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

Philippe Rigollet

Journal of Machine Learning Research X (200X) X-X Submitted XX/XX; Published XX/XX

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

Philippe Rigollet RIGOLLET@MATH.GATECH.EDU
School of Mathematics
Georgia Institute of Technology
Atlanta, GA 30332-0160, U.S.A

Editor: Gábor Lugosi

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 (2000) 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.

Keywords: 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 mainly two approaches to solve this problem. The first one consists in using the unlabeled data to reduce the *complexity* of the problem in a broad sense. For instance, assume 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 set of candidate classifiers (see, e.g., Balcan and Blum, 2005; Blum and Mitchell, 1998). Unlabeled data can also be used to reduce the dimension of the problem, which is another way to reduce complexity. For example, in Belkin and Niyogi (2004), it is assumed that the data actually live on a submanifold of low dimension.

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 (Hartigan, 1975) definition of clusters, namely the connected components of the density level sets. However, they use a parametric–usually mixture–model to estimate the under-

lying density which can be far from reality. Moreover, no generalization error bounds are available for such methods. In the same spirit, Tipping (1999) and Rattray (2000) 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 Zhu (2005). 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 Bousquet et al., 2004; Chapelle and Zien, 2005; Chapelle et al., 2006).

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 (d'Alché Buc et al., 2001; Hertz et al., 2004) or SVM (Bousquet et al., 2004; Chapelle and Zien, 2005), 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., Seeger (2000); Zhu (2005); Chapelle et al. (2006). 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 supervised 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 Section 2. The main result of this section is Proposition 2.1 which essentially states that the effect of unlabeled data on the rates of convergence cannot be observed on the whole excess-risk. We therefore introduce the *cluster excess-risk* which corresponds to a part of the excess-risk that is interesting for this problem. In Section 3, we study the population case where the clusters are perfectly known, to get an idea of our target. Indeed, such a population case corresponds in some way to the case where the amount of unlabeled data is infinite. Section 4 contains the main result: after having defined the clusters in terms of density level sets, we propose an algorithm for which we derive rates of convergence for the cluster excess-risk as a measure of performance. An example of consistent density level set estimators is given in Section 5. Section 6 is devoted to a discussion on the choice of λ as well as possible implementations and improvements. Proofs of the results are gathered in Section 7.

Notation. Throughout the paper, we denote positive constants by c_j. 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 $\alpha > 0, C > 0$. Moreover, we write $u_p = o(v_p)$, if there exists a non negative sequence $(\varepsilon_p)_p$ that tends to 0 when p tends to infinity and such that $|u_p| \leq \varepsilon_p |v_p|$. 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 the sample $(X_1, Y_1), \ldots, (X_n, Y_n)$. We denote hereafter by $\mathbb{P}_n$ the corresponding probability.

In many applications, a large amount of unlabeled data is available together with 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 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 $m \gg n$. Most existing theoretical studies of supervised classification use empirical processes theory (Devroye et al., 1996; Vapnik, 1998; van de Geer, 2000; Boucheron et al., 2005) to obtain rates of convergence for the excess-risk that are polynomial in n. Typically these rates are of the order $O(1/\sqrt{n})$ and can be as small as $\widetilde{O}(1/n)$ under some low noise assumptions (cf., e.g., Tsybakov, 2004; Audibert and Tsybakov, 2007). However, simulations indicate that much faster rates should be attainable when unlabeled data is used to identify homogeneous clusters. Of course, 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) (see, e.g. Zhang and Oles, 2000). Seeger (2000) 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.

1. The notation is adapted to the present framework.

(*i*) Assume $P[Y = Y'|X, X' \in C] \geq P[Y \neq Y'|X, X' \in C]$, where C is a cluster.

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

Assume for the moment that we know what the clusters are, so that we do not have to define them in terms of density level sets. This will be done in Section 4. Let $T_1, T_2, \ldots,$ be a countable family of subsets of $\mathcal{X}$. We now make the assumption that the T_j's are clusters of homogeneous data.

Cluster Assumption (CA) Let $T_1, T_2, \ldots,$ be a collection of measurable sets (clusters) such that $T_j \subset \mathcal{X}, j = 1, 2, \ldots$ Then the function $x \in \mathcal{X} \mapsto \mathbb{1}\{\eta(x) \geq 1/2\}$ takes a constant value on each of the $T_j, j = 1, 2, \ldots$.

It is not hard to see that the cluster assumption **(CA)** is equivalent to the following assumption.

> Let $T_j, j = 1, 2, \ldots,$ be a collection of measurable sets such that $T_j \subset \mathcal{X}$, $j = 1, 2, \ldots$ Then, for any $j = 1, 2, \ldots,$ we have
>
> $$P[Y = Y'|X, X' \in T_j] \geq P[Y \neq Y'|X, X' \in T_j]\,.$$

A question remains: what happens outside of the clusters? Define the union of the clusters,

$$\mathcal{C} = \bigcup_{j \geq 1} T_j \tag{1}$$

and assume that we are in the problematic case, $P_X(\mathcal{C}^c) > 0$ such that the question makes sense. Since the cluster assumption **(CA)** says nothing about what happens outside of the set $\mathcal{C}$, we can only perform supervised classification on $\mathcal{C}^c$. Consider 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 Devroye et al., 1996),

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

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 $\mathcal{C}^c$, any reasonable classifier should be based on the sample $\mathbb{X}_l$ so that $\hat{g}_{n,m}(x) = \hat{g}_n(x),\ \forall\, x \in \mathcal{C}^c$, and we have

$$\mathcal{E}(\hat{g}_{n,m}) \geq \mathbb{E}_n \int_{\mathcal{C}^c} |2\eta(x) - 1| \mathbb{1}_{\{\hat{g}_n(x) \neq g^\star(x)\}} \mathrm{d}P_X(x)\,. \tag{2}$$

Since we assumed $P_X(\mathcal{C}^c) \neq 0$, the RHS of (2) is bounded from below by the optimal rates of convergence that appear in supervised classification.

The previous heuristics can be stated more formally as follows. Recall that the distribution P of the random couple (X, Y) is completely characterized by the couple (P_X, η) where P_X is the marginal distribution of X and η is the regression function of Y on X. In the following proposition, we are interested in a class of distributions with cylinder form, i.e. a class $\mathcal{D}$ that can be decomposed as $\mathcal{D} = \mathcal{M} \times \Xi$ where $\mathcal{M}$ is a fixed class of marginal distributions on $\mathcal{X}$ and Ξ is a fixed class of regression functions on $\mathcal{X}$ with values in $[0, 1]$.

Proposition 2.1 *Fix $n, m \geq 1$ and let $\mathcal{C}$ be a measurable subset of $\mathcal{X}$. Let $\mathcal{M}$ be a class of marginal distributions on $\mathcal{X}$ and let Ξ be a class of regression functions. Define the class of distributions $\mathcal{D}$ as $\mathcal{D} = \mathcal{M} \times \Xi$. Then, for any marginal distribution $P_X^0 \in \mathcal{M}$, we have*

$$\inf_{T_n} \sup_{\eta \in \Xi} \mathbb{E}_n \int_{\mathcal{C}^c} |2\eta - 1| \mathbb{1}_{\{T_n \neq g^\star\}} \mathrm{d}P_X^0 \leq \inf_{T_{n,m}} \sup_{P \in \mathcal{D}} \mathbb{E}_{n,m} \int_{\mathcal{C}^c} |2\eta - 1| \mathbb{1}_{\{T_{n,m} \neq g^\star\}} \mathrm{d}P_X \,, \tag{3}$$

where $\inf_{T_{n,m}}$ denotes the infimum over all classifiers based on the pooled sample $(\mathbb{X}_l, \mathbb{X}_u)$ and $\inf_{T_n}$ denotes the infimum over all classifiers based only on the labeled sample $\mathbb{X}_l$.

