



# Minimax semi-supervised set-valued approach to multi-class classification

Evgenii Chzhen, Christophe Denis, Mohamed Hebiri

## ► To cite this version:

Evgenii Chzhen, Christophe Denis, Mohamed Hebiri. Minimax semi-supervised set-valued approach to multi-class classification. Bernoulli, 2021, 10.3150/20-BEJ1313 . hal-02112918

**HAL Id: hal-02112918**

**<https://hal.science/hal-02112918>**

Submitted on 27 Apr 2019

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Minimax semi-supervised confidence sets for multi-class classification

Evgenii Chzhen<sup>\*</sup>, Christophe Denis and Mohamed Hebiri

*Université Paris-Est – Marne-la-Vallée  
Cité Descartes, Btiment Copernic  
5 boulevard Descartes  
77454 Marne-la-Vallée cedex 2  
e-mail: [evgenii.chzhen@u-pem.fr](mailto:evgenii.chzhen@u-pem.fr)*

[christophe.denis@u-pem.fr](mailto:christophe.denis@u-pem.fr)

[mohamed.hebiri@u-pem.fr](mailto:mohamed.hebiri@u-pem.fr)

**Abstract:** In this work we study the semi-supervised framework of confidence set classification with controlled expected size in minimax settings. We obtain semi-supervised minimax rates of convergence under the margin assumption and a Hölder condition on the regression function. Besides, we show that if no further assumptions are made, there is no supervised method that outperforms the semi-supervised estimator proposed in this work. We establish that the best achievable rate for any supervised method is  $n^{-1/2}$ , even if the margin assumption is extremely favorable. On the contrary, semi-supervised estimators can achieve faster rates of convergence provided that sufficiently many unlabeled samples are available. We additionally perform numerical evaluation of the proposed algorithms empirically confirming our theoretical findings.

**MSC 2010 subject classifications:** Primary 62G05; secondary 62G30, 62H05, 68T10.

**Keywords and phrases:** multi-class classification, confidence sets, minimax optimality, semi-supervised classification.

## 1. Introduction

Let  $K \geq 2$  and  $(X, Y) \in \mathbb{R}^d \times [K] := \{1, \dots, K\}$  be a random couple distributed according to a distribution  $\mathbb{P}$  on  $\mathbb{R}^d \times [K]$ , where  $X \in \mathbb{R}^d$  is seen as the feature vector and  $Y \in [K]$  as the class. This problem falls within the scope of the multi-class setting where the goal is to predict the label  $Y$  for a given feature. Commonly, prediction is performed by a classifier that outputs a single label. However, in the confidence set framework, the objective differs: we aim at predicting a *set* of labels instead of a *single* one. This problem has been studied in a few works, and we consider in this contribution the setup put forward by [Denis and Hebiri \(2017\)](#). The essential feature of their perspective is the control of the size of confidence sets in expectation. While they provided a procedure to build confidence sets based on Empirical Risk Minimization (ERM) and estab-

---

<sup>\*</sup>This work was partially supported by “Labex Bézout” of Université Paris-Est

lished upper bounds, the present work aims at giving a general analysis of the confidence problem in the minimax sense.

### 1.1. Problem statement

All along the paper, we denote by  $\mathbb{P}_X$  the marginal distribution of  $X \in \mathbb{R}^d$  and by  $p(\cdot) := (p_1(\cdot), \dots, p_K(\cdot))^\top$  the regression function defined for all  $k \in [K]$  and all  $x \in \mathbb{R}^d$  as  $p_k(x) := \mathbb{P}(Y = 1|X = x)$ . For any sets  $A, A' \subset [K]$  we denote by  $A \triangle A'$  their symmetric difference. We assume that two data samples  $\mathcal{D}_n, \mathcal{D}_N$  are available. The first sample  $\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n$  consists of  $n \in \mathbb{N}$  *i.i.d.* copies of  $(X, Y) \in \mathbb{R}^d \times [K]$  and the second sample  $\mathcal{D}_N = \{X_i\}_{i=n+1}^{n+N}$  consist of  $N \in \mathbb{N}$  *i.i.d.* copies of  $X \in \mathbb{R}^d$ .

A confidence set classifier  $\Gamma$  is a measurable function from  $\mathbb{R}^d$  to  $2^{[K]} := \{A : A \subset [K]\}$ , that is,  $\Gamma : \mathbb{R}^d \rightarrow 2^{[K]}$  and we denote by  $\Upsilon$  the set of all such functions. For any confidence set  $\Gamma : \mathbb{R}^d \rightarrow 2^{[K]}$  we define its error and its information as

$$\underbrace{P(\Gamma) = \mathbb{P}(Y \notin \Gamma(X))}_{\text{error}}, \quad \underbrace{I(\Gamma) = \mathbb{E}_{\mathbb{P}_X} |\Gamma(X)|}_{\text{information}},$$

respectively, where  $\mathbb{E}_{\mathbb{P}_X}$  stands for the expectation *w.r.t.* the marginal distribution of  $X \in \mathbb{R}^d$  and  $|\Gamma(x)|$  is the cardinal of  $\Gamma$  at  $x \in \mathbb{R}^d$ .

For a fixed integer  $\beta \in [K]$  a  $\beta$ -Oracle confidence set  $\Gamma_\beta^*$  is defined as

$$\Gamma_\beta^* \in \arg \min \{P(\Gamma) : \Gamma \in \Upsilon \text{ s.t. } I(\Gamma) = \beta\}.$$

The set  $\{\Gamma \in \Upsilon : I(\Gamma) = \beta\}$  is always non-empty, as it always contain those confidence sets whose cardinal is equals to  $\beta$  for every  $x \in \mathbb{R}^d$ .

The description of  $\beta$ -Oracle confidence set in general situation might be complicated. Hence, we introduce the following mild assumption, which allows to obtain an explicit expression.

**Assumption 1.1** (Continuity of CDF). *For all  $k \in [K]$  the cumulative distribution function (CDF)  $F_{p_k}(\cdot) := \mathbb{P}_X(p_k(X) \leq \cdot)$  of  $p_k(X)$  is continuous on  $(0, 1)$ .*

**Proposition 1.2** ( $\beta$ -Oracle confidence set). *Fix  $\beta \in [K-1]$ , and let the function  $G : [0, 1] \rightarrow [0, K]$  be defined for all  $t \in [0, 1]$  as*

$$G(t) := \sum_{k=1}^K (1 - F_{p_k}(t)) = \sum_{k=1}^K \mathbb{P}_X(p_k(X) > t),$$

*then under Assumption 1.1 a  $\beta$ -Oracle confidence set  $\Gamma_\beta^*$  can be obtained as*

$$\Gamma_\beta^*(x) = \{k \in [K] : p_k(x) \geq G^{-1}(\beta)\}, \quad (1.1)$$

*where we denote by  $G^{-1}$  the generalized inverse of  $G$  defined for all  $\beta \in [0, K]$  as*

$$G^{-1}(\beta) := \inf \{t \in [0, 1] : G(t) \leq \beta\}.$$

**Proposition 1.3.** *Assume that Assumption 1.1 is fulfilled, then the  $\beta$ -Oracle defined in Eq. (1.1) is a minimizer of the following risk*

$$R_\beta(\Gamma) = P(\Gamma) + G^{-1}(\beta) I(\Gamma) . \quad (1.2)$$

These propositions have been proven in (Denis and Hebiri, 2017, Proposition 4 and Proposition 7). Consequently, the accuracy of a confidence set  $\Gamma$  can be for instance quantified according its excess risk

$$R_\beta(\Gamma) - R_\beta(\Gamma_\beta^*) = \sum_{k=1}^K \mathbb{E}_{\mathbb{P}_X} \left[ |p_k(X) - G^{-1}(\beta)| \mathbf{1}_{\{k \in \Gamma(X) \Delta \Gamma_\beta^*(X)\}} \right] .$$

The statistical learning problem is then to estimate  $\Gamma_\beta^*$  given the data sample  $\mathcal{D}_n$  and  $\mathcal{D}_N$ . The formulation in Eq. (1.1) of the  $\beta$ -Oracle appears to be closely related to the level set estimation problem (Hartigan, 1987; Polonik, 1995; Tsybakov, 1997; Rigollet and Vert, 2009). Hence at first sight, the introduction of an unlabeled sample may be surprising. However, in our setup the estimation of the  $\beta$ -Oracle does not only rely on the regression function but also on the threshold  $G^{-1}(\beta)$  which is *unknown* beforehand and can be estimated in a semi-supervised way (Denis and Hebiri, 2017). To fix these ideas, we give some examples of possible estimation procedures of  $\Gamma_\beta^*$ .

### 1.2. Confidence set estimators

An estimator  $\hat{\Gamma}$  is a measurable function that maps any given data samples into a confidence set classifier. We shall distinguish two types of estimators: supervised and semi-supervised whose formal definition is provided below.

**Definition 1.4** (Supervised and semi-supervised estimators). *A measurable mapping*

$$\hat{\Gamma} : \bigcup_{n, N \in \mathbb{N}} (\mathbb{R}^d \times [K])^n \times (\mathbb{R}^d)^N \rightarrow \Upsilon ,$$

*is called a supervised estimator if for any  $n, N \in \mathbb{N}$  and any data samples  $\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n$ ,  $\mathcal{D}_N = \{X_i\}_{i=n+1}^{n+N}$ , and  $\mathcal{D}'_N = \{X'_i\}_{i=n+1}^{n+N}$  it holds that*

$$\hat{\Gamma}(x; \mathcal{D}_n, \mathcal{D}_N) = \hat{\Gamma}(x; \mathcal{D}_n, \mathcal{D}'_N), \quad \text{a.e. } x \in \mathbb{R}^d \text{ w.r.t. the Lebesgue measure .}$$

*Otherwise the estimator is called semi-supervised. In the sequel, for the simplicity of notation we write  $\hat{\Gamma}(x)$  instead of  $\hat{\Gamma}(x; \mathcal{D}_n, \mathcal{D}_N)$  where no ambiguity is present.*

Intuitively, the supervised estimators do not take into account the information that is provided by the unlabeled sample. Besides, if we denote by  $\hat{\Upsilon}$  the set of all estimators, Definition 1.4 generates a natural partition of  $\hat{\Upsilon}$  into two disjoint sets: the *supervised* estimators  $\hat{\Upsilon}_{\text{SE}}$  and the *semi-supervised* estimators  $\hat{\Upsilon}_{\text{SSE}}$ .

Hereafter, we provide three different examples of estimation procedures which are the core of our study. All these methods rely on *plug-in* principle.

- *Top- $\beta$  procedure.* This method is the most intuitive estimator in the considered context. It is a supervised procedure, that is, based only on  $\mathcal{D}_n$ . Let consider an estimator  $\hat{p}$  of the regression function  $p$ . Let  $(\hat{p}_{\sigma_k(X)})_{k \in [K]}$  be the order statistic associated to  $\hat{p}(X)$ , such that for all  $x \in \mathbb{R}^d$  we have  $\hat{p}_{\sigma_1(x)}(x) \geq \dots \geq \hat{p}_{\sigma_K(x)}(x)$ . A top- $\beta$  confidence set is then defined as

$$\hat{\Gamma}_{\text{top}}(x) = \{\sigma_1(x), \dots, \sigma_\beta(x)\}, \quad \forall x \in \mathbb{R}^d. \quad (1.3)$$

- *Supervised procedure.* Formally, in this type of methods, we only care about  $\mathcal{D}_n$  (we forget about  $\mathcal{D}_N$ ). We split  $\mathcal{D}_n$  into two independent samples such that  $\mathcal{D}_n = \mathcal{D}_{\lfloor n/2 \rfloor} \cup \mathcal{D}_{\lceil n/2 \rceil}$ . Based on the first sample  $\mathcal{D}_{\lfloor n/2 \rfloor}$ , we consider an estimator  $\hat{p}$  of the regression function  $p$ . Furthermore, we define

$$\hat{G}(\cdot) = \frac{1}{\lfloor n/2 \rfloor} \sum_{i \in \mathcal{D}_{\lfloor n/2 \rfloor}} \sum_{k=1}^K \mathbf{1}_{\{\hat{p}_k(X_i) \geq \cdot\}},$$

and one type of supervised estimator is then defined as follows

$$\hat{\Gamma}_{\text{SE}}(x) = \left\{ k \in [K] : \hat{p}_k(x) \geq \hat{G}^{-1}(\beta) \right\}, \quad \forall x \in \mathbb{R}^d. \quad (1.4)$$

Interestingly, conditional on the data sample  $\mathcal{D}_{\lfloor n/2 \rfloor}$ , the definition of the estimator  $\hat{G}$  does not involves the labels associated to  $\mathcal{D}_{\lceil n/2 \rceil}$ . As a consequence, we can naturally consider a semi-supervised version of this estimator.

- *Semi-supervised procedure.* Based on  $\mathcal{D}_n$ , we consider an estimator  $\hat{p}$  of the regression function  $p$ . Furthermore, we define

$$\hat{G}(\cdot) = \frac{1}{N} \sum_{i \in \mathcal{D}_N} \sum_{k=1}^K \mathbf{1}_{\{\hat{p}_k(X_i) \geq \cdot\}},$$

and one type of semi-supervised estimator is then defined as follows

$$\hat{\Gamma}_{\text{SSE}}(x) = \left\{ k \in [K] : \hat{p}_k(x) \geq \hat{G}^{-1}(\beta) \right\}, \quad \forall x \in \mathbb{R}^d. \quad (1.5)$$

One can note that these procedures are based on a preliminary estimator of  $p$  built from  $\mathcal{D}_n$ , that is, all of them are plug-in type procedures. However, these procedures differ by the construction of the output set. The top- $\beta$  procedure and the supervised procedure rely only on the labeled data while the semi-supervised estimator takes advantage of the information provided by the unlabeled data. The top- $\beta$  procedure is the simplest among them, it naturally satisfies  $|\hat{\Gamma}(x)| = \beta$  for all  $x \in \mathbb{R}^d$ . At the same time, the others are more involved and can have different cardinals for different values of  $x \in \mathbb{R}^d$ . Nevertheless, for the other two procedures one can guarantee  $\mathbb{I}(\hat{\Gamma}) \approx \beta$ .

These examples give a rise to natural questions which form the core our theoretical study and which are summarized below.

1. The first question is the statistical performance of these plug-in procedures which is assessed through rates of convergence and their optimality in the minimax sense.
2. The second question focuses on the benefit of the semi-supervised approach. Roughly speaking, are there situations where the semi-supervised approach outperforms the supervised one and how can it be quantified?
3. The third question concentrates on the reason why it is more relevant for this problem to consider more involved estimators than the simple top- $\beta$  method.

