

Now consider

$$\mathbb{E}[(f^*(X) - \widehat{f}_{\mathcal{D},n}(X))^2] = \sum_{D \in \mathcal{D}} \mathbb{E}[(f^*(X) - \widehat{f}_{\mathcal{D},n}(X))^2 \mathbf{1}_{X \in D}] P(D).$$

We will establish the result by taking expectation over $n_D \sim$Binomial$(n, P(D))$ (if $P(D) = O(\log n/n)$, we simply use the fact that the excess risk is bounded), and summing over all decision sets recalling that $|\mathcal{D}|$ is a finite constant. Consider two cases:

1. If $P(D) > \frac{28 \log n}{3n}$,

$$\begin{aligned}
&\mathbb{E}[(f^*(X) - \widehat{f}_{\mathcal{D},n}(X))^2 \mathbf{1}_{X \in D}] P(D) \\
&= \mathbb{E}[\mathbb{E}[(f^*(X) - \widehat{f}_{\mathcal{D},n}(X))^2 \mathbf{1}_{X \in D} | n_D]] P(D) \\
&\leq \mathbb{E}\left[C \left(\frac{n_D}{\log n_D}\right)^{-\frac{2\alpha}{d+2\alpha}}\right] P(D) \\
&= \sum_{n_D=0}^{n} C \left(\frac{n_D}{\log n_D}\right)^{-\frac{2\alpha}{d+2\alpha}} P(n_D) P(D) \\
&\leq C \sum_{n_D=0}^{n} \left(\frac{n_D}{\log n}\right)^{-\frac{2\alpha}{d+2\alpha}} P(n_D) P(D) \\
&\leq \frac{C}{(\log n)^{-\frac{2\alpha}{d+2\alpha}}} \left[\sum_{n_D=0}^{\lceil nP(D)/2 \rceil - 1} n_D^{-\frac{2\alpha}{d+2\alpha}} P(n_D) + \sum_{n_D=\lceil nP(D)/2 \rceil}^{n} n_D^{-\frac{2\alpha}{d+2\alpha}} P(n_D)\right] P(D) \\
&\leq \frac{C}{(\log n)^{-\frac{2\alpha}{d+2\alpha}}} \left[P(n_D \leq nP(D)/2) + (nP(D)/2)^{-\frac{2\alpha}{d+2\alpha}}\right] P(D) \\
&\leq \frac{C}{(\log n)^{-\frac{2\alpha}{d+2\alpha}}} \left[e^{-\frac{3nP(D)}{28}} P(D) + 2n^{-\frac{2\alpha}{d+2\alpha}} P(D)^{\frac{d}{d+2\alpha}}\right] \\
&\leq \frac{C}{(\log n)^{-\frac{2\alpha}{d+2\alpha}}} \left[\frac{1}{n} + 2n^{-\frac{2\alpha}{d+2\alpha}}\right] \\
&= O\left(\left(\frac{n}{\log n}\right)^{-\frac{2\alpha}{d+2\alpha}}\right)
\end{aligned}$$

The second last step follows since

$$\begin{aligned} P(n_D \leq nP(D)/2) &= P(nP(D) - n_D \geq nP(D)/2) \\ &= P\left(\sum_{i=1}^{n} P(D) - \mathbf{1}_{X_i \in D} \geq nP(D)/2\right) \\ &= P\left(\sum_{i=1}^{n} Z_i \geq nP(D)/2\right) \\ &\leq \exp\left\{\frac{(nP(D)/2)^2/2}{nP(D)(1-P(D)) + nP(D)/6}\right\} \leq e^{-\frac{3nP(D)}{28}}. \end{aligned}$$

The last step follows using Bernstein's inequality since for $Z_i = P(D) - \mathbf{1}_{X_i \in D}$, we have that $|Z_i| \leq 1$ and var$(Z_i) = P(D)(1 - P(D))$.

2. If $P(D) \leq \frac{28 \log n}{3n}$, we have

$$\mathbb{E}[(f^*(X) - \widehat{f}_{\mathcal{D},n}(X))^2 \mathbf{1}_{X \in D}]P(D) \ \leq \ 4M^2 P(D) = O\left(\frac{\log n}{n}\right).$$

Thus, it follows that since $|\mathcal{D}| \leq 2^K$

$$\mathbb{E}[(f^*(X) - \widehat{f}_{\mathcal{D},n}(X))^2] = O\left(\left(\frac{n}{\log n}\right)^{-\frac{2\alpha}{d+2\alpha}}\right).$$

And using Corollary 1,

$$\mathbb{E}[(f^*(X) - \widehat{f}_{m,n}(X))^2] = O\left(\left(\frac{n}{\log n}\right)^{-\frac{2\alpha}{d+2\alpha}} + \frac{1}{m} + n\left(\frac{m}{(\log m)^2}\right)^{-1/d}\right).$$

If $m \gg n^{2d}$, then $1/m + n(m/(\log m)^2)^{-1/d} = O((n/\log n)^{-1})$ and we get an upper bound of $O\left((n/\log n)^{-\frac{2\alpha}{d+2\alpha}}\right)$ on the performance of the semi-supervised learner.