The main consequence of Proposition 2.1 is that even when the cluster assumption **(CA)** is valid the unlabeled data are useless to improve the rates of convergence. If the class $\mathcal{M}$ is reasonably large and satisfies $P_X^0(\mathcal{C}^c) > 0$, the left hand side in (3) can be bounded from below by the minimax rate of convergence with respect to n, over the class $\mathcal{D}$. Indeed a careful check of the proofs of minimax lower bounds reveals that they are constructed using a single marginal P_X^0 that is well chosen. These rates are typically of the order $n^{-\alpha}, 0 < \alpha \leq 1$ (see e.g. Mammen and Tsybakov (1999); Tsybakov (2004); Audibert and Tsybakov (2007) and Boucheron et al. (2005) 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 *cluster excess-risk* of a classifier $\hat{g}_{n,m}$ defined by

$$\mathcal{E}_{\mathcal{C}}(\hat{g}_{n,m}) \triangleq \mathbb{E}_{n,m} \int_{\mathcal{C}} |2\eta(x) - 1| \mathbb{1}_{\{\hat{g}_{n,m}(x) \neq g^\star(x)\}} \mathrm{d}P_X(x) \,. \tag{4}$$

We will therefore focus on this measure of performance. The cluster excess-risk can also be expressed in terms of an excess-risk. To observe it, define the set $\mathcal{G}_{\mathcal{C}}$ of all classifiers restricted to $\mathcal{C}$:

$$\mathcal{G}_{\mathcal{C}} = \big\{ g : \mathcal{C} \to \{0,1\},\ g \text{ measurable} \big\} \,.$$

The performance of a classifier $g \in \mathcal{G}_{\mathcal{C}}$ is measured by the average classification error on $\mathcal{C}$

$$R(g) = P\big(g(X) \neq Y\big) = P\big(g(X) \neq Y, X \in \mathcal{C}\big)$$

A minimizer of $R(\cdot)$ over $\mathcal{G}_{\mathcal{C}}$ is given $g^\star_{|\mathcal{C}}(x) = \mathbb{1}_{\{\eta(x) \geq 1/2\}}$, $x \in \mathcal{C}$, i.e., the restriction of the Bayes classifier to $\mathcal{C}$. Now it can be easily shown that for any classifier $g \in \mathcal{G}_{\mathcal{C}}$ we have,

$$R(g) - R(g^\star_{|\mathcal{C}}) = \int_{\mathcal{C}} |2\eta(x) - 1| \mathbb{1}_{\{g(x) \neq g^\star_{|\mathcal{C}}(x)\}} \mathrm{d}P_X(x) \,. \tag{5}$$

Taking expectations on both sides of (5) with $g = \hat{g}_{n,m}$, it follows that

$$\mathbb{E}_{n,m} R(\hat{g}_{n,m}) - R(g^\star_{|\mathcal{C}}) = \mathcal{E}_{\mathcal{C}}(\hat{g}_{n,m}) \,.$$

Therefore, cluster excess-risk equals the excess-risk of classifiers in $\mathcal{G}_{\mathcal{C}}$. In the sequel, we only consider classifiers $\hat{g}_{n,m} \in \mathcal{G}_{\mathcal{C}}$, i.e., classifiers that are defined on $\mathcal{C}$.

We now propose a method to obtain good upper bounds on the cluster excess-risk, taking advantage of the cluster assumption **(CA)**. 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 clusters

Consider the ideal situation where the family $T_1, T_2, \ldots$, is known and we observe only the labeled sample $\mathbb{X}_l = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$. Define

$$\mathcal{C} = \bigcup_{j \geq 1} T_j \, .$$

Under the cluster assumption **(CA)**, 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 $\delta_j \geq 0$, $\delta_j \leq 1$ by

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

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

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

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

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

The following theorem gives rates of convergence for the cluster excess-risk of the classifier $\hat{g}_n$ under **(CA)** that can be exponential in n under a mild additional assumption.

Theorem 3.1 *Let $T_j, j \geq 1$ be a family of measurable sets that satisfy Assumption* **(CA)**. *Then, the classifier $\hat{g}_n$ defined above satisfies*

$$\mathcal{E}_{\mathcal{C}}(\hat{g}_n) \leq 2 \sum_{j \geq 1} \delta_j e^{-n\delta_j^2/2} \, . \tag{6}$$

Moreover, if there exists $\delta > 0$ such that $\delta = \inf_j \{\delta_j : \delta_j > 0\}$, we obtain an exponential rate of convergence:

$$\mathcal{E}_{\mathcal{C}}(\hat{g}_n) \leq 2e^{-n\delta^2/2} \, . \tag{7}$$

In a different framework, Castelli and Cover (1995, 1996) have proved that exponential rates of convergence were attainable for semi-supervised classification. A rapid overview of the proof shows that the rate of convergence $e^{-n\delta^2/2}$ cannot be improved without further assumption. It will be our target in semi-supervised classification. However, we need estimators of the clusters $T_j, j = 1, 2, \ldots$. In the next section we provide the main result on semi-supervised learning, that is when the clusters are unknown but we can estimate them using the unlabeled sample $\mathbb{X}_u$.

4. Main result

We now deal with a more realistic case where the clusters $T_1, T_2, \ldots$, are unknown and we have to estimate them using the unlabeled sample $\mathbb{X}_u = \{X_1, \ldots, X_m\}$. We begin by giving a definition of the clusters in terms of density level sets. In this section, we assume that $\mathcal{X}$ has finite Lebesgue measure.

4.1 Definition of the clusters

Following Hartigan (1975), we propose a definition of clusters that is also compatible with the expression "regions of relatively high P_X" proposed by Seeger (2000).

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

$$\Gamma(\lambda) = \{x \in \mathcal{X} : p(x) \geq \lambda\} \ . \tag{8}$$

On these sets, the density is relatively high. The cluster assumption involves also a notion of connectedness of a set. For any $C \subset \mathcal{X}$, define the binary relation $\mathcal{R}$ on any set C as follows: two points $x, y \in C$ satisfy $x\mathcal{R}y$ if and only if there exists a continuous map $f : [0,1] \to C$, such that $f(0) = x$ and $f(1) = y$. If $x\mathcal{R}y$, we say that x and y are *pathwise connected*. It can be easily show that $\mathcal{R}$ is an equivalence relation and its classes of equivalence are called *connected components* of C. At this point, in view of the formulation of the cluster assumption, it is very tempting to define the clusters as the connected components of C. However, this definition suffers from two major flaws:

1. a connected set cannot be defined up to a set of null Lebesgue measure. Indeed, consider for example the case $d = 1$ and $C = [0,1]$. This set is obviously connected (take the map f equal to the identity on $[0,1]$) but the set $\widetilde{C} = C \setminus \{1/2\}$ is not connected anymore even though C and $\widetilde{C}$ only differ by a set of null Lebesgue measure. In our setup we want to impose connectedness on certain subsets of the λ-level set of the density p which is actually defined up to a set of null Lebesgue measure. Figure 1 (left) is an illustration of a set with one connected component whereas it is desirable to have two clusters.

2. There is no scale consideration in this definition of clusters. When two clusters are too close to each other in a certain sense, we wish identify them as a single cluster. In Figure 1 (right), the displayed set has two connected components whereas we wish to identify only one cluster.

To fix the first flaw, we introduce the following notions. Let $\mathcal{B}(z,r)$ be the d-dimensional closed ball of center $z \in \mathbb{R}^d$ and radius $r > 0$, defined by

$$\mathcal{B}(z,r) = \Big\{x \in \mathbb{R}^d : \|z - x\| \leq r\Big\} \ ,$$

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

Definition 4.1 *Fix $r_0 \geq 0$ and let $\bar{d}$ be an integer such that $\bar{d} \geq d$. We say that a measurable set $C \subset \mathcal{X}$ is* r_0-standard *if for any $z \in C$ and any $0 \leq r \leq r_0$, we have*

$$\mathrm{Leb}_d\big(\mathcal{B}(z,r) \cap C\big) \geq c_0 r^{\bar{d}} \ . \tag{9}$$

We now comment upon this definition.

Remark 4.1 *The definition of a standard set has been introduced by Cuevas and Fraiman (1997). This definition ensures that the set C has no "flat" parts which allows to exclude pathological cases such as the one presented on the left hand side of Figure 1.*

Remark 4.2 *The constant c_0 may depend on r_0 and this avoids large-scale shape considerations. Indeed, if the set C is bounded, then for any $z \in C$, $\mathrm{Leb}_d\big(\mathcal{B}(z,r) \cap C\big) = \mathrm{Leb}_d\big(C\big)$ for $r \geq r_0$ where r_0 is the diameter of C. Thus for C to be r_0-standard, we have to impose at least that $c_0 \leq \mathrm{Leb}_d\big(C\big) r_0^{-\bar{d}}$.*

Remark 4.3 *The case $\bar{d} > d$ allows us to include a wide variety of shapes in this definition. Consider the following example where $d = 2$:*

$$C_\delta = \big\{(x,y) \,:\, -1 \leq x \leq 1,\, 0 \leq y \leq |x|^\delta\big\}, \quad \delta > 0$$

Fix $r \leq \sqrt{2}$ and consider the point $z = (0,0)$. It holds

$$\mathrm{Leb}_d\big(\mathcal{B}(z,r) \cap C_\delta\big) \geq \int_{-r'}^{r'} \min(|x|^\delta, r')\mathrm{d}x\,, \quad \text{where } r' = \frac{r}{\sqrt{2}}\,.$$

For any $|x| \leq r' \leq 1$, we have $|x|^\delta \geq |x|^{(\delta \vee 1)}$ and $|x|^{(\delta\vee 1)} \leq |r'|^{(\delta\vee 1)} \leq r'$. Thus

$$\mathrm{Leb}_d\big(\mathcal{B}(z,r) \cap C_\delta\big) \geq \int_{-r'}^{r'} |x|^{(\delta\vee 1)}\mathrm{d}x = 2(r')^{(\delta\vee 1)+1}\,.$$

We conclude that (9) is satisfied at $z = (0,0)$ for $\bar{d} = (\delta \vee 1) + 1$. However, notice that

$$\mathrm{Leb}_d\big(\mathcal{B}(z,r) \cap C_\delta\big) \leq \int_{-r}^{r} |x|^{\delta}\mathrm{d}x = 2r^{(\delta+1)}\,.$$

Thus (9) is not satisfied at $z = (0,0)$ when $\bar{d} = d$, if $\delta > 1$.

