Generalization error bounds in semi-supervised classification under the cluster assumption

Philippe Rigollet

Generalization error bounds in semi-supervised classification under the cluster assumption

Philippe Rigollet *

April 11, 2006

Abstract

We consider semi-supervised classification when part of the available data is unlabeled. These unlabeled data can be useful for the classification problem when we make an assumption relating the behavior of the regression function to that of the marginal distribution. Seeger [18] proposed the well-known *cluster assumption* as a reasonable one. We propose a mathematical formulation of this assumption and a method based on density level sets estimation that takes advantage of it to achieve fast rates of convergence both in the number of unlabeled examples and the number of labeled examples.

Key Words: Semi-supervised learning, statistical learning theory, classification, cluster assumption, generalization bounds.

1 Introduction

Semi-supervised classification has been of growing interest over the past few years and many methods have been proposed. The methods try to give an answer to the question: "How to improve classification accuracy using unlabeled data together with the labeled data?". Unlabeled data can be used in different ways depending on the assumptions on the model. There are two types of assumptions. The first one consists in assuming that we have a set of potential classifiers and we want to aggregate them. In that case, unlabeled data is used to measure the *compatibility* between the classifiers and reduces the complexity of the resulting classifier (see, e.g., [3], [4]). The second approach is the one that we use here. It assumes that the data contains clusters that have homogeneous labels and the unlabeled observations are used to identify these clusters. This is the so-called *cluster assumption*. This idea can be put in practice in several ways giving rise to various methods. The simplest is the one presented here: estimate the clusters, then label each cluster uniformly. Most of these methods use Hartigan's [11] definition of clusters, namely the connected components of the density level sets. However, they use a parametric (usually mixture) model to estimate the underlying density which can be far from reality. Moreover, no generalization error bounds are available for such methods. In the same spirit, [20] and [17] propose methods that learn a distance using unlabeled data in order to have intra-cluster distances smaller than inter-clusters distances. The whole family of graph-based methods aims also at using unlabeled data to learn the distances between points. The edges of the graphs reflect the proximity between points. For a detailed survey on graph methods we refer to [23].

*Laboratoire de Probabilités et Modèles aléatoires, UMR 7599, Université Paris 6, case 188, 4, pl. Jussieu, F-75252 Paris Cedex 5, France. Email: rigollet@ccr.jussieu.fr. Part of this work was done when the author was visiting researcher at Department of Statistics, University of California, Berkeley, CA 94720-1776, USA (Fund by France-Berkeley fund).

Finally, we mention kernel methods, where unlabeled data are used to build the kernel. Recalling that the kernel measures proximity between points, such methods can also be viewed as learning a distance using unlabeled data (see [6], [7], [8]).

The cluster assumption can be interpreted in another way, i.e., as the requirement that the decision boundary has to lie in low density regions. This interpretation has been widely used in learning since it can be used in the design of standard algorithms such as Boosting [1], [13] or SVM [6], [7], which are closely related to kernel methods mentioned above. In these algorithms, a greater penalization is given to decision boundaries that cross a cluster. For more details, see, e.g., [18], [23], [8]. Although most methods make, sometimes implicitly, the cluster assumption, no formulation in probabilistic terms has been provided so far. The formulation that we propose in this paper remains very close to its original text formulation and allows to derive generalization error bounds. We also discuss what can and cannot be done using unlabeled data. One of the conclusions is that considering the whole excess-risk is too ambitious and we need to concentrate on a smaller part of it to observe the improvement of semi-supervised classification over standard classification.

Outline of the paper. After describing the model, we formulate the cluster assumption and discuss why and how it can improve classification performance in the next section. In Section 3, we study the population case when the marginal density p is known, to get an idea of our target. Indeed, such a population case corresponds in some way to the case when the amount of unlabeled data is infinite. Section 4 contains the main result: we propose an algorithm for which we derive rates of convergence for the λ-thresholded excess-risk as a measure of performance. An exemple of consistent density level set estimators is given in Section 5. Section 6 is devoted to discussion on the choice of λ and possible improvements. Proofs of the results are gathered in Section 7.

Notation. Throughout the paper, we denote by c_j positive constants. We write Γ^c for the complement of the set Γ. For two sequences $(u_p)_p$ and $(v_p)_p$ (in that paper, p will be m or n), we write $u_p = O(v_p)$ if there exists a constant $C > 0$ such that $u_p \leq Cv_p$ and we write $u_p = \widetilde{O}(v_p)$ if $u_p \leq C(\log p)^\alpha v_p$ for some constants $\alpha > 0, C > 0$. Thus, if $u_p = \widetilde{O}(v_p)$, we have $u_p = o(v_p p^\beta)$, for any $\beta > 0$.

2 The model

Let (X, Y) be a random couple with joint distribution P, where $X \in \mathcal{X} \subset \mathbb{R}^d$ is a vector of d features and $Y \in \{0, 1\}$ is a label indicating the class to which X belongs. The distribution P of the random couple (X, Y) is completely determined by the pair (P_X, η) where P_X is the marginal distribution of X and η is the regression function of Y on X, i.e., $\eta(x) \triangleq P(Y = 1|X = x)$. The goal of classification is to predict the label Y given the value of X, i.e., to construct a measurable function $g : \mathcal{X} \to \{0, 1\}$ called a *classifier*. The performance of g is measured by the average classification error

$$R(g) \triangleq P\left(g(X) \neq Y\right)$$

A minimizer of the risk $R(g)$ over all classifiers is given by the *Bayes classifier* $g^\star(x) = \mathbb{1}_{\{\eta(x) \geq 1/2\}}$, where $\mathbb{1}_{\{\cdot\}}$ denotes the indicator function. Assume that we have a sample of n observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ that are independent copies of (X, Y). An empirical classifier is a random function $\hat{g}_n : \mathcal{X} \to \{0, 1\}$ constructed on the basis of the sample $(X_1, Y_1), \ldots, (X_n, Y_n)$. Since $g^\star$ is the best possible classifier, we measure the performance of an empirical classifier $\hat{g}_n$ by its *excess-risk*

$$\mathcal{E}(\hat{g}_n) = \mathbb{E}_n R(\hat{g}_n) - R(g^\star)\,,$$

where $\mathbb{E}_n$ denotes the expectation with respect to the joint distribution of $(X_1, Y_1), \ldots, (X_n, Y_n)$. We denote hereafter by $\mathbb{P}_n$ the corresponding probability.

In many applications, a large amount of unlabeled data is available as well as a small set of labeled data $(X_1, Y_1), \ldots, (X_n, Y_n)$ and the goal of semi-supervised classification is to use of the unlabeled data to improve the performance of classifiers. Thus, we observe two independent samples $\mathbb{X}_l = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ and $\mathbb{X}_u = \{X_{n+1}, \ldots, X_{n+m}\}$, where n is rather small and typically $m \gg n$. It is well known that in order to make use of the additional unlabeled observations, we have to make an assumption on the dependence between the marginal distribution of X and the joint distribution of (X, Y). Seeger [18] formulated the rather intuitive *cluster assumption* as follows[1]

> Two points $x, x' \in \mathcal{X}$ should have the same label y if there is a path between them which passes only through regions of relatively high P_X.