If $|\gamma| < C_o(m/(\log m)^2)^{-1/d}$, the decision sets are not discernable using unlabeled data and the target regression function is piecewise Hölder-α smooth on each p-connected set. As shown in [17], for piecewise Hölder-α functions, the proposed estimator achieves an error bound of $\max(n^{-2\alpha/(2\alpha+d)}, n^{-1/d})$. Also, notice that the number of resulting p-connected sets cannot be more than $|\mathcal{D}|$ since the procedure can miss detecting where the marginal density jumps, however with high probability it will not declare two points to be p-connected when the marginal density does not jump between them. Thus, the number of p-connected sets is also a finite constant. Using similar analysis as above, an overall error bound of $\max(n^{-2\alpha/(2\alpha+d)}, n^{-1/d})$ follows, which scales as $n^{-1/d}$ when $d \geq 2\alpha/(2\alpha - 1)$.

■

7.3.2 Supervised Learning Lower Bound

Consider the single cluster class $\mathcal{P}'_{XY}$ with supp$(p_X) = [0,1]^d$. For this class, it is known [18] that there exists a constant $c > 0$ such that

$$\inf_{f_n} \sup_{\mathcal{P}'_{XY}} \mathbb{E}[(f^*(X) - f_n(X))^2] \geq cn^{-2\alpha/(d+2\alpha)}.$$

Notice that $\mathcal{P}'_{XY} \subset \mathcal{P}_{XY}(\gamma)$ for all γ. Therefore, we get:

$$\inf_{f_n} \sup_{\mathcal{P}_{XY}(\gamma)} \mathbb{E}[(f^*(X) - f_n(X))^2] \geq cn^{-2\alpha/(d+2\alpha)}.$$

If $\gamma < c_o n^{-1/d}$, where $c_o > 0$ is a constant, we derive a tighter lower bound of $cn^{-1/d}$. Thus, we will have

$$\inf_{f_n} \sup_{\mathcal{P}_{XY}(\gamma)} \mathbb{E}[(f^*(X) - f_n(X))^2] \geq cn^{-1/d}.$$

To establish the tighter lower bound of $cn^{-1/d}$, we use the following theorem based on Assouad's lemma (adapted from Theorem 2.10 (iii) in [19]).

Theorem 2. *Let $\Omega = \{0,1\}^q$, the collection of binary vectors of length q. Let $\mathcal{P}_\Omega = \{P^\omega, \omega \in \Omega\}$ be the corresponding collection of 2^q probability measures associated with each vector. Also let $H(\cdot,\cdot)$ denote the Hellinger distance between two distributions, and $\rho(\cdot,\cdot)$ denotes the Hamming distance between two binary vectors. If $H^2(P^{\omega'}, P^\omega) \leq \kappa < 2, \forall \omega, \omega' \in \Omega : \rho(\omega,\omega') = 1$, then*

$$\inf_{\hat{\omega}} \max_{\omega \in \Omega} \mathbb{E}_\omega[\rho(\hat{\omega},\omega)] \geq \frac{q}{2}(1 - \sqrt{\kappa(1-\kappa/4)})$$

We will construct such a collection of joint probability distributions $\mathcal{P}_\Omega \subseteq \mathcal{P}_{XY}(\gamma)$ satisfying Theorem 2 with $q = \ell^{d-1}$, where $\ell = \lceil c_6 n^{1/d} \rceil$, $c_6 > 0$ is a constant. Notice that $\mathbb{E}[(f^*(X) - f_n(X))^2] = \mathbb{E}[R(f^*, f_n)]$, where $R(f^*, f_n)$ denotes the mean square error

$$R(f^*, f_n) = \int (f^*(x) - f_n(x))^2 p(x) dx.$$

Since the mean square error is not symmetric, we will first relate it to a semi-distance $d(\cdot,\cdot)$ defined as follows:

$$d^2(f, f_n) = \int (f^*(x) - f_n(x))^2 dx.$$

For $f^* \equiv f^\omega$ and $f_n \equiv f^{\hat{\omega}}$, we will show that the mean square error and semi-distance are related as follows:

$$R(f^\omega, f^{\hat{\omega}}) \geq b\,[d^2(f^\omega, f^{\hat{\omega}}) - 4M^2\gamma]. \tag{2}$$