### 1.3. Minimax estimation

For a given family  $\mathcal{P}$  of joint distributions on  $\mathbb{R}^d \times [K]$ , a given estimator  $\hat{\Gamma} \in \hat{\mathbf{Y}}$ , and fixed integers  $K \geq 2$ ,  $\beta \in [K]$ ,  $n, N \in \mathbb{N}$  we are interested in the following maximal risks

$$\Psi_{n,N}^H(\hat{\Gamma}; \mathcal{P}) := \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{(\mathcal{D}_n, \mathcal{D}_N)} \mathbb{E}_{\mathbb{P}_X} \left| \hat{\Gamma}(X) \triangle \Gamma_\beta^*(X) \right| \quad (\text{Hamming risk}) ,$$

$$\Psi_{n,N}^E(\hat{\Gamma}; \mathcal{P}) := \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{(\mathcal{D}_n, \mathcal{D}_N)} R_\beta(\hat{\Gamma}) - R_\beta(\Gamma_\beta^*) \quad (\text{Excess risk}) ,$$

$$\Psi_{n,N}^D(\hat{\Gamma}; \mathcal{P}) := \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{(\mathcal{D}_n, \mathcal{D}_N)} \left[ \left| P(\hat{\Gamma}) - P(\Gamma_\beta^*) \right| + \left| \beta - I(\hat{\Gamma}) \right| \right] \quad (\text{Discrepancy}) ,$$

where  $\mathbb{E}_{(\mathcal{D}_n, \mathcal{D}_N)}$  denotes the expectation *w.r.t.*  $\mathbb{P}^{\otimes n} \otimes \mathbb{P}_X^{\otimes N}$ . These maximal risks are arising in a natural way in the context of the confidence set estimation with controlled expected size. The risk  $\Psi_{n,N}^H(\hat{\Gamma}; \mathcal{P})$  corresponds to the estimation of the  $\beta$ -Oracle through the Hamming distance. The second risks is directly connected with Proposition 1.2, which gives a description of the  $\beta$ -Oracle as a minimizer of  $R_\beta(\cdot)$ . As the goal in this problem is to construct a procedure  $\hat{\Gamma}$  that exhibits a low error  $P(\hat{\Gamma})$  and low cardinal discrepancy  $|\beta - I(\hat{\Gamma})|$ , it is natural to consider  $\Psi_{n,N}^D(\hat{\Gamma}; \mathcal{P})$  which is composed of both.

Finally, we are in position to define the notion of the minimax rate. The minimax rate in this context is not only determined by the family of distributions  $\mathcal{P}$  but also by the family of estimators  $\hat{\mathbf{\Gamma}} \subset \hat{\mathbf{Y}}$  that we consider.

**Definition 1.5** (Minimax rate of convergence). *For a given family  $\mathcal{P}$  of joint distributions on  $\mathbb{R}^d \times [K]$  and a given family of estimators  $\hat{\mathbf{\Gamma}} \subset \hat{\mathbf{Y}}$  the minimax rates are defined as*

$$\Psi_{n,N}^\square(\hat{\mathbf{\Gamma}}; \mathcal{P}) := \inf_{\hat{\Gamma} \in \hat{\mathbf{\Gamma}}} \Psi_{n,N}^\square(\hat{\Gamma}; \mathcal{P}) ,$$

where  $\square$  is H, E or D.

The main families of estimators that we study are the *supervised*  $\hat{\mathbf{Y}}_{\text{SE}}$  and the *semi-supervised*  $\hat{\mathbf{Y}}_{\text{SSE}}$  estimators. Obviously, since  $\hat{\mathbf{Y}} = \hat{\mathbf{Y}}_{\text{SE}} \cup \hat{\mathbf{Y}}_{\text{SSE}}$  and  $\hat{\mathbf{Y}}_{\text{SE}} \cap \hat{\mathbf{Y}}_{\text{SSE}} = \emptyset$ , we have the following relation

$$\Psi_{n,N}^\square(\hat{\mathbf{Y}}; \mathcal{P}) = \Psi_{n,N}^\square(\hat{\mathbf{Y}}_{\text{SE}}; \mathcal{P}) \bigwedge \Psi_{n,N}^\square(\hat{\mathbf{Y}}_{\text{SSE}}; \mathcal{P}) .$$

As a consequence, a lower and an upper bounds on  $\Psi_{n,N}^\square(\hat{\Upsilon}_{\text{SE}}; \mathcal{P})$ ,  $\Psi_{n,N}^\square(\hat{\Upsilon}_{\text{SSE}}; \mathcal{P})$  yield the bounds on the minimax rate over all estimators.

#### 1.4. Related works

Confidence set approach for classification was pioneered by Vovk (2002a,b); Vovk, Gammernan and Shafer (2005) by the means of conformal prediction theory. They rely on non-conformity measures which are based on some pattern recognition methods, and develop an asymptotic theory. In this work, we consider a statistical perspective of confidence set classification and put our focus on non-asymptotic minimax theory.

The problem of confidence set multi-class classification has strong ties with the binary classification with reject option, also known as binary classification with abstention in machine learning literature. In the binary classification with rejection, a classifier is allowed to output some special symbol, which indicates the rejection. Such type of classifiers can be seen as confidence sets, which are allowed to output  $\emptyset$  or  $\{0, 1\}$  and are interpreted as reject. This line of research was initiated by Chow (1957, 1970) in the context of information retrieval, where a predefined cost of rejection was considered. An extensive statistical study of this framework was carried in (Herbei and Wegkamp, 2006; Bartlett and Wegkamp, 2008; Wegkamp and Yuan, 2011).

Instead of considering a fixed cost for rejection, which might be too restrictive, one may define two entities: probability of rejection and the probability of missclassification. In the spirit of conformal prediction, Lei (2014) aims at minimizing the probability rejection provided a fixed upper bound on the probability of missclassification. In contrast, Denis and Hebiri (2015) consider a reversed problem of minimizing the probability of missclassification given a fixed upper bound on the probability of rejection.

Once the multi-class classification is considered, there are several possible ways to extend the binary case: the confidence set approach and the rejection approach. The reject counterpart is a more studied and known version, though it lacks statistical analysis. To the best of our knowledge the only work which provides statistical guarantees is (Ramaswamy, Tewari and Agarwal, 2018).

As for the confidence set approach there are again two possibilities, similar to the binary case. The one that is considered in this work was proposed by Denis and Hebiri (2017), where the authors analyse an ERM algorithm and derive oracle inequalities under the margin assumption (Tsybakov, 2004). More specifically, they consider a convex surrogate of the error  $P(\cdot)$  which relies on a convex real valued loss function  $\phi$ . For a suitable choice of the convex function  $\phi$  they show that, under Assumption 1.1, their  $\beta$ -Oracle satisfies

$$\Gamma_\beta^*(\cdot) = \left\{ k \in [K] : f_k^*(\cdot) \geq G_{f^*}^{-1}(\beta) \right\} ,$$

where the function  $f^*$  depends on  $\phi$  and the value of  $G_{f^*}^{-1}(\beta)$  is defined similarly to the present manuscript. They propose a two step estimation procedure of the

$\beta$ -Oracle set. Based on the ERM algorithm, they first estimate  $f^*$  and in the second step, they estimate the threshold  $G_{f^*}^{-1}(\beta)$  with an unlabeled sample. This procedure is in the same spirit as the semi-supervised procedure (1.5). Under mild assumptions, they provide an upper bound on the excess risk and obtain a rate of convergence of order  $(n/\log n)^{-\alpha/(\alpha+s)} + N^{-1/2}$ , with  $s$  being a parameter that depends on the function  $\phi$  and  $\alpha$  being the margin parameter. Note that this rate is slower than the rate obtained in the standard classification framework.

The conformal prediction theory (Vovk, Gammerman and Shafer, 2005) suggests to minimize the information level with a fixed budget on the error level. Statistical properties of this framework were considered in the work of Sadinle, Lei and Wasserman (2018). Their objective is formulated for some  $a \in (0, 1)$  as

$$\Gamma_a^* \in \arg \min \{I(\Gamma) : \Gamma \in \Upsilon \text{ s.t. } P(\Gamma) \leq a\} \text{ ,}$$

and such a confidence set is called a least ambiguous confidence set with bounded error rate. The authors show that under Assumption 1.1 this oracle set can be described as a thresholding of the regression function

$$\Gamma_a^*(\cdot) = \{k \in [K] : p_k(\cdot) \geq t_a\} \text{ ,}$$

where the threshold  $t_a$  is defined as

$$t_a = \sup \left\{ t \in [0, 1] : \sum_{k=1}^L \mathbb{P}(p_k(X) \geq t | Y = k) \mathbb{P}(Y = k) \geq 1 - a \right\} \text{ .}$$

Notice that this framework is very similar to (Denis and Hebiri, 2017) in the treatment of the Bayes optimal confidence set, as in both cases they are obtained via thresholding of the posterior distribution of the labels. Sadinle, Lei and Wasserman (2018) also proceed in two steps as here, that is, they first estimate the posterior distribution  $p_k(\cdot)$  for all  $k \in [K]$  and estimate the threshold  $t_a$  after. However, they require the second *labeled* dataset for the estimator of  $t_a$ , due to the presence of  $\mathbb{P}(Y = k)$ , the marginal distribution of the labels. Besides, their theoretical analysis is carried out under a different set of assumptions on the joint distribution  $\mathbb{P}$ . Apart from the standard margin assumption, they require a so-called detectability, that is, they require that the upper bound in the margin assumption is tight. Under these assumptions they provide an upper bound on the Hamming excess risk and obtain a rate of convergence of order  $\mathcal{O}((n/\log n)^{-1/2})$ .

Interestingly, both approaches can be encompassed into the constrained estimation framework (Anbar, 1977; Lepskii, 1990; Brown and Low, 1996), where one would like to construct an estimator with some prescribed properties. These properties are typically reflected by the form of the risk which in our case is the discrepancy measure, that is, the sum of error and information discrepancies. Thus, both frameworks of Sadinle, Lei and Wasserman (2018); Denis and Hebiri (2017) can be seen as an extension of the constrained estimation to the classification problems. From the modeling point of view, we believe that the two



frameworks can co-exist nicely and a particular choice depends on the considered application. The major difference between the present work and those by [Denis and Hebiri \(2017\)](#) and [Sadinle, Lei and Wasserman \(2018\)](#) is the minimax analysis which we provide here and our treatment of semi-supervised techniques.

As already pointed out, the confidence set estimation problem is closely related to the level set estimation setup ([Hartigan, 1987](#); [Polonik, 1995](#); [Tsybakov, 1997](#); [Rigollet and Vert, 2009](#)). This problem focuses on the estimation of a level set defined as

$$\Gamma_p(\lambda) = \{x \in \mathbb{R}^d : p(x) \geq \lambda\},$$

where  $p$  is the density of the observations and  $\lambda > 0$  is some fixed value. Given a sample  $X_1, \dots, X_n$  distributed according the density  $p$  the goal is to estimate  $\Gamma_p(\lambda)$ . In ([Rigollet and Vert, 2009](#)), the authors study plug-in density level set estimators through the measure of symmetric differences and *the excess mass*. In confidence set estimation the measure of symmetric differences is the Hamming risk whereas *the excess mass* is the excess risk. They show that kernel based estimators are optimal in the minimax sense over a Hölder class of densities and under a margin type assumption ([Polonik, 1995](#); [Tsybakov, 2004](#)). In particular, they derive fast rates of convergence, that is faster than  $n^{-1/2}$ , for *the excess mass*. In the level set estimation problem, the threshold  $\lambda$  is chosen beforehand; whereas in our work, the threshold  $G^{-1}(\beta)$  depends on the distribution of the data which makes the statistical analysis more difficult.

On the other part, the confidence set estimation problem is directly related to the standard classification settings. This problem has been widely studied from a theoretical point of view in the binary classification framework. [Audibert and Tsybakov \(2007\)](#) study the statistical performance of plug-in classification rules under assumptions which involve the smoothness of the regression function and the margin condition. In particular, they derive fast rates of convergence for plug-in classifiers based on local polynomial estimators ([Stone, 1977](#); [Tsybakov, 1986](#); [Audibert and Tsybakov, 2007](#)) and show their optimality in the minimax sense. One of the aim of present work is to extend these results to the confidence set classification framework.

Another part of our work is to provide a comparison between supervised and semi-supervised procedures. Semi-supervised methods are studied in several papers ([Vapnik, 1998](#); [Rigollet, 2007](#); [Singh, Nowak and Zhu, 2009](#); [Bellec et al., 2018](#)) and references therein. A simple intuition can be provided on whether one should or not expect a superior performance of the semi-supervised approach. Imagine a situation when the unlabeled sample  $\mathcal{D}_N$  is so large that one can approximate  $\mathbb{P}_X$  up to any desired precision, then, if the optimal decision is independent of  $\mathbb{P}_X$ , the semi-supervised estimators are not to be considered superior over the supervised estimation. This is the case in a lot of classical problems of statistics, where the inference is solely governed by the behavior of the conditional distribution  $\mathbb{P}_{Y|X}$  (for instance regression or binary classification). The situation might be different once the optimal decision relies on the marginal distribution  $\mathbb{P}_X$ . In this case, as suggested by our findings, the semi-supervised

approach might or not outperform the supervised one even in the context of the same problem. Similar conclusions were stated by [Singh, Nowak and Zhu \(2009\)](#) in the context of learning under the cluster assumption ([Rigollet, 2007](#)).

### 1.5. Main contributions

Bellow we summarize our contributions.

- Our results focus on the case where the regression  $p$  belongs to a Hölder class and satisfy the margin condition. Under these assumptions, we establish lower bounds on the minimax rates, defined in [Section 1.3](#) in the confidence set framework.
- As important consequences of our results, we first show that top- $\beta$  type procedures are in general inconsistent. Furthermore, by providing a rigorous definition of the semi-supervised and supervised estimators, we describe the situations when the semi-supervised estimation should be considered superior to its supervised counterpart. Interestingly, our analysis suggests that these regimes are governed by the interplay of the family of distributions and by the considered measure of performance. Besides, we show that in our settings supervised procedures cannot achieve fast rates, that is, its rate cannot be faster than  $n^{-1/2}$ . In contrast, some other classical settings ([Audibert and Tsybakov, 2007](#); [Rigollet and Vert, 2009](#); [Herbei and Wegkamp, 2006](#)) allow to achieve faster rates for supervised methods.
- We provide supervised and semi-supervised estimation procedures, which are optimal or optimal up to an extra logarithmic factor. Importantly, our results show that semi-supervised plug-in procedure based on local polynomial estimators can achieve fast rates, provided that the size of the unlabeled samples is large enough.
- Finally, we perform a numerical evaluation of the proposed plug-in algorithms against the top- $\beta$  counterparts. This part supports our theoretical results and empirically demonstrates the reason to consider more involved procedures.

### 1.6. Organization of the paper

The paper is organized as follow. In [Section 2](#), we put some additional notation and introduce the family of distributions  $\mathcal{P}$  that we consider. [Section 3](#) is devoted to the lower bounds on the minimax rates and their implications. In [Section 4](#) we introduce the proposed algorithm, establish upper bounds for it, and evaluate its numerical performance. We conclude this paper by [Sections 5](#) and [6](#) where we discuss and sum-up our results.

## 2. Class of confidence sets

First let us introduce some generic notation that is used throughout this work. For two numbers  $a, a' \in \mathbb{R}$  we denote by  $a \vee a'$  (resp.  $a \wedge a'$ ) the maximum (resp. minimum) between  $a$  and  $a'$ . For a positive real number  $a$  we denote by  $\lfloor a \rfloor$  (resp.  $\lceil a \rceil$ ) the largest (resp. the smallest) non-negative integer that is less than or equal (resp. greater than or equal) to  $a$ . The standard Euclidean norm of a vector  $x \in \mathbb{R}^d$  is denoted by  $\|x\|$  and the standard Lebesgue measure is denoted by  $\text{Leb}(\cdot)$ . A Euclidean ball centered at  $x \in \mathbb{R}^d$  of radius  $r > 0$  is denoted by  $\mathcal{B}(x, r)$ . For an arbitrary Borel measure  $\mu$  on  $\mathbb{R}^d$  that is absolutely continuous w.r.t. the Lebesgue measure we denote by  $\text{supp}(\mu)$  its support, that is, the set where the Radon-Nikodym derivative of  $\mu$  w.r.t.  $\text{Leb}$  is strictly positive. For a vector function  $p : \mathbb{R}^d \mapsto \mathbb{R}^K$  and a Borel measure  $\mu$  on  $\mathbb{R}^d$  we define the infinity norm of  $p$  as

$$\|p\|_{\infty, \mu} := \inf \left\{ C \geq 0 : \max_{k \in [K]} |p_k(x)| \leq C, \text{ a.e. } x \in \mathbb{R}^d \text{ w.r.t. } \mu \right\}.$$

In this work  $C$  or its lower-cased versions always refer to some constants which might differ from line to line. Importantly, all these constants are independent of  $n, N$  but could depend on  $K, d$  and other parameters which are assumed to be fixed. Before introducing the families of distributions  $\mathcal{P}$  that are considered in this work we need the following definitions.