To overcome the scale problem described in the second flaw, we introduce the notion of s_0-separated sets.

Define the pseudo-distance distance d_∞, between two sets C_1 and C_2 by

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

We say that two sets C_1, C_2, are *s_0-separated* if $d_\infty(C_1, C_2) > s_0$, for some $s_0 \geq 0$. More generally, we say that the sets $C_1, C_2, \ldots$ are *mutually s_0-separated* if for any $j \neq j'$, C_j and $C_{j'}$ are s_0-separated. On the right hand side of Figure 1, we show an example of two sets that are not s_0-separated for a reasonable s_0. In that particular example, if s_0 is sufficiently small, we would like to identify a single cluster.

We now define s_0-connectedness which is a weaker version of connectedness in the form of a binary relation

Definition 4.2 *Fix $s > 0$ and let $\overset{s}{\underset{C}{\longleftrightarrow}}$ be the binary relation defined on $C \subset \mathcal{X}$ as follows: two points $x, y \in C$ satisfy $x \overset{s}{\underset{C}{\longleftrightarrow}} y$ if and only if there exists a piecewise constant map $f : [0,1] \to C$ such that $f(0) = x$ and $f(1) = y$ and such that f has a* finite *number of jumps that satisfy $\|f(t_+) - f(t_-)\| \leq s$ for any $t \in [0,1]$, where*

$$f(t_+) = \lim_{\substack{\theta \to t \\ \theta > t}} f(\theta) \quad \text{and} \quad f(t_-) = \lim_{\substack{\theta \to t \\ \theta < t}} f(\theta)\,.$$

If $x \overset{s}{\underset{C}{\longleftrightarrow}} y$, we say that x and y are s-connected.

Note that x and y are *s-connected* if and only if there exists $z_1, \ldots, z_n \in C$ such that $\|x - z_1\| \leq s$, $\|y - z_n\| \leq s$ and $\|z_i - z_{i+1}\| \leq s$ for any $j = 1, \ldots, n-1$. In other words, there exists a finite sequence of points in C that links x to y and such that two consecutive points in this sequence have distance smaller that s.

Lemma 4.1 *Fix $s > 0$, then the binary relation $\overset{s}{\underset{C}{\longleftrightarrow}}$ is an equivalence relation and C can be partitioned into its classes of equivalence. The classes of equivalence of $\overset{s}{\underset{C}{\longleftrightarrow}}$ are called* s-connected components of C.

In the next proposition we prove that given a certain scale $s > 0$, it is possible split a r_0-standard and closed set C into a unique partition that is a coarser than the partition defined by the connected components of C and that this partition is finite for such sets.

Proposition 4.1 *Fix $r_0 > 0, s > 0$ and assume that C is a r_0-standard and closed set. Then there exists a unique partition $C_1, \ldots C_J$, $J \geq 1$, of C such that*

- *for any $j = 1, \ldots, J$ and any $x, y \in C_j$, we have $x \overset{s}{\underset{C}{\longleftrightarrow}} y$,*
- *the sets $C_1, \ldots, C_J$ are mutually s-separated.*

Remark 4.4 *In what follows we assume that the scale $s = s_0$ is fixed by the statistician. It should be fixed depending on a priori considerations about the scale of the problem. Actually, in the proof of Proposition 4.3, we could even assume that $s_0 = 1/(3 \log m)$, which means that we can have the scale depend on the number of observations. This is consistent with the fact that the finite number of unlabeled observations allows us to have only a blurred vision of the clusters. In this case, we are not able to differentiate between two clusters that are too close to each other but our vision becomes clearer and clearer as m tends to infinity.*

We now formulate the cluster assumption when the clusters are defined in terms of density level sets. In the rest of the section, fix $\lambda > 0$ and let Γ denote the λ-level set of the density p. We also assume in what follows that Γ is closed which is the case if the density p is continuous for example.

Strong Cluster Assumption (SCA) Fix $s_0 > 0$ and $r_0 > 0$ and assume that Γ admits a version that is r_0-standard and closed. Denote by $T_1, \ldots, T_J$ the s_0-connected components of this version of Γ. Then the function $x \in \mathcal{X} \mapsto \mathbb{1}_{\{\eta(x) \geq 1/2\}}$ takes a constant value on each of the $T_j, j = 1, \ldots J$.

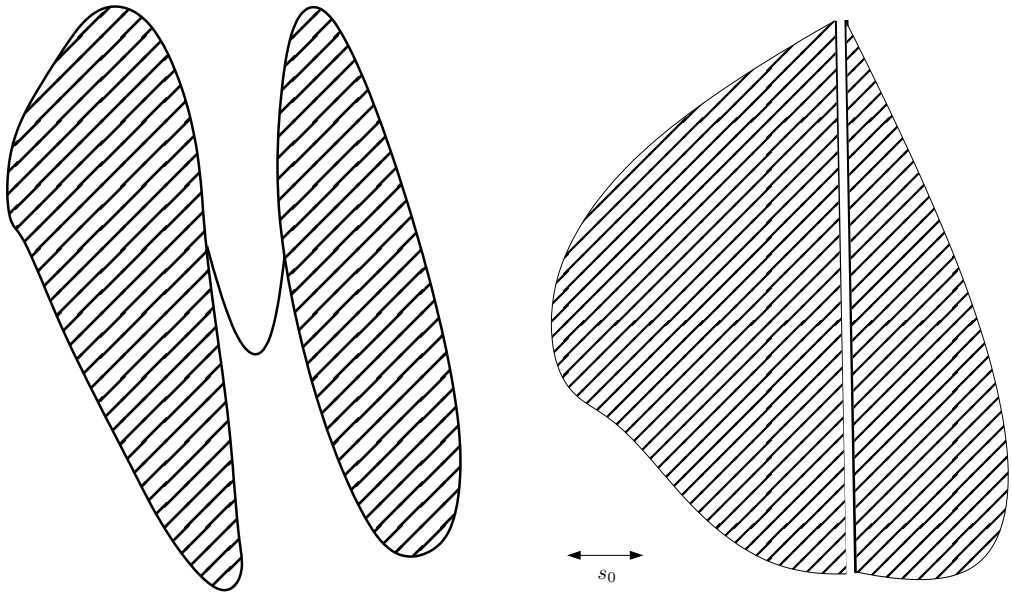

Figure 1: A set that is not r_0-standard for any r_0 (left). A set that has two connected components but only one s_0-connected components (right).

4.2 Estimation of the clusters

Assume that p is uniformly bounded by a constant $L(p)$ and that $\mathcal{X}$ is bounded. 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 we use the sample $\mathbb{X}_u$ to construct an estimator $\hat{G}_m$ of Γ satisfying

$$\mathbb{E}_m\big[\mathrm{Leb}_d(\hat{G}_m \triangle \Gamma)\big] \to 0, \quad m \to +\infty\,, \tag{10}$$

where $\triangle$ is the sign for the symmetric difference. We call such estimators *consistent* estimators of Γ. Recall that we are interested in identifying the s_0-connected components $T_1, \dots, T_J$ of Γ. That is, we seek a partition of $\hat{G}_m$, denoted here by $\hat{H}_1, \dots, \hat{H}_{J'}$ such that for any $j = 1, \dots, J$, $\hat{H}_j$ is a consistent estimator of T_j and $\mathbb{E}_m\big[\mathrm{Leb}_d(\hat{H}_j)\big] \to 0$ for $j > J$. From Proposition 4.1, we know that for any $1 \le j, j' \le J$, $j \neq j'$, we have $d_\infty(T_j, T_{j'}) > s_0$. Let $\bar{s} > s_0$ be defined by

$$\bar{s} = \min_{j \neq j'} d_\infty(T_j, T_{j'})\,. \tag{11}$$

To define the partition $\hat{H}_1, \dots, \hat{H}_{J'}$, it is therefore natural to use a suitable reordering of the $(s_0 + u_m)$-connected components of $\hat{G}_m$, where u_m is a positive sequence that tends to 0 as m tends to infinity. Since the measure of performance $\mathbb{E}_m\big[\mathrm{Leb}_d(\hat{G}_m \triangle \Gamma)\big]$ is defined up to a set of null Lebesgue measure it may be the case that even an estimator $\hat{G}_m$ that satisfies $\mathbb{E}_m\big[\mathrm{Leb}_d(\hat{G}_m \triangle \Gamma)\big] = 0$ has only one (s_0+u_m)-connected components whereas Γ has several s_0-connected components. This happens for example in the case where $\hat{G}_m = \Gamma \cup R$ where R is a set of thin ribbons with null Lebesgue measure that link the s_0-connected components of Γ to each other (see Figure 1, left). If $\hat{G}_m$ were r_0-standard, such configurations would

not occur. To have $\hat{G}_m$ more "standard", we apply the following *clipping* transformation: define the set

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

In the sequel, we will only consider the clipped version of $\hat{G}_m$ defined by $\tilde{G}_m = \hat{G}_m \setminus \mathrm{Clip}(\hat{G}_m)$. For any $x \in \tilde{G}_m$, we have

$$\mathrm{Leb}_d\big(\hat{G}_m \cap \mathcal{B}(x, (\log m)^{-1})\big) > \frac{(\log m)^{-d}}{m^\alpha}\,.$$

However, this is not enough to ensure that the union of several s_0-connected components of Γ is not estimated by a single $(s_0 + u_m)$-connected component of $\tilde{G}_m$ due to the magnitude of random fluctuations of $\tilde{G}_m$ around Γ.