This assumption, in its raw formulation cannot be exploited in the probabilistic model since (i) the labels are random variables Y, Y' so that the expression "should have the same label" is meaningless unless η takes values in $\{0, 1\}$ and (ii) it is not clear what "regions of relatively high P_X" are. To match the probabilistic framework, we propose the following modifications

(i) $P[Y = Y'|X, X' \text{connected}] \geq P[Y \neq Y'|X, X' \text{connected}]$, where "connected" means that there is the path between X and X' which passes only through regions of relatively high P_X.

(ii) Define "regions of relatively high P_X" in terms of *density level sets*.

We now need to precise the term *relatively high density*. Assume that P_X admits a density p with respect to the Lebesgue measure on $\mathbb{R}^d$ denoted hereafter by Leb_d. For a fixed $\lambda > 0$, the λ-level set of the density p is defined by

$$\Gamma(\lambda) \triangleq \{x \in \mathcal{X} : p(x) \geq \lambda\} \,. \tag{2.1}$$

We are now in position to give a precise definition of the cluster assumption.

Cluster Assumption CA(λ): Fix $\lambda > 0$ and assume that the density level set $\Gamma = \Gamma(\lambda)$ has a countable number of connected components $T_j = T_j(\lambda),\ j = 1, 2, \ldots$. Then the function $x \in \mathcal{X} \mapsto \mathbb{1}\{\eta(x) \geq 1/2\}$ takes a constant value on each of the $T_j, j = 1, 2, \ldots$.

Note that density level sets have the monotonicity property: $\lambda \geq \lambda'$, implies $\Gamma(\lambda) \subset \Gamma(\lambda')$. In terms of the cluster assumption, it means that when λ decreases to 0, the assumption CA(λ) becomes more restrictive. As a result, the parameter λ can be considered as a level of confidence characterizing to which extent the cluster assumption is valid for the distribution P and its choice is left to the user. For more details on the choice of λ, see Section 6.

A question remains: what happens outside of the set $\Gamma(\lambda)$? Assume that we are in the problematic case, $P_X(\Gamma^c) = C > 0$ such that the question makes sense. Since the cluster assumption says nothing about what happens outside of the set Γ, we can only perform supervised classification on Γ^c. Consider now a classifier $\hat{g}_{n,m}$ built from labeled and unlabeled samples $(\mathbb{X}_l, \mathbb{X}_u)$ pooled together. The excess-risk of $\hat{g}_{n,m}$ can be written (see [10])

$$\mathcal{E}(\hat{g}_{n,m}) = \mathbb{E}_{n,m} \int_{\mathcal{X}} |2\eta(x) - 1| \mathbb{1}_{\{\hat{g}_{n,m}(x) \neq g^*(x)\}} p(x) \mathrm{d}x \,,$$

where $\mathbb{E}_{n,m}$ denotes the expectation with respect to the pooled sample $(\mathbb{X}_l, \mathbb{X}_u)$. We denote hereafter by $\mathbb{P}_{n,m}$ the corresponding probability. Since, the unlabeled sample is of no help to classify points in Γ^c, any reasonable classifier should be based on the sample $\mathbb{X}_l$ so that $\hat{g}_{n,m}(x) = \hat{g}_n(x),\ \forall\, x \in \Gamma^c$, and we have

$$\mathcal{E}(\hat{g}_{n,m}) = \mathcal{E}(\hat{g}_n) \geq \mathbb{E}_n \int_{\Gamma^c} |2\eta(x) - 1| \mathbb{1}_{\{\hat{g}_n(x) \neq g^*(x)\}} p(x) \mathrm{d}x \,. \tag{2.2}$$

[1] the notation is adapted to the present framework

Since we assumed $P_X(\Gamma^c) = C > 0$, the RHS of (2.2) is bounded from below by the optimal rates of convergence that appear in supervised classification. These rates are typically of the order $n^{-\alpha}, 1/2 \leq \alpha \leq 1$ (see e.g. [15], [22], [2] and [5] for a comprehensive survey). Thus, unlabeled data do not improve the rate of convergence of this part of the excess-risk. To observe the effect of unlabeled data on the rates of convergence, we have to consider the λ-*thresholded excess-risk* of a classifier $\hat{g}_{n,m}$ defined by

$$\mathcal{E}_\lambda(\hat{g}_{n,m}) \triangleq \mathbb{E}_{n,m} \int_{\Gamma(\lambda)} |2\eta(x) - 1| \mathbb{I}_{\{\hat{g}_{n,m}(x) \neq g^\star(x)\}} p(x) \mathrm{d}x \,. \tag{2.3}$$

We will therefore focus on this measure of performance. Note that for such a measure, we only need to consider classifiers $\hat{g}_{n,m}$ that are defined on Γ.

We now propose a method to obtain good upper bounds on this quantity, taking advantage of the cluster assumption. The idea is to estimate the regions where the sign of $(\eta - 1/2)$ is constant and make a majority vote on each region.

3 Results for known marginal distribution

Consider the ideal situation where the density p is known and we observe only the labeled sample $\mathbb{X}_l = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$. Fix $\lambda > 0$ and assume that $\Gamma = \Gamma(\lambda)$ has a countable number of connected components:

$$\Gamma = \bigsqcup_{j \geq 1} T_j \,,$$

where the $T_j = T_j(\lambda)$ are non empty disjoint connected sets. Under the cluster assumption CA(λ), the function $x \mapsto \eta(x) - 1/2$ has constant sign on each T_j. Thus a simple and intuitive method for classification is to perform a majority vote on each T_j.

For any $j \geq 1$, define $\delta_j = \delta_j(\lambda) \geq 0$, $\delta_j \leq 1$ by

$$\delta_j \triangleq \int_{T_j} |2\eta(x) - 1| p(x) \mathrm{d}x \,.$$

The following assumption characterizes how far is η from $1/2$ on every connected component T_j.

Global Margin Assumption GMA(λ): There exists $\delta > 0$ such that, for any $j \geq 1$, either $\delta_j = 0$ or $\delta_j \geq \delta$.

Since $\sum_j \delta_j \leq 1$, a direct consequence of the GMA is that only a finite number of δ_j are positive. The GMA assumption imposes that, on average over T_j, the regression function η is away from $1/2$ for any $j \geq 1$ such that $\delta_j > 0$. It describes the global behavior of η on each connected component T_j as opposed to the standard margin assumption formulated in [15] and [22] which we will call here *local margin assumption* (LMA). Assumption LMA characterizes the local behavior of η in a neighborhood of $1/2$. In [2], it is stated as follows

Local Margin Assumption LMA: There exist constants $C_0 > 0$ and $\alpha \geq 0$ such that

$$P_X \left(0 < |2\eta(X) - 1| \leq t\right) \leq C_0 t^\alpha, \quad \forall\, t \geq 0 \,.$$

It is straightforward that when there is only a finite number of connected components $T_j, j = 1, \ldots, J$ with non-zero Lebesgue measure, GMA is a consequence of LMA. However we will see in our analysis

that the rates of convergence depend crucially on the value of $\delta > 0$, $j = 1, 2, \ldots$, while deriving GMA from LMA yields a δ depending on C_0. For this reason, it is natural to introduce GMA instead of using the well known but less flexible LMA.

