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 *

September 21, 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 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.

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 mainly two approaches to solve this problem. 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., Balcan and Blum, 2005; Blum and Mitchell, 1998). 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 underlying 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

*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).

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 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 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 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 when the clusters are perfectly 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: 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 discussion on the choice of λ and possible 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 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 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, 2005). However, simulations indicate that much faster rates should be attainable when the 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). 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

(i) $P[Y = Y'|X, X' \in C] \geq P[Y \neq Y'|X, X' \in C]$, 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 (CA1) 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 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 **(CA1)** 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

$$P[Y = Y'|X, X' \in T_j] \geq P[Y \neq Y'|X, X' \in T_j]\,.$$

[1] the notation is adapted to the present framework

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

$$\mathcal{C} = \bigcup_{j \geq 1} T_j \tag{2.1}$$

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 **(CA1)** 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),

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

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 $\mathcal{C}^c$, 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

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

Since we assumed $P_X(\mathcal{C}^c) \neq 0$, the RHS of (2.2) is bounded from below by the optimal rates of convergence that appear in supervised classification.

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. Thus, any class of distributions $\mathcal{D}$ can be decomposed as $\mathcal{D} = \mathcal{M} \times \Xi$ where $\mathcal{M}$ is a class of marginal distributions on $\mathcal{X}$ and Ξ is a 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 $\mathcal{X}$. Assume that the class $\mathcal{D} = \mathcal{M} \times \Xi$ is such that for any $\eta \in \Xi$ and any $x \in \mathcal{C}^c$ the value of $\eta(x)$ is independent of P_X. Then, for any marginal distribution $P_X^0 \in \mathcal{M}$, we have*

$$\inf_{T_n} \sup_{\eta \in \Xi} \mathbb{E}_n \int_{\mathcal{C}^c} |2\eta - 1| \mathbb{1}_{\{T_n \neq g^\star\}} \mathrm{d}P_X^0 \leq \inf_{T_{n,m}} \sup_{P \in \mathcal{D}} \mathbb{E}_{n,m} \int_{\mathcal{C}^c} |2\eta - 1| \mathbb{1}_{\{T_{n,m} \neq g^\star\}} \mathrm{d}P_X \,, \tag{2.3}$$

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 **(CA1)** 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 (2.3) 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 (2005) 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

$$\mathcal{E}_{\mathcal{C}}(\hat{g}_{n,m}) \triangleq \mathbb{E}_{n,m} \int_{\mathcal{C}} |2\eta(x) - 1| \mathbb{1}_{\{\hat{g}_{n,m}(x) \neq g^\star(x)\}} p(x) \mathrm{d}x \,. \tag{2.4}$$

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

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

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

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

A minimizer of $R_{\mathcal{C}}(\cdot)$ over $\mathcal{G}$ 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 $g \in \mathcal{G}$ we have,

$$R_{\mathcal{C}}(g) - R_{\mathcal{C}}(g^\star_{|\mathcal{C}}) = \int_{\mathcal{C}} |2\eta(x) - 1| \mathbb{1}_{\{g(x)\neq g^\star_{|\mathcal{C}}(x)\}} p(x) \mathrm{d}x \,. \tag{2.5}$$

Taking expectations on both sides of (2.5) with $g = \hat{g}_{n,m}$, it follows that

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

Therefore, cluster excess-risk equals the excess-risk of classifiers in $\mathcal{G}$. In the sequel, we only consider classifiers $\hat{g}_{n,m} \in \mathcal{G}$, 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 **(CA1)**. 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 **(CA1)**, 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

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

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 $\mathcal{C}$ by

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

The following theorem gives an exponential rate of convergence for the cluster excess-risk of the classifier $\hat{g}_n$ under **(CA1)**.

Theorem 3.1 *Let $T_j, j \geq 1$ be a family of measurable sets that satisfy Assumption* **(CA1)**. *Then, the classifier $\hat{g}_n$ defined above satisfies*

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

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

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

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 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 $\mathcal{X}$ is *connected* (see definition below) and 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

$$\Gamma(\lambda) = \{x \in \mathcal{X} : p(x) \geq \lambda\}\,. \tag{4.1}$$

On these sets, the density is relatively high. The cluster assumption involves also a notion of connectedness of a set. A set $C \subset \mathbb{R}^d$ is said to be *connected* (or pathwise connected) if, for any $x, x' \in C$, there exists a continuous map $f : [0,1] \to C$, such that $f(0) = x$ and $f(1) = x'$. A direct consequence of this definition is that 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 no more connected 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. To overcome this problem, we introduce the following notions.