To ensure componentwise consistency, we make assumptions on the estimator $\hat{G}_m$. Note that the performance of a density level set estimator $\hat{G}_m$ is measured by the quantity

$$\mathbb{E}_m\big[\mathrm{Leb}_d(\hat{G}_m \triangle \Gamma)\big] = \mathbb{E}_m\big[\mathrm{Leb}_d(\hat{G}_m^c \cap \Gamma)\big] + \mathbb{E}_m\big[\mathrm{Leb}_d(\hat{G}_m \cap \Gamma^c)\big]\,. \tag{12}$$

For some estimators, such as the offset plug-in density level sets estimators presented in Section 5, we can prove that the dominant term in the RHS of (12) is $\mathbb{E}_m\big[\mathrm{Leb}_d(\hat{G}_m^c \cap \Gamma)\big]$. It yields that the probability of having Γ included in the consistent estimator $\hat{G}_m$ is negligible. We now give a precise definition of such estimators.

Definition 4.3 *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

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

The following proposition ensures that the clipped version of an estimator that is consistent from inside is also consistent from inside at the same rate.

Proposition 4.2 *Fix $\alpha > 0, s_0 > 0$ and let (u_m) be a positive sequence. Assume that $\mathcal{X}$ is bounded and let $\hat{G}_m$ be an estimator of Γ that is consistent from inside at rate $m^{-\alpha}$. Then, the clipped estimator $\tilde{G}_m = \hat{G}_m \setminus \mathrm{Clip}(\hat{G}_m)$ is also consistent from inside a rate $m^{-\alpha}$ and has a finite number $\tilde{K}_m \leq \mathrm{Leb}_d(\mathcal{X})m^\alpha$ of $(s_0 + u_m)$-connected components that have Lebesgue measure greater than or equal to $m^{-\alpha}$. Moreover, the $(s_0 + u_m)$-connected components of $\tilde{G}_m$ are mutually $(s_0 + \theta u_m)$-separated for any $\theta \in (0, 1)$.*

We are now in position to define the estimators of the s_0-connected components of Γ. Define $s_m = s_0 + (3 \log m)^{-1}$ and denote by $\tilde{H}_1, \ldots, \tilde{H}_{\tilde{K}_m}$ the s_m-connected components of $\tilde{G}_m$ that have Lebesgue measure greater than or equal to $m^{-\alpha}$. The number $\tilde{K}_m$ depends on $\mathbb{X}_u$ and is therefore random but bounded from above by the deterministic quantity $\mathrm{Leb}_d(\mathcal{X})m^\alpha$.

Let $\mathcal{J}$ be a subset of $\{1,\dots,J\}$. Define $\kappa(j) = \{k = 1,\dots,\tilde{K}_m : \tilde{H}_k \cap T_j \neq \emptyset\}$ and let $D(\mathcal{J})$ be the event on which the sets $\kappa(j), j \in \mathcal{J}$ are reduced to singletons $\{k(j)\}$ that are disjoint, i.e.,

$$\begin{aligned} D(\mathcal{J}) &= \Big\{\kappa(j) = \{k(j)\}, k(j) \neq k(j'),\ \forall\, j, j' \in \mathcal{J}, j \neq j'\Big\} \\ &= \Big\{\kappa(j) = \{k(j)\},\ (T_j \cup \tilde{H}_{k(j)}) \cap (T_{j'} \cup \tilde{H}_{k(j')}) = \emptyset,\ \forall\, j, j' \in \mathcal{J}, j \neq j'\Big\}\,. \end{aligned} \tag{13}$$

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 $D(\{1,\dots,J\})$ has asymptotically overwhelming probability. The following proposition ensures that $D(\mathcal{J})$ has large enough probability.

Proposition 4.3 *Fix $r_0 > 0$ and $s_0 \geq (3\log m)^{-1}$. Assume that there exists a version of Γ that is r_0-standard and closed. Then, denoting by J the number of s_0-connected components if Γ, for any $\mathcal{J} \subset \{1,\dots,J\}$, we have*

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

where $\mathcal{D}(\mathcal{J})$ is defined in (13).

4.3 Labeling the clusters

From the strong cluster assumption **(SCA)** the clusters are homogeneous regions. To estimate the clusters, we apply the method described above that consists in estimating the s_m-connected components of the clipped estimator $\tilde{G}_m$ and keep only those that have Lebesgue measure greater than or equal to $m^{-\alpha}$. Then we make a majority vote on each homogeneous region. It yields the following procedure.

Three-step procedure

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 s_m-connected components of $\tilde{G}_m = \hat{G}_m \setminus \mathrm{Clip}(\hat{G}_m)$ (clipping step) that have Lebesgue measure greater than or equal to $m^{-\alpha}$.

3. Assign a single label to each estimated homogeneous region by a majority vote on labeled data.

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 $\alpha > 0$ and let $\hat{G}_m$ be an estimator of the density level set Γ, that is consistent from inside at rate $m^{-\alpha}$. For any $1 \leq k \leq \tilde{K}_m$, define the random variable

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

where $\tilde{H}_k$ is obtained by Step 2 of the three-step procedure. Denote by $\tilde{g}^k_{n,m}$ the function $\tilde{g}^k_{n,m}(x) = \mathbb{I}_{\{Z^k_{n,m}>0\}}$ for all $x \in \tilde{H}_k$ and consider the classifier defined on $\mathcal{X}$ by

$$\tilde{g}_{n,m}(x) = \sum_{k=1}^{\tilde{K}_m} \tilde{g}^k_{n,m}(x) \mathbb{I}_{\{x \in \tilde{H}_k\}}\,, \quad x \in \mathcal{X}\,. \tag{14}$$

Note that the classifier $\tilde{g}_{n,m}$ assigns 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 cluster 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 reasonable 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., Herbei and Wegkamp, 2006). Remark that when there is only a finite number of clusters, there exists $\delta > 0$ such that

$$\delta = \min_{j=1,\dots,J} \left\{\delta_j \,:\, \delta_j > 0\right\}. \tag{15}$$

Theorem 4.1 *Fix $\alpha > 0$ and assume that* **(SCA)** *holds. Consider an estimator $\hat{G}_m$ of Γ, based on $\mathbb{X}_u$ that is consistent from inside at rate $m^{-\alpha}$. Then, the classifier $\tilde{g}_{n,m}$ defined in* (14) *satisfies*

$$\mathcal{E}_\Gamma\left(\tilde{g}_{n,m}\right) \leq \widetilde{O}\left(\frac{m^{-\alpha}}{1-\theta}\right) + \sum_{j=1}^{J} \delta_j e^{-n(\theta\delta_j)^2/2} \leq \widetilde{O}\left(\frac{m^{-\alpha}}{1-\theta}\right) + e^{-n(\theta\delta)^2/2}\,, \tag{16}$$

for any $0 < \theta < 1$ and where $\delta > 0$ is defined in (15).

Note that, since we often have $m \gg n$, the first term in the RHS of (16) 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 clusters are completely 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 use the unlabeled sample $\mathbb{X}_u$ of size m to construct an estimator $\hat{G}_m$ of $\Gamma = \Gamma(\lambda) = \{x \in \mathcal{X} : p(x) \geq \lambda\}$, that is consistent from inside at rate $m^{-\alpha}$ for some $\alpha > 0$ that should be as large as possible. A simple and intuitive way to achieve this goal is to use *plug-in estimators* of Γ defined by

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

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

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

where $\ell > 0$ is an offset. Clearly, we have $\tilde{\Gamma}_\ell \subset \hat{\Gamma}$. Keeping in mind that we want estimators that are consistent from inside we are going to consider sufficiently large offset $\ell = \ell(m)$.

Plug-in rules is not the only choice for density level set estimation. Direct methods such as empirical excess mass maximization (see, e.g., Polonik, 1995; Tsybakov, 1997; Steinwart et al., 2005) are also popular. One advantage of plug-in rules over direct methods is that 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. In Section 6, we briefly describe a possible implementation based on existing software that makes use of kernel or nearest neighbors density estimators. To conclude this discussion, remark that the greater flexibility of plug-in rules may result in a poorer learning performance and even though we do not discuss any implementation based on direct methods, it may well be the case that the latter perform better in practice. However, it is not our intent to propose here the best clustering algorithm or the best density level set estimator and we present a simple proof of convergence for offset plug-in rules only for the sake of completeness.