We now define our classifier based on the sample $\mathbb{X}_l$. For any $j \geq 1$, define the random variable

$$Z_n^j \triangleq \sum_{i=1}^{n} (2Y_i - 1)\, \mathbb{1}_{\{X_i \in T_j\}}\,,$$

and denote by $\hat{g}_n^j$ the function $\hat{g}_n^j(x) = \mathbb{1}_{\{Z_n^j > 0\}}$ for all $x \in T_j$. Consider the classifier defined on Γ by

$$\hat{g}_n(x) = \sum_{j \geq 1} \hat{g}_n^j(x) \mathbb{1}_{\{x \in T_j\}}, \quad x \in \Gamma\,.$$

The following theorem gives exponential rates of convergence for the classifier $\hat{g}_n$ under CA(λ).

Theorem 3.1 *Fix $\lambda > 0$ and assume that CA(λ) holds. Then, the classifier $\hat{g}_n$ satisfies*

$$\mathcal{E}_\lambda(\hat{g}_n) \leq 2 \sum_{j \geq 1} \delta_j e^{-n\delta_j^2/2}\,. \tag{3.1}$$

Moreover, if GMA(λ) holds, inequality (3.1) *reduces to*

$$\mathcal{E}_\lambda(\hat{g}_n) \leq 2 e^{-n\delta^2/2}\,. \tag{3.2}$$

A rapid overview of the proof shows that the rate of convergence $\mathrm{e}^{-n\delta^2/2}$ cannot be improved without further assumption. It will be our target in semi-supervised classification. However, we need estimators of the connected components $T_j, j \geq 1$. In the next section we provide the main result on semi-supervised learning, that is when the density p is unknown but we can estimate it using the unlabeled sample $\mathbb{X}_u$.

4 Main result

We now deal with a more realistic case where the density p is unknown and so are the density level sets which have to be estimated using the unlabeled sample $\mathbb{X}_u = \{X_1, \ldots, X_m\}$. Fix $\lambda > 0$ and assume that $\Gamma = \Gamma(\lambda)$ has a countable number of connected components:

$$\Gamma = \bigsqcup_{j \geq 1} T_j\,,$$

where the $T_j = T_j(\lambda)$ are non empty disjoint connected sets.

4.1 Density level set estimation

Assume that the density p is uniformly bounded by a constant $L(p)$ and that $\mathrm{Leb}_d(\mathcal{X}) < \infty$, where Leb_d denotes the Lebesgue measure on $\mathbb{R}^d$. Denote by $\mathbb{P}_m$ and $\mathbb{E}_m$ respectively the probability and the expectation w.r.t the sample $\mathbb{X}_u$ of size m. Assume that for any $\lambda > 0$, we use the sample $\mathbb{X}_u$ to construct an estimator $\hat{G}_m = \hat{G}_m(\lambda)$ of $\Gamma = \Gamma(\lambda)$ satisfying

$$\mathbb{E}_m\big[\mathrm{Leb}_d(\hat{G}_m \triangle \Gamma)\big] \to 0, \quad m \to +\infty. \tag{4.1}$$

We call such estimators *consistent* estimators of Γ. However, the connected components of a consistent estimator of Γ are not in general consistent estimators of the connected components of Γ. To

ensure componentwise consistency, we have to make assumptions on the connected component of Γ and those of $\hat{G}$.

Let $\mathcal{B}(x,r)$ be the d-dimensional closed ball of center $x \in \mathbb{R}^d$ and radius $r > 0$, defined by

$$\mathcal{B}(x,r) \triangleq \left\{ y \in \mathbb{R}^d : \|y - x\| \leq r \right\} ,$$

where $\|\cdot\|$ denotes the Euclidean norm in $\mathbb{R}^d$.

Definition 4.1 *Fix $r_0 \geq 0$ and $c_0 > 0$. We say that a set $C \subset \mathbb{R}^d$ is* r_0-connected *if for any $x, x' \in C$, there exists a continuous map $f : [0,1] \to C$ such that $f(0) = x, f(1) = x'$ and for any $t \in [0,1]$ and any $r \leq r_0$, we have*

$$\mathrm{Leb}_d\big(\mathcal{B}(f(t), r) \cap C\big) \geq c_0 r^d .$$

A 0-connected set is simply called connected or pathwise connected.

This definition ensures that Γ has no flat parts which allows to exclude pathological cases such as the one presented on the left of Figure 1. Now, define the distance d_∞, between two closed connected sets C_1 and C_2 by

$$d_\infty(C_1, C_2) = \min_{\substack{x \in C_1 \\ y \in C_2}} \|x - y\|$$

We say that a collection of connected sets $C_1, C_2, \ldots,$ is s_0-*separated* if $d_\infty(C_j, C_{j'}) \geq s_0,\ \forall j \neq j'$ for some $s_0 \geq 0$. If the connected components of Γ are not s_0-separated for some $s_0 > 0$, cases such as the one presented on Figure 1 (right) could arise. In that case, two connected components and therefore two clusters are identified which is obviously not desirable. Therefore, the cluster assumption should not hold for that particular level λ but it might hold for some $\lambda' \neq \lambda$.

Note that the performance of a density level estimator $\hat{G}_m$ is measured by the quantity

$$\mathbb{E}_m\big[\mathrm{Leb}_d(\hat{G}_m \triangle \Gamma)\big] = \mathbb{E}_m\big[\mathrm{Leb}_d(\hat{G}_m^c \cap \Gamma)\big] + \mathbb{E}_m\big[\mathrm{Leb}_d(\hat{G}_m \cap \Gamma^c)\big] . \tag{4.2}$$

For some estimators, such as the penalized plug-in density level sets estimators presented in Section 5, we can prove that the dominant term in the RHS of (4.2) is $\mathbb{E}_m\big[\mathrm{Leb}_d(\hat{G}_m^c \cap \Gamma)\big]$. This ensures that with high probability the estimator $\hat{G}_m$ is included in Γ. We now give a precise definition of such estimators.

Definition 4.2 *Let $\hat{G}_m$ be an estimator of Γ and fix $\alpha > 0$. We say that the estimator $\hat{G}_m$ is* consistent from inside *at rate $m^{-\alpha}$ if it satisfies*

$$\mathbb{E}_m\big[\mathrm{Leb}_d(\hat{G}_m \triangle \Gamma)\big] = \widetilde{O}(m^{-\alpha})$$

and

$$\mathbb{E}_m\big[\mathrm{Leb}_d(\hat{G}_m \cap \Gamma^c)\big] = \widetilde{O}(m^{-2\alpha})$$

For fixed $\alpha > 0, \lambda > 0$, let $\hat{G}_m \subset \mathcal{X}$ be a consistent from inside estimator of Γ at rate $m^{-\alpha}$. We begin by clipping $\hat{G}_m$ in the following manner. Define the set

$$\mathrm{Clip}(\hat{G}_m) = \Big\{ x \in \hat{G}_m : \mathrm{Leb}_d\big(\hat{G}_m \cap \mathcal{B}(x, (\log m)^{-1})\big) \leq \frac{(\log m)^{-d}}{m^\alpha} \Big\}.$$