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

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

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

Definition 4.1 *Fix $r_0 > 0$, $c_0 > 0$ and let $\bar{d}$ be an integer such that $\bar{d} \geq d$. Let C be a measurable subset of $\mathcal{X}$. For two points $x, x' \in C$, we say that x* is r_0-connected to x' in C *and we write $x \xleftrightarrow[C]{r_0} x'$ if there exists a continuous map $f : [0,1] \to \mathcal{X}$ such that $f(0) = x, f(1) = x'$ and for any $t \in [0,1]$ and any $0 < r \leq r_0$, we have*

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

Moreover, we say that C is a standard *set, if for any $x \in C$, we have $x \xleftrightarrow[C]{r_0} x$.*

Remark that 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 also that the path f takes values in $\mathcal{X}$ which is connected, so that removing sets of null Lebesgue measure from C does not affect the r_0-connectedness of its elements, contrary to the usual notion of connectedness defined above.

When C is standard, the following lemma holds.

Lemma 4.1 *If the set C is standard, then the binary relation $\overset{r_0}{\underset{C}{\longleftrightarrow}}$ is an equivalence relation and C can be partitioned into its classes of equivalence.*

Before considering the classes of equivalence of the relation $\overset{r_0}{\underset{C}{\longleftrightarrow}}$, for some set $C \subset \mathcal{X}$ we make sure that there is only a finite number of them. To that end, 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 $C_1, C_2 \ldots,$ are s_0-*separated* if $d_\infty(C_1, C_2) \geq s_0$, for some $s_0 \geq 0$. On the right hand side of Figure 1, we show an example of two sets that are not s_0-separated.

Proposition 4.1 *Fix $r_0 > 0, s_0 > 0$ and assume that C is a standard set such that the classes of equivalence of the relation $\overset{r_0}{\underset{C}{\longleftrightarrow}}$ are two by two s_0-separated. Then there exists a partition $C_1, \ldots C_J$ of C, where the C_j are such that*

- *For any $j = 1, \ldots, J$ and any $x, x' \in C_j$, we have $x \overset{r_0}{\underset{C}{\longleftrightarrow}} x'$ and*
- *For any $j \neq j'$ and any $x \in C_j, x' \in C_{j'}$, x is not r_0-connected to x' in C.*

We call $C_1, \ldots, C_J$ the r_0-connected components *of C.*

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.

Cluster Assumption (CA2) Fix $s_0 > 0, r_0 > 0, c_0 > 0$ and assume that Γ admits a version that is standard and such that the classes of equivalence of the relation $\overset{r_0}{\underset{\Gamma}{\longleftrightarrow}}$ are two by two s_0-separated. Denote by $T_1, \ldots, T_J$ the r_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 $T_j, j = 1, \ldots J$.

4.2 Estimation of the clusters

Assume that p is uniformly bounded by a constant $L(p)$ and that $\mathrm{Leb}_d(\mathcal{X}) < \infty$. 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

$$\mathbb{E}_m\big[\mathrm{Leb}_d(\hat{G}_m \triangle \Gamma)\big] \to 0, \quad m \to +\infty\,, \tag{4.2}$$

where $\triangle$ is the sign for the symmetric difference. We call such estimators *consistent* estimators of Γ. However, for any $r_0 > 0$, the r_0-connected components of a consistent estimator of Γ are not in general consistent estimators of the r_0-connected components of Γ. To ensure componentwise

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

$$\mathbb{E}_m\big[\mathrm{Leb}_d(\hat{G}_m \triangle \Gamma)\big] = \mathbb{E}_m\big[\mathrm{Leb}_d(\hat{G}_m^c \cap \Gamma)\big] + \mathbb{E}_m\big[\mathrm{Leb}_d(\hat{G}_m \cap \Gamma^c)\big]\,. \tag{4.3}$$

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 (4.3) 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 4.2 *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 a fixed $\alpha > 0$, let $\hat{G}_m \subset \mathcal{X}$ be an estimator of Γ that is consistent from inside at rate $m^{-\alpha}$ and recall that we want to estimate the r_0-connected components of Γ. To this end, we apply the following transformations to the estimator $\hat{G}_m$:

1. **Clipping** In this step we remove some elements from $\hat{G}_m$ and obtain a clipped set $\tilde{G}_m$. Since $\hat{G}_m$ is an estimator of Γ that is consistent from inside, it ensures that any connected subset of $\tilde{G}_m$ is included in one of the r_0-connected components of Γ except on an event of negligible probability. In other words, it ensures that we do not estimate the union of two r_0-connected components of Γ by a single r_0-connected component of $\hat{G}_m$.