The next assumption has been introduced in Polonik (1995). It is an analog of the margin assumption formulated in Mammen and Tsybakov (1999) and Tsybakov (2004) but for arbitrary level λ in place of $1/2$.

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 $c^\star > 0$ such that, for all $\varepsilon > 0$,*

$$\mathrm{Leb}_d \{x \in \mathcal{X} : |f(x) - \lambda| \leq \varepsilon\} \leq c^\star \varepsilon^\gamma \,.$$

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 offset 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 offset 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 based on the sample $\mathbb{X}_u$ of size $m \geq 1$ 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*

$$\sup_{p \in \mathcal{P}} \mathbb{P}_m \left(|\hat{p}_m(x) - p(x)| \geq \delta \right) \leq c_1 e^{-c_2 m^a \delta^2} \,, \; m^{-a/2} < \delta < \Delta \,. \tag{17}$$

Assume further that p has γ-exponent at level λ for any $p \in \mathcal{P}$ and that the offset ℓ is chosen as

$$\ell = \ell(m) = m^{-\frac{a}{2}} \log m \,. \tag{18}$$

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

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

$$\hat{p}_m^K(x) = \frac{1}{mh^d} \sum_{i=n+1}^{n+m} K\left(\frac{X_i - x}{h}\right), \quad x \in \mathcal{X} \,, \tag{19}$$

where $h > 0$ is the bandwidth parameter and $K : \mathbb{R}^d \to \mathbb{R}$ 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 (17) 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 (see Rigollet and Vert, 2006).

6. Discussion

We proposed a formulation of the cluster assumption in probabilistic terms. This formulation relies on Hartigan's (Hartigan, 1975) definition of clusters but it can be modified to match other definitions of clusters.

We also proved that there is no hope to improve the classification performance outside of these clusters. Based on these remarks, we defined the cluster excess-risk on which we observe the effect of unlabeled data. Finally we proved that when we have consistent estimators of the clusters, it is possible to achieve exponential rates of convergence for the cluster 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. Indeed, 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 **(SCA)** concerns bigger and bigger sets $\Gamma(\lambda)$ and in that sense, it becomes more and 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.

The choice of λ can be made by fixing $P_X(\mathcal{C})$, where $\mathcal{C}$ is defined in (1), the probability of the rejection region. We refer to Cuevas et al. (2001) 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 (15). 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 when $n \ll m$.

A possible implementation of the ideas presented in this paper can be designed using existing clustering software such as DBSCAN (Ester et al., 1996) (and its algorithmic improvement called OPTICS (Ankerst et al., 1999)) or *runt pruning* (Stuetzle, 2003). These three algorithms implement clustering using a definition of clusters that involves density level sets and a certain notion of connectedness. The idea is to use these algorithms on the pooled sample of instances $(X_1, \ldots, X_{n+m}, X)$, where X is the new instance to be classified. As a result every instance will be affected to a cluster by the chosen algorithm. The label for X is then predicted using a majority vote on the labeled instances that are affected to the same cluster as X. Observe that unlike the method described in the paper, the clusters depend on the labeled instances $(X_1, \ldots, X_n)$. Proceeding so allows us to use directly existing clustering algorithms without any modification. Since all three algorithms are distance based, we could run them only on unlabeled instance and then affect each labeled instance and the new instance to the same cluster as its nearest neighbor. However, if we assume that $m \gg n$, incorporating labeled instances will not significantly affect the resulting clusters.

We now describe more precisely why these algorithms produce estimated clusters that are related to the s_m-connected components of a plug-in estimator of the density level

set. Each algorithm has instances $(X_1, \dots, X_m)$ and several parameters described below as inputs. Note that these clustering algorithms will affect every instance to a cluster. This can be transformed into our framework by removing clusters that contain only one instance.

- DBSCAN has two input parameters: a real number $\varepsilon > 0$ and and integer $M \geq 1$. The basic version of this algorithm proceeds as follows. For a given instance X_i, let $J_\varepsilon(i) \subset \{1, \dots, m\}$ be the set of indexes $j \neq i$ such that $\|X_j - X_i\| \leq \varepsilon$. If $\mathrm{card}(J_\varepsilon(i)) \geq M$ then all instances $X_j, j \in J_\varepsilon(i)$ are affected to the same cluster as X_i and the procedure is repeated with each $X_j, j \in J_\varepsilon(i)$. Otherwise a new cluster is defined and the procedure is repeated with another instance.

  Observe first that the instances X_j that satisfy $\|X_j - X_i\| \leq \varepsilon$ are ε-connected to X_i. Also, define the kernel density estimator $\hat{p}_m$ by:

$$\hat{p}_m(x) = \frac{1}{m\varepsilon^d} \sum_{j=1}^{m} K\big(\frac{x - X_i}{\varepsilon}\big)\,,$$

  where $K : \mathbb{R}^d \to \mathbb{R}$ is defined by $K(x) = \mathbb{1}_{\{\|x\| \leq 1\}}$ for any $x \in \mathbb{R}^d$. Then $\mathrm{card}(J_\varepsilon(i)) \geq M$ is equivalent to $\hat{p}_m(X_i) \geq \frac{M+1}{m\varepsilon^d}$. Thus, if we chose $s_0 = \varepsilon - (3 \log m)^{-1}$ and $\lambda + \ell(m) = \frac{M+1}{m\varepsilon^d}$, we see that DBSCAN implements our method. Conversely, for given λ and s_0, we can derive the parameters ε and M such that DBSCAN implements our method.

- OPTICS is a modification of DBSCAN that allows the user to compute in an efficient fashion all cluster partitions for different $\varepsilon \leq \varepsilon_0$ for some user specified $\varepsilon_0 > 0$. The user still has to input the chosen value for ε so that from our point of view, the two algorithms are the same.

- Both of the previous algorithms suffer from a major drawback that is inherent to our definition of cluster based on a global level when determining the density level sets. Indeed, in many real data sets, some clusters can only be identified using several density levels. Stuetzle (2003) recently described an algorithm called *runt pruning* that is free from this drawback. Since, it does not implement our method, we do not describe the algorithm in detail but mention it because it implements a more suitable definition of clusters that is also based on connectedness and density level sets. In particular it resolves the problem of choosing λ. It uses a nearest neighbor density estimator as a running horse and uses a single input parameter that corresponds to the scale s_0.

This paper is an attempt to give a proper mathematical framework for the cluster assumption proposed in Seeger (2000). As mentioned above, the definition of clusters we use here is one among several available and it could be interesting to modify the formulation of the cluster assumption to match other definitions of cluster. In particular, the definition of cluster as s_0-connected components of the λ-level set of the density leaves the problem of choosing λ correctly.

Acknowledgments. The author is most indebted to anonymous referees that contributed to a significant improvement of the paper through their questions and comments.

7. Proofs

This section contains proofs of the results presented in the paper.

7.1 Proof of Proposition 2.1

Since the distribution of the unlabeled sample $\mathbb{X}_u$ does not depend on η, we have for any marginal distribution P_X,

$$\begin{aligned}\sup_{\eta\in\Xi}\mathbb{E}_{n,m}\int_{\mathcal{C}^c}|2\eta-1|\mathbb{1}_{\{T_{n,m}\neq g^\star\}}\mathrm{d}P_X &= \sup_{\eta\in\Xi}\mathbb{E}_m\mathbb{E}_n\Big[\int_{\mathcal{C}^c}|2\eta-1|\mathbb{1}_{\{T_{n,m}\neq g^\star\}}\mathrm{d}P_X\big|\mathbb{X}_u\Big]\\ &= \mathbb{E}_m\sup_{\eta\in\Xi}\mathbb{E}_n\Big[\int_{\mathcal{C}^c}|2\eta-1|\mathbb{1}_{\{T_{n,m}\neq g^\star\}}\mathrm{d}P_X\big|\mathbb{X}_u\Big]\\ &\geq \inf_{T_n}\sup_{\eta\in\Xi}\mathbb{E}_n\int_{\mathcal{C}^c}|2\eta-1|\mathbb{1}_{\{T_n\neq g^\star\}}\mathrm{d}P_X\,,\end{aligned}$$

where in the last inequality, we used the fact that conditionally on $\mathbb{X}_u$, the classifier $T_{n,m}$ only depends on $\mathbb{X}_l$ and can therefore be written T_n.