**Assumption 2.1** ( $\alpha$ -margin assumption). *We say that the distribution  $\mathbb{P}$  of the pair  $(X, Y) \in \mathbb{R}^d \times [K]$  satisfies  $\alpha$ -margin assumption if there exists  $C_1 > 0$  and  $t_0 \in (0, 1)$  such that for every positive  $t \leq t_0$*

$$\mathbb{P}_X (0 < |p_k(X) - G^{-1}(\beta)| \leq t) \leq C_1 t^\alpha.$$

Let us point out an important consequence of Assumption 1.1. We have that the condition

$$\mathbb{P}_X (|p_k(X) - G^{-1}(\beta)| \leq t) \leq C_1 t^\alpha,$$

for all  $t \in [0, t_0]$  is equivalent to Assumption 2.1. Indeed, since the random variables  $p_k(X)$ 's cannot concentrate at a constant level, in particular at  $G^{-1}(\beta)$ . Moreover, again due to the continuity Assumption 1.1 we have

$$\lim_{t \rightarrow +0} \mathbb{P}_X (|p_k(X) - G^{-1}(\beta)| \leq t) = 0,$$

thus the  $\alpha$ -margin Assumption 2.1 specifies the rate of this convergence. Finally, the restriction of the range of  $t$  to  $[0, t_0]$  in  $\alpha$ -margin Assumption 2.1 does not affect its global behavior as for all  $t \in [0, 1]$

$$\mathbb{P}_X (0 < |p_k(X) - G^{-1}(\beta)| \leq t) \leq c_1 t^\alpha, \quad \text{with } c_1 = C_1 \vee t_0^{-\alpha}.$$

Let  $c_0$  and  $r_0$  be two positive constants. We say that a Borel set  $A \subset \mathbb{R}^d$  is a  $(c_0, r_0)$ -regular set if

$$\text{Leb}(A \cap \mathcal{B}(x, r)) \geq c_0 \text{Leb}(\mathcal{B}(x, r)), \quad \forall r \in (0, r_0], \forall x \in A.$$

**Definition 2.2** (Strong density). *We say that the probability measure  $\mathbb{P}_X$  on  $\mathbb{R}^d$  satisfies the  $(\mu_{\min}, \mu_{\max}, c_0, r_0)$ -strong density assumption if it is supported on a compact  $(c_0, r_0)$ -regular set  $A \subset \mathbb{R}^d$  and has a density  $\mu$  w.r.t. the Lebesgue measure such that  $\mu(x) = 0$  for all  $x \in \mathbb{R}^d \setminus A$  and*

$$0 < \mu_{\min} \leq \mu(x) \leq \mu_{\max} < \infty, \quad \forall x \in A .$$

**Definition 2.3** (Hölder class, [Tsybakov \(2008\)](#)). *We say that a function  $h : \mathbb{R}^d \rightarrow \mathbb{R}$  is  $(\gamma, L)$ -Hölder for  $\gamma > 0$  and  $L > 0$  if  $h$  is  $\lfloor \gamma \rfloor$  times continuously differentiable and  $\forall x, x' \in \mathbb{R}^d$  we have*

$$|h(x') - h_x(x')| \leq L \|x - x'\|^\gamma ,$$

where  $h_x(\cdot)$  is the Taylor polynomial of degree  $\lfloor \gamma \rfloor$  of  $h(\cdot)$  at the point  $x \in \mathbb{R}^d$ . Consequently, the set of all functions from  $\mathbb{R}^d$  to  $\mathbb{R}$  satisfying the above conditions is called  $(\gamma, L, \mathbb{R}^d)$ -Hölder and is denoted by  $\mathcal{H}(\gamma, L, \mathbb{R}^d)$ .

**Definition 2.4.** *We denote by  $\mathcal{P}(L, \gamma, \alpha)$  a set of joint distributions on  $\mathbb{R}^d \times [K]$  which satisfies the following conditions*

- the marginal  $\mathbb{P}_X$  satisfies the  $(\mu_{\min}, \mu_{\max}, c_0, r_0)$ -strong density,
- for all  $k \in [K]$  the  $k^{\text{th}}$  regression function  $p_k(\cdot) = \mathbb{P}(Y = k | X = \cdot)$  belongs to the  $(\gamma, L, \mathbb{R}^d)$ -Hölder class, that is  $p_k \in \mathcal{H}(\gamma, L, \mathbb{R}^d)$  for all  $k \in [K]$ ,
- for all  $k \in [K]$  the regression function  $p_k$  satisfy the  $(C_1, \alpha, \beta)$ -Margin assumption,
- for all  $k \in [K]$ , the cumulative distribution function  $F_{p_k}$  of  $p_k(X)$  is continuous.

The family of distributions  $\mathcal{P}(L, \gamma, \alpha)$  is similar to the one considered in ([Audibert and Tsybakov, 2007](#)) in the context of binary classification. The only major difference is the continuity Assumption 1.1, which does not allow to reuse in a straightforward way their construction for lower bounds.

### 3. Lower bounds

The main results in the present work are the lower bounds we provide in this section. In particular, we establish in Section 3.1 the inconsistency of top- $\beta$  procedures (see Eq. (1.3) for a definition of the method). Therefore more elaborate methods are required in this framework. As pointed out in the introduction, we distinguish two types of estimators: *supervised* and *semi-supervised* for which we provide lower bounds in Section 3.2. The obtained rates highlight the benefit of the semi-supervised approach in the context of the confidence set classification.

Before considering the lower bounds, let us first display connection between the different minimax rates. Such links are used in the proofs of the lower bounds.

**Proposition 3.1.** *Let  $\Gamma$  be a measurable function from  $\mathbb{R}^d$  to  $2^{[K]}$ ,  $\beta \in [k]$  and assume that Assumption 1.1 is fulfilled, then*

$$\begin{aligned} P(\Gamma) - P(\Gamma_\beta^*) &= R_\beta(\Gamma) - R_\beta(\Gamma_\beta^*) + G^{-1}(\beta) (\beta - I(\Gamma)) \quad , \\ R_\beta(\Gamma) - R_\beta(\Gamma_\beta^*) &= \sum_{k=1}^K \mathbb{E}_{\mathbb{P}_X} \left[ |p_k(X) - G^{-1}(\beta)| \mathbf{1}_{\{k \in \Gamma(X) \Delta \Gamma_\beta^*(X)\}} \right] \quad . \end{aligned}$$

Furthermore, if additionally Assumption 2.1 is satisfied with  $\alpha > 0$ , then there exist  $C > 0$  which depends only on  $K, \alpha, C_1$  such that for any pair of confidence set classifiers  $\Gamma, \Gamma'$  it holds that

$$\mathbb{E}_{\mathbb{P}_X} \left| \Gamma(X) \Delta \Gamma'(X) \right| \leq C (R_\beta(\Gamma) - R_\beta(\Gamma'))^{\alpha/(\alpha+1)} \quad . \quad (3.1)$$

**Proposition 3.2.** *For any  $K \geq 2$ ,  $\beta \in [K]$  and  $n, N \in \mathbb{N}$  the following relation between minimax rates holds:*

$$\Psi_{n,N}^H(\hat{\Gamma}; \mathcal{P}) \geq \Psi_{n,N}^D(\hat{\Gamma}; \mathcal{P}) \geq \Psi_{n,N}^E(\hat{\Gamma}; \mathcal{P}) \quad .$$

Proposition 3.1, and in particular Eq. (3.1) gives an easy way to establish a lower bound on  $\Psi_{n,N}^E(\hat{\Gamma}; \mathcal{P})$  via a lower bound on the Hamming distance  $\Psi_{n,N}^H(\hat{\Gamma}; \mathcal{P})$ . However, this approach does not allow to get  $(N+n)^{-1/2}$  (resp.  $n^{-1/2}$ ) part of the rate in the lower bound of  $\Psi_{n,N}^E(\hat{\Gamma}_{\text{SSE}}; \mathcal{P})$  (resp.  $\Psi_{n,N}^E(\hat{\Gamma}_{\text{SE}}; \mathcal{P})$ ). Besides, Proposition 3.2 allows to prove a lower bound on the discrepancy  $\Psi_{n,N}^D(\hat{\Gamma}; \mathcal{P})$  with the correct rate via the lower bound on the excess risk  $\Psi_{n,N}^E(\hat{\Gamma}; \mathcal{P})$ .

### 3.1. Inconsistency of the top- $\beta$ procedure

Before stating our results on the supervised and the semi-supervised estimators, we discuss another interesting class of confidence sets, which might be a natural choice at the first sight. We consider estimators which consists of  $\beta$  classes at every point  $x \in \mathbb{R}^d$  since such estimators naturally satisfy  $I(\hat{\Gamma}) = \beta$ . Let us denote by  $\hat{\Gamma}_\beta$  the set of all estimators  $\hat{\Gamma}$  such that  $|\hat{\Gamma}(x)| = \beta$  for all  $x \in \mathbb{R}^d$ , that is,

$$\hat{\Gamma}_\beta = \left\{ \hat{\Gamma} \in \hat{\Gamma} : |\hat{\Gamma}(x)| = \beta, \text{ a.e. } x \in \mathbb{R}^d \text{ w.r.t. Lebesgue} \right\} \quad .$$

Despite an obvious restriction on the cardinal of the confidence sets, the family of estimators  $\hat{\Gamma}_\beta$  is rather broad. Indeed, every procedure which estimates the regression functions  $p_k(\cdot)$ 's and includes the top  $\beta$  scores as the output are included in  $\hat{\Gamma}_\beta$ . The nature of the estimator can also be different, that is, the estimates could be based on the ERM, non-parametric or parametric approaches. Clearly, the family  $\hat{\Gamma}_\beta$  is neither included in  $\hat{\Gamma}_{\text{SE}}$  nor in  $\hat{\Gamma}_{\text{SSE}}$  and has a non-trivial intersection with both. The next result states that there is no uniformly consistent estimator  $\hat{\Gamma} \in \hat{\Gamma}_\beta$  over the family of distributions  $\mathcal{P}(L, \gamma, \alpha)$ .

**Proposition 3.3.** *Assume that  $K \geq 4$ ,  $\beta \in [\lfloor K/2 \rfloor - 1]$  and  $\beta \geq 2$ , then for all  $n, N \in \mathbb{N}$  we have*

$$\Psi_{n,N}^E(\hat{\Upsilon}_\beta; \mathcal{P}(L, \gamma, \alpha)) \geq \frac{\beta - 1}{4K}.$$

The proof builds an explicit construction of a distribution  $\mathbb{P}$  whose  $\beta$ -Oracle satisfies  $|\Gamma_\beta^*(x)| > \beta$  for all  $x$  in some  $A \subset \mathbb{R}^d$  with  $\mathbb{P}_X(A) > 0$ . Indeed, if such a distribution exists then there is no estimator in  $\hat{\Upsilon}_\beta$  that would consistently estimate this  $\beta$ -Oracle. The negative result established in Proposition 3.3 is rather instructive by itself as it advocates that a more involved estimation procedure ought to be constructed.

### 3.2. Supervised vs semi-supervised estimation

Clearly, estimators which achieve the infimum in the minimax rates are either supervised or semi-supervised, thus a lower bound on  $\Psi_{n,N}^\square(\hat{\Upsilon}_{SE}; \mathcal{P})$  together with a lower bound on  $\Psi_{n,N}^\square(\hat{\Upsilon}_{SSE}; \mathcal{P})$  yield a lower bound on  $\Psi_{n,N}^\square(\hat{\Upsilon}; \mathcal{P})$ . However, a lower bound on  $\Psi_{n,N}^\square(\hat{\Upsilon}; \mathcal{P})$  does not discriminate between the supervised and the semi-supervised estimators.

**Theorem 3.4** (Supervised estimation). *Let  $K \leq 3$ ,  $\beta \in [\lfloor K/2 \rfloor - 1]$ . If  $2\alpha \lceil \frac{\gamma}{2} \rceil \leq d$ , then there exist constants  $c, c', c'' > 0$  such that for all  $n, N \in \mathbb{N}$*

$$\begin{aligned} \Psi_{n,N}^H(\hat{\Upsilon}_{SE}; \mathcal{P}(L, \gamma, \alpha)) &\geq c \left( n^{-\frac{\alpha\gamma}{2\gamma+d}} \bigvee n^{-1/2} \right), \\ \Psi_{n,N}^E(\hat{\Upsilon}_{SE}; \mathcal{P}(L, \gamma, \alpha)) &\geq c' \left( n^{-\frac{(1+\alpha)\gamma}{2\gamma+d}} \bigvee n^{-1/2} \right), \\ \Psi_{n,N}^D(\hat{\Upsilon}_{SE}; \mathcal{P}(L, \gamma, \alpha)) &\geq c'' \left( n^{-\frac{(1+\alpha)\gamma}{2\gamma+d}} \bigvee n^{-1/2} \right). \end{aligned}$$

Based on this results we observe that the lower bound for the Hamming risk  $\Psi_{n,N}^H$  is slower than those for the other risks. It is even more significant that the best rate that a supervised estimator can achieve for all of the risks is  $n^{-1/2}$  even if the margin assumption holds. This is the major difference with the classical settings where the value of threshold is known (such as classification and level set estimation). Indeed, under the same assumptions on the family of distributions, besides the continuity Assumption 1.1, the minimax rate in those frameworks is  $n^{-(1+\alpha)\gamma/(2\gamma+d)}$  as proved for instance in (Audibert and Tsybakov, 2007; Rigollet and Vert, 2009). Next theorem deals with semi-supervised procedures and displays another behavior.

**Theorem 3.5** (Semi-supervised estimation). *Let  $K \geq 3$ ,  $\beta \in [\lfloor K/2 \rfloor - 1]$ . If*

$2\alpha \lceil \frac{\gamma}{2} \rceil \leq d$ , then there exist constants  $c, c', c'' > 0$  such that for all  $n, N \in \mathbb{N}$

$$\begin{aligned}\Psi_{n,N}^H(\hat{\Upsilon}_{\text{SSE}}; \mathcal{P}(L, \gamma, \alpha)) &\geq c \left( n^{-\frac{\alpha\gamma}{2\gamma+d}} \bigvee (n+N)^{-1/2} \right) , \\ \Psi_{n,N}^E(\hat{\Upsilon}_{\text{SSE}}; \mathcal{P}(L, \gamma, \alpha)) &\geq c' \left( n^{-\frac{(1+\alpha)\gamma}{2\gamma+d}} \bigvee (n+N)^{-1/2} \right) , \\ \Psi_{n,N}^D(\hat{\Upsilon}_{\text{SSE}}; \mathcal{P}(L, \gamma, \alpha)) &\geq c'' \left( n^{-\frac{(1+\alpha)\gamma}{2\gamma+d}} \bigvee (n+N)^{-1/2} \right) .\end{aligned}$$

First, observe that the lower bound for the Hamming distance is, as in the supervised setting, worse than for the other measures of performance. However there is a major difference with the supervised case: as compared to Theorem 3.4, it is possible for a semi-supervised estimator to achieve rates that are faster than  $n^{-1/2}$  if the size of the unlabeled dataset  $N \in \mathbb{N}$  is large enough. In particular, when we consider  $\Psi_{n,N}^E$  or  $\Psi_{n,N}^D$  the following relations are necessary to get fast rates

$$(n+N)^{-1/2} = o\left(n^{-(1+\alpha)\gamma/(2\gamma+d)}\right), \quad n^{-(1+\alpha)\gamma/(2\gamma+d)} = o(n^{-1/2}) .$$

In this case, we recover the same fast rates as in the classical settings of classification and level set estimation. It suggests that the lack of knowledge of the threshold  $G^{-1}(\beta)$  does not alter the quality of estimation for the semi-supervised procedure, provided that  $N$  is sufficiently large. Next corollary makes these observations clearer.