2. **Merging** In this step we want to prevent ourselves from estimating a single r_0-connected component of Γ by several closer and closer disjoint connected sets. The idea used here is to estimate the r_0-connected components of Γ by the connected components of the clipped set $\tilde{G}_m$. When two connected components of $\tilde{G}_m$ are too close we merge them by taking their union.

We now describe the clipping step in more details. Define the set

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

Since for sufficiently large m, we have $(\log m)^{-d} m^{-\alpha} \leq r_0$ eventually, $\mathrm{Clip}(\hat{G}_m)$ is such that none of its elements is r_0-connected to itself in $\hat{G}_m$. In the sequel, we will only consider the clipped version of $\hat{G}_m$ defined by $\tilde{G}_m = \hat{G}_m \setminus \mathrm{Clip}(\hat{G}_m)$.

Proposition 4.2 *Fix $\alpha > 0$. Assume that $\mathrm{Leb}_d(\mathcal{X}) < \infty$ 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 $\tilde{J}_m$ of connected components.*

Denote by $\tilde{T}_1, \ldots, \tilde{T}_{\tilde{J}_m}$ the connected components of $\tilde{G}_m$, where $\tilde{J}_m$ is the number of connected components of $\tilde{G}_m$. This number depends on $\mathbb{X}_u$ and is therefore random.

We now describe the merging step. For simplicity we present it in terms of a recursive pseudo-algorithm. For any $j = 1, \ldots, \tilde{J}_m$, define the set of integers

$$\mathcal{N}(j) = \big\{k \in \{1, \ldots, \tilde{J}_m\} : d_\infty(\tilde{T}_j, \tilde{T}_k) \leq 2(\log m)^{-1}\big\}\,,$$

Consider the following pseudo-algorithm.

MERGING

- Initialize: $\mathcal{Z} = \{1, \ldots, \tilde{J}_m\}$, $j = 0$, $k = 0$.
- While $\mathcal{Z} \neq \emptyset$, do:

$$\begin{aligned} j &\leftarrow \min(\mathcal{Z}), \\ k &\leftarrow k + 1, \\ \tilde{H}_k &= \bigcup_{l \in \mathcal{N}(j)} \tilde{T}_j, \\ \mathcal{Z} &\leftarrow \mathcal{Z} \setminus \mathcal{N}(j), \end{aligned}$$

- $\tilde{K}_m = k$.

Remark that since $j \in \mathcal{N}(j)$, the pseudo-algorithm stops after at most $\tilde{J}_m$ iterations. The family of sets $\tilde{H}_1, \ldots, \tilde{H}_{\tilde{K}_m}$ is such that $d_\infty(\tilde{H}_k, \tilde{H}_{k'}) > 2(\log m)^{-1}$, $\forall\, k \neq k'$. The $\tilde{H}_k$'s correspond to the estimators of the r_0-connected components of Γ. The next proposition states that that the $\tilde{H}_k$'s are consistent estimators of the r_0-connected components of $\Gamma(\lambda)$.

Let $\mathcal{J}$ be a subset of $\{1, \ldots, J\}$. Define $\kappa(j) = \{k = 1, \ldots, \tilde{K}_m : \tilde{H}_k \cap T_j \neq \emptyset\}$ 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.,

$$\begin{aligned} D(\mathcal{J}) &= \Big\{ \kappa(j) = \{k(j)\}, k(j) \neq k(j'),\ \forall\, j, j' \in \mathcal{J}, j \neq j' \Big\} \\ &= \Big\{ \kappa(j) = \{k(j)\},\ (T_j \cup \tilde{H}_{k(j)}) \cap (T_{j'} \cup \tilde{H}_{k(j')}) = \emptyset,\ \forall\, j, j' \in \mathcal{J}, j \neq j' \Big\}. \end{aligned} \tag{4.4}$$

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 $D(\{1, \ldots, J\})$ has asymptotically overwhelming probability. The following proposition gives an upper bound on the probability of the complement of the event $D(\mathcal{J})$.