Note that $\mathrm{Leb}_d(\mathcal{X}) < \infty$ yields

$$\mathrm{Leb}_d\big(\mathrm{Clip}(\hat{G}_m)\big) = \widetilde{O}(m^{-\alpha})$$

and therefore the clipped set $\tilde{G}_m = \hat{G}_m \setminus \mathrm{Clip}(\hat{G}_m)$ is also consistent from inside at rate $m^{-\alpha}$. We now use only $\tilde{G}_m$. It is straightforward that $\tilde{G}_m$ can be decomposed into a finit number $\tilde{J}_m$ of connected components. We write for simplicity

$$\tilde{G}_m = \bigsqcup_{l\geq 1} \tilde{T}_l \,, \tag{4.3}$$

where $\tilde{T}_l$ depends on m and λ. Denote by $\tilde{H}_k, k = 1, 2, \ldots$, the family of sets such that

$$\bigsqcup_{l\geq 1} \tilde{T}_l = \bigsqcup_{k\geq 1} \tilde{H}_k \,, \tag{4.4}$$

and $d_\infty(\tilde{H}_k, \tilde{H}_{k'}) > 2(\log m)^{-1}$, $\forall\, k \neq k'$. It is not hard to see that the sets $\tilde{H}_k$ are uniquely defined from $\tilde{T}_1, \tilde{T}_2, \ldots$. Let $\mathcal{J}$ be a subset of $\mathbb{N}^\star = \{1, 2, \ldots\}$. Define $\kappa(j) = \{k : \tilde{H}_k \cap T_j \neq \emptyset\}$ and let $D(\mathcal{J})$ be the event on which the sets $\kappa(j), j \in \mathcal{J}$ are reduced to singletons $\{k(j)\}$ which are disjoint, i.e.,

$$\begin{aligned} D(\mathcal{J}) &\triangleq \Big\{\kappa(j) = \{k(j)\}, k(j) \neq k(j'),\ \forall\, j, j' \in \mathcal{J}, j \neq j'\Big\} \\ &= \Big\{\kappa(j) = \{k(j)\},\ (T_j \cup \tilde{H}_{k(j)}) \cap (T_{j'} \cup \tilde{H}_{k(j')}) = \emptyset,\ \forall\, j, j' \in \mathcal{J}, j \neq j'\Big\}\,. \end{aligned} \tag{4.5}$$

In other words, on the event $D(\mathcal{J})$, there is a one-to-one correspondence between the collection $\{T_j\}_{j\in\mathcal{J}}$ and the collection $\big\{\{\tilde{H}_k\}_{k\in\kappa(j)}\big\}_{j\in\mathcal{J}}$. Componentwise convergence of $\tilde{G}_m$ to Γ, is ensured when $D(\mathbb{N}^\star)$ has asymptotically overwhelming probability. The following proposition gives an upper bound on the probability of the complementary of $D(\mathcal{J})$ under certain conditions including the finiteness of $\mathcal{J}$.

Proposition 4.1 *Fix $r_0 > 0, s_0 > 0$ and let $\mathcal{J}$ be a subset of $\{1, 2, \ldots\}$. Assume that $\{T_j\}_{j\in J}$ is a s_0 separated collection of r_0-connected sets. Then, if $\hat{G}_m$ is an estimator of Γ that is consistent from the inside at rate $m^{-\alpha}$, we have*

$$\mathbb{P}_m\big(D^c(\mathcal{J})\big) = \widetilde{O}\big(m^{-\alpha}\big)\,.$$

The r_0-connectedness of all $T_j, j \in \mathcal{J}$ and $\mathrm{Leb}_d(\mathcal{X}) < \infty$ entails that $\mathcal{J}$ is necessarily finite. Nevertheless, the number of connected components of Γ can be infinite as long as there is only a finite number of them for which $\delta_j = \int_{T_j} |2\eta - 1| \mathrm{d}P_X > 0$.

To estimate the homogeneous regions, we will simply estimate the connected components of Γ. In addition, when two connected components T_j and $T_{j'}$ are close with respect to the distance d_∞, we merge[2] them into the same homogeneous region.

It yields the following pseudo-algorithm.

Pseudo-Algorithm

1. Use the unlabeled data $\mathbb{X}_u$ to construct an estimator $\hat{G}_m$ of Γ that is consistent from inside at rate $m^{-\alpha}$.
2. Define homogeneous regions as the unions of the connected components of $\tilde{G}_m = \hat{G}_m \setminus \mathrm{Clip}(\hat{G}_m)$ that are closer than $2(\log m)^{-1}$ for the distance d_∞, accordinf to (4.3) and (4.4).
3. Assign a single label to each estimated homogeneous region by majority vote on labeled data.

[2]Merging two sets means here replacing them by their union

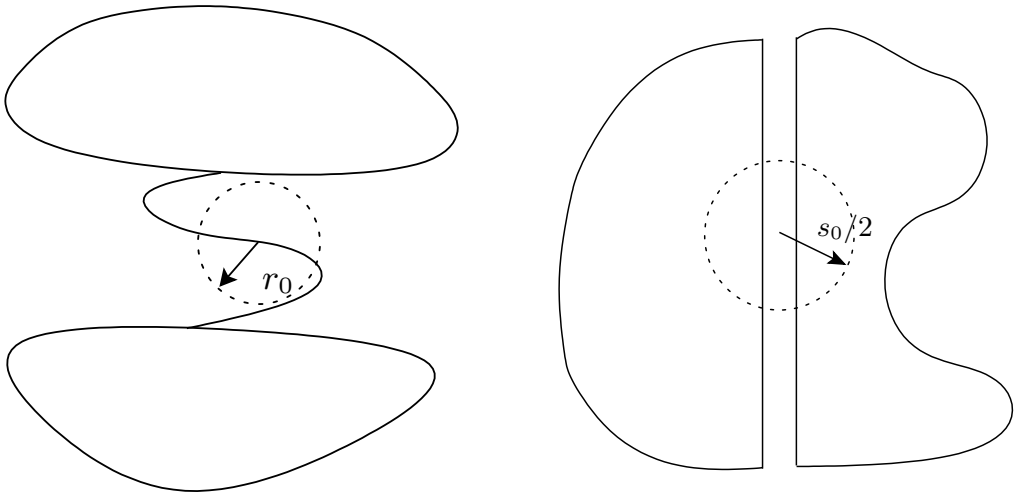

Figure 1: Set that is 0-connected but not r_0-connected for any $r_0 > 0$ (left) and non-separated connected components (right).