We will then show the following lower bound on the semi-distance in terms of the Hamming distance:

$$d^2(f^\omega, f^{\hat{\omega}}) \geq c_7 \ell^{-d} \rho(\hat{\omega}, \omega) \tag{3}$$

where $c_7 > 0$ is a constant. Thus, we will have

$$\begin{aligned}
\inf_{f_n} \sup_{\mathcal{P}_{XY}(\gamma)} \mathbb{E}[(f^*(X) - f_n(X))^2] &= \inf_{f_n} \sup_{\mathcal{P}_{XY}(\gamma)} \mathbb{E}[R(f^*, f_n)] \\
&\geq \inf_{f^{\hat{\omega}}} \sup_{\mathcal{P}_\Omega} \mathbb{E}[R(f^\omega, f^{\hat{\omega}})] = \inf_{\hat{\omega}} \sup_{\omega \in \Omega} \mathbb{E}_\omega[R(f^\omega, f^{\hat{\omega}})] \\
&\geq b\left(\inf_{\hat{\omega}} \sup_{\omega \in \Omega} \mathbb{E}_\omega[d^2(f^\omega, f^{\hat{\omega}})] - 4M^2\gamma \right) \\
&\geq b\left(c_7 \ell^{-d} \inf_{\hat{\omega}} \sup_{\omega \in \Omega} \mathbb{E}_\omega[\rho(\omega, \hat{\omega})] - 4M^2\gamma \right) \\
&\geq b\left(c_7 \ell^{-d} \frac{q}{2}(1 - \sqrt{\kappa(1-\kappa/4)}) - 4M^2\gamma \right) \\
&\geq b\left(\frac{c_7}{2c_6}(1 - \sqrt{\kappa(1-\kappa/4)}) - 4M^2 c_o \right) n^{-1/d}
\end{aligned}$$

where the last step follows since $q = \ell^{d-1}$, $\ell = \lceil c_6 n^{1/d} \rceil$ and $\gamma < c_o n^{-1/d}$. Thus, there exists $c_o \equiv c_o(c_6, c_7, M, \kappa)$, for which we obtain the desired lower bound of $cn^{-1/d}$, where $c > 0$ is a constant.

We now construct $\mathcal{P}_\Omega \subseteq \mathcal{P}_{XY}(\gamma)$ along the lines of standard minimax construction that satisfies Theorem 2 with $q = \ell^{d-1}$, $\ell = \lceil c_6 n^{1/d} \rceil$, and Equations. (2) and (3). We construct the elements (p^ω, f^ω) of our collection as follows. Let $x = (\tilde{x}, x_d) \in [0,1]^d$, where $\tilde{x} \in [0,1]^{d-1}$ and $x_d \in [0,1]$. Define

$$\tilde{x}_{\tilde{j}} = \frac{\tilde{j} - 1/2}{\ell} \quad \text{and} \quad \eta_{\tilde{j}}(\tilde{x}) = \frac{L}{\ell}\zeta(\ell(\tilde{x} - \tilde{x}_{\tilde{j}}))$$

where $\tilde{j} \in \{1, \ldots, \ell\}^{d-1}$ and $\zeta > 0$ is a Lipschitz function with Lipschitz constant 1, and $\text{supp}(\zeta) = (-1/2, 1/2)^{d-1}$. Now define

$$g_\omega(\tilde{x}) = \sum_{\tilde{j} \in \{1,\ldots,\ell\}^{d-1}} \omega_{\tilde{j}} \eta_{\tilde{j}}(\tilde{x})$$

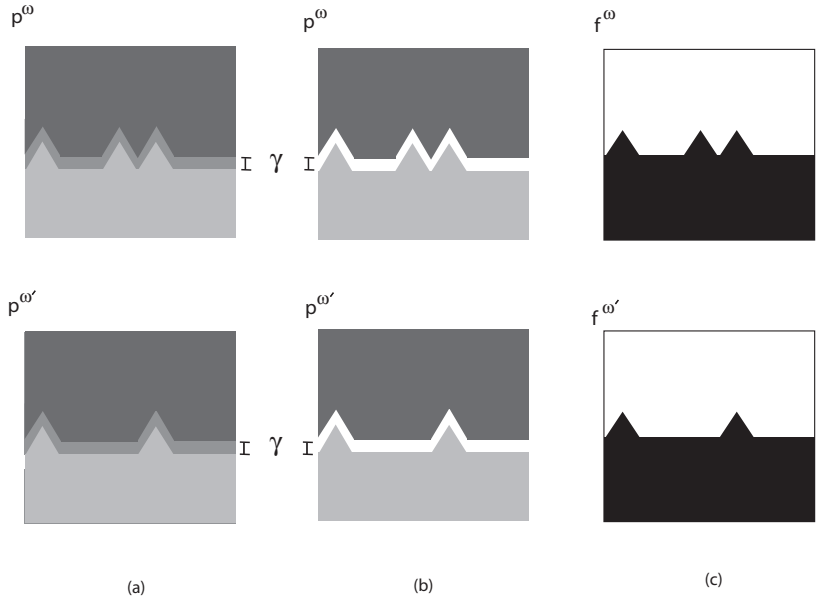

Figure 3: Examples of two sets of marginal density functions $p^{\omega}, p^{\omega'}$ for (a) $\gamma < 0$, (b) $\gamma > 0$ and regression functions $f^{\omega}, f^{\omega'}$ used for minimax construction.