Proposition 4.3 *Fix $r_0 > 0, s_0 > 0$ and assume that there exists a version of Γ that admits a decomposition into a number $J \geq 1$ of r_0-connected components $\Gamma = \bigcup_{j=1}^{J} T_j$ where the $\{T_j\}_{j=1,\ldots,J}$ are two by two s_0-separated. Consider an estimator $\hat{G}_m$ of Γ that is consistent from inside at rate $m^{-\alpha}$. Denote by $\{\tilde{H}_k\}_{k=1,\ldots,\tilde{K}_m}$ the family of sets obtained by the clipping and merging steps described above. Then, for any $\mathcal{J} \subset \{1, \ldots, J\}$, 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 (4.4).

4.3 Labeling the clusters

To estimate the homogeneous regions, we will simply estimate the connected components of Γ and apply the clipping and merging steps described above. Then we make a majority vote on each homogeneous region. It yields the following procedure.

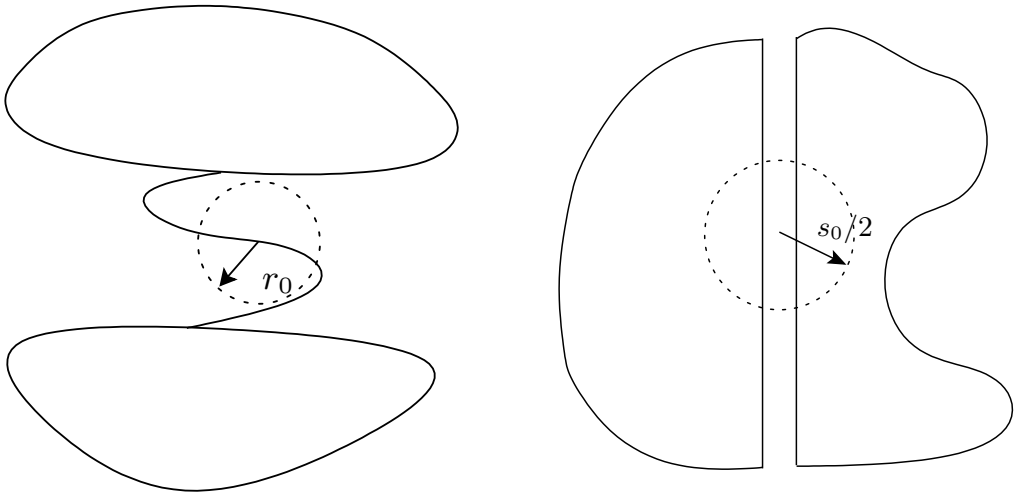

Figure 1: Set that is 0-connected but not r_0-connected for any $r_0 > 0$ (left) and non-separated connected components (right).

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 unions of the connected components of $\tilde{G}_m = \hat{G}_m \setminus \mathrm{Clip}(\hat{G}_m)$ (clipping step) that are closer than $2(\log m)^{-1}$ for the distance d_∞ using pseudo-algorithm MERGING.
3. Assign a single label to each estimated homogeneous region by 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 the clipping and merging steps defined in the previous subsection. 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

$$\tilde{g}_{n,m}(x) = \sum_{k=1}^{\tilde{K}_m} \tilde{g}^k_{n,m}(x) \mathbb{1}_{\{x \in \tilde{H}_k\}}, \quad x \in \mathcal{X}\,. \tag{4.5}$$

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

$$\delta = \min_{j=1,\dots,J} \delta_j \,. \tag{4.6}$$

Theorem 4.1 *Fix $\alpha > 0, r_0 > 0$ and assume that* **(CA2)** *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* (4.5) *satisfies*

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

for any $0 < \theta < 1$ and where $\delta > 0$ is defined in (4.6).

Note that, since we often have $m \gg n$, the first term in the RHS of (4.7) 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 *penalized 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 a penalization. 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 penalization $\ell = \ell(m)$.

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.

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 $\varepsilon > 0$,*

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

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 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 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*

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

Assume further that p has γ-exponent at level λ for any $p \in \mathcal{P}$ 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 $m^{-\frac{\gamma a}{2}}$ for any $p \in \mathcal{P}$.