This method translates into two distinct error terms, one term in m and another term in n. We apply our three-step procedure to build a classifier $\tilde{g}_{n,m}$ based on the pooled sample $(\mathbb{X}_l, \mathbb{X}_u)$. Fix $\lambda > 0, \alpha > 0$ and let $\hat{G}_m$ be an estimator of the density level set $\Gamma = \{p \geq \lambda\}$, that is consistent from inside with rate $m^{-\alpha}$. For any $k \geq 1$, define the random variable

$$Z^k_{n,m} \triangleq \sum_{i=1}^{n} (2Y_i - 1)\, \mathbb{1}_{\{X_i \in \tilde{H}_k\}}\,,$$

where $\tilde{H}_k$ is defined in (4.4). Denote by $\tilde{g}^k_{n,m}$ the function $\tilde{g}^k_{n,m}(x) = \mathbb{1}_{\{Z^k_{n,m} > 0\}}$ for all $x \in \tilde{H}_k$ and consider the classifier defined on $\mathcal{X}$ by

$$\tilde{g}_{n,m} \triangleq \sum_{k \geq 1} \tilde{g}^k_{n,m}(x) \mathbb{1}_{\{x \in \tilde{H}_k\}}, \quad x \in \mathcal{X}\,. \tag{4.6}$$

Note that the classifier $\tilde{g}_{n,m}$ assigns the label 0 to any x outside of $\tilde{G}_m$. This is a notational convention and we can assign any value to x on this set since we are only interested in the λ-thresholded excess-risk. Nevertheless, it is more appropriate to assign a label referring to a rejection, e.g., the values "2"or "R" (or any other value different from $\{0, 1\}$). The rejection meaning that this point should be classified using labeled data only. However, when the amount of labeled data is too small, it might be more reasonnable not to classify this point at all. This modification is of particular interest in the context of classification with a rejection option when the cost of rejection is smaller than the cost of misclassification (see, e.g., [12]).

Theorem 4.1 *Fix $\lambda > 0, \alpha > 0, r_0 > 0$ and assume that CA(λ) holds. Consider an estimator $\hat{G}_m$ based on $\mathbb{X}_u$ that is consistent from inside with rate $m^{-\alpha}$. Then if the connected components of $\Gamma(\lambda)$ are r_0-connected and s_0-separated, the classifier $\tilde{g}_{n,m}$ defined in* (4.6) *satisfies*

$$\mathcal{E}_\lambda\left(\tilde{g}_{n,m}\right) \leq \widetilde{O}\left(\frac{m^{-\alpha}}{1-\theta}\right) + \sum_{j \geq 1} \delta_j\, e^{-n(\theta\delta_j)^2/2}\,, \tag{4.7}$$

for any $0 < \theta < 1$. *Moreover, if GMA*(λ) *holds, inequality* (4.7) *reduces to*

$$\mathcal{E}_\lambda\left(\tilde{g}_{n,m}\right) \leq \widetilde{O}\left(\frac{m^{-\alpha}}{1-\theta}\right) + e^{-n(\theta\delta)^2/2}\,. \tag{4.8}$$

Note that, since we often have $m \gg n$, the first term in the RHS of (4.7) and (4.8) can be considered negligible so that we achieve an exponential rate of convergence in n which is almost the same (up to the constant θ in the exponent) as in the case where the density p is known. The constant θ seems to be natural since it balances the two terms.

5 Plug-in rules for density level sets estimation

Fix $\lambda > 0$ and recall that our goal is to estimate the connected components $T_j = T_j(\lambda)$, $j = 1, 2, \ldots$, of $\Gamma = \Gamma(\lambda) = \{x \in \mathcal{X} : p(x) \geq \lambda\}$, using the unlabeled sample $\mathbb{X}_u$ of size m. A simple and intuitive way to achieve this goal is to use *plug-in estimators* of Γ defined by

$$\hat{\Gamma} = \hat{\Gamma}(\lambda) \triangleq \{x \in \mathcal{X} : \hat{p}_m(x) \geq \lambda\}\ ,$$

where $\hat{p}_m$ is some estimator of p. A straightforward generalization are the *penalized plug-in estimators* of $\Gamma(\lambda)$, defined by

$$\tilde{\Gamma}_\ell = \tilde{\Gamma}_\ell(\lambda) \triangleq \{x \in \mathcal{X} : \hat{p}_m(x) \geq \lambda + \ell\}\ ,$$

where $\ell > 0$ is a penalization. Clearly $\tilde{\Gamma}_\ell \subset \hat{\Gamma}$. Therefore the connected components of $\tilde{\Gamma}_\ell$ are farther from each other than those of $\hat{\Gamma}$. Keeping in mind that we want estimators that are consistent from inside we are going to consider sufficiently large penalization $\ell = \ell(m)$.

Plug-in rules have a practical advantage over direct methods such as empirical excess mass maximization (see, e.g., [16], [21], [19]). Once we have an estimator $\hat{p}_m$, we can compute the whole collection $\{\tilde{\Gamma}_\ell(\lambda), \lambda > 0\}$, which might be of interest for the user who wants to try several values of λ. Note also that a wide range of density estimators is available in usual software. A density estimator can be parametric, typically based on a mixture model, or nonparametric such as histograms or kernel density estimators.

Definition 5.1 *For any* $\lambda, \gamma \geq 0$, *a function* $f : \mathcal{X} \to \mathbb{R}$ *is said to have* γ*-exponent at level* λ *if there exists a constant* $c_0 > 0$ *such that, for all* $\varepsilon > 0$,

$$\mathrm{Leb}_d\left\{|f(X) - \lambda| \leq \varepsilon\right\} \leq c_0 \varepsilon^\gamma\,.$$

It is an analog of the local margin assumption but for arbitrary level λ in place of $1/2$. When $\gamma > 0$ it ensures that the function f has no flat part at level λ.

The next theorem gives fast rates of convergence for penalized plug-in rules when $\hat{p}_m$ satisfies an exponential inequality and p has γ-exponent at level λ. Moreover, it ensures that when the penalization ℓ is suitably chosen, the plug-in estimator is consistent from inside.

Theorem 5.1 *Fix* $\lambda > 0, \gamma > 0$ *and* $\Delta > 0$. *Let* $\hat{p}_m$ *be an estimator of the density* p *such that* $P_X(\hat{p}_m(X) \geq \lambda) \leq C$, $\mathbb{P}_m$*-almost surely for some positive constant* C *and let* $\mathcal{P}$ *be a class of densities on* $\mathcal{X}$. *Assume that there exist positive constants* c_1, c_2 *and* $a \leq 1$, *such that for* P_X*-almost all* $x \in \mathcal{X}$, *we have*

$$\sup_{p \in \mathcal{P}} \mathbb{P}_m\left(|\hat{p}_m(x) - p(x)| \geq \delta\right) \leq c_1 e^{-c_2 m^a \delta^2}\ ,\ m^{-a/2} < \delta < \Delta\,. \tag{5.1}$$

Assume further that p *has* γ*-exponent at level* λ *and that the penalty* ℓ *is chosen as*

$$\ell = \ell(m) = m^{-\frac{a}{2}} \log m\,. \tag{5.2}$$

Then the plug-in estimator $\tilde{\Gamma}_\ell$ *is consistent from inside at rate* $m^{-\frac{\gamma a}{2}}$.

Consider a kernel density estimator $\hat{p}_m^K$ based on the sample $\mathbb{X}_u$ defined by

$$\hat{p}_m^K(x) \triangleq \frac{1}{mh^d} \sum_{i=n+1}^{n+m} K\left(\frac{X_i - x}{h}\right), \quad x \in \mathcal{X}\,, \tag{5.3}$$

where $h > 0$ is the bandwidth parameter and $K : \mathcal{X} \to \mathbb{R}$ is a kernel. If p is assumed to have Hölder smoothness parameter $\beta > 0$ and if K and h are suitably chosen, it is a standard exercise to prove inequality of type (5.1) with $a = 2\beta/(2\beta + d)$. In that case, it can be shown that the rate $m^{-\frac{\gamma a}{2}}$ is optimal in a minimax sense.