**Corollary 3.6.** *Assume that the rates in Theorem 3.5 (resp. Theorem 3.4) are minimax, that is, there exist a confidence set  $\hat{\Upsilon}_{\text{SSE}}$  (resp.  $\hat{\Upsilon}_{\text{SE}}$ ) that achieves these rates. Regarding  $\Psi_{n,N}^E$  and  $\Psi_{n,N}^D$  the following conclusions hold*

- *There is no semi-supervised estimator that achieves faster rate than  $\hat{\Upsilon}_{\text{SE}}$  if:*

$$\begin{cases} \frac{(1+\alpha)\gamma}{2\gamma+d} \leq 1/2 \\ N \in \mathbb{N} \end{cases} \quad \text{or} \quad \begin{cases} \frac{(1+\alpha)\gamma}{2\gamma+d} > 1/2 \\ N = \mathcal{O}(n) \end{cases} .$$

- *The rate of  $\hat{\Upsilon}_{\text{SSE}}$  is faster than the rate of any supervised estimator if:*

$$\frac{(1+\alpha)\gamma}{2\gamma+d} > 1/2 \quad \text{and} \quad n = o(N) .$$

*Moreover, if there exists  $\rho > 0$  such that  $n^{1+\rho} = o(N)$ , then the rate of  $\hat{\Upsilon}_{\text{SSE}}$  is polynomially faster than  $n^{-1/2}$ .*

- *The rate of  $\hat{\Upsilon}_{\text{SSE}}$  is fast similarly to the classical frameworks if*

$$\frac{(1+\alpha)\gamma}{2\gamma+d} > 1/2 \quad \text{and} \quad N = \Omega\left(n^{\frac{2(1+\alpha)\gamma}{2\gamma+d}}\right) .$$

Clearly, similar observation is true for the Hamming risk  $\Psi_{n,N}^H$ ; however the regime when improvement is possible thanks to semi-supervised approaches is narrowed as  $n^{-(1+\alpha)\gamma/(2\gamma+d)} = o(n^{-\alpha\gamma/(2\gamma+d)})$ . We summarize Corollary 3.6 in Table 3.2.

$\frac{(1+\alpha)\gamma}{2\gamma+d}$	$N, n$	SE rate	SSE rate	SSE > SE
$\leq \frac{1}{2}$	$N \in \mathbb{N}, n \in \mathbb{N}$	$n^{-\frac{(1+\alpha)\gamma}{2\gamma+d}}$	$n^{-\frac{(1+\alpha)\gamma}{2\gamma+d}}$	NO
$> \frac{1}{2}$	$N = \mathcal{O}(n)$	$n^{-\frac{1}{2}}$	$n^{-\frac{1}{2}}$	NO
$> \frac{1}{2}$	$n = o(N)$	$n^{-\frac{1}{2}}$	$N^{-\frac{1}{2}} \vee n^{-\frac{(1+\alpha)\gamma}{2\gamma+d}}$	YES
$> \frac{1}{2}$	$N = \Omega\left(n^{\frac{2(1+\alpha)\gamma}{2\gamma+d}}\right)$	$n^{-\frac{1}{2}}$	$n^{-\frac{(1+\alpha)\gamma}{2\gamma+d}}$	YES

TABLE 1

This table summarizes observations of Corollary 3.6 for  $\Psi_{n,N}^E$  and  $\Psi_{n,N}^D$ . Depending on the relations between  $\alpha, \gamma, d$  and  $N, n$  the semi-supervised approach can significantly improve the rates of convergence.

Essentially, the above results suggest that the advantage of the semi-supervised approaches over the supervised ones depends not only on the underlying family of distributions  $\mathcal{P}$  but also on the metric that is considered. Yet, necessary and sufficient conditions that must be imposed in general on the problem and the metric so that the semi-supervised estimation provably improve upon the supervised one remain an open problem.

A final remark we could make before going further concerns the assumption on the parameters  $\alpha$  and  $\gamma$ . The condition  $2\alpha\lceil\frac{\gamma}{2}\rceil \leq d$  in the lower bounds is slightly more restrictive than the conditions given in (Audibert and Tsybakov, 2007) (they have  $\alpha\gamma \leq d$ ). We believe that this is an artifact of our proof and could be avoided with a finer choice of hypotheses. Simple modifications of the lower bound of Audibert and Tsybakov (2007) do not work in our settings because their hypotheses are not satisfying Assumption 1.1. In contrast, the construction of Rigollet and Vert (2009) satisfies<sup>1</sup> Assumption 1.1 but their lower bound is limited by the condition  $\alpha\gamma \leq 1$ , that is, it does not cover the fast rates as long as the dimension  $d > 2$ .

### 3.3. Sketch of the proof

In order to prove the lower bounds of Theorems 3.4, 3.5 we actually prove two separate lower bounds on the minimax rates. The two lower bounds that we prove are naturally connected with the proposed two-step estimator in Eq. (1.5), that is, the first lower bound is connected with the problem of non-parametric estimation of  $p_k$  for all  $k \in [K]$  and the second describes the estimation of the unknown threshold  $G^{-1}(\beta)$ .

In particular, the first lower bound is closely related to the one provided in (Audibert and Tsybakov, 2007; Rigollet and Vert, 2009), however, the continuity Assumption 1.1 makes the proof more involved and results in a final construction of hypotheses that differs significantly. This part of our lower bound

<sup>1</sup>Modified properly to fit the classification framework.



relies on Fano's inequality in the form of Birgé (2005). The second lower bound is based on two hypotheses testing and is derived by constructing two different marginal distributions of  $X \in \mathbb{R}^d$  which are sufficiently close and a fixed regression function  $p(\cdot)$ . Crucially, these marginal distributions admit two different values of threshold  $G^{-1}(\beta)$  and thus two different  $\beta$ -Oracle. In this part we make use of Pinsker's inequality, see for instance (Tsybakov, 2008).

In order to discriminate the supervised and the semi-supervised procedures we make use of Definition 1.4. Notice that every supervised procedure thanks to Definition 1.4 is not "sensitive" to the expectation taken *w.r.t.* the unlabeled dataset  $\mathcal{D}_N$ , that is, randomness is only induced by the labeled dataset  $\mathcal{D}_n$ . This strategy allows to eliminate the dependence of the lower bound on the size of the unlabeled dataset  $\mathcal{D}_N$  for supervised procedures. Informally, the lower bound on  $\Psi_{n,N}^\square(\hat{\Upsilon}_{\text{SE}}; \mathcal{P})$  is obtained from the lower bound on  $\Psi_{n,N}^\square(\hat{\Upsilon}_{\text{SSE}}; \mathcal{P})$  by setting  $N = 0$ .

#### 4. Upper bounds

In this section, we show that we can build confidence set estimators that achieve, up to a logarithmic factor, the lower bounds stated in Theorems 3.4-3.5. In other words, those estimators are *nearly* optimal in the minimax sense. To come straight to the point, we delay the construction of the estimators to Section 4.1 and their properties to Section 4.2, and focus right now on their upper bounds.

**Theorem 4.1** (Supervised estimation). *Let  $K \in \mathbb{N}$ ,  $\beta \in [K - 1]$ , then there exists a supervised estimator  $\hat{\Gamma}_{\text{SE}} \in \hat{\Upsilon}_{\text{SE}}$  and a constant  $C > 0$  such that for all  $n, N \in \mathbb{N}$  we have*

$$\begin{aligned} \Psi_{n,N}^{\text{H}}(\hat{\Gamma}_{\text{SE}}; \mathcal{P}(L, \gamma, \alpha)) &\leq C \left( n^{-\frac{\alpha\gamma}{2\gamma+d}} \bigvee n^{-1/2} \right) , \\ \Psi_{n,N}^{\text{E}}(\hat{\Gamma}_{\text{SE}}; \mathcal{P}(L, \gamma, \alpha)) &\leq C' \left( \left( \frac{n}{\log n} \right)^{-\frac{(1+\alpha)\gamma}{2\gamma+d}} \bigvee n^{-1/2} \right) , \\ \Psi_{n,N}^{\text{D}}(\hat{\Gamma}_{\text{SE}}; \mathcal{P}(L, \gamma, \alpha)) &\leq C'' \left( \left( \frac{n}{\log n} \right)^{-\frac{(1+\alpha)\gamma}{2\gamma+d}} \bigvee n^{-1/2} \right) . \end{aligned}$$

**Theorem 4.2** (Semi-supervised estimation). *Let  $K \in \mathbb{N}$ ,  $\beta \in [K - 1]$ , then there exists a semi-supervised estimator  $\hat{\Gamma}_{\text{SSE}} \in \hat{\Upsilon}_{\text{SSE}}$  and constants  $C, C', C'' > 0$*

such that for all  $n, N \in \mathbb{N}$  we have

$$\begin{aligned}\Psi_{n,N}^H(\hat{\Gamma}_{\text{SSE}}; \mathcal{P}(L, \gamma, \alpha)) &\leq C \left( n^{-\frac{\alpha\gamma}{2\gamma+d}} \bigvee (n+N)^{-1/2} \right), \\ \Psi_{n,N}^E(\hat{\Gamma}_{\text{SSE}}; \mathcal{P}(L, \gamma, \alpha)) &\leq C' \left( \left( \frac{n}{\log n} \right)^{-\frac{(1+\alpha)\gamma}{2\gamma+d}} \bigvee (n+N)^{-1/2} \right), \\ \Psi_{n,N}^D(\hat{\Gamma}_{\text{SSE}}; \mathcal{P}(L, \gamma, \alpha)) &\leq C'' \left( \left( \frac{n}{\log n} \right)^{-\frac{(1+\alpha)\gamma}{2\gamma+d}} \bigvee (n+N)^{-1/2} \right).\end{aligned}$$

We show here that the lower bounds of Theorems 3.4-3.5 are achievable. In particular, in the case of Hamming risk, the upper bounds are optimal; whereas for the Excess risk and the Discrepancy, the upper bounds fit the lower bounds up to a logarithmic factor. Thus, the comments we made in Corollary 3.6 are correct. Let us mention that the presence of the logarithmic factor in these upper bounds is due to  $\ell_\infty$ -norm estimation (see Lemma 4.5).

Hamming risk as a measure of performance was considered in the settings of [Sadinle, Lei and Wasserman \(2018\)](#). They also establish upper bounds for this measure, though do not assess their optimality. Besides, as we already mentioned, [Denis and Hebiri \(2017\)](#) provide an upper bound on the excess risk in the context of ERM. Let us point out, that the comparison with these two works is not fair as the assumptions and even frameworks under which we and they formulate results are different.

#### 4.1. Construction of the estimators

Building estimators  $\hat{\Gamma}_{\text{SE}}$  and  $\hat{\Gamma}_{\text{SSE}}$  that reach the rates in the former upper bounds involves a preliminary estimators  $\hat{p}_k$  of the regression functions  $p_k$ ,  $k \in [K]$ . These estimators  $\hat{p}_k$  are constructed using an arbitrary half  $\mathcal{D}_{\lfloor n/2 \rfloor}$  of the labeled dataset  $\mathcal{D}_n$  and they satisfy the following assumptions.

**Assumption 4.3** (Exponential concentration). *There exist estimators  $\hat{p}_k$  for all  $k \in [K]$  based on  $\mathcal{D}_{\lfloor n/2 \rfloor}$  and positive constants  $C_1, C_2$  such that for all  $k \in [K]$  and all  $n \geq 2$  we have for all  $\delta > 0$*

$$\sup_{\mathbb{P} \in \mathcal{P}(L, \gamma, \alpha)} \mathbb{P}^{\otimes \lfloor n/2 \rfloor} (|\hat{p}_k(x) - p_k(x)| \geq \delta) \leq C_1 \exp \left( -C_2 n^{\frac{2\gamma}{2\gamma+d}} \delta^2 \right),$$

for almost all  $x \in \mathbb{R}^d$  w.r.t.  $\mathbb{P}_X$ .

**Assumption 4.4** (Continuity of CDF). *For all  $k \in [K]$  the cumulative distribution function  $F_{\hat{p}_k}(t) := \mathbb{P}_X(\hat{p}_k(X) \leq t)$  of  $\hat{p}_k(X)$  is almost surely  $\mathbb{P}^{\otimes \lfloor n/2 \rfloor}$  continuous on  $(0, 1)$ .*

First let us point out that Assumption 4.3 induces that there exists a constant  $C > 0$  such that for all  $n \geq 2$  and all  $\alpha > 0$

$$\sup_{\mathbb{P} \in \mathcal{P}(L, \gamma, \alpha)} \mathbb{E}_{\mathcal{D}_{\lfloor n/2 \rfloor}} \|p - \hat{p}\|_{\infty, \mathbb{P}_X}^{1+\alpha} \leq C \left( \frac{n}{\log n} \right)^{-\frac{(1+\alpha)\gamma}{2\gamma+d}}.$$

Assumption 4.3 is commonly used in the statistical community when we deal with rates of convergence in the classification settings (Audibert and Tsybakov, 2007; Lei, 2014; Sadinle, Lei and Wasserman, 2018). It is for instance satisfied by the locally polynomial estimator (Stone, 1977; Tsybakov, 1986; Audibert and Tsybakov, 2007). Assumption 4.4 can always be satisfied by slightly processing any estimator  $\hat{p}$ . Indeed, assume Assumption 4.4 fails to be satisfied by some estimator  $\hat{p}$ . It means that there exists a subset of  $\mathbb{R}^d$  of non-zero measure such that at least one  $\hat{p}_k$ , with  $k \in [K]$ , is constant on this set. Then, if we add a *deterministic* continuous function of a sufficiently bounded variation<sup>2</sup> to  $\hat{p}$  such regions can no longer exist.