7.2 Proof of Theorem 3.1

We can decompose $\mathcal{E}_{\mathcal{C}}(\hat{g}_n)$ into

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

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

$$\begin{aligned}\int_{T_j}|2\eta(x)-1|\mathbb{1}_{\{\hat{g}_n^j(x)\neq g^\star(x)\}}p(x)\mathrm{d}x &= \mathbb{1}_{\{Z_n^j\leq 0\}}\int_{T_j}(2\eta(x)-1)p(x)\mathrm{d}x\\ &\leq \delta_j\mathbb{1}_{\left\{|\delta_j-\frac{Z_n^j}{n}|\geq\delta_j\right\}}\,.\end{aligned}\tag{20}$$

Taking expectation $\mathbb{E}_n$ on both sides of (20) we get

$$\begin{aligned}\mathbb{E}_n\int_{T_j}|2\eta(x)-1|\mathbb{1}_{\{\hat{g}_n^j(x)\neq g^\star(x)\}}p(x)\mathrm{d}x &\leq \delta_j\mathbb{P}_n\Big[\big|\delta_j-\frac{Z_n^j}{n}\big|\geq\delta_j\Big]\\ &\leq 2\delta_j e^{-n\delta_j^2/2}\,,\end{aligned}\tag{21}$$

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

7.3 Proof of Lemma 4.1

The binary relation $\overset{s}{\underset{C}{\longleftrightarrow}}$ is an equivalence relation if it satisfies reflexivity, symmetry and transitivity.

To prove, reflexivity, consider the trivial constant path $f(t) = x$ for all $t \in [0,1]$. We immediately obtain that $x \overset{s}{\underset{C}{\longleftrightarrow}} x$.

To prove symmetry, fix $x, y \in C$ such that $x \overset{s}{\underset{C}{\longleftrightarrow}} y$ and denote by f_1 the piecewise constant map with n_1 jumps that satisfies $f_1(0) = x$, $f_1(1) = y$ and $\|f_1(t_+) - f_1(t_-)\| \leq s$. It is not difficult to see that the map $\tilde{f}_1$ defined by $\tilde{f}_1(t) = f_1(1-t)$ for any $t \in [0,1]$ is piecewise constant with n_1 jumps, satisfies $\tilde{f}_1(0) = y$, $\tilde{f}_1(1) = x$ and $\|\tilde{f}_1(t_+) - \tilde{f}_1(t_-)\| \leq s$ for any $t \in [0,1]$, so that $y \overset{s}{\underset{C}{\longleftrightarrow}} x$.

To prove transitivity, let $z \in C$ be such that $y \overset{s}{\underset{C}{\longleftrightarrow}} z$ and let f_2 be a piecewise constant map with n_2 jumps that satisfies $f_2(0) = y$, $f_2(1) = z$ and $\|f_2(t_+), f_2(t_-)\| < s$ for any $t \in [0,1]$. Let now $f : [0,1] \to \mathcal{X}$ be the map defined by:

$$f(t) = \begin{cases} f_1(2t) & \text{if } t \in [0, 1/2] \\ f_2(2t-1) & \text{if } t \in [1/2, 1]\,. \end{cases}$$

This map is obviously piecewise constant with $n_1 + n_2$ jumps and satisfies $f(0) = x, f(1) = z$. Moreover, for any $t \in [0,1]$, f satisfies $\|\tilde{f}(t_+), \tilde{f}(t_-)\| \leq s$.

Thus $\overset{s}{\underset{C}{\longleftrightarrow}}$ is an equivalence relation and C can be partitioned into its classes of equivalence.

7.4 Proof of Proposition 4.1

From Lemma 4.1, we know that $\overset{s}{\underset{C}{\longleftrightarrow}}$ is an equivalence relation and C can be partitioned into its classes of equivalence denoted by $C_1, C_2, \ldots$. The classes of equivalences $C_1, C_2, \ldots$ obviously satisfy the first point of Proposition 4.1 from the very definition of a class of equivalence.

To check the second point, remark first that since C is a closed set, each $C_j, j \geq 1$ is also a closed set. Indeed, fix some $j \geq 1$ and let $(x_n, n \geq 1)$ be a sequence of points in C_j that converges to $\bar{x}$. Since C is closed, we have $\bar{x} \in C$ so there exists $j' \geq 1$ such that $\bar{x} \in C_{j'}$. If $j \neq j'$, then $\|x_n - \bar{x}\| > s$ for any $n \geq 1$ which contradicts the fact that x_n converges to $\bar{x}$. Therefore, $\bar{x} \in C_j$ and C_j is closed. Then let C_j and $C_{j'}$, be two classes of equivalence such that $d_\infty(C_j, C_{j'}) \leq s$. Using the fact that C_j and $C_{j'}$ are closed sets, we conclude that there exist $x \in C_j$ and $x' \in C_{j'}$ such that $\|x - x'\| \leq s$ and hence that $x \overset{s}{\underset{C}{\longleftrightarrow}} x'$. Thus $C_j = C_{j'}$ and we conclude that for any $C_j, C_{j'}, j \neq j'$, we have $d_\infty(C_j, C_{j'}) > s$ and the C_j are mutually s-separated.

We now prove that the decomposition is finite. Since the C_j are mutually s-separated, for any $1 \leq j \leq k$, for any $x_j \in C_j$, the Euclidean balls $\mathcal{B}(x_j, s/3)$ are disjoint. Using the facts that $\mathcal{X}$ is bounded and that C is r_0-standard we obtain,

$$\infty > \mathrm{Leb}_d(\mathcal{X}) \geq \sum_{j=1}^{k} \mathrm{Leb}_d\big[\mathcal{B}(x_j, s/3) \cap \mathcal{X}\big] \geq \sum_{j=1}^{k} \mathrm{Leb}_d\big[\mathcal{B}(x_j, s/3) \cap C\big] \geq ck\,,$$

for a positive constant c. Thus we proved the existence of a finite partition

$$C = \bigcup_{j=1}^{J} C_j\,.$$

It remains to prove that this partition is unique. To this end, we make use of the fundamental theorem of equivalence relations (see, e.g., Dummit and Foote, 1991, Prop. 2, page 3) which states that any partition of C corresponds to the classes of equivalences of a unique equivalence relation. Let $\mathcal{P}' = \{C'_1, \ldots, C'_{J'}\}$ be a partition of C that satisfies the two points of Proposition 4.1 and denote by $\mathcal{R}'$ the corresponding equivalence relation. We now prove that $\overset{s}{\underset{C}{\longleftrightarrow}} \equiv \mathcal{R}'$. From the first point of Proposition 4.1, we easily conclude that if $x\mathcal{R}'y$ then $x \overset{s}{\underset{C}{\longleftrightarrow}} y$. Now if we choose $x, y \in C$ such that $x\mathcal{R}'y$ does not hold, then there exist $j \neq j'$ such that $x \in C'_j$ and $y \in C'_{j'}$. If we had $x \overset{s}{\underset{C}{\longleftrightarrow}} y$, it would hold $d_\infty(C'_j, C'_{j'}) \leq s$ which contradicts the second point of Proposition 4.1 so $x \overset{s}{\underset{C}{\longleftrightarrow}} y$ does not hold. As a consequence we have proved that for any $x, y \in C$, $x\mathcal{R}y$ if and only if $x \overset{s}{\underset{C}{\longleftrightarrow}} y$ and the two relations are the same so as their classes of equivalence. This allows us to conclude that $\mathcal{P}' = \mathcal{P}$.

7.5 Proof of Proposition 4.2

Consider a regular grid $\mathcal{G}$ on $\mathbb{R}^d$ with step size $1/\log(m)$ and observe that the Euclidean balls of centers in $\tilde{\mathcal{G}} = \mathcal{G} \cap \mathrm{Clip}(\hat{G}_m)$ and radius $\sqrt{d}/\log(m)$ cover the set $\mathrm{Clip}(\hat{G}_m)$. Since $\mathcal{X}$ is bounded, there exists a constant $c_1 > 0$ such that $\mathrm{card}\{\tilde{\mathcal{G}}\} = c_1(\log m)^d$. Therefore

$$\mathrm{Leb}_d(\mathrm{Clip}(\hat{G}_m)) \leq \sum_{x \in \tilde{\mathcal{G}}} \mathrm{Leb}_d\big(\mathcal{B}(x, \sqrt{d}/\log(m)) \cap \hat{G}_m\big) \leq \frac{c_2(\log m)^{d-\bar{d}}}{m^\alpha}\,,$$

for some positive constant c_2. Therefore, the rate of convergence $\tilde{G}_m$ is the same as that of $\hat{G}_m$. Observe also that $\tilde{G}_m \subset \hat{G}_m$, so that $\tilde{G}_m$ is also consistent from inside.

Assume that $\tilde{G}_m$ can be decomposed in at least a number k of $(s_0 + u_m)$-connected components, $\tilde{H}_1, \ldots, \tilde{H}_k$ with Lebesgue measure greater than or equal to $m^{-\alpha}$. It holds

$$\infty > \mathrm{Leb}_d(\mathcal{X}) \geq \sum_{j=1}^{k} \mathrm{Leb}_d(\tilde{T}_j) \geq k m^{-\alpha}\,,$$

Therefore, the number of $(s_0 + u_m)$-connected components of $\tilde{G}_m$ with Lebesgue measure greater than or equal to $m^{-\alpha}$ is at most $\mathrm{Leb}_d(\mathcal{X})m^\alpha$.