6 Discussion

We proposed a formulation of the cluster assumption in probabilistic terms. This formulation relies on Hartigan's [11] definition of clusters but it can be modified to match other definitions of clusters in the following way.

> Consider a collection of r_0-connected and s_0-separated sets (clusters) $T_j, j = 1, 2, \ldots$.
> Then the function $x \mapsto (\eta(x) - 1/2)$ has constant sign on each T_j.

We also proved that there is no hope to improve the classification performance outside of these clusters. Based on these remarks, we defined the λ-thresholded excess-risk which can be easily generalized to the setup of general clusters defined above. Finally we proved that when we have consistent estimators of the clusters, it is possible to achieve exponential rates of convergence for the λ-thresholded excess-risk. The theory developed here can be extended to any definition of clusters as long as they can be consistently estimated.

Note that our definition of clusters is parametrized by λ which is left to the user, depending on his trust in the cluster assumption. The choice of λ can be made by fixing $P_X(\Gamma^c)$, the probability of the rejection region. We refer to [9] for more details. Note that data-driven choices of λ could be easily derived if we impose a condition on the purity of the clusters, i.e. if we are given the δ in the global margin assumption. Such a choice could be made by decreasing λ until the level of purity is attained. However, any data-driven choice of λ has to be made using the labeled data. It would therefore yield much worse bounds.

General open problems are: applying the cluster assumption to other definitions of clusters and study the whole excess-risk in the framework of semi-supervised classification with a rejection option.

7 Appendix: proofs

7.1 Proof of Theorem 3.1

Using the decomposition of Γ into its connected components, we can decompose $\mathcal{E}_\lambda(\hat{g}_n)$ into

$$\mathcal{E}_\lambda(\hat{g}_n) = \mathbb{E}_n \sum_{j \geq 1} \int_{T_j} |2\eta(x) - 1| \mathbb{1}_{\{\hat{g}_n^j(x) \neq g^\star(x)\}} p(x) \mathrm{d}x\,.$$

Fix $j \in \{1, 2, \ldots\}$ and assume w.l.o.g. that $\eta \geq 1/2$ on T_j. It yields $g^\star(x) = 1,\ \forall\, x \in T_j$, and since $\hat{g}_n$ is also constant on T_j, we get

$$\begin{aligned} \int_{T_j} |2\eta(x) - 1| \mathbb{1}_{\{\hat{g}_n^j(x) \neq g^\star(x)\}} p(x) \mathrm{d}x &= \mathbb{1}_{\{Z_n^j \leq 0\}} \int_{T_j} (2\eta(x) - 1) p(x) \mathrm{d}x \\ &\leq \delta_j \mathbb{1}_{\left\{|\delta_j - \frac{Z_n^j}{n}| \geq \delta_j\right\}}\,, \end{aligned} \tag{7.1}$$

Taking expectation $\mathbb{E}_n$ on both sides of (7.1) we get

$$\begin{aligned}\mathbb{E}_n \int_{T_j} |2\eta(x)-1| \mathbb{1}_{\{\hat{g}_n^j(x) \neq g^\star(x)\}} p(x) \mathrm{d}x &\leq \delta_j \mathbb{P}_n \Big[\big| \delta_j - \frac{Z_n^j}{n} \big| \geq \delta_j \Big] \\ &\leq 2\delta_j \mathrm{e}^{-n\delta_j^2/2} \,,\end{aligned} \tag{7.2}$$

where we used Hoeffding's inequality to get the last inequality. Summing now over j yields the theorem.

7.2 Proof of Proposition 4.1

Define $m_0 \triangleq \exp(1/(r_0 \wedge s_0))$. Since the connected components T_j are r_0-connected, there is only a finite number $J \geq 1$ of them. We simply denote $D(\mathcal{J})$ by D. For any $j = 1, \ldots, J$, the r_0 connectedness of T_j yields on the one hand,

$$\begin{aligned} A_1(j) \triangleq \{\mathrm{card}[\kappa(j)] = 0\} &\subset \big\{ \mathrm{Leb}_d\big[\tilde{G}_m \triangle \Gamma\big] > \lambda c (\log m)^{-d} \big\} \,, \\ A_2(j) \triangleq \{\mathrm{card}[\kappa(j)] \geq 2\} &\subset \big\{ \mathrm{Leb}_d\big[\tilde{G}_m \triangle \Gamma\big] > \lambda c (\log m)^{-d} \big\} \,. \end{aligned}$$

The previous inclusions are illustrated in Figure 2.

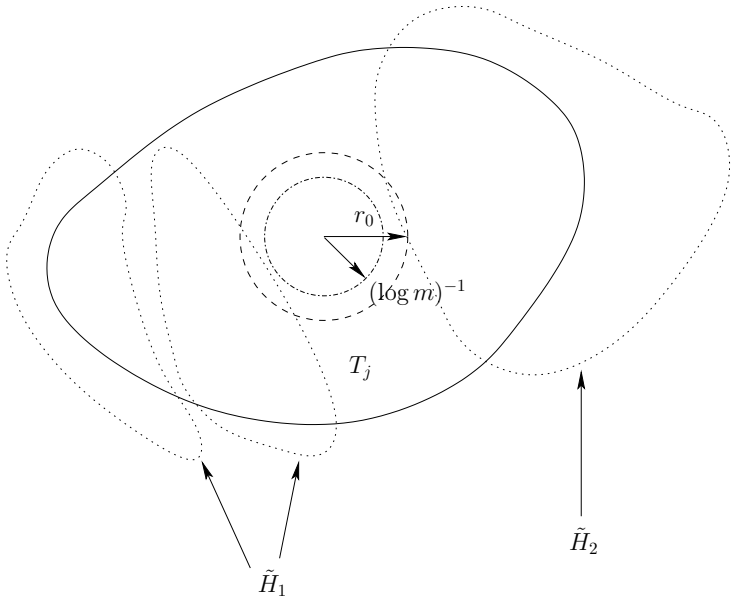

Figure 2: By construction, $\tilde{H}_1$ and $\tilde{H}_2$ are separated by a ball of radius $(\log m)^{-1}$, which is included in $\mathcal{B}(x, r_0)$ when $m \geq m_0$. So if $\{1,2\} \subset \kappa(j)$ or $\kappa(j) = \emptyset$, this ball is included in in $\tilde{\Gamma}_\ell \triangle \Gamma$.