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

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

where $h > 0$ is the bandwidth parameter and $K : \mathbb{R}^d \to \mathbb{R}$ 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 (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 (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 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 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 **(CA2)** 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 (2.1), 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 (4.6). 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 $n \ll m$.

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 r_0-connected components of the λ-level set of the density leaves the problem of choosing λ correctly.

7 Proofs

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

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

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 bound. Summing now over j yields the theorem.

7.3 Proof of Lemma 4.1

We have to prove three points: reflexivity, symmetry and transitivity. Reflexivity is obvious from the definition of a standard set. Next, remark that if $x \xleftrightarrow[C]{r_0} x'$, there exists a continuous map $f_1 : [0,1] \to \mathcal{X}$ such that $f_1(0) = x, f_1(1) = x'$ and for any $t \in [0,1]$ and any $0 < r \leq r_0$, we have

$$\mathrm{Leb}_d\big(\mathcal{B}(f_1(t), r) \cap C\big) \geq c_0 r^{\bar{d}}\,.$$

To prove symmetry, it is sufficient to consider the continuous map $\tilde{f}_1$ defined by $\tilde{f}_1(t) = f_1(1-t)$ for any $t \in [0,1]$. We now prove transitivity. Assume also that $x' \xleftrightarrow[C]{r_0} x''$, i.e., there exists a continuous map $f_2 : [0,1] \to \mathcal{X}$ such that $f_2(0) = x', f_2(1) = x''$ and for any $t \in [0,1]$ and any $0 < r \leq r_0$, we have

$$\mathrm{Leb}_d\big(\mathcal{B}(f_2(t), r) \cap C\big) \geq c_0 r^{\bar{d}}\,.$$

Define now the map $f : [0,1] \to \mathcal{X}$ 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 continuous on $[0,1]$ and satisfies $f(0) = x, f(1) = x''$. Moreover, for any $t \in [0,1]$, we have

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

The second assertion in the lemma is trivial.

7.4 Proof of Proposition 4.1

From Lemma 4.1, we know that C can be decomposed into is classes of equivalence. Fix $k \geq 1$ and assume that there is at least k classes of equivalence that we denote by $C_1, \ldots, C_k$. Since the classes are assumed to be s_0-separated, for any $1 \leq j \leq k$, for any $x_j \in C_j$, the Euclidean balls $\mathcal{B}(x_j, s/2)$ are disjoint. Thus

$$\infty > \mathrm{Leb}_d(\mathcal{X}) \geq \sum_{j=1}^{k} \mathrm{Leb}_d\big[\mathcal{B}(x_j, s_0/2) \cap \mathcal{X}\big] \geq \sum_{j=1}^{k} \mathrm{Leb}_d\big[\mathcal{B}(x_j, s_0/2) \cap C\big] \geq ck\,,$$

for a positive constant c. Thus we must have a finite decomposition.

7.5 Proof of Proposition 4.2

Consider a regular grid $\mathcal{G}$ on $\mathbb{R}^d$ with step size $1/\log(m)$ and let $c_1 > 0$ be a constant such that the Euclidean balls of centers in $\tilde{\mathcal{G}} = \mathcal{G} \cap \mathrm{Clip}(\hat{G}_m)$ cover the set $\hat{G}_m$. Since $\mathrm{Leb}_d(\mathcal{X}) < \infty$, there exists a constant $c_2 > 0$ such that $\mathrm{card}\{\tilde{\mathcal{G}}\} \leq c_2(\log m)^d$. Therefore

$$\mathrm{Leb}_d(\mathrm{Clip}(\hat{G}_m)) \leq \sum_{x \in \tilde{\mathcal{G}}} \mathrm{Leb}_d\big(\mathcal{B}(x, 1/\log(m)) \cap \hat{G}_m\big) \leq \frac{c_2(\log m)^{\bar{d}-d}}{m^{\alpha}}\,.$$

So the rate of convergence $\tilde{G}_m$ is the same as that of $\hat{G}_m$. We conclude the proof by observing that $\tilde{G}_m \subset \hat{G}_m$, so that $\tilde{G}_m$ is also consistent from inside.