To prove that the $(s_0 + u_m)$-connected components of $\tilde{G}_m$ are mutually s_0-separated, let $\tilde{T}_1 \neq \tilde{T}_2$ be two $(s_0 + u_m)$-connected components of $\tilde{G}_m$ and fix $x_1 \in \tilde{T}_1$, $x_2 \in \tilde{T}_2$. We have $\|x_1 - x_2\| > s_0 + u_m$, otherwise $\tilde{T}_1 = \tilde{T}_2$. Thus $d_\infty(\tilde{T}_1, \tilde{T}_2) \geq s_0 + u_m > s_0 + \theta u_m$ for any $u_m > 0, \theta \in (0, 1)$. Thus two $(s_0 + u_m)$-connected components of $\tilde{G}_m$ are $(s_0 + \theta u_m)$-separated for any $\theta \in (0, 1)$.

7.6 Proof of Proposition 4.3

Define $m_0 = e^{\frac{1}{3(r_0 \wedge s_0)}}$ and denote $D(\mathcal{J})$ by D. Remark that

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

where

$$A_1(j) = \{\mathrm{card}[\kappa(j)] = 0\}$$
$$A_2(j) = \{\mathrm{card}[\kappa(j)] \geq 2\}$$
$$A_3(j) = \bigcup_{j' \neq j} \{\kappa(j) \cap \kappa(j') \neq \emptyset\}\,.$$

In words, $A_1(j)$ is the event on which T_j is estimated by none of the $(\tilde{H}_k)_k$, $A_2(j)$ is the event on which T_j is estimated by at least two different elements of the collection $(\tilde{H}_k)_k$ and $A_3(j)$ is the event on which T_j is estimated by an element of the collection $(\tilde{H}_k)_k$ that also estimates another $T_{j'}$ from the collection $(T_j)_j$.

For any $j = 1, \ldots, J$, we have

$$A_1(j) = \{\mathrm{card}[\kappa(j)] = 0\} \subset \{T_j \subset \tilde{G}_m \triangle \Gamma\} \subset \{\mathcal{B}(x, r) \cap T_j \subset \tilde{G}_m \triangle \Gamma\}\,,$$

for any $x \in T_j$ and $r > 0$. Remark that from Proposition 4.1, the T_j are mutually s_0-separated so we have $\mathcal{B}(x, r) \cap T_j = \mathcal{B}(x, r) \cap \Gamma$ for any $r \leq s_0$. Thus, for any $m \geq m_0$, it holds $(3 \log m)^{-1} \leq s_0 \wedge r_0$ and

$$A_1(j) \subset \{\mathrm{Leb}_d[\mathcal{B}(x, (3\log m)^{-1}) \cap T_j] \leq \mathrm{Leb}_d[\tilde{G}_m \triangle \Gamma]\} \subset \{\mathrm{Leb}_d[\tilde{G}_m \triangle \Gamma] \geq c_0 (3 \log m)^{-\bar{d}}\}\,,$$

where in the last inclusion we used the fact that Γ is r_0-standard.

We now treat $A_2(j)$. Assume without loss of generality that $\{1, 2\} \subset \kappa(j)$. On $A_2(j)$, there exist $x_1 \in T_j \cap \tilde{H}_1$, $x_n \in T_j \cap \tilde{H}_2$ and a sequence $x_2, \ldots, x_{n-1} \in T_j$ such that $\|x_j - x_{j+1}\| \leq s_0$. Observe now that from Proposition 4.2, we have $\|x_1 - x_n\| > s_0 \geq (3 \log m)^{-1}$ for $m \geq m_0$. Therefore the integer

$$j^\star = \min\left\{j \,:\, 2 \leq j \leq n,\, \exists\, z \in \tilde{H}_1 \,\mathrm{s.t.}\, \|x_j - z\| > (3 \log m)^{-1}\right\},$$

is well defined. Moreover, there exists $z_0 \in \tilde{H}_1$ such that $\|x_{j^\star - 1} - z_0\| \leq (3 \log m)^{-1}$. Now, if there exists $z \in \tilde{H}_k$, for some $k \in \{2, \ldots, \tilde{K}_m\}$, such that $\|x_{j^\star} - z\| \leq (3 \log m)^{-1}$, then

$$d_\infty(\tilde{H}_1, \tilde{H}_k) \leq \|z_0 - x_{j^\star - 1}\| + \|x_{j^\star - 1} - x_{j^\star}\| + \|x_{j^\star} - z\| \leq s_0 + 2(3 \log m)^{-1}\,.$$

This contradicts the conclusion of Proposition 4.2 which states that $d_\infty(\tilde{H}_1, \tilde{H}_k) > s_0 + \theta(\log m)^{-1}$ for any $k = 2, \ldots, \tilde{K}_m$ in particular when $\theta = 2/3$. Therefore we obtain that on $A_2(j)$ there exists $x_{j^\star} \in T_j$ such that

$$\mathcal{B}(x_{j^\star}, (3 \log m)^{-1}) \cap \tilde{G}_m = \emptyset\,.$$

It yields

$$\begin{aligned} A_2(j) &\subset \left\{\mathcal{B}(x_{j^\star}, (3 \log m)^{-1}) \cap T_j \subset \tilde{G}_m \triangle \Gamma\right\} \\ &\subset \left\{\mathrm{Leb}_d\big[\tilde{G}_m \triangle \Gamma\big] > c_0 (3 \log m)^{-\bar{d}}\right\}, \end{aligned}$$

where in the second inclusion used the fact that $\mathcal{B}(x_{j^\star}, r) \cap T_j = \mathcal{B}(x_{j^\star}, r) \cap \Gamma$ for any $r \leq s_0$ and that Γ is r_0-standard.

We now consider the event $A_3(j)$. Assume without loss of generality that $j = 1$ and let k be such that $k \in \kappa(1) \cap \kappa(j')$ for some $j' \in \{2, \dots, J\}$. On $A_3(1)$, there exist $y_1 \in T_1 \cap \tilde{H}_k$, $y_n \in T_{j'} \cap \tilde{H}_k$ and a sequence $y_2, \dots, y_{n-1} \in \tilde{H}_k$ such that $\|y_j - y_{j+1}\| \leq s_m$.

Observe now that from Proposition 4.1, we have $\|y_1 - y_n\| > s_0 \geq (3 \log m)^{-1}$ for $m \geq m_0$. Therefore the integer

$$j^\sharp = \min \left\{ j \,:\, 2 \leq j \leq n, \, \exists \, z \in T_1 \text{ s.t. } \|y_j - z\| > (3 \log m)^{-1} \right\},$$

is well defined. Moreover, there exists $z_1 \in T_1$ such that $\|y_{j^\sharp - 1} - z_1\| \leq (3 \log m)^{-1}$. Now, if there exists $z \in T_{j'}$ for some $j' \in \{2, \dots, J\}$ such that $\|y_{j^\sharp} - z\| \leq (3 \log m)^{-1}$, then

$$d_\infty(T_1, T_{j'}) \leq \|y_{j^\sharp - 1} - z_1\| + \|y_{j^\sharp - 1} - y_{j^\sharp}\| + (3 \log m)^{-1} \leq s_0 + (\log m)^{-1} < \bar{s} \,,$$

for sufficiently large m and where $\bar{s}$ is defined in (11). This contradicts the definition of $\bar{s}$ which implies that $d_\infty(T_1, T_{j'}) \geq \bar{s}$ for any $j \in \{2, \dots, J\}$. Therefore we obtain that on $A_3(1)$ there exists $y_{j^\sharp} \in \tilde{H}_k$ such that $\mathcal{B}(y_{j^\sharp}, (3 \log m)^{-1}) \subset \Gamma^c$. It yields

$$A_3(1) \subset \left\{ \mathrm{Leb}_d(\tilde{G}_m \cap \Gamma^c) \geq \mathrm{Leb}_d(\tilde{G}_m \cap \mathcal{B}(y_{j^\sharp}, (3 \log m)^{-1})) \right\}.$$

Since $y_{j^\sharp} \in \tilde{G}_m \subset \hat{G}_m$, we have $\mathrm{Leb}_d(\hat{G}_m \cap \mathcal{B}(y_{j^\sharp}, (3 \log m)^{-1})) \geq m^{-\alpha} (3 \log m)^{-d}$. On the other hand, we have

$$\begin{aligned}
\mathrm{Leb}_d(\tilde{G}_m \cap \mathcal{B}(y_{j^\sharp}, (3 \log m)^{-1})) &= \mathrm{Leb}_d(\hat{G}_m \cap \mathcal{B}(y_{j^\sharp}, (3 \log m)^{-1})) \\
&\qquad - \mathrm{Leb}_d(\mathrm{Clip}(\hat{G}_m) \cap \mathcal{B}(y_{j^\sharp}, (3 \log m)^{-1})) \\
&\geq m^{-\alpha} (3 \log m)^{-d} - \mathrm{Leb}_d(\hat{G}_m \cap \Gamma^c) \\
&\geq m^{-\alpha} (3 \log m)^{-d} - c_3 m^{-1.1\alpha} \\
&\geq c_4 m^{-\alpha} (\log m)^{-d} \,,
\end{aligned}$$

where we used the fact that $\hat{G}_m$ is consistent from inside at rate $m^{-\alpha}$. Hence,

$$A_3(j) = \bigcup_{j' \neq j} \{\kappa(j) \cap \kappa(j') \neq \emptyset\} \subset \left\{ \mathrm{Leb}_d(\tilde{G}_m \cap \Gamma^c) \geq c_5 m^{-\alpha} (\log m)^{-d} \right\}.$$