Then $g_\omega(\cdot)$ is a Lipschitz function with Lipschitz constant L. Now define for $\omega \in \Omega$

$$p^{\omega}(x) \quad = \quad a p_1^{\omega}(x) + (1-a) p_2^{\omega}(x),$$

where $a \leq 1/2$, $p_1^{\omega}(x)$ is uniform and supported over $C_1^{\omega} = \left\{x \in [0,1]^d : x_d \geq \frac{1}{2} + \frac{\gamma}{2} + g_\omega(\tilde{x})\right\}$ and $p_2^{\omega}(x)$ is uniform and supported over $C_2^{\omega} = \left\{x \in [0,1]^d : x_d \leq \frac{1}{2} - \frac{\gamma}{2} + g_\omega(\tilde{x})\right\}$. Therefore, the margin is equal to γ. And

$$f^{\omega}(x) \quad = \quad \frac{a p_1^{\omega}(x) m_1(x) + (1-a) p_2^{\omega}(x) m_2(x)}{p^{\omega}(x)} \mathbf{1}_{\{p^{\omega}(x) \neq 0\}} - M \mathbf{1}_{\{p^{\omega}(x) = 0\}},$$

where $m_1(x) = M$ and $m_2(x) = -M$. Let Y be continuous and bounded, and also assume that $p_1^{\omega}(Y|X=x), p_2^{\omega}(Y|X=x) \leq W$, where $W > 0$ is a constant. This implies that

$$p^{\omega}(Y|X=x) = \frac{a p_1^{\omega}(x) p_1^{\omega}(Y|X=x) + (1-a) p_2^{\omega}(x) p_2^{\omega}(Y|X=x)}{p^{\omega}(x)} \leq \frac{2BW}{ab} = \frac{p_{\max} W}{p_{\min}}.$$

Figure 3 shows examples of two marginal density functions $p^{\omega}, p^{\omega'}$ for positive and negative margin, and corresponding regression functions $f^{\omega}, f^{\omega'}$.

Notice that the component densities are supported on compact, connected sets, are Hölder-α smooth for any α, and are bounded from above and below by $b \leq 1$ and $B \geq 4$. To see the latter, notice that

$$p_1^{\omega}(x) = \frac{1}{\text{vol}(C_1^{\omega})} = \frac{1}{\frac{1}{2} - \frac{\gamma}{2} - \int g_\omega(\tilde{x}) d\tilde{x}}, p_2^{\omega}(x) = \frac{1}{\text{vol}(C_2^{\omega})} = \frac{1}{\frac{1}{2} - \frac{\gamma}{2} + \int g_\omega(\tilde{x}) d\tilde{x}}.$$

The lower bound follows since $\text{vol}(C_1^{\omega}), \text{vol}(C_1^{\omega}) \leq 1$, and the upper bound follows since

$$\text{vol}(C_1^{\omega}) \geq \text{vol}(C_1^{\omega}) = \frac{1}{2} - \frac{\gamma}{2} - \int g_\omega(\tilde{x}) d\tilde{x} > \frac{1}{2} - \frac{c_o}{2} n^{-1/d} - \frac{L \|\zeta\|_1}{2 c_6} n^{-1/d} \geq 1/4.$$

Here the second last step follows since

$$\int g_\omega(\tilde{x}) d\tilde{x} = \sum_{\tilde{i} \in \{1,\ldots,\ell\}^{d-1}} \omega_{\tilde{i}} \eta_{\tilde{i}}(\tilde{x}) \leq \ell^{d-1} L \|\zeta\|_1 \ell^{-d} = L \|\zeta\|_1 \ell^{-1} \leq \frac{L \|\zeta\|_1}{2 c_6} n^{-1/d},$$

and the last step holds for $n \equiv n(c_o, c_6, d, L, \|\zeta\|_1)$ large enough. Further, the support sets of the component densities have Lipschitz boundaries with Lipschitz constant L. The component regression functions are uniformly bounded between $-M$ and M, and are Hölder-α smooth for any α. Thus $\mathcal{P}_\Omega \subseteq \mathcal{P}_{XY}(\gamma)$.

We first establish (2).

$$\begin{aligned} R(f^\omega, f^{\hat{\omega}}) &= \int (f^\omega(x) - f^{\hat{\omega}}(x))^2 p^\omega(x) dx \\ &\geq b\left[\int (f^\omega(x) - f^{\hat{\omega}}(x))^2 \mathbf{1}_{\{p^\omega(x)\neq 0\}} dx\right] \\ &\geq b\left[\int (f^\omega(x) - f^{\hat{\omega}}(x))^2 dx - \int (f^\omega(x) - f^{\hat{\omega}}(x))^2 \mathbf{1}_{\{p^\omega(x)=0\}} dx\right] \\ &\geq b\left[d^2(f^\omega, f^{\hat{\omega}}) - 4M^2\gamma\right] \end{aligned}$$

Next, we establish (3). We will consider two cases:
If $\gamma > 0$,

$$\begin{aligned} d^2(f^\omega, f^{\hat{\omega}}) &= \int (f^\omega(x) - f^{\hat{\omega}}(x))^2 dx \\ &= 4M^2 \sum_{\tilde{i}\in\{1,\ldots,\ell\}^{d-1}} |\omega_{\tilde{i}} - \hat{\omega}_{\tilde{i}}|^2 \int_{[0,1]^{d-1}} \eta_{\tilde{i}}(\tilde{x}) d\tilde{x} = 4M^2 L\|\zeta\|_1 \ell^{-d} \rho(\hat{\omega}, \omega) \end{aligned}$$