On the other hand, $\kappa(j) \cap \kappa(j') \neq \emptyset$ for some $j' \neq j$ when either (i) $\exists\, l$ s.t. $\tilde{T}_l \cap T_j \neq \emptyset$, $\tilde{T}_l \cap T_{j'} \neq \emptyset$ or (ii) $\exists\, l \neq l'$ s.t. $\tilde{T}_l \cap T_j \neq \emptyset$, $\tilde{T}_{l'} \cap T_{j'} \neq \emptyset$ and $d_\infty(\tilde{T}_l, \tilde{T}_{l'}) < 2(\log m)^{-1}$. Both cases yield the existence of $x \in \Gamma^c \cap \tilde{G}_m$ such that $\mathcal{B}(x, (\log m)^{-1}) \subset \Gamma^c$ for $m \geq m_0$. Therefore

$$\mathrm{Leb}_d(\tilde{G}_m \cap \Gamma^c) \geq \mathrm{Leb}_d(\tilde{G}_m \cap \mathcal{B}(x, (\log m)^{-1}))$$

By construction of $\tilde{G}_m$, we have $\mathrm{Leb}_d(\mathcal{B}(x,(\log m)^{-1}) \cap \tilde{G}_m) \geq m^{-\alpha}(\log m)^{-d}$. Hence

$$A_3(j) \triangleq \bigcup_{j' \neq j} \{\kappa(j) \cap \kappa(j') \neq \emptyset\} \subset \left\{\mathrm{Leb}_d(\tilde{G}_m \cap \Gamma^c) \geq m^{-\alpha}(\log m)^{-d}\right\}$$

Both cases are illustrated in Figure 3.

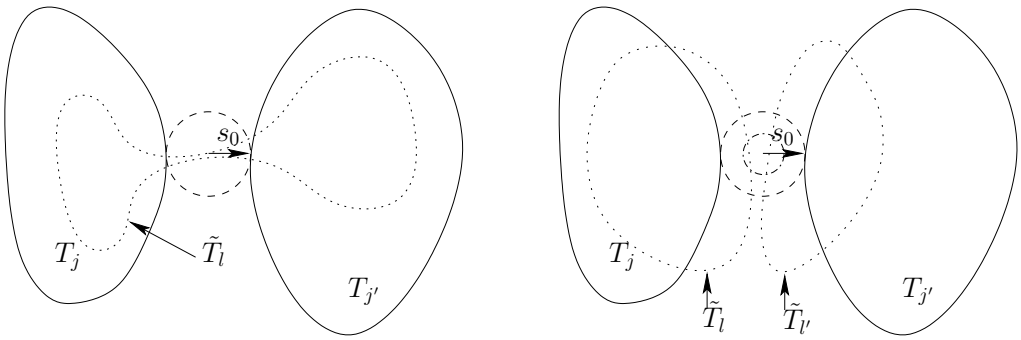

Figure 3: Case (i) (left) and case (ii) (right).

Now, since

$$D^c = \bigcup_{j=1}^{J} A_1(j) \cup A_2(j) \cup A_3(j),$$

we get

$$\mathbb{P}_m(D^c) \leq \mathbb{P}_m\left\{\mathrm{Leb}_d\left[\tilde{G}_m \triangle \Gamma\right] > \lambda c(\log m)^{-d}\right\} + \mathbb{P}_m\left\{\mathrm{Leb}_d(\tilde{G}_m \cap \Gamma^c) \geq m^{-\alpha}(\log m)^{-d}\right\}.$$

Using the Markov inequality for both terms we obtain

$$\mathbb{P}_m\left\{\mathrm{Leb}_d\left[\tilde{G}_m \triangle \Gamma\right] > \lambda c(\log m)^{-d}\right\} = \widetilde{O}\left(m^{-\alpha}\right).$$

and

$$\mathbb{P}_m\left\{\mathrm{Leb}_d(\tilde{G}_m \cap \Gamma^c) \geq m^{-\alpha}(\log m)^{-d}\right\} = \widetilde{O}\left(m^{-\alpha}\right)$$

where we used the fact that $\tilde{G}_m$ is consistent from inside with rate $m^{-\alpha}$. It yields the statement of the proposition.

7.3 Proof of Theorem 4.1

The λ-thresholded excess-risk $\mathcal{E}_\lambda(\tilde{g}_{n,m})$ can be decomposed w.r.t the event D and its complement. It yields

$$\mathcal{E}_\lambda(\tilde{g}_{n,m}) \leq \mathbb{E}_m\left[\mathbb{1}_D \mathbb{E}_n\Big(\int_\Gamma |2\eta(x)-1| \mathbb{1}_{\{\tilde{g}_{n,m}(x) \neq g^\star(x)\}} p(x)\mathrm{d}x \Big| \mathbb{X}_u\Big)\right] + \mathbb{P}_m\left(D^c\right)$$

We now treat the first term of the RHS of the above inequality, i.e., on the event D. Fix $j \in \{1,2,\ldots\}$ and assume w.l.o.g. that $\eta \geq 1/2$ on T_j. Simply write Z^k for $Z^k_{m,n}$. By definition of D, there is a one-to-one correspondence between the collection $\{T_j\}_j$ and the collection $\{\tilde{H}_k\}_k$. We denote by

$\tilde{H}_j$ the unique element of $\{\tilde{H}_k\}_k$ such that $\tilde{H}_j \cap T_j \neq \emptyset$. On D, for any $j \geq 1$, we have,

$$\begin{aligned}
&\mathbb{E}_n\Big(\int_{T_j} |2\eta(x) - 1| \mathbb{1}_{\{\tilde{g}^j_{n,m}(x) \neq g^\star(x)\}} p(x) \mathrm{d}x \Big| \mathbb{X}_u \Big) \\
&\qquad \leq \int_{T_j \setminus \tilde{G}_m} (2\eta - 1) \mathrm{d}P_X + \mathbb{E}_n\Big(\mathbb{1}_{\{Z^j \leq 0\}} \int_{T_j \cap \tilde{H}_j} (2\eta - 1) \mathrm{d}P_X \Big| \mathbb{X}_u \Big) \\
&\qquad \leq L(p) \mathrm{Leb}_d(T_j \setminus \tilde{G}_m) + \delta_j \mathbb{P}_n\big(Z^j \leq 0 | \mathbb{X}_u \big)
\end{aligned}$$

On the event D, For any $0 < \theta < 1$, it holds

$$\begin{aligned}
\mathbb{P}_n\big(Z^j \leq 0 | \mathbb{X}_u \big) &= \mathbb{P}_n\big(\int_{T_j} (2\eta - 1) \mathrm{d}P_X - Z^j \geq \delta_j | \mathbb{X}_u \big) \\
&\leq \mathbb{P}_n\big(\big| Z^j - \int_{\tilde{H}_j} (2\eta - 1) \mathrm{d}P_X \big| \geq \theta \delta_j | \mathbb{X}_u \big) \\
&\qquad + \mathbb{1}_{\left\{ P_X\left[T_j \triangle \tilde{H}_j \right] \geq (1-\theta)\delta_j \right\}} .
\end{aligned}$$

Using Hoeffding's inequality to control the first term, we get

$$\mathbb{P}_n\big(Z^j \leq 0 | \mathbb{X}_u \big) \leq 2\mathrm{e}^{-n(\theta\delta_j)^2/2} + \mathbb{1}_{\left\{ P_X\left[T_j \triangle \tilde{H}_j \right] \geq (1-\theta)\delta_j \right\}} .$$

Taking expectations, and summing over j, the λ-thresholded excess-risk is upper bounded by

$$\mathcal{E}_\lambda(\tilde{g}_{n,m}) \leq \frac{2L(p)}{1-\theta} \mathbb{E}_m\Big[\mathrm{Leb}_d(\Gamma \triangle \tilde{G}_m) \Big] + 2 \sum_{j \geq 1} \delta_j \mathrm{e}^{-n(\theta\delta_j)^2/2} + \mathbb{P}_m\left(D^c \right) ,$$

where we used the fact that on D,

$$\sum_{j \geq 1} \mathrm{Leb}_d\big[T_j \triangle \tilde{H}_j \big] \leq \mathrm{Leb}_d\big[\Gamma \triangle \tilde{G}_m \big] .$$

From Proposition 4.1, we have $\mathbb{P}_m\left(D^c \right) = \widetilde{O}\left(m^{-\alpha} \right)$ and $\mathbb{E}_m\Big[\mathrm{Leb}_d(\Gamma \triangle \tilde{G}_m) \Big] = \widetilde{O}\left(m^{-\alpha} \right)$ and the theorem is proved.