7.6 Proof of Proposition 4.3

Define $m_0 = \exp(1/(r_0 \wedge s_0))$ and denote $D(\mathcal{J})$ by D. For any $j = 1, \ldots, J$, the r_0 connectedness of T_j yields on the one hand,

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

The previous inclusions are illustrated in Figure 2.

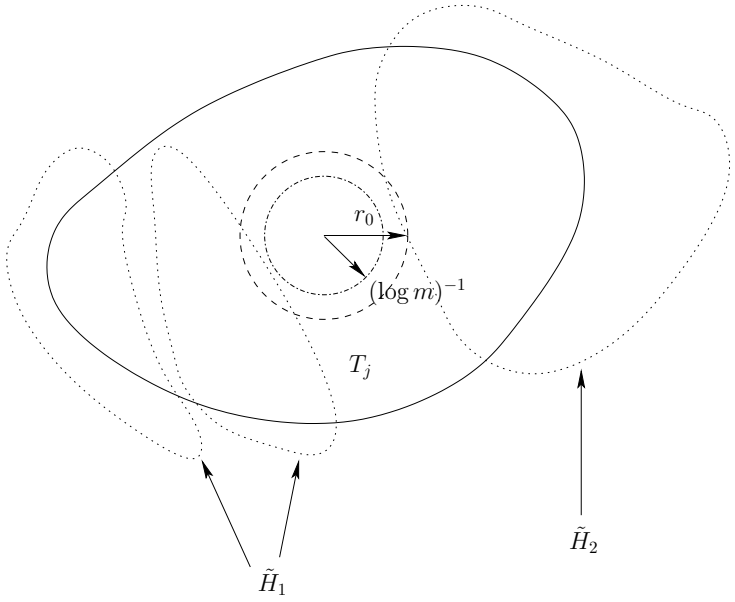

Figure 2: By construction, $\tilde{H}_1$ and $\tilde{H}_2$ are separated by a ball of radius $(\log m)^{-1}$, which is included in $\mathcal{B}(x, r_0)$ when $m \geq m_0$. So if $\{1, 2\} \subset \kappa(j)$ or $\kappa(j) = \emptyset$, this ball is included in in $\tilde{G}_m \triangle \Gamma$.

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

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

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

$$\begin{aligned} \mathrm{Leb}_d(\tilde{G}_m \cap \mathcal{B}(x, \frac{1}{\log m})) &= \mathrm{Leb}_d(\hat{G}_m \cap \mathcal{B}(x, \frac{1}{\log m})) - \mathrm{Leb}_d(\mathrm{Clip}(\hat{G}_m) \cap \mathcal{B}(x, \frac{1}{\log m})) \\ &\geq m^{-\alpha}(\log m)^{-d} - \mathrm{Leb}_d(\hat{G}_m \cap \Gamma^c) \\ &\geq m^{-\alpha}(\log m)^{-d} - c_3 m^{-2\alpha} \\ &\geq c_4 m^{-\alpha}(\log m)^{-d}, \end{aligned}$$

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

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

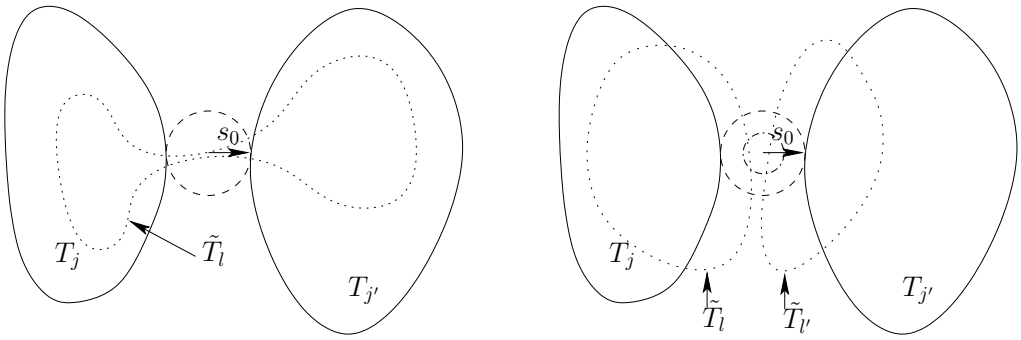

Figure 3: Case (i) (left) and case (ii) (right).