If $\gamma \leq 0$,

$$\begin{aligned} d^2(f^\omega, f^{\hat{\omega}}) &= \int (f^\omega(x) - f^{\hat{\omega}}(x))^2 dx \\ &\geq \min\left(\frac{\frac{2M(1-a)}{\text{vol}(C_2^\omega)}}{\frac{a}{\text{vol}(C_1^\omega)} + \frac{(1-a)}{\text{vol}(C_2^\omega)}}, \frac{\frac{2M(1-a)}{\text{vol}(C_2^{\hat{\omega}})}}{\frac{a}{\text{vol}(C_1^{\hat{\omega}})} + \frac{(1-a)}{\text{vol}(C_2^{\hat{\omega}})}}\right)^2 \cdot \\ &\quad \sum_{\tilde{i}\in\{1,\ldots,\ell\}^{d-1}} |\omega_{\tilde{i}} - \hat{\omega}_{\tilde{i}}|^2 \int_{[0,1]^{d-1}} \eta_{\tilde{i}}(\tilde{x}) d\tilde{x} \\ &\geq 4M^2 a^2 \min\left(\frac{\text{vol}(C_1^\omega)}{\text{vol}(C_2^\omega)}, \frac{\text{vol}(C_1^{\hat{\omega}})}{\text{vol}(C_2^{\hat{\omega}})}\right)^2 L\|\zeta\|_1 \ell^{-d} \rho(\omega, \hat{\omega}) \\ &\geq 4M^2 \frac{a^2}{16} L\|\zeta\|_1 \ell^{-d} \rho(\omega, \hat{\omega}) \end{aligned}$$

The second step follows from the definition of f^ω and p^ω, and the third step follows since $a \leq 1/2 \Rightarrow 1 - a \geq a$ and since $\text{vol}(C_1^\omega) \leq \text{vol}(C_2^\omega)$ for all $\omega \in \Omega$. To see the last step, observe that for all $\omega \in \Omega$,

$$\frac{\text{vol}(C_1^\omega)}{\text{vol}(C_2^\omega)} = \frac{\frac{1}{2} - \frac{\gamma}{2} - \int g_\omega(\tilde{x}) d\tilde{x}}{\frac{1}{2} - \frac{\gamma}{2} + \int g_\omega(\tilde{x}) d\tilde{x}} \geq \frac{\frac{1}{2} - \int g_\omega(\tilde{x}) d\tilde{x}}{\frac{1}{2} - \frac{\gamma}{2} + \int g_\omega(\tilde{x}) d\tilde{x}} \geq \frac{\frac{1}{2} - L\|\zeta\|_1 \ell^{-1}}{1 + L\|\zeta\|_1 \ell^{-1}} \geq \frac{1}{4}.$$

Here the third step follows since $\gamma \geq -1$ and $\int g_\omega(\tilde{x}) d\tilde{x} \leq L\|\zeta\|_1 \ell^{-1}$, and the last step follows for $n \equiv n(c_6, d, L, \|\zeta\|_1)$ large enough since $\ell = \lceil c_6 n^{1/d} \rceil$. Therefore, for all γ, we have

$$d^2(f^\omega, f^{\hat{\omega}}) \geq 4M^2 \frac{a^2}{16} L\|\zeta\|_1 \ell^{-d} \rho(\omega, \hat{\omega}) =: c_7 \ell^{-d} \rho(\omega, \hat{\omega}).$$

Thus, (3) is satisfied.

Now we only need to show that the condition of Theorem 2 is met, that is, $H^2(P^{\omega'}, P^\omega) \leq \kappa < 2$, $\forall \omega, \omega' \in \Omega : \rho(\omega, \omega') = 1$. Observe that

$$\begin{aligned} H^2(P^{\omega'}, P^\omega) &= H^2(P^{\omega'}(\{X_1, Y_1\}_{i=1}^n), P^\omega(\{X_1, Y_1\}_{i=1}^n)) \\ &= 2\left(1 - \prod_{i=1}^n \left(1 - \frac{H^2(P^{\omega'}(X_i, Y_i), P^\omega(X_i, Y_i))}{2}\right)\right) \end{aligned}$$

We now evaluate

$$\begin{aligned} H^2(P^{\omega'}(X_i,Y_i),P^{\omega}(X_i,Y_i)) &= \int(\sqrt{p^{\omega'}(X_i,Y_i)}-\sqrt{p^{\omega}(X_i,Y_i)})^2 \\ &= \int(\sqrt{p_X^{\omega'}(X_i)p_{Y|X}^{\omega'}(Y_i|X_i)}-\sqrt{p_X^{\omega}(X_i)p_{Y|X}^{\omega}(Y_i|X_i)})^2 \end{aligned}$$