7.4 Proof of Theorem 5.1

Recall that

$$\tilde{\Gamma}_\ell \triangle \Gamma = \left(\tilde{\Gamma}_\ell \cap \Gamma^c \right) \sqcup \left(\tilde{\Gamma}_\ell^c \cap \Gamma \right) .$$

We begin by the first term. We have

$$\tilde{\Gamma}_\ell \cap \Gamma^c = \big\{ x \in \mathcal{X} : \hat{p}_m(x) \geq \lambda + \ell, p(x) < \lambda \big\} \subset \big\{ x \in \mathcal{X} : |\hat{p}_m(x) - p(x)| \geq \ell \big\} .$$

The Fubini theorem yields

$$\mathbb{E}_m\big[\mathrm{Leb}_d(\tilde{\Gamma}_\ell \cap \Gamma^c) \big] \leq \mathrm{Leb}_d(\mathcal{X}) \sup_{x \in \mathcal{X}} \mathbb{P}_m\left[|\hat{p}_m(x) - p(x)| \geq \ell \right] \leq c_3 \mathrm{e}^{-c_2 m^a \ell^2} ,$$

where the last inequality is obtained using (5.1) and $c_3 = c_1 \mathrm{Leb}_d(\mathcal{X}) > 0$. Taking ℓ as in (5.2) yields for $m \geq \exp(\gamma a / c_2)$,

$$\mathbb{E}_m\big[\mathrm{Leb}_d(\tilde{\Gamma}_\ell \cap \Gamma^c) \big] \leq c_3 m^{-\gamma a}. \tag{7.3}$$

We now prove that $\mathbb{E}_m\big[\mathrm{Leb}_d(\tilde{\Gamma}_\ell \cap \Gamma^c)\big] = \widetilde{O}\big(m^{-\frac{\gamma a}{2}}\big)$. Consider the following decomposition where we drop the dependence in x for notational convenience,

$$\tilde{\Gamma}_\ell^c \cap \Gamma = B_1 \cup B_2,$$

where

$$B_1 = \big\{\hat{p}_m < \lambda + \ell, p \geq \lambda + 2\ell\big\} \subset \big\{|\hat{p}_m - p| \geq \ell\big\}$$

and

$$B_2 = \big\{\hat{p}_m < \lambda + \ell, \lambda \leq p(x) < \lambda + 2\ell\big\} \subset \big\{|p - \lambda| \leq \ell\big\}.$$

Using (5.1) and (5.2) in the same fashion as above we get $\mathbb{E}_m\big[\mathrm{Leb}_d(B_1)\big] = \widetilde{O}\big(m^{-\frac{\gamma a}{2}}\big)$. The term corresponding to B_2 is controlled using the γ-exponent of density p at level λ. Indeed, we have

$$\mathrm{Leb}_d(B_2) \leq c_0 \ell^\gamma \leq c_0 (\log m)^\gamma m^{-\frac{\gamma a}{2}} = \widetilde{O}\big(m^{-\frac{\gamma a}{2}}\big)$$

The previous upper bounds for B_1 and B_2 together with (7.3) yield the consistency from inside.

References

[1] d'Alché Buc F., Grandvalet Y. and Ambroise, C. Semi-supervised MarginBoost. In *NIPS*, 2002.

[2] Audibert, J.-Y. and Tsybakov, A. Fast learning rates for plug-in classifiers under the margin condition. Manuscript, 2005.

[3] Balcan, M.F. and Blum, A. A PAC-style Model for Learning from Labeled and Unlabeled Data. In *COLT*, 2005.

[4] Blum, A. and Mitchel, T. Combining Labeled and Unlabeled Data with Co-Training. In *COLT*, 1998.

[5] Boucheron, S., Bousquet, 0. and Lugosi, M. Theory of classification: some recent advances. *ESAIM Probability & Statistics*, **9** (2005) 323-375.

[6] Bousquet, O., Chapelle, O. and Hein, M.: Measure Based Regularization. In *NIPS*, 2004.

[7] Chapelle, O. and A. Zien: Semi-Supervised Classification by Low Density Separation. *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics* (2005) 57-64.

[8] Chapelle, O., Schölkopf, B. and Zien, A.: *Semi-Supervised Learning*. MIT Press, Cambridge, MA (*in press*) (2006)

[9] Cuevas, A., Febrero, M., and Fraiman, R. Cluster analysis: a further approach based on density estimation. *Comput. Statist. Data Anal.* **36** (2001) 441-459.

[10] Devroye, L., Györfi, L., Lugosi, G.: *A Probabilistic Theory of Pattern Recognition*. Springer, N.Y. (1996).

[11] Hartigan, J. A. *Clustering algorithms*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, New-York London Sydney, 1975.

[12] Herbei, R. and Wegkamp, M.: Classification with rejection option. Manuscript, 2005.

[13] Hertz, T., Bar-Hillel, A. and Weinshall, D. Boosting Margin-Based Distance Functions for Clustering. In *Proceedings of 21st International Conference on Machine Learning (ICML) 2004*, 2004.

[14] Korostelev, A. P. and Tsybakov, A. B. *Minimax theory of image reconstruction*. Lecture Notes in Statistics, **82**. Springer-Verlag, New York, 1993.

[15] Mammen, E. and Tsybakov, A. B. Smooth discrimination analysis. *Ann. Statist.*, **27** (1999) 1808-1829

[16] Polonik, W. Measuring mass concentrations and estimating density contour clusters—an excess mass approach. *Ann. Statist.* **23** (1995) 855-881.

[17] Rattray, M. A model-based distance for clustering. In *Proceedings of IJCNN*, 2000.

[18] Seeger, M.: Learning with labeled and unlabeled data. Technical Report, 2000.

[19] Steinwart, I., Hush, D. and Scovel, C. Density level detection is classification. In *NIPS*, 2005.

[20] Tipping, M. E. Deriving cluster analytic distance functions from Gaussian mixture models. In *Proceedings of the Ninth International Conference on Artificial Neural Networks*, **2** (1999) 815-820.

[21] Tsybakov, A. B. On nonparametric estimation of density level sets. *Ann. Statist.* **25** (1997) 948-969.

[22] Tsybakov, A. B. Optimal aggregation of classifiers in statistical learning. *Ann. Statist.*, **32** (2004) 135-166.

[23] Zhu, X. (2005). *Semi-supervised learning with graphs*. Doctoral dissertation, Carnegie Mellon University, 2005.