Both cases are illustrated in Figure 3.
Now, since

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

we get

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

Using the Markov inequality for both terms we obtain

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

and

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

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

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

We now treat the first term of the RHS of the above inequality, i.e., on the event D. Fix $j \in \{1,\dots,J\}$ 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 $j = 1,\dots,J$, we have,

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

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

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

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

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

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

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

where we used the fact that on D,

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

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

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

The Fubini theorem yields

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

where the last inequality is obtained using (5.1) and $c_6 = c_1\mathrm{Leb}_d(\mathcal{X}) > 0$. Taking ℓ as in (5.2) yields for $m \ge \exp(\gamma a/c_2)$,

$$\mathbb{E}_m\big[\mathrm{Leb}_d(\tilde{\Gamma}_\ell \cap \Gamma^c)\big] \le c_6 m^{-\gamma a}. \tag{7.3}$$

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

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

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

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

The previous upper bounds for $\mathrm{Leb}_d(B_1)$ and $\mathrm{Leb}_d(B_2)$ together with (7.3) yield the consistency from inside.

References

Audibert, J.-Y. and Tsybakov, A. (2005). Fast learning rates for plug-in classifiers under the margin condition. Tech. rep., Laboratoire de Probabilités et Modèles Aléatoires de Paris 6. To appear in the Annals of Statistics. Available at http://hal.ccsd.cnrs.fr/ccsd-00005882.

Balcan, M. F. and Blum, A. (2005). A pac-style model for learning from labeled and unlabeled data. In P. Auer and R. Meir, eds., *COLT*, *Lecture Notes in Computer Science*, vol. 3559. Springer, pp. 111–126.

Blum, A. and Mitchell, T. M. (1998). Combining labeled and unlabeled sata with co-training. In *COLT*. pp. 92–100.

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

Bousquet, O., Chapelle, O., and Hein, M. (2004). Measure based regularization. In L. S. Thrun, S. and B. Schlkopf, eds., *NIPS*, vol. 16. MIT Press, Cambridge, MA USA.

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

Chapelle, O. and Zien, A. (2005). Semi-supervised classification by low density separation. In *NIPS*. pp. 57–64.

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

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

d'Alché Buc, F., Grandvalet, Y., and Ambroise, C. (2001). Semi-supervised marginboost. In T. G. Dietterich, S. Becker, and Z. Ghahramani, eds., *NIPS*. MIT Press, pp. 553–560.

Devroye, L., Györfi, L., and Lugosi, G. (1996). *A probabilistic theory of pattern recognition*, *Applications of Mathematics (New York)*, vol. 31. Springer-Verlag, New York.

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

Herbei, R. and Wegkamp, M. (2006). Classification with rejection option. *Canad. J. Statist.* To appear.

Hertz, T., Bar-Hillel, A., and Weinshall, D. (2004). Boosting margin based distance functions for clustering. In *ICML '04: Proceedings of the twenty-first international conference on Machine learning*. ACM Press, New York, NY, USA, p. 50.

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

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

Rattray, M. (2000). A model-based distance for clustering. In *Proc. of the IEEE-INNS-ENNS Int. Joint Conf. on Neural Networks*. IEEE Computer Society Press, pp. IV–13. ISBN 0769506216.

Rigollet, P. and Vert, R. (2006). Fast rates for plug-in estimators of density level sets. Tech. rep., Laboratoire de Probabilités et Modèles Aléatoires de Paris 6.

Seeger, M. (2000). Learning with labeled and unlabeled data. Tech. rep., Institute for ANC, Edinburgh, UK. Http://www.dai.ed.ac.uk/ seeger/papers.html.

Steinwart, I., Hush, D., and Scovel, C. (2005). Density level detection is classification. In L. K. Saul, Y. Weiss, and L. Bottou, eds., *Advances in Neural Information Processing Systems 17*. MIT Press, Cambridge, MA.

Tipping, M. (1999). Deriving cluster analytic distance functions from Gaussian mixture models. In *ICANN*.

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

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

van de Geer, S. A. (2000). *Applications of empirical process theory*, *Cambridge Series in Statistical and Probabilistic Mathematics*, vol. 6. Cambridge University Press, Cambridge.

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

Zhu, X. (2005). Semi-supervised learning literature survey. Tech. Rep. 1530, Computer Sciences, University of Wisconsin-Madison. Http://www.cs.wisc.edu/∼jerryzhu/pub/ssl_survey.pdf.