Recall that $p_{Y|X}^{\omega}(Y_i|X_i) \leq p_{\max}W/p_{\min}$. Since $\rho(\omega,\omega')=1$, let $\tilde{j}$ denote the index for which $\omega_{\tilde{j}} \neq \omega'_{\tilde{j}}$ and without loss of generality, assume that $\omega_{\tilde{j}}=1$ and $\omega'_{\tilde{j}}=0$. Also let $B_{\tilde{j}}=\{x:\tilde{x}\in(\tilde{x}_{\tilde{j}}-\frac{1}{2\ell},\tilde{x}_{\tilde{j}}+\frac{1}{2\ell})\}$. We will evaluate the Hellinger integral over 4 different regions: (Here we use $\pm$ or $\mp$ to denote that the top sign is for the case $\gamma>0$ and bottom sign is for the case $\gamma\leq 0$)
First consider

$$\begin{aligned} A_1 := \{x:\tilde{x}\in B_{\tilde{j}}, \quad & 1/2\pm\gamma/2 \leq x_d < 1/2\pm\gamma/2+g_\omega(\tilde{x}), \\ & 1/2\mp\gamma/2 < x_d \leq \min(1/2\mp\gamma/2+g_\omega(\tilde{x}),1/2\pm\gamma/2)\} \end{aligned}$$

Since $p_X^{\omega'}(X_i),p_X^{\omega}(X_i)\in[b,B]$ and $p_{Y|X}^{\omega'}(Y_i|X_i),p_{Y|X}^{\omega}(Y_i|X_i)\in[0,p_{\max}W/p_{\min}]$, for this region, we bound the argument of the integral by $Bp_{\max}W/p_{\min}$.

$$\begin{aligned} \int_{A_1}(\sqrt{p_X^{\omega'}(X_i)p_{Y|X}^{\omega'}(Y_i|X_i)}-\sqrt{p_X^{\omega}(X_i)p_{Y|X}^{\omega}(Y_i|X_i)})^2 &\leq \frac{Bp_{\max}W}{p_{\min}}\int_{A_1}dx \\ &\leq \frac{Bp_{\max}W}{p_{\min}}2\int\eta_{\tilde{j}}d\tilde{x} \\ &= \frac{2Bp_{\max}W}{p_{\min}}L\|\zeta\|_1\ell^{-d} \\ &\leq \frac{2Bp_{\max}WL\|\zeta\|_1}{p_{\min}(2c_6)^d}n^{-1} \end{aligned}$$

For $x\notin A_1$, notice that $p_{Y|X}^{\omega'}(Y_i|X_i)=p_{Y|X}^{\omega}(Y_i|X_i)\leq p_{\max}W/p_{\min}$, therefore we have:

$$\begin{aligned} &\int_{x\notin A_1}(\sqrt{p_X^{\omega'}(X_i)p_{Y|X}^{\omega'}(Y_i|X_i)}-\sqrt{p_X^{\omega}(X_i)p_{Y|X}^{\omega}(Y_i|X_i)})^2 \\ &\qquad\leq \frac{p_{\max}W}{p_{\min}}\int_{x\notin A_1}\left(\sqrt{p_X^{\omega'}(X_i)}-\sqrt{p_X^{\omega}(X_i)}\right)^2 \end{aligned}$$

We now evaluate the latter integral over three regions: Before that, we set up some results that will be used in all these cases.

$$|\text{vol}(C_1^{\omega})-\text{vol}(C_1^{\omega'})|,|\text{vol}(C_2^{\omega})-\text{vol}(C_2^{\omega'})| \leq \int\eta_{\tilde{j}}d\tilde{x}=L\|\zeta\|_1\ell^{-d}\leq\frac{L\|\zeta\|_1}{(2c_6)^d}n^{-1},$$

Also, we establish that

$$\text{vol}(C_1^{\omega}),\text{vol}(C_1^{\omega'}),\text{vol}(C_2^{\omega}),\text{vol}(C_2^{\omega'})\geq 1/4.$$

For this, observe that for $n\equiv n(c_o,c_6,d,L,\|\zeta\|_1)$ large enough

$$\begin{aligned} \text{vol}(C_1^{\omega'})\geq\text{vol}(C_1^{\omega}) &= 1/2-\gamma/2-\int g_\omega(\tilde{x})>1/2-\frac{c_o}{2}n^{-1/d}-\frac{L\|\zeta\|_1}{2c_6}n^{-1/d}\geq 1/4 \\ \text{vol}(C_2^{\omega})\geq\text{vol}(C_2^{\omega'}) &= 1/2-\gamma/2+\int g_{\omega'}(\tilde{x})\geq 1/2-\frac{c_o}{2}n^{-1/d}\geq 1/4 \end{aligned}$$