Since, the threshold level  $G^{-1}(\beta)$  is not known beforehand, it ought to be estimated using data. A straightforward estimator of this threshold can be constructed using the unlabeled dataset  $\mathcal{D}_N$ . To make our presentation mathematically correct we introduce the following notation  $\mathcal{D}_n = \mathcal{D}_{\lfloor n/2 \rfloor} \cup \mathcal{D}_{\lceil n/2 \rceil}$ , where  $\mathcal{D}_{\lfloor n/2 \rfloor}$  is the dataset used to build the estimators  $\hat{p}_k$  for  $k \in [K]$ . Now, all the labels are removed from  $\mathcal{D}_{\lceil n/2 \rceil}$ , that is it consists of  $\lceil n/2 \rceil$  *i.i.d.* samples from  $\mathbb{P}_X$ . The supervised and semi-supervised estimators of  $G(\cdot)$  are defined as

$$\hat{G}_{\text{SE}}(\cdot) = \frac{1}{\lceil n/2 \rceil} \sum_{X \in \mathcal{D}_{\lceil n/2 \rceil}} \sum_{k=1}^K \mathbf{1}_{\{\hat{p}_k(X) > \cdot\}} ,$$

$$\hat{G}_{\text{SSE}}(\cdot) = \frac{1}{\lceil n/2 \rceil + N} \sum_{X \in \mathcal{D}_N \cup \mathcal{D}_{\lceil n/2 \rceil}} \sum_{k=1}^K \mathbf{1}_{\{\hat{p}_k(X) > \cdot\}} ,$$

respectively. Finally, we are in position to define  $\hat{\Gamma}_{\text{SE}}$  and  $\hat{\Gamma}_{\text{SSE}}$  as

$$\hat{\Gamma}_{\text{SE}}(x) = \left\{ k \in [K] : \hat{p}_k(x) \geq \hat{G}_{\text{SE}}^{-1}(\beta) \right\} ,$$

$$\hat{\Gamma}_{\text{SSE}}(x) = \left\{ k \in [K] : \hat{p}_k(x) \geq \hat{G}_{\text{SSE}}^{-1}(\beta) \right\} ,$$

for all  $x \in \mathbb{R}^d$ . Note that  $\hat{\Gamma}_{\text{SE}}$  is clearly supervised in the sense of Definition 1.4, as it is independent of the unlabeled sample  $\mathcal{D}_N$ . In contrast,  $\hat{\Gamma}_{\text{SSE}}$  is semi-supervised, since we can find two samples  $\mathcal{D}_N$  and  $\mathcal{D}'_N$  which induce different confidence sets. To show that the estimators introduced in this section satisfy the statements of Theorems 4.1-4.2 we refine the proof technique used in (Denis and Hebiri, 2017). That is, we introduce an intermediate quantity

$$\tilde{G}(\cdot) := \sum_{k=1}^K \mathbb{P}_X (\hat{p}_k(X) > \cdot) ,$$

and the associated confidence set, which we refer to as the pseudo Oracle confidence set given for all  $x \in \mathbb{R}^d$  by

$$\tilde{\Gamma}(x) := \left\{ k \in [K] : \hat{p}_k(x) \geq \tilde{G}^{-1}(\beta) \right\} .$$

---

<sup>2</sup>It is sufficient to make sure that adding the function preserves its statistical properties, that is, Assumption 4.3

The confidence set  $\tilde{\Gamma}$  assumes knowledge of the marginal distribution  $\mathbb{P}_X$  and is seen as an idealized version of both  $\hat{\Gamma}_{\text{SE}}$  and  $\hat{\Gamma}_{\text{SSE}}$ , note however, that the pseudo Oracle  $\tilde{\Gamma}$  is not an estimator.

#### 4.2. Properties of the plug-in confidence sets

An important step of our analysis is the following lemma, that bounds the difference between  $\tilde{G}^{-1}(\beta)$  and  $G^{-1}(\beta)$ .

**Lemma 4.5** (Upper bound on the thresholds). *Let Assumption 1.1 be satisfied, then for all  $\beta \in [K]$*

$$\left| G^{-1}(\beta) - \tilde{G}^{-1}(\beta) \right| \leq \|p - \hat{p}\|_{\infty, \mathbb{P}_X}, \quad \text{almost surely } \mathbb{P}^{\otimes n} \otimes \mathbb{P}_X^{\otimes N}.$$

The proof of Lemma 4.5 uses elementary properties of the generalized inverse functions which are provided in Appendix. Besides, let us mention, that the difference  $|G^{-1}(\beta) - \tilde{G}^{-1}(\beta)|$  resembles the Wasserstein infinity distance which gives an alternative approach to prove Lemma 4.5, see (Bobkov and Ledoux, 2016). Lemma 4.5 explains the extra  $\log n$  factor that appears in the upper bound, as the minimax estimation in sup norm contains the  $\log n$  factor, see for instance (Stone, 1982; Tsybakov, 2008). Another important property of the introduced estimators  $\hat{\Gamma}_{\text{SE}}$  and  $\hat{\Gamma}_{\text{SSE}}$  is obtained via Assumption 4.4. It describes the deviation of the information of  $\hat{\Gamma}_{\text{SE}}$  and  $\hat{\Gamma}_{\text{SSE}}$  from the desired level  $\beta$ .

**Proposition 4.6** (Denis and Hebiri (2017)). *Let  $\hat{p}_k$  for all  $k \in [K]$  be arbitrary estimators of the regression functions constructed using  $\mathcal{D}_{\lfloor n/2 \rfloor}$  that satisfies Assumption 4.4, then there exist constants  $C, C' > 0$  such that for all  $n, N \in \mathbb{N}$  it holds that*

$$\begin{aligned} \mathbb{E}_{(\mathcal{D}_n, \mathcal{D}_N)} \left| \beta - \mathbb{I}(\hat{\Gamma}_{\text{SE}}) \right| &\leq C n^{-1/2}, \\ \mathbb{E}_{(\mathcal{D}_n, \mathcal{D}_N)} \left| \beta - \mathbb{I}(\hat{\Gamma}_{\text{SSE}}) \right| &\leq C' (N + n)^{-1/2}. \end{aligned}$$

Note that if  $\hat{p}_k$  satisfies Assumption 4.4 for all  $k \in [K]$ , then  $\beta = \mathbb{I}(\tilde{\Gamma})$ . This simple fact is a step in the proof of Proposition 4.6. Finally, combination of Lemma 4.5, Proposition 4.6, Assumption 4.3 with the peeling argument used in (Audibert and Tsybakov, 2007, Lemma 3.1) yields the results of Theorems 4.1-4.2.

#### 4.3. Simulation study

The goal of this part is to numerically address the following points.

- 1) Is it more advantageous to go outside of the *classical* multi-class classification settings and consider the confidence set framework? To respond to this question we compute the Bayes optimal multi-class classifier and view it as a confidence set with one label. We compare this Bayes rule with the

$K = 10$		
$\beta$	$\beta$ -Oracle	top- $\beta$ Oracle
2	0.05 (0.01)	0.09 (0.01)
5	0.00 (0.00)	0.01 (0.00)

$K = 100$		
$\beta$	$\beta$ -Oracle	top- $\beta$ Oracle
2	0.39 (0.01)	0.42 (0.01)
5	0.20 (0.01)	0.22 (0.01)
10	0.09 (0.01)	0.11 (0.01)
20	0.03 (0.01)	0.04 (0.01)

TABLE 2

For each of the  $B = 100$  repetitions and each model, we derive the estimated errors  $P_M$  of the  $\beta$ -Oracle and of the top- $\beta$  Oracle w.r.t.  $\beta$ . We compute the means and standard deviations (between parentheses) over the  $B = 100$  repetitions. Top: the data are generated according to  $K = 10$  – Bottom: the data are generated according to  $K = 100$ .

$\beta$ -Oracle in terms of the error  $P(\cdot)$  using various values of  $\beta \in [K]$  and  $K \in \mathbb{N}$ .

- 2) How does the  $\beta$ -Oracle confidence set compares to another "Oracle" (top- $\beta$  Oracle) which simply includes classes corresponding to the largest values of  $p_k(\cdot)$ 's?
- 3) Does the proposed plug-in approach indeed gives a good approximation of the  $\beta$ -Oracle through the error  $P(\cdot)$  and the information  $I(\cdot)$ ?
- 4) Despite demonstrating the minimax inconsistency of the top- $\beta$  approach, we wonder whether in some scenarios it can achieve a comparable performance against our semi-supervised plug-in procedure.

We consider two simulation schemes depending on the parameter  $K \in \{10, 100\}$ . For each  $K$ , we generate  $(X, Y)$  according to a mixture model. More precisely,

- i) the label  $Y$  follows uniform distribution on  $[K]$ ;
- ii) conditional on  $Y = k$ , the feature  $X$  is generated according to a multivariate gaussian distribution with mean  $\mu_k \in \mathbb{R}^{10}$  and identity covariance matrix.

For each  $k \in [K]$ , the vectors  $\mu_k$  are *i.i.d.* realizations of uniform distribution

$\beta$	$K = 10$	$K = 100$
2	2.00 (0.03)	2.00 (0.03)
5	5.00 (0.08)	5.00 (0.06)
10	.	10.00 (0.13)
20	.	20.02 (0.31)

TABLE 3

For each of the  $B = 100$  repetitions and each model, we derive the estimated information levels  $I_M$  of the  $\beta$ -Oracle set w.r.t.  $\beta$ . We compute the means and standard deviations (in parentheses) over the  $B = 100$  repetitions. Left: the data are generated according to  $K = 10$  – Right: the data are generated according to  $K = 100$ .

on  $[0, 4]^{10}$ . For this distribution, we have

$$p_k(X) = \frac{f_k(X)}{\sum_{j=1}^K f_j(X)},$$

where for each  $k \in [K]$ ,  $f_k(X)$  is the density function of a multivariate gaussian distribution with mean parameter  $\mu_k$  and identity covariance matrix.

For each  $K$ , the *missclassification* error of the *classical* multi-class classification Bayes rule is evaluated based on a sufficiently large dataset. It is valued at 0.22 and at 0.60 for  $K = 10$  and for  $K = 100$  respectively. These values are relatively high, which suggests that confusion is induced by the large number of classes. Hence, it is reasonable to apply the confidence set approach to this problem. In the sequel, we aim at providing the estimation of the error of the  $\beta$ -Oracle. To this end, for  $\beta \in \{2, 5, 10, 20\}$  and each  $K$ , we repeat  $B$  times the following steps.

- i) simulate two datasets  $\mathcal{D}_N$  and  $\mathcal{D}_M$  with  $N = 10000$  and  $M = 1000$ ;
- ii) based on  $\mathcal{D}_N$ , we compute the empirical counterpart of  $G$  and provide an approximation of the  $\beta$ -Oracle  $\Gamma_\beta^*$  given in Eq. (1.1) (we recall that this step requires a dataset which contains only unlabeled features);
- iii) finally, over  $\mathcal{D}_M$ , we compute the empirical counterparts  $P_M$  (of  $P(\Gamma_\beta^*)$ ) and  $I_M$  (of  $I(\Gamma_\beta^*)$ ).

From this estimates, we compute the mean and the standard deviation of  $P_M$  and  $I_M$ . Tables 2 and 3 present values of the error and of the information which are achieved by the  $\beta$ -Oracle and by the top- $\beta$  Oracle.

We now move towards the construction of our semi-supervised plug-in estimators  $\hat{\Gamma}_{\text{SSE}}$ . For each  $K$  and each  $\beta$ , we evaluate the performance of  $\hat{\Gamma}_{\text{SSE}}$  according to three different estimations of the regression function: the  $\hat{p}_k$ 's are based on random forests, softmax regression and deep learning procedures. Let us point out, that for random forests and softmax regression algorithms, the

$K = 10$						
$\beta$	$\hat{\Gamma}_{\text{SSE}}$			$\text{top-}\beta$		
	rforest	softmax reg	deep learn	rforest	softmax reg	deep learn
2	0.09 (0.01)	0.06 (0.01)	0.09 (0.01)	0.13 (0.01)	0.10 (0.01)	0.13 (0.02)
5	0.01 (0.00)	0.00 (0.00)	0.01 (0.00)	0.02 (0.00)	0.01 (0.00)	0.02 (0.00)
$K = 100$						
$\beta$	$\hat{\Gamma}_{\text{SSE}}$			$\text{top-}\beta$		
	rforest	softmax reg	deep learn	rforest	softmax reg	deep learn
2	0.48 (0.02)	0.93 (0.01)	0.46 (0.02)	0.51 (0.01)	0.96 (0.01)	0.48 (0.02)
5	0.30 (0.02)	0.85 (0.02)	0.25 (0.02)	0.31 (0.01)	0.90 (0.01)	0.27 (0.01)
10	0.17 (0.01)	0.75 (0.02)	0.12 (0.01)	0.18 (0.01)	0.80 (0.01)	0.14 (0.01)
20	0.07 (0.01)	0.59 (0.02)	0.04 (0.01)	0.09 (0.01)	0.61 (0.02)	0.06 (0.01)

TABLE 4

For each of the  $B = 100$  repetitions and for each model, we derive the estimated errors  $\mathbf{P}$  of three different  $\hat{\Gamma}_{\text{SSE}}$ 's w.r.t.  $\beta$ . We compute the means and standard deviations (in parentheses) over the  $B = 100$  repetitions. For each  $\beta$  and for each  $N$ , the  $\hat{\Gamma}_{\text{SSE}}$ 's, as well as the top procedures are based on, from left to right, **rforest**, **softmax reg** and **deep learn**, which are respectively the random forest, the softmax regression and the deep learning methods. Top: the data are generated according to  $K = 10$  – Bottom: the data are generated according to  $K = 100$ .

random variables  $\hat{p}_k(X)$  appear to be not continuous. Hence Assumption 4.4 is violated. To alleviate this issue, we add to  $\hat{p}_k(X)$  an independent small perturbation  $|\mathcal{N}(0, e^{-10})|$  for simplicity. The evaluation of the performance of  $\hat{\Gamma}_{\text{SSE}}$  relies on the following steps

- i) simulate three datasets  $\mathcal{D}_n$ ,  $\mathcal{D}_N$  and  $\mathcal{D}_M$ ;
- ii) based on  $\mathcal{D}_n$ , we compute the estimators  $\hat{p}_k$  of  $p_k$  according to the considered procedure;
- iii) based on  $\mathcal{D}_N$  and  $\hat{p}_k$  we compute the function  $\hat{G}$  and the estimator  $\hat{\Gamma}_{\text{SSE}}$  as in Eq. (1.5) (we recall that this step requires a dataset which contains only unlabeled features);
- iv) finally, we compute over  $\mathcal{D}_M$  the empirical counterpart of  $\mathbf{P}$  and of  $\mathbf{I}$  for the considered  $\hat{\Gamma}_{\text{SSE}}$ .

Again, during these experiments, we compute means and standard deviations. The parameters  $K, n, N$  are fixed as follows: for  $K = 10$ , we fix  $n = 1000$  and  $N \in \{100, 10000\}$ ; for  $K = 100$  we fix  $n = 10000$  and  $N \in \{100, 10000\}$ . Finally, the size of  $\mathcal{D}_M$  is fixed to  $M = 1000$ . The results are illustrated in Tables 4 and 5.

$K = 10$							
$N = 100$				$N = 10000$			
$\beta$	rforest	softmax reg	deep learn	rforest	softmax reg	deep learn	
2	2.01 (0.09)	2.01 (0.10)	2.02 (0.11)	2.00 (0.02)	2.00 (0.03)	2.00 (0.03)	
5	5.02 (0.18)	4.99 (0.20)	5.00 (0.21)	5.00 (0.06)	5.00 (0.08)	5.00 (0.07)	

$K = 100$							
$N = 100$				$N = 10000$			
$\beta$	rforest	softmax reg	deep learn	rforest	softmax reg	deep learn	
2	2.02 (0.10)	2.09 (0.43)	2.01 (0.09)	2.00 (0.03)	2.02 (0.15)	2.00 (0.02)	
5	4.97 (0.15)	5.27 (0.70)	5.01 (0.24)	5.00 (0.04)	5.01 (0.27)	5.00 (0.07)	
10	9.98 (0.24)	10.02 (1.00)	10.02 (0.42)	10.01 (0.09)	10.05 (0.32)	10.00 (0.16)	
20	20.08 (0.48)	19.74 (0.98)	20.11 (0.85)	20.00 (0.16)	20.01 (0.36)	20.01 (0.28)	

TABLE 5

For each of the  $B = 100$  repetitions and for each model, we derive the estimated information levels  $\hat{\Gamma}$  of three different  $\hat{\Gamma}_{\text{SSE}}$ 's w.r.t.  $\beta$  and the sample size  $N$ . We compute the means and standard deviations (in parentheses) over the  $B = 100$  repetitions. For each  $\beta$  and each  $N$ , the  $\hat{\Gamma}_{\text{SSE}}$ 's are based on, from left to right, *rforest*, *softmax reg* and *deep learn*, which are respectively the random forest, the softmax regression and the deep learning procedures. Top: the data are generated according to  $K = 10$  – Bottom: the data are generated according to  $K = 100$ .

As benchmark for the continuation of our experiments, the classical *missclassification* errors of the multi-class classifiers based on random forests, softmax regression and deep learning methods are valued respectively at 0.28, 0.24, 0.29 for  $K = 10$ , and at 0.65, 0.98 0.63 for  $K = 100$ .