Combining the results for $A_1(j)$, $A_2(j)$ and $A_3(j)$, we have

$$\mathbb{P}_m(D^c) \leq \mathbb{P}_m\left\{ \mathrm{Leb}_d\left[\tilde{G}_m \triangle \Gamma\right] > c_0 (3 \log m)^{-\bar{d}} \right\} + \mathbb{P}_m\left\{ \mathrm{Leb}_d(\tilde{G}_m \cap \Gamma^c) \geq c_5 m^{-\alpha} (\log m)^{-d} \right\}.$$

Using the Markov inequality for both terms we obtain

$$\mathbb{P}_m\left\{ \mathrm{Leb}_d\left[\tilde{G}_m \triangle \Gamma\right] > c_0 (\log m)^{-\bar{d}} \right\} = \widetilde{O}\left(m^{-\alpha}\right),$$

and

$$\mathbb{P}_m\left\{ \mathrm{Leb}_d(\tilde{G}_m \cap \Gamma^c) \geq c_5 m^{-\alpha} (\log m)^{-d} \right\} = \widetilde{O}\left(m^{-\alpha}\right),$$

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

7.7 Proof of Theorem 4.1

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

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

We now treat the first term of the RHS of the above inequality, i.e., on the event D. Fix $j \in \{1,\dots,J\}$ and assume w.l.o.g. that $\eta \ge 1/2$ 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 $j = 1,\dots,J$, we have,

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

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

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

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

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

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

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

where we used the fact that on D,

$$\sum_{j=1}^J \mathrm{Leb}_d\big[T_j \triangle \tilde{H}_j\big] \le \mathrm{Leb}_d\big[\Gamma \triangle \tilde{G}_m\big] .$$

From Proposition 4.3, 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.

7.8 Proof of Theorem 5.1

Recall that

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

We begin by the first term. We have

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

The Fubini theorem yields

$$\mathbb{E}_m\left[\mathrm{Leb}_d(\tilde{\Gamma}_\ell \cap \Gamma^c)\right] \leq \mathrm{Leb}_d(\mathcal{X}) \sup_{x \in \mathcal{X}} \mathbb{P}_m\left[|\hat{p}_m(x) - p(x)| \geq \ell\right] \leq c_6 e^{-c_2 m^a \ell^2} ,$$

where the last inequality is obtained using (17) and $c_6 = c_1 \mathrm{Leb}_d(\mathcal{X}) > 0$. Taking ℓ as in (18) yields for $m \geq \exp(\gamma a / c_2)$,

$$\mathbb{E}_m\left[\mathrm{Leb}_d(\tilde{\Gamma}_\ell \cap \Gamma^c)\right] \leq c_6 m^{-\gamma a}. \tag{22}$$

We now prove that $\mathbb{E}_m\left[\mathrm{Leb}_d(\tilde{\Gamma}_\ell \cap \Gamma^c)\right] = \widetilde{O}\left(m^{-\frac{\gamma a}{2}}\right)$. 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

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

and

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

Using (17) and (18) in the same fashion as above we get $\mathbb{E}_m\left[\mathrm{Leb}_d(B_1)\right] = \widetilde{O}\left(m^{-\gamma a}\right)$. The term corresponding to B_2 is controlled using the γ-exponent of density p at level λ. Indeed, we have

$$\mathrm{Leb}_d(B_2) \leq c^\star \ell^\gamma = c^\star (\log m)^\gamma m^{-\frac{\gamma a}{2}} = \widetilde{O}\left(m^{-\frac{\gamma a}{2}}\right) .$$

The previous upper bounds for $\mathrm{Leb}_d(B_1)$ and $\mathrm{Leb}_d(B_2)$ together with (22) yield the consistency from inside.

References

M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander. Optics: Ordering points to identify the clustering structure. In *Proceedings ACM SIGMOD International Conference on Management of Data*, pages 49–60, 1999.

J.-Y. Audibert and A. Tsybakov. Fast learning rates for plug-in classifiers. *Ann. Statist.*, 34(2), 2007.

M.-F. Balcan and A. Blum. A pac-style model for learning from labeled and unlabeled data. In *Proceedings of the Eighteenth Annual Conference on Learning Theory*, pages 111–126, 2005.

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

A. Blum and T. M. Mitchell. Combining labeled and unlabeled sata with co-training. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, pages 92–100, 1998.

S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification: a survey of some recent advances. *ESAIM Probab. Stat.*, 9:323–375 (electronic), 2005.

O. Bousquet, O. Chapelle, and M. Hein. Measure based regularization. In *Advances in Neural Information Processing Systems*, volume 16, 2004.

V. Castelli and T. M. Cover. On the exponential value of labeled samples. *Pattern Recogn. Lett.*, 16(1):105–111, 1995.

V. Castelli and T. M. Cover. The relative value of labeled and unlabeled samples in pattern recognition with an unknown mixing parameter. *IEEE Trans. Inform. Theory*, 42(6, part 2):2102–2117, 1996.

O. Chapelle and A. Zien. Semi-supervised classification by low density separation. In *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, pages 57–64, 2005.

O. Chapelle, B. Schlkopf, and A. Zien. *Semi-Supervised Learning*. MIT Press, Cambridge, Massachusetts, 2006.

A. Cuevas and R. Fraiman. A plug-in approach to support estimation. *Ann. Statist.*, 25 (6):2300–2312, 1997.

A. Cuevas, M. Febrero, and R. Fraiman. Cluster analysis: a further approach based on density estimation. *Comput. Statist. Data Anal.*, 36(4):441–459, 2001.

F. d'Alché Buc, Y. Grandvalet, and C. Ambroise. Semi-supervised marginboost. In *Advances in Neural Information Processing Systems*, volume 14, pages 553–560, 2001.

L. Devroye, L. Györfi, and G. Lugosi. *A probabilistic theory of pattern recognition*. Springer-Verlag, New York, 1996.

D. S. Dummit and R. M. Foote. *Abstract algebra*. Prentice Hall Inc., Englewood Cliffs, New Jersey, 1991.

M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pages 226–231, 1996.

J. H. Hartigan. *Clustering Algorithms*. John Wiley & Sons, Inc., New York, 1975.

R. Herbei and M. Wegkamp. Classification with rejection option. *Canad. J. Statist.*, 34(4): 709–721, 2006.

T. Hertz, A. Bar-Hillel, and D. Weinshall. Boosting margin based distance functions for clustering. In *Proceedings of the twenty-first international conference on Machine learning*, 2004.

E. Mammen and A. B. Tsybakov. Smooth discrimination analysis. *Ann. Statist.*, 27(6): 1808–1829, 1999.

W. Polonik. Measuring mass concentrations and estimating density contour clusters—an excess mass approach. *Ann. Statist.*, 23(3):855–881, 1995.

M. Rattray. A model-based distance for clustering. In *Proc. of the IEEE-INNS-ENNS Int. Joint Conf. on Neural Networks*, pages 13–16, 2000.

P. Rigollet and R. Vert. Fast rates for plug-in estimators of density level sets. Technical Report 1102, Laboratoire de Probabilités et Modèles Aléatoires de Paris 6, 2006. URL `http://hal.ccsd.cnrs.fr/ccsd-00114180`.

M. Seeger. Learning with labeled and unlabeled data. Technical report, Institute for ANC, Edinburgh, UK, 2000. URL `http://www.dai.ed.ac.uk/~seeger/papers.html`.

I. Steinwart, D. Hush, and C. Scovel. A classification framework for anomaly detection. *J. Mach. Learn. Res.*, 6:211–232, 2005.

W. Stuetzle. Estimating the cluster type of a density by analyzing the minimal spanning tree of a sample. *J. Classification*, 20(1):25–47, 2003.

M. Tipping. Deriving cluster analytic distance functions from Gaussian mixture models. In *Proceedings of the Ninth International Conference on Neural Networks*, pages 815–820, 1999.

A. B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Ann. Statist.*, 32 (1):135–166, 2004.

A. B. Tsybakov. On nonparametric estimation of density level sets. *Ann. Statist.*, 25(3): 948–969, 1997.

S. A. van de Geer. *Applications of empirical process theory*. Cambridge University Press, Cambridge, 2000.

V. N. Vapnik. *Statistical Learning Theory*. Wiley, New-York, 1998.

T. Zhang and F. J. Oles. A probability analysis on the value of unlabeled data for classification problems. In *Proceedings of the seventeenth International Conference on Machine Learning*, 2000.

X. Zhu. Semi-supervised learning literature survey. Technical Report 1530, Computer Sciences, University of Wisconsin-Madison, 2005. URL `http://www.cs.wisc.edu/~jerryzhu/pub/ssl_survey.pdf`.