We are now ready to consider the three regions:

$$A_2 := \{x:x_d\geq 1/2\pm\gamma/2+g_\omega(\tilde{x})\}$$

Notice that

$$\begin{aligned}\int_{A_2}\left(\sqrt{p_X^{\omega'}(X_i)}-\sqrt{p_X^{\omega}(X_i)}\right)^2 &= \int_{A_2}\left(\sqrt{\frac{a}{\text{vol}(C_1^{\omega'})}}-\sqrt{\frac{a}{\text{vol}(C_1^{\omega})}}\right)^2\\ &\leq \frac{a}{4}\int_{A_2}\left(\frac{1}{\text{vol}(C_1^{\omega'})}-\frac{1}{\text{vol}(C_1^{\omega})}\right)^2\\ &\leq \frac{1}{4}\int_{A_2}\left(\frac{|\text{vol}(C_1^{\omega})-\text{vol}(C_1^{\omega'})|}{\text{vol}(C_1^{\omega'})\text{vol}(C_1^{\omega})}\right)^2 \leq 4\frac{L^2\|\zeta\|_1^2}{(2c_6)^{2d}}n^{-2}\end{aligned}$$

The second step follows since $2 \leq \sqrt{1/\text{vol}(C_1^{\omega'})}+\sqrt{1/\text{vol}(C_1^{\omega})}$.

$$A_3 := \{x : x_d \leq 1/2 \mp \gamma/2 + g_{\omega'}(\tilde{x})\}$$

Notice that

$$\begin{aligned}\int_{A_3}\left(\sqrt{p_X^{\omega'}(X_i)}-\sqrt{p_X^{\omega}(X_i)}\right)^2 &= \int_{A_3}\left(\sqrt{\frac{1-a}{\text{vol}(C_2^{\omega'})}}-\sqrt{\frac{1-a}{\text{vol}(C_2^{\omega})}}\right)^2\\ &\leq \frac{1-a}{4}\int_{A_3}\left(\frac{1}{\text{vol}(C_2^{\omega'})}-\frac{1}{\text{vol}(C_2^{\omega})}\right)^2\\ &\leq \frac{1}{4}\int_{A_3}\left(\frac{|\text{vol}(C_2^{\omega})-\text{vol}(C_2^{\omega'})|}{\text{vol}(C_2^{\omega'})\text{vol}(C_2^{\omega})}\right)^2 \leq 4\frac{L^2\|\zeta\|_1^2}{(2c_6)^{2d}}n^{-2}\end{aligned}$$

The second step follows since $2 \leq \sqrt{1/\text{vol}(C_2^{\omega'})}+\sqrt{1/\text{vol}(C_2^{\omega})}$.

$$\begin{aligned}A_4 = \{x : \tilde{x} \notin B_{\tilde{j}}, \quad & 1/2 \mp \gamma/2 + g_{\omega'}(\tilde{x}) < x_d < 1/2 \pm \gamma/2 + g_{\omega'}(\tilde{x}),\\ \tilde{x} \in B_{\tilde{j}} \quad & \min(1/2 \mp \gamma/2 + g_{\omega}(\tilde{x}), 1/2 \pm \gamma/2) < x_d < 1/2 \pm \gamma/2\}\end{aligned}$$

If $\gamma > 0$, Notice that

$$\int_{A_4}\left(\sqrt{p_X^{\omega'}(X_i)}-\sqrt{p_X^{\omega}(X_i)}\right)^2 = 0$$

If $\gamma \leq 0$, then

$$\begin{aligned}\int_{A_4}\left(\sqrt{p_X^{\omega'}(X_i)}-\sqrt{p_X^{\omega}(X_i)}\right)^2 &= \int_{A_4}\left(\sqrt{\frac{a}{\text{vol}(C_1^{\omega'})}+\frac{1-a}{\text{vol}(C_2^{\omega'})}}-\sqrt{\frac{a}{\text{vol}(C_1^{\omega})}+\frac{1-a}{\text{vol}(C_2^{\omega})}}\right)^2\\ &\leq \frac{1}{4}\int_{A_4}\left(\frac{a}{\text{vol}(C_1^{\omega'})}+\frac{1-a}{\text{vol}(C_2^{\omega'})}-\frac{a}{\text{vol}(C_1^{\omega})}-\frac{1-a}{\text{vol}(C_2^{\omega})}\right)^2\\ &\leq \frac{1}{4}\int_{A_4}\left(\left|\frac{a}{\text{vol}(C_1^{\omega'})}-\frac{a}{\text{vol}(C_1^{\omega})}\right|+\left|\frac{1-a}{\text{vol}(C_2^{\omega'})}-\frac{1-a}{\text{vol}(C_2^{\omega})}\right|\right)^2\\ &\leq \frac{1}{4}\int_{A_4}\left(\frac{|\text{vol}(C_1^{\omega})-\text{vol}(C_1^{\omega'})|}{\text{vol}(C_1^{\omega'})\text{vol}(C_1^{\omega})}+\frac{|\text{vol}(C_2^{\omega})-\text{vol}(C_2^{\omega'})|}{\text{vol}(C_2^{\omega'})\text{vol}(C_2^{\omega})}\right)^2\\ &\leq 16\frac{L^2\|\zeta\|_1^2}{(2c_6)^{2d}}n^{-2}\end{aligned}$$

The second step follows since $2 \leq \sqrt{\frac{a}{\text{vol}(C_1^{\omega'})}+\frac{1-a}{\text{vol}(C_2^{\omega'})}}+\sqrt{\frac{a}{\text{vol}(C_1^{\omega})}+\frac{1-a}{\text{vol}(C_2^{\omega})}}$.