Turning to Table 2 we confirm the intuition that the error of the  $\beta$ -Oracle decreases as the value of the parameter  $\beta$  increases. Nevertheless, for moderate values of  $\beta$ , compared to  $K$ , we obtain a satisfactory improvement compared to *standard* multi-class classification Bayes rule. For instance, when  $K = 10$  and  $\beta = 2$  the error of the 2-Oracle confidence set is 0.05, whereas the Bayes classifier has 0.22; likewise, when  $K = 100$  and  $\beta = 5$  the the classification error decreases from 0.60 to 0.20. Table 2 shows that the top- $\beta$  Oracle is slightly outperformed by the  $\beta$ -Oracle in terms of the error, but still performs well.

From Tables 3 and 5, we observe that the approximation of the information is reasonably good and it gets better with  $N$  the number of unlabeled data. Besides, Tables 2 and 4 demonstrate that our algorithm is sensitive to the choice of the underlying estimator  $\hat{p}_k$ . Indeed, when  $\hat{p}_k$  is estimated via the softmax regression, our algorithm fails to give a good approximation to the error of the



$\beta$ -Oracle.

Table 4 provides similar conclusions regarding  $\hat{\Gamma}_{\text{SSE}}$ , though, unlike the theoretical quantities, there are more scenarios where our method is better than its top- $\beta$  counterpart. Let us point out, that for  $K = 100$  methods that are based on the softmax regression perform poorly in this setup.

## 5. Discussions

### 5.1. Around continuity Assumption 1.1

The bedrock of this paper is Assumption 1.1. Based on it, we ensure that the  $\beta$ -Oracle confidence set given by Eq. (1.1) is indeed of information  $\beta$ . On top of that, the explicit formulation of excess risk in Proposition 3.1 relies on the continuity of function  $G(\cdot)$ . Should Assumption 1.1 fail to be satisfied, then there might be no  $\beta$ -Oracle given by thresholding on some level  $\theta \in (0, 1)$ . Indeed, assume Assumption 1.1 is not satisfied but one can build a  $\beta$ -Oracle having the form  $\Gamma_\beta^*(\cdot) = \{k \in [K] : p_k(\cdot) > \theta\}$  with some  $\theta$ , then

$$\beta = I(\Gamma_\beta^*) = G(\theta) .$$

However, without the continuity, the function  $G(\cdot)$  is not surjective and therefore, the equation  $G(\theta) = \beta$  may have no solution, which contradicts the fact that  $I(\Gamma_\beta^*) = \beta$ . Therefore, the settings without the continuity of  $G(\cdot)$  deserve a separate study. Let us also point out that the continuity assumption implies that the  $\beta$ -Oracle can also be defined as

$$\Gamma_\beta^* \in \arg \min \{P(\Gamma) : \Gamma \in \Upsilon \text{ s.t. } I(\Gamma) \leq \beta\} ,$$

where the inequality is used in place of the equality. Indeed, under continuity assumption thanks to Propositions 1.3 and 3.1 we have for all confidence sets  $\Gamma$  such that  $I(\Gamma) \leq \beta$

$$P(\Gamma) - P(\Gamma_\beta^*) = \underbrace{R_\beta(\Gamma) - R_\beta(\Gamma_\beta^*)}_{\geq 0} + \underbrace{G^{-1}(\beta)(\beta - I(\Gamma))}_{\geq 0} ,$$

which implies that the  $\beta$ -Oracle  $\Gamma_\beta^*$  is a minimizer.

### 5.2. Around Lipschitz continuity of $G^{-1}(\cdot)$

Under the assumptions needed in this work, and in particular the continuity assumption we showed two important facts: i) no supervised approach can achieve fast rates, that is, faster than  $n^{-1/2}$ ; ii) some semi-supervised approaches can achieve fast rate.

One might wonder whether extra assumptions on the problem allow a supervised method to get faster rates than  $n^{-1/2}$ . We give to this question a partial

answer following the recent work of [Bobkov and Ledoux \(2016\)](#) and more precisely their Theorem 5.11. Applying this result to our framework, we can state that there exists a positive constant  $c$  such that

$$\mathbb{E}_{\mathcal{D}_N} |G^{-1}(\beta) - G_N^{-1}(\beta)| \leq c \text{Lip}(G^{-1})N^{-1/2} ,$$

where  $\text{Lip}(G^{-1})$  is the Lipschitz constant of  $G^{-1}(\cdot)$  and  $G_N^{-1}(\cdot)$  is the generalized inverse of

$$G_N(\cdot) = \frac{1}{N} \sum_{X \in \mathcal{D}_N} \sum_{k=1}^K \mathbf{1}_{\{p_k(X) > \cdot\}} .$$

If, on top of the above, one can show that for any  $\alpha > 0$  and for some positive constant  $c'$

$$\mathbb{E}_{\mathcal{D}_N} |G^{-1}(\beta) - G_N^{-1}(\beta)|^{1+\alpha} \leq c' \text{Lip}^{1+\alpha}(G^{-1})N^{-(1+\alpha)/2} ,$$

then under Lipschitz continuity of  $G^{-1}(\cdot)$ , we can prove that

$$\Psi_{n,N}^{\text{E}}(\hat{\Gamma}_{\square}; \mathcal{P}(L, \gamma, \alpha)) \lesssim \left( \frac{n}{\log n} \right)^{-\frac{(1+\alpha)\gamma}{2\gamma+d}} ,$$

where  $\square$  stands for SE or SSE. This would illustrate that both  $\hat{\Gamma}_{\text{SE}}$  and  $\hat{\Gamma}_{\text{SSE}}$  are statistically equivalent under Lipschitz condition on  $G^{-1}(\cdot)$ , that is, both reach the same rate and the impact of the unlabeled data  $\mathcal{D}_N$  is negligible. We plan to further investigate the influence of this Lipschitz condition on the minimax rates of convergence in our future works. Since in the present contribution we do not impose this assumption on  $G^{-1}(\cdot)$ , the upper bound of [Bobkov and Ledoux \(2016\)](#) is not applicable and we had to rely on a different approach.

### 5.3. Around extra logarithm

Theorems 3.5 and 4.2 demonstrate that for the excess risk and the Discrepancy, the upper and the lower bounds differ by a logarithmic factor. As we have already pointed out, this factor appears in the upper bounds due to Lemma 4.5 which relates the difference between two thresholds to the infinity norm. One might hope that if we manage to replace the infinity norm by any other  $\ell_q$ -norm on the right hand side of the inequality in Lemma 4.5 this logarithm can be eliminated. Unfortunately, it appears that this bound is actually tight, in a sense that one can construct a distribution  $\mathbb{P}$  and an estimator  $\hat{p}_k$  for all  $k \in [K]$  such that an equality is achieved in Lemma 4.5. These arguments suggest that the obtained upper bound should be optimal. They also imply that the lower bounds could be further refined to get an extra logarithmic factor. Let us also mention that the continuity Assumption 1.1 in combination with the margin Assumption 2.1 are main obstacles that did not allow us to provide better lower bounds. Nevertheless, our proofs are already involved and our results allow to make non trivial conclusions even without going into the details concerning the logarithms.

## 6. Conclusion

In this work we have studied the minimax settings of confidence set multi-class classification. First of all, following previous works we have shown that a top- $\beta$  type procedure is inconsistent in our settings and more involved estimators should be proposed. Besides, we have demonstrated that no supervised estimator can achieve rates that are faster than  $n^{-1/2}$ , which stays in contrast with other classical settings. Additionally, we have shown that fast rates are achievable for semi-supervised techniques provided that the size of the unlabeled sample is large enough. Consecutively, we have established that our lower bounds are either optimal or nearly optimal by providing a supervised and a semi-supervised estimators which are tractable in practice. Our future works shall be focused on the Lipschitz condition of  $G^{-1}(\cdot)$  discussed in Section 5.2, in particular, we want to understand how this extra assumption affects our lower bounds.

## References

- ANBAR, D. (1977). A Modified Robbins-Monro Procedure Approximating the Zero of a Regression Function from Below. *Ann. Statist.* **5** 229–234.
- AUDIBERT, J. and TSYBAKOV, A. (2007). Fast learning rates for plug-in classifiers. *Ann. Statist.* **35** 608–633.
- BARTLETT, P. and WEGKAMP, M. (2008). Classification with a reject option using a hinge loss. *J. Mach. Learn. Res.* **9** 1823–1840.
- BELLEÇ, P., DALALYAN, A., GRAPPIN, E. and PARIS, Q. (2018). On the prediction loss of the lasso in the partially labeled setting. *Electron. J. Statist.* **12** 3443–3472.
- BIRGÉ, L. (2005). A new lower bound for multiple hypothesis testing. *IEEE Trans. Inform. Theory* **51**.
- BOBKOV, S. and LEDOUX, M. (2016). One-dimensional empirical measures, order statistics and Kantorovich transport distances. to appear in the Memoirs of the Amer. Math. Soc.
- BROWN, L. and LOW, M. (1996). A constrained risk inequality with applications to nonparametric functional estimation. *Ann. Statist.* **24** 2524–2535.
- CHOW, C. (1957). An optimum character recognition system using decision functions. *IRE Transactions on Electronic Computers* **4** 247–254.
- CHOW, C. (1970). On optimum error and reject trade-off. *IEEE Trans. Inform. Theory* **16** 41–46.
- DE FINETTI, B. (1972). *Probability, induction and statistics. The art of guessing.* John Wiley & Sons, London-New York-Sydney Wiley Series in Probability and Mathematical Statistics.
- DE FINETTI, B. (1974). *Theory of probability: a critical introductory treatment. Vol. 1.* John Wiley & Sons, London-New York-Sydney.
- DENIS, C. and HEBIRI, M. (2015). Consistency of plug-in confidence sets for classification in semi-supervised learning. preprint.
- DENIS, C. and HEBIRI, M. (2017). Confidence Sets with Expected Sizes for Multiclass Classification. *J. Mach. Learn. Res.* **18(102)** 1–28.

- GILBERT, E. (1952). A comparison of signalling alphabets. *The Bell system technical journal* **31** 504–522.
- HARTIGAN, J. A. (1987). Estimation of a Convex Density Contour in Two Dimensions. *J. Amer. Statist. Assoc.* **82** 267–270.
- HERBEI, R. and WEGKAMP, M. (2006). Classification with reject option. *Canad. J. Statist.* **34** 709–721.
- HOEFFDING, W. (1963). Probability Inequalities for Sums of Bounded Random Variables. *J. Amer. Statist. Assoc.* **58** 13–30.
- LEI, J. (2014). Classification with confidence. *Biometrika* **101** 755–769.
- LEPSKII, O. (1990). Asymptotic minimax estimation with prescribed properties. *Theory Probab. Appl.* **34** 604–615.
- POLONIK, W. (1995). Measuring Mass Concentrations and Estimating Density Contour Clusters-An Excess Mass Approach. *Ann. Statist.* **23** 855–881.
- RAMASWAMY, H., TEWARI, A. and AGARWAL, S. (2018). Consistent algorithms for multiclass classification with an abstain option. *Electron. J. Stat.* **12** 530–554.
- RIGOLLET, P. (2007). Generalization error bounds in semi-supervised classification under the cluster assumption. *J. Mach. Learn. Res.* **8** 1369–1392.
- RIGOLLET, P. and VERT, R. (2009). Optimal rates for plug-in estimators of density level sets. *Bernoulli* **4** 1154–1178.
- SADINLE, M., LEI, J. and WASSERMAN, L. (2018). Least Ambiguous Set-Valued Classifiers With Bounded Error Levels. *J. Amer. Statist. Assoc.* 1–12.
- SINGH, A., NOWAK, R. and ZHU, J. (2009). Unlabeled data: Now it helps, now it doesn't. In *NIPS* 1513–1520.
- STONE, C. (1977). Consistent nonparametric regression. *Ann. Statist.* 595–620.
- STONE, C. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* 1040–1053.
- TSYBAKOV, A. (1986). Robust reconstruction of functions by the local-approximation method. *Problemy Peredachi Informatsii* **22** 69–84.
- TSYBAKOV, A. B. (1997). On nonparametric estimation of density level sets. *Ann. Statist.* **25** 948–969.
- TSYBAKOV, A. (2004). Optimal aggregation of classifiers in statistical learning. *Ann. Statist.* **32** 135–166.
- TSYBAKOV, A. (2008). *Introduction to Nonparametric Estimation. Springer Ser. Statist.* Springer New York.
- VAN DER VAART, A. (1998). *Asymptotic statistics. Camb. Ser. Stat. Probab. Math.* **3**. Cambridge University Press, Cambridge.
- VAPNIK, V. (1998). *Statistical learning theory. Adaptive and Learning Systems for Signal Processing, Communications, and Control.* John Wiley & Sons Inc., New York. A Wiley-Interscience Publication.
- VARSHAMOV, R. (1957). Estimate of the number of signals in error correcting codes. *Dokl. Akad. Nauk SSSR* **117** 739–741.
- VOVK, V. (2002a). Asymptotic optimality of transductive confidence machine. In *Algorithmic learning theory. Lecture Notes in Comput. Sci.* **2533** 336–350. Springer, Berlin.
- VOVK, V. (2002b). On-line confidence machines are well-calibrated. In *Pro-*

- ceedings of the Forty-Third Annual Symposium on Foundations of Computer Science* 187–196. CA. IEEE Computer Society, Los Alamitos.
- VOVK, V., GAMMERMAN, A. and SHAFER, G. (2005). *Algorithmic learning in a random world*. Springer, New York.
- WEGKAMP, M. and YUAN, M. (2011). Support vector machines with a reject option. *Bernoulli* **17** 1368–1385.

## Appendix A: Technical results

Here we provide proofs for our result. This Appendix is composed of the following part: in Appendix A we introduce some technical results used for our proofs; Appendix B is devoted to the proofs of the upper bounds; Appendix C provides with the proofs of our main lower bounds; finally, in Appendix D we prove the inconsistency of top- $\beta$  approaches.

In this section we gather several technical results which are used to derive the contributions of this work. Let us start by introducing notation used in the appendix. Given any two probability measures  $\mathbb{P}_1, \mathbb{P}_2$  on some space measurable space  $(\mathcal{X}, \mathcal{A})$  the KullbackLeibler divergence between  $\mathbb{P}_1$  and  $\mathbb{P}_2$  is defined as

$$\text{KL}(\mathbb{P}_1, \mathbb{P}_2) := \begin{cases} \int_{\mathcal{X}} \log \left( \frac{d\mathbb{P}_1}{d\mathbb{P}_2} \right) d\mathbb{P}_1, & \text{supp}(\mathbb{P}_1) \subset \text{supp}(\mathbb{P}_2) \\ +\infty, & \text{otherwise} \end{cases}, \quad (\text{A.1})$$

and the total variation distance is defined as

$$\text{TV}(\mathbb{P}_1, \mathbb{P}_2) := \sup_{A \in \mathcal{A}} |\mathbb{P}_1(A) - \mathbb{P}_2(A)|. \quad (\text{A.2})$$

We start with Fano's inequality in the form proved by (Birgé, 2005).