Therefore, we get that

$$\begin{aligned}H^2(P^{\omega'}(X_i,Y_i),P^{\omega}(X_i,Y_i)) &\leq \frac{2Bp_{\max}WL\|\zeta\|_1}{p_{\min}(2c_6)^d}n^{-1}+24\frac{p_{\max}WL^2\|\zeta\|_1^2}{p_{\min}(2c_6)^{2d}}n^{-2}\\ &\leq \frac{4Bp_{\max}WL\|\zeta\|_1}{p_{\min}(2c_6)^d}n^{-1} =: c_8 n^{-1}\end{aligned}$$

where the second step holds for $n \equiv n(c_6, d, L, \|\zeta\|_1)$ large enough. And $c_8 > 0$ is a constant.

$$H^2(P^{\omega'}, P^{\omega}) \leq 2\left(1 - \left(1 - \frac{c_8}{2}n^{-1}\right)^n\right) \leq 2(1 - e^{-c_8/2}) =: \kappa$$

where the second step holds for $n \equiv n(c_8)$ large enough. Thus, the conditions of Theorem 2 are met and we have established the desired lower bounds for supervised learning.

■

References

[1] Zhu, X.: Semi-supervised learning literature survey. Technical Report 1530, Computer Sciences Department, University of Wisconsin-Madison. URL http://www.cs.wisc.edu/∼jerryzhu/pub/ssl_survey.pdf (2005)

[2] Balcan, M.F., Blum, A.: A PAC-style model for learning from labeled and unlabeled data. In: 18th Annual Conference on Learning Theory, COLT. (2005)

[3] Kääriäinen, M.: Generalization error bounds using unlabeled data. In: 18th Annual Conference on Learning Theory, COLT. (2005)

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

[5] Lafferty, J., Wasserman, L.: Statistical analysis of semi-supervised regression. In: Advances in Neural Information Processing Systems 20, NIPS. (2008) 801–808

[6] Ben-David, S., Lu, T., Pal, D.: Does unlabeled data provably help? worst-case analysis of the sample complexity of semi-supervised learning. In: 21st Annual Conference on Learning Theory, COLT. (2008)

[7] Niyogi, P.: Manifold regularization and semi-supervised learning: Some theoretical analyses. Technical Report TR-2008-01, Computer Science Department, University of Chicago. URL http://people.cs.uchicago.edu/∼niyogi/papersps/ssminimax2.pdf (2008)

[8] Seeger, M.: Learning with labeled and unlabeled data. Technical report, Institute for ANC, Edinburgh, UK. URL http://www.dai.ed.ac.uk/∼seeger/papers.html (2000)

[9] Castelli, V., Cover, T.M.: On the exponential value of labeled samples. Pattern Recognition Letters **16**(1) (1995) 105–111

[10] Castelli, V., Cover, T.M.: The relative value of labeled and unlabeled samples in pattern recognition. IEEE Transactions on Information Theory **42**(6) (1996) 2102–2117

[11] Bickel, P.J., Li, B.: Local polynomial regression on unknown manifolds. In: IMS Lecture NotesMonograph Series, Complex Datasets and Inverse Problems: Tomography, Networks and Beyond. Volume 54. (2007) 177–186

[12] Korostelev, A.P., Tsybakov, A.B.: Minimax Theory of Image Reconstruction. Springer, NY (1993)

[13] Korostelev, A., Nussbaum, M.: The asymptotic minimax constant for sup-norm loss in non-parametric density estimation. Bernoulli **5(6)** (1999) 1099–1118

[14] Chapelle, O., Zien, A.: Semi-supervised classification by low density separation. In: Tenth International Workshop on Artificial Intelligence and Statistics. (2005) 57–64

[15] Singh, A., Scott, C., Nowak, R.: Adaptive hausdorff estimation of density level sets. Technical Report ECE-07-06, ECE Deparment, University of Wisconsin - Madison. URL http://www.cae.wisc.edu/∼singh/TR_Hausdorff.pdf (2007)

[16] Singh, A., Nowak, R., Scott, C.: Adaptive hausdorff estimation of density level sets. In: 21st Annual Conference on Learning Theory, COLT. (2008)

[17] Castro, R., Willett, R., Nowak, R.: Faster rates in regression via active learning. Technical Report ECE-05-03, ECE Department, University of Wisconsin - Madison. URL http://www.ece.wisc.edu/∼nowak/ECE-05-03.pdf (2005)

[18] Stone, C.J.: Optimal rates of convergence for nonparametric estimators. The Annals of Statistics **8(6)** (1980) 1348–1360

[19] Tsybakov, A.B.: Introduction a l'estimation non-parametrique. Springer, Berlin Heidelberg (2004)