**Lemma A.1** (Fano's inequality (Birgé, 2005)). *Let  $\{\mathbb{P}_i\}_{i=0}^m$  be a finite family of probability measures on  $(\mathcal{X}, \mathcal{A})$  and let  $\{A_i\}_{i=0}^m$  be a finite family of disjoint events such that  $A_i \in \mathcal{A}$  for each  $i = 0, \dots, m$ . Then,*

$$\min_{i \in \{0, 1, \dots, m\}} \mathbb{P}_i(A_i) \leq \left( 0.71 \bigvee \frac{\frac{1}{m} \sum_{i=1}^m \text{KL}(\mathbb{P}_i, \mathbb{P}_0)}{\log(m+1)} \right).$$

**Lemma A.2** (Pinsker's inequality). *Given any two probability measures  $\mathbb{P}_1, \mathbb{P}_2$  on some measurable space  $(\mathcal{X}, \mathcal{A})$  we have*

$$\text{TV}(\mathbb{P}_1, \mathbb{P}_2) \leq \sqrt{\frac{1}{2} \text{KL}(\mathbb{P}_1, \mathbb{P}_2)}.$$

**Lemma A.3** (Hoeffding's inequality (Hoeffding, 1963)). *Let  $b > 0$  be a real number, and  $N$  be a positive integer. Let  $X_1, \dots, X_N$  be  $N$  random variables having values in  $[0, b]$ , then*

$$\mathbb{P} \left( \left| \frac{1}{N} \sum_{i=1}^N (X_i - \mathbb{E}[X_i]) \right| \geq t \right) \leq 2 \exp \left( -\frac{2Nt^2}{b^2} \right), \quad \forall t > 0.$$

**Proposition A.4** (Properties of the generalized inverse). *Let  $X \in \mathbb{R}^d$  and  $\mathbb{P}_X$  be a Borel measure on  $\mathbb{R}^d$ , let  $p : \mathbb{R}^d \rightarrow [0, 1]^K$  be a vector function, we define for all  $t \in [0, 1]$  and all  $\beta \in (0, K)$*

$$G(t) := \sum_{k=1}^K \mathbb{P}_X(p_k(X) > t), \quad G^{-1}(\beta) := \inf \{t \in [0, 1] : G(t) \leq \beta\} .$$

Then,

- for all  $t \in (0, 1)$  and  $\beta \in (0, K)$  we have  $G^{-1}(\beta) \leq t \iff G(t) \leq \beta$ .
- if for all  $k \in [K]$  the mappings  $t \mapsto \mathbb{P}_X(p_k(X) > t)$  are continuous on  $(0, 1)$ , then
  - for all  $\beta \in (0, K)$  we have  $G(G^{-1}(\beta)) = \beta$  .

The next result is an analogue of the classical inverse transform theorem ([van der Vaart, 1998](#), Lemma 21.1) and was already established by [Denis and Hebiri \(2017\)](#).

**Lemma A.5.** *Let  $\varepsilon$  distributed from a uniform distribution on  $[K]$  and  $Z_1, \dots, Z_K$ ,  $K$  real valued random variables independent from  $\varepsilon$ , such that the function  $t \mapsto H(t)$  defined as*

$$H(t) := \frac{1}{K} \sum_{k=1}^K \mathbb{P}(Z_k \leq t) ,$$

*is continuous. Consider random variable  $Z = \sum_{k=1}^K Z_k \mathbf{1}_{\{\varepsilon=k\}}$  and let  $U$  be distributed according to the uniform distribution on  $[0, 1]$ . Then*

$$H(Z) \stackrel{\mathcal{L}}{=} U \text{ and } H^{-1}(U) \stackrel{\mathcal{L}}{=} Z ,$$

where  $H^{-1}$  denotes the generalized inverse of  $H$ .

*Proof.* First we note that for every  $t \in [0, 1]$ ,  $\mathbb{P}(H(Z) \leq t) = \mathbb{P}(Z \leq H^{-1}(t))$ . Moreover, we have

$$\begin{aligned} \mathbb{P}(H(Z) \leq t) &= \sum_{k=1}^K \mathbb{P}(Z \leq H^{-1}(t), \varepsilon = k) \\ &= \frac{1}{K} \sum_{k=1}^K \mathbb{P}(Z_k \leq H^{-1}(t)) \quad (\text{with } \varepsilon \text{ independent of } Z) \\ &= H(H^{-1}(t)) \\ &= t \quad (\text{with } H \text{ continuous}) . \end{aligned}$$

To conclude the proof, we observe that

$$\begin{aligned}\mathbb{P}(H^{-1}(U) \leq t) &= \mathbb{P}(U \geq H(t)) = \frac{1}{K} \sum_{k=1}^K \mathbb{P}(Z_k \leq t) \\ &= \sum_{k=1}^K \mathbb{P}(Z_k \leq t, \varepsilon = k) = \mathbb{P}(Z \leq t) .\end{aligned}$$

□

## Appendix B: Upper bounds

In this section we prove Theorems 4.1 and 4.2. It will be clear from our analysis that the proof of Theorem 4.1 follows directly from Theorem 4.2 by setting  $N = 0$  in the statement of Theorem 4.2. Thus, in this section for simplicity we omit the subscript SSE from  $\hat{\Gamma}_{\text{SSE}}$ . Recall that our dataset consists of three parts  $\mathcal{D}_{\lfloor n/2 \rfloor}, \mathcal{D}_{\lceil n/2 \rceil}, \mathcal{D}_N$ . The set  $\mathcal{D}_{\lfloor n/2 \rfloor}$  is used to construct an estimator  $\hat{p}$  of the regression function  $p$ , that is,  $\hat{p}$  is independent from both  $\mathcal{D}_{\lceil n/2 \rceil}, \mathcal{D}_N$ . The other two sets  $\mathcal{D}_{\lceil n/2 \rceil}, \mathcal{D}_N$  are used in a semi-supervised manner to estimate the threshold, that is, we erase the labels from  $\mathcal{D}_{\lceil n/2 \rceil}$ . Let  $\beta \in [K - 1]$ , and also recall the definition of the proposed semi-supervised estimator for a given  $x \in \mathbb{R}^d$

$$\hat{\Gamma}(x) = \left\{ k \in [K] : \hat{p}_k(x) \geq \hat{G}^{-1}(\beta) \right\} ,$$

with  $\hat{p}_k(x)$  satisfying Assumptions 4.4, 4.3 for all  $k \in [K]$ . Moreover,  $\hat{G}^{-1}(\beta)$  defined as the generalized inverse of

$$\hat{G}(t) = \frac{1}{\lfloor n/2 \rfloor + N} \sum_{X \in \mathcal{D}_N \cup \mathcal{D}_{\lceil n/2 \rceil}} \sum_{k=1}^K \mathbf{1}_{\hat{p}_k(X) > t} ,$$

where  $t \in [0, 1]$ . Additionally, recall that the  $\beta$ -Oracle is given as

$$\Gamma_{\beta}^*(x) = \left\{ k \in [K] : p_k(x) \geq G^{-1}(\beta) \right\} , \quad (\text{B.1})$$

where  $G^{-1}(\cdot)$  is the generalized inverse of

$$G(t) := \sum_{k=1}^K \mathbb{P}(p_k(X) \geq t) .$$

Lastly, let us re-introduce an idealized version  $\tilde{\Gamma}$  of the proposed estimator  $\hat{\Gamma}$  which 'knows' the marginal distribution  $\mathbb{P}_X$  of the feature vector  $X \in \mathbb{R}^d$  as

$$\tilde{\Gamma}(x) = \left\{ k \in [K] : \hat{p}_k(x) \geq \tilde{G}^{-1}(\beta) \right\} ,$$

with  $\tilde{G} := \sum_{k=1}^K \mathbb{P}_X(\hat{p}_k(X) > t)$ , conditionally on the data. The following result is needed to relate the threshold  $\tilde{G}^{-1}(\beta)$  of  $\tilde{\Gamma}$  to the true value of the threshold  $G^{-1}(\beta)$ .

**Lemma B.1** (Upper-bound on the thresholds). *Let  $X \in \mathbb{R}^d$  and  $\mathbb{P}_X$  be a Borel measure on  $\mathbb{R}^d$ . For two vector functions  $p, \hat{p} : \mathbb{R}^d \rightarrow [0, 1]^K$ , we define*

$$G(\cdot) := \sum_{k=1}^K \mathbb{P}_X(p_k(X) > \cdot), \quad \tilde{G}(\cdot) := \sum_{k=1}^K \mathbb{P}_X(\hat{p}_k(X) > \cdot) .$$

*If for all  $k \in [K]$  the mapping  $t \mapsto \mathbb{P}_X(p_k(X) > t)$  is continuous on  $(0, 1)$ , then for every  $\beta \in (0, K)$*

$$\left| G^{-1}(\beta) - \tilde{G}^{-1}(\beta) \right| \leq \|\hat{p} - p\|_{\infty, \mathbb{P}_X} .$$

*Proof.* The proof of this result is very similar to the proof of (Bobkov and Ledoux, 2016, Theorem 2.12). We start by defining the following quantity

$$h^* = \inf \left\{ h \geq 0 : \forall t \in [0, 1] \tilde{G}(t + h) \leq G(t) \leq \tilde{G}(t - h) \right\} .$$

Due to the definition of  $h^*$  we have that for all  $t \in [0, 1]$

$$\tilde{G}(t + h^*) \leq G(t) \leq \tilde{G}(t - h^*) ,$$

that is, applying Proposition A.4 to the second inequality we get for all  $t \in [0, 1]$

$$t - h^* \leq \tilde{G}^{-1}(G(t)) ,$$

thus, for  $t = G^{-1}(\beta)$  with  $\beta \in (0, K)$  thanks to Proposition A.4 we get

$$G^{-1}(\beta) - \tilde{G}^{-1}(\beta) \leq h^* .$$

The inequality  $\tilde{G}^{-1}(\beta) - G^{-1}(\beta) \leq h^*$  is obtained in the same way. Thus, we have proved that

$$\left| G^{-1}(\beta) - \tilde{G}^{-1}(\beta) \right| \leq h^* .$$

Finally, notice that for all  $t \in [0, 1]$

$$\begin{aligned} \underbrace{\sum_{k=1}^K \mathbb{P}_X(\hat{p}_k(X) > t + \|\hat{p} - p\|_{\infty, \mathbb{P}_X})}_{\tilde{G}(t + \|\hat{p} - p\|_{\infty, \mathbb{P}_X})} &\leq \underbrace{\sum_{k=1}^K \mathbb{P}_X(p_k(X) > t)}_{G(t)} \\ &\leq \underbrace{\sum_{k=1}^K \mathbb{P}_X(\hat{p}_k(X) > t - \|\hat{p} - p\|_{\infty, \mathbb{P}_X})}_{\tilde{G}(t - \|\hat{p} - p\|_{\infty, \mathbb{P}_X})} , \end{aligned}$$

where we used the fact that for all  $k \in [K]$

$$\begin{aligned} \mathbb{P}_X(\hat{p}_k(X) > t + |\hat{p}_k(X) - p_k(X)|) &\leq \mathbb{P}_X(p_k(X) > t) \\ &\leq \mathbb{P}_X(\hat{p}_k(X) > t - |\hat{p}_k(X) - p_k(X)|) , \end{aligned}$$

and  $\mathbb{P}_X(|\hat{p}_k(X) - p_k(X)| \leq \|\hat{p} - p\|_{\infty, \mathbb{P}_X}) = 1$ . Therefore by definition of  $h^*$ , we can write  $h^* \leq \|\hat{p} - p\|_{\infty, \mathbb{P}_X}$  and we conclude.  $\square$



We are in position to prove Theorem 4.2, let us point out that the most difficult part in Theorem 4.2 is the upper-bound on the excess risk. The upper-bound on the discrepancy follows the same arguments as the ones we use for the excess-risk.

**Excess risk and discrepancy:** to upper-bound the excess risk we first separate it into two parts as

$$R_\beta(\hat{\Gamma}) - R_\beta(\Gamma_\beta^*) = \underbrace{\left( R_\beta(\tilde{\Gamma}) - R_\beta(\Gamma_\beta^*) \right)}_{R_1} + \underbrace{\left( R_\beta(\hat{\Gamma}) - R_\beta(\tilde{\Gamma}) \right)}_{R_2} .$$

Recall that thanks to Proposition 3.1 we have

$$R_1 = \sum_{k=1}^K \mathbb{E} \left[ |p_k(X) - G^{-1}(\beta)| \mathbf{1}_{\{k \in \tilde{\Gamma}(X) \triangle \Gamma_\beta^*(X)\}} \right] .$$

Moreover, let us point out that if some  $k \in \tilde{\Gamma}(X) \triangle \Gamma_\beta^*(X)$  then either

$$\begin{cases} p_k(X) - G^{-1}(\beta) \geq 0 \\ \hat{p}_k(X) - \tilde{G}^{-1}(\beta) < 0 \end{cases} \quad \text{or} \quad \begin{cases} p_k(X) - G^{-1}(\beta) < 0 \\ \hat{p}_k(X) - \tilde{G}^{-1}(\beta) \geq 0 \end{cases} ,$$

holds. Thus on the event  $k \in \tilde{\Gamma}(X) \triangle \Gamma_\beta^*(X)$  we have

$$\begin{aligned} |p_k(X) - G^{-1}(\beta)| &\leq |\hat{p}_k(X) - p_k(X) + G^{-1}(\beta) - \tilde{G}^{-1}(\beta)| \\ &\leq |\hat{p}_k(X) - p_k(X)| + |G^{-1}(\beta) - \tilde{G}^{-1}(\beta)| . \end{aligned}$$

Therefore, for  $R_1$  using Lemma B.1 and the observations above we can write

$$\begin{aligned} R_1 &\leq \sum_{k=1}^K \mathbb{E} \left[ |p_k(X) - G^{-1}(\beta)| \mathbf{1}_{\{|p_k(X) - G^{-1}(\beta)| \leq |\hat{p}_k(X) - p_k(X)| + |G^{-1}(\beta) - \tilde{G}^{-1}(\beta)|\}} \right] \\ &\leq \sum_{k=1}^K \mathbb{E} \left[ |p_k(X) - G^{-1}(\beta)| \mathbf{1}_{\{|p_k(X) - G^{-1}(\beta)| \leq 2\|\hat{p} - p\|_{\infty, \mathbb{P}_X}\}} \right] \\ &\leq \sum_{k=1}^K \mathbb{E} \left[ 2\|\hat{p} - p\|_{\infty, \mathbb{P}_X} \mathbf{1}_{\{|p_k(X) - G^{-1}(\beta)| \leq 2\|\hat{p} - p\|_{\infty, \mathbb{P}_X}\}} \right] \\ &= 2\|\hat{p} - p\|_{\infty, \mathbb{P}_X} \sum_{k=1}^K \mathbb{P}_X \left( |p_k(X) - G^{-1}(\beta)| \leq 2\|\hat{p} - p\|_{\infty, \mathbb{P}_X} \right) , \end{aligned}$$

finally, using the margin Assumption 2.1 we get almost surely data

$$R_1 \leq c_1 2^{1+\alpha} K \|\hat{p} - p\|_{\infty, \mathbb{P}_X}^{1+\alpha} .$$

Integrating over the data from both sides and using Assumption 4.3 we get

$$\mathbb{E}_{(\mathcal{D}_n, \mathcal{D}_N)} R_1 \leq c_1 C_2 2^{1+\alpha} K \left( \frac{n}{\log n} \right)^{-\frac{(1+\alpha)\gamma}{2\gamma+d}} .$$

For  $R_2$  the following trivial upper-bound holds

$$\begin{aligned}
R_2 &= \left( P(\hat{\Gamma}) - P(\tilde{\Gamma}) \right) + G^{-1}(\beta) \left( I(\hat{\Gamma}) - I(\tilde{\Gamma}) \right) \\
&= \sum_{k=1}^K \mathbb{E} \left( p_k(X) - G^{-1}(\beta) \right) \left( \mathbf{1}_{\{k \in \tilde{\Gamma}(X)\}} - \mathbf{1}_{\{k \in \hat{\Gamma}(X)\}} \right) \\
&\leq \sum_{k=1}^K \mathbb{E} \underbrace{\left| \mathbf{1}_{\{k \in \tilde{\Gamma}(X)\}} - \mathbf{1}_{\{k \in \hat{\Gamma}(X)\}} \right|}_{\mathbb{E} |\tilde{\Gamma}(X) \Delta \hat{\Gamma}(X)|} \\
&= \sum_{k=1}^K \mathbb{E} \left| \mathbf{1}_{\{\hat{p}_k(X) \geq \hat{G}^{-1}(\beta)\}} - \mathbf{1}_{\{\tilde{p}_k(X) \geq \tilde{G}^{-1}(\beta)\}} \right|,
\end{aligned} \tag{B.2}$$

now, thanks to the first property of Proposition A.4 we can write

$$\begin{aligned}
R_2 &\leq \sum_{k=1}^K \mathbb{E} \left| \mathbf{1}_{\{\hat{G}(\hat{p}_k(X)) \leq \beta\}} - \mathbf{1}_{\{\tilde{G}(\tilde{p}_k(X)) \leq \beta\}} \right| \\
&\leq \sum_{k=1}^K \mathbb{P}_X \left( \left| \hat{G}(\hat{p}_k(X)) - \tilde{G}(\tilde{p}_k(X)) \right| \geq \left| \tilde{G}(\tilde{p}_k(X)) - \beta \right| \right)
\end{aligned}$$

To finish our proof we make use of the peeling technique of (Audibert and Tsybakov, 2007, Lemma 3.1). That is, we define for  $\delta > 0$  and  $k \in [K]$

$$\begin{aligned}
A_0^k &= \left\{ \left| \tilde{G}(\tilde{p}_k(X)) - \beta \right| \leq \delta \right\} \\
A_j^k &= \left\{ 2^{j-1}\delta < \left| \tilde{G}(\tilde{p}_k(X)) - \beta \right| \leq 2^j\delta \right\}, \quad j \geq 1.
\end{aligned}$$

Since, for every  $k \in [K]$ , the events  $(A_j^k)_{j \geq 0}$  are mutually exclusive, we deduce

$$\begin{aligned}
\sum_{k=1}^K \mathbb{P}_X \left( \left| \hat{G}(\hat{p}_k(X)) - \tilde{G}(\tilde{p}_k(X)) \right| \geq \left| \tilde{G}(\tilde{p}_k(X)) - \beta \right| \right) &= \\
\sum_{k=1}^K \sum_{j \geq 0} \mathbb{P}_X \left( \left| \hat{G}(\hat{p}_k(X)) - \tilde{G}(\tilde{p}_k(X)) \right| \geq \left| \tilde{G}(\tilde{p}_k(X)) - \beta \right|, A_j^k \right).
\end{aligned} \tag{B.3}$$

Now, we consider  $\varepsilon$  uniformly distributed on  $[K]$  independent of the data and  $X$ . Conditional on the data and under Assumption 4.4, we apply Lemma A.5 with  $Z_k = \hat{p}_k(X)$ ,  $Z = \sum_{k=1}^K Z_k \mathbf{1}_{\{\varepsilon=k\}}$  and then obtain that  $\tilde{G}(Z)$  is uniformly distributed on  $[0, K]$ . Therefore, for all  $j \geq 0$  and  $\delta > 0$ , we deduce

$$\frac{1}{K} \sum_{k=1}^K \mathbb{P}_X \left( \left| \tilde{G}(\tilde{p}_k(X)) - \beta \right| \leq 2^j\delta \right) = \mathbb{P}_X \left( \left| \tilde{G}(Z) - \beta \right| \leq 2^j\delta \right) \leq \frac{2^{j+1}\delta}{K}.$$

Hence, for all  $j \geq 0$ , we obtain

$$\sum_{k=1}^K \mathbb{P}_X(A_j^k) \leq \sum_{k=1}^K \mathbb{P}_X \left( |\tilde{G}(\hat{p}_k(X)) - \beta| \leq 2^j \delta \right) \leq 2^{j+1} \delta . \quad (\text{B.4})$$

Next, we observe that for all  $j \geq 1$

$$\begin{aligned} \sum_{k=1}^K \mathbb{P}_X \left( |\hat{G}(\hat{p}_k(X)) - \tilde{G}(\hat{p}_k(X))| \geq |\tilde{G}(\hat{p}_k(X)) - \beta|, A_j^k \right) &\leq \\ \sum_{k=1}^K \mathbb{P}_X \left( |\hat{G}(\hat{p}_k(X)) - \tilde{G}(\hat{p}_k(X))| \geq 2^{j-1} \delta, A_j^k \right) &. \end{aligned} \quad (\text{B.5})$$

Thus, we obtain that

$$R_2 \leq \sum_{k=1}^K \sum_{j \geq 0} \mathbb{P}_X \left( |\hat{G}(\hat{p}_k(X)) - \tilde{G}(\hat{p}_k(X))| \geq 2^{j-1} \delta, A_j^k \right) ,$$

almost surely data. Integrating from both sides with respect to the data we get

$$\begin{aligned} \mathbb{E}_{(\mathcal{D}_n, \mathcal{D}_N)} R_2 &\leq \sum_{k=1}^K \sum_{j \geq 0} \mathbb{E}_{(\mathcal{D}_n, \mathcal{D}_N)} \mathbb{P}_X \left( |\hat{G}(\hat{p}_k(X)) - \tilde{G}(\hat{p}_k(X))| \geq 2^{j-1} \delta, A_j^k \right) \\ &= \sum_{k=1}^K \sum_{j \geq 0} \mathbb{E}_{(\mathcal{D}_{\lfloor n/2 \rfloor}, \mathcal{D}_{\lceil n/2 \rceil}, \mathcal{D}_N, X \sim \mathbb{P}_X)} \mathbf{1}_{\{|\hat{G}(\hat{p}_k(X)) - \tilde{G}(\hat{p}_k(X))| \geq 2^{j-1} \delta\}} \mathbf{1}_{\{A_j^k\}} . \end{aligned}$$

recall that the function  $\mathbf{1}_{\{A_j^k\}}$  for all  $j \geq 0$  and  $k \in [K]$  is independent from  $\mathcal{D}_{\lceil n/2 \rceil}, \mathcal{D}_N$ , thus we can write

$$\begin{aligned} &\mathbb{E}_{(\mathcal{D}_{\lfloor n/2 \rfloor}, \mathcal{D}_{\lceil n/2 \rceil}, \mathcal{D}_N, X \sim \mathbb{P}_X)} \mathbf{1}_{\{|\hat{G}(\hat{p}_k(X)) - \tilde{G}(\hat{p}_k(X))| \geq 2^{j-1} \delta\}} \mathbf{1}_{\{A_j^k\}} = \\ &\mathbb{E}_{(\mathcal{D}_{\lfloor n/2 \rfloor}, X \sim \mathbb{P}_X)} \mathbb{E}_{(\mathcal{D}_{\lceil n/2 \rceil}, \mathcal{D}_N)} \left[ \mathbf{1}_{\{|\hat{G}(\hat{p}_k(X)) - \tilde{G}(\hat{p}_k(X))| \geq 2^{j-1} \delta\}} \middle| \mathcal{D}_{\lfloor n/2 \rfloor}, X \right] \mathbf{1}_{\{A_j^k\}} \end{aligned}$$

Now, since conditional on  $(\mathcal{D}_{\lfloor n/2 \rfloor}, X)$ ,  $\hat{G}(\hat{p}_k(X))$  is an empirical mean of *i.i.d.* random variables of common mean  $\tilde{G}(\hat{p}_k(X)) \in [0, K]$ , we deduce from Hoeffding's inequality that

$$\mathbb{E}_{(\mathcal{D}_{\lceil n/2 \rceil}, \mathcal{D}_N)} \left[ \mathbf{1}_{\{|\hat{G}(\hat{p}_k(X)) - \tilde{G}(\hat{p}_k(X))| \geq 2^{j-1} \delta\}} \middle| \mathcal{D}_{\lfloor n/2 \rfloor}, X \right] \leq 2e^{-\frac{(N + \lceil n/2 \rceil) \delta^2 2^{2j-1}}{K^2}} .$$

Therefore, treating  $A_0^k$  separately, we get from inequalities of Eqs. (B.3), (B.4), and (B.5)

$$\mathbb{E}_{(\mathcal{D}_n, \mathcal{D}_N)} R_2 \leq 2\delta + \delta \sum_{j \geq 1} 2^{j+2} \exp \left( -\frac{(N + \lceil n/2 \rceil) \delta^2 2^{2j-1}}{K^2} \right) .$$

Finally, choosing  $\delta = \frac{K}{\sqrt{N + \lceil n/2 \rceil}}$  in the above inequality, we finish the proof.

**Hamming risk:** here we provide an upper bound on the Hamming risk. First, by the triangle inequality we can write for the proposed estimator  $\hat{\Gamma}$  and the pseudo Oracle  $\beta$  set  $\tilde{\Gamma}$

$$\begin{aligned} \mathbb{E}_{(\mathcal{D}_n, \mathcal{D}_N)} \mathbb{E}_{X \sim \mathbb{P}_X} \left| \hat{\Gamma}(X) \Delta \Gamma_\beta^*(X) \right| &\leq \mathbb{E}_{(\mathcal{D}_n, \mathcal{D}_N)} \mathbb{E}_{X \sim \mathbb{P}_X} \left| \tilde{\Gamma}(X) \Delta \Gamma_\beta^*(X) \right| \\ &\quad + \mathbb{E}_{(\mathcal{D}_n, \mathcal{D}_N)} \mathbb{E}_{X \sim \mathbb{P}_X} \left| \hat{\Gamma}(X) \Delta \tilde{\Gamma}(X) \right|. \end{aligned}$$

Notice that for the term  $\mathbb{E}_{(\mathcal{D}_n, \mathcal{D}_N)} \mathbb{E}_{X \sim \mathbb{P}_X} \left| \hat{\Gamma}(X) \Delta \tilde{\Gamma}(X) \right|$  we can re-use the proof technique used for the term  $R_2$  in Eq. (B.2). Thus, it remain to upper-bound the term  $\mathbb{E}_{(\mathcal{D}_n, \mathcal{D}_N)} \mathbb{E}_{X \sim \mathbb{P}_X} \left| \tilde{\Gamma}(X) \Delta \Gamma_\beta^*(X) \right|$ . The proof on this part closely follows the machinery used in Denis and Hebiri (2017), however, let us mention that they used this method to obtain a bound on the Discrepancy which leads to a sub-optimal rate. Nevertheless, their approach gives a correct rate if instead of the Discrepancy we bound the Hamming distance. For the sake of completeness we write the principal parts of the proof here.

First of all, by the definition of sets  $\Gamma_\beta^*$  and  $\tilde{\Gamma}$  we can write for  $(*) = \mathbb{E}_{X \sim \mathbb{P}_X} \left| \tilde{\Gamma}(X) \Delta \Gamma_\beta^*(X) \right|$

$$(*) = \sum_{k=1}^K \mathbb{E}_{X \sim \mathbb{P}_X} \left| \mathbf{1}_{\{\hat{p}_k(X) \geq \tilde{G}^{-1}(\beta)\}} - \mathbf{1}_{\{p_k(X) \geq G^{-1}(\beta)\}} \right|,$$

Now if  $\hat{p}_k(X) \geq \tilde{G}^{-1}(\beta)$  and  $p_k(X) < G^{-1}(\beta)$  we can have the following situations

- if  $\tilde{G}^{-1}(\beta) > G^{-1}(\beta)$ , then  $|p_k(X) - G^{-1}(\beta)| \leq |\hat{p}_k(X) - p_k(X)|$ ;
- if  $\tilde{G}^{-1}(\beta) \leq G^{-1}(\beta)$ , then either  $|p_k(X) - G^{-1}(\beta)| \leq |\hat{p}_k(X) - p_k(X)|$  or  $\hat{p}_k(X) \in (\tilde{G}^{-1}(\beta), G^{-1}(\beta))$ ;

Similar conditions are satisfied if  $\hat{p}_k(X) < \tilde{G}^{-1}(\beta)$  and  $p_k(X) \geq G^{-1}(\beta)$ . Using

the above arguments we can upper-bound  $(*)$  as

$$\begin{aligned}
(*) &\leq \sum_{k=1}^K \mathbb{P}_X \left( |p_k(X) - G^{-1}(\beta)| \leq |\hat{p}_k(X) - p_k(X)| \right) \\
&\quad + \mathbf{1}_{\{\tilde{G}^{-1}(\beta) \leq G^{-1}(\beta)\}} \sum_{k=1}^K \mathbb{P}_X \left( \tilde{G}^{-1}(\beta) < \hat{p}_k(X) < G^{-1}(\beta) \right) \\
&\quad + \mathbf{1}_{\{G^{-1}(\beta) < \tilde{G}^{-1}(\beta)\}} \sum_{k=1}^K \mathbb{P}_X \left( G^{-1}(\beta) < \hat{p}_k(X) < \tilde{G}^{-1}(\beta) \right) \\
&= \sum_{k=1}^K \mathbb{P}_X \left( |p_k(X) - G^{-1}(\beta)| \leq |\hat{p}_k(X) - p_k(X)| \right) \\
&\quad + \left| \tilde{G} \left( \tilde{G}^{-1}(\beta) \right) - \tilde{G} \left( G^{-1}(\beta) \right) \right|.
\end{aligned}$$

Thanks to the continuity Assumption 4.4 on the estimator and the continuity Assumption 1.1 on the distribution we clearly have  $\tilde{G} \left( \tilde{G}^{-1}(\beta) \right) = \beta = G \left( G^{-1}(\beta) \right)$ . Moreover, we can write

$$\begin{aligned}
|\tilde{G} \left( \tilde{G}^{-1}(\beta) \right) - \tilde{G} \left( G^{-1}(\beta) \right)| &= \left| G \left( G^{-1}(\beta) \right) - \tilde{G} \left( G^{-1}(\beta) \right) \right| \\
&\leq \sum_{k=1}^K \mathbb{E}_{X \sim \mathbb{P}_X} \left| \mathbf{1}_{\{\hat{p}_k(X) \geq G^{-1}(\beta)\}} - \mathbf{1}_{\{p_k(X) \geq G^{-1}(\beta)\}} \right| \\
&\leq \sum_{k=1}^K \mathbb{P}_X \left( |p_k(X) - G^{-1}(\beta)| \leq |\hat{p}_k(X) - p_k(X)| \right).
\end{aligned}$$

Thus, our bound reads as

$$(*) \leq 2 \sum_{k=1}^K \mathbb{P}_X \left( |p_k(X) - G^{-1}(\beta)| \leq |\hat{p}_k(X) - p_k(X)| \right).$$

Finally, in order to upper-bound the term above one can use the peeling argument of Audibert and Tsybakov (2007) applied with the exponential concentration inequality provided by Assumption 4.3. This part of the proof we omit here and refer the reader to Denis and Hebiri (2017) or to Audibert and Tsybakov (2007) for a complete result.

Let us emphasize that the argument above is only possible due to the continuity Assumptions 1.1, 4.4 on the distribution and the estimator respectively.

### Appendix C: Proof of the lower bounds

This section is devoted to the proof of the lower bounds provided by Theorems 3.4-3.5. Before proceeding to the proofs let us briefly sketch the high-level

strategy used in this work. In order to prove the lower bounds of Theorems 3.4-3.5 we actually prove to separate lower bounds on the minimax risk. Clearly, if some non-negative quantity is lower-bounded by two different values, therefore it is lower-bounded by the maximum between the two. The two lower bounds that we prove are naturally connected with the proposed two-steps estimator, that is, the first lower bound is connected with the problem of non-parametric estimation of  $p_k$  for all  $k \in [K]$  and the second describes the estimation of the unknown threshold  $G^{-1}(\beta)$ .

In particular, the first lower bound is closely related to the one provided in (Audibert and Tsybakov, 2007; Rigollet and Vert, 2009), though, crucially the continuity Assumption 1.1 makes the proof more involved. The second lower bound is based on two hypotheses testing and is derived by constructing two different marginal distributions of  $X \in \mathbb{R}^d$  and a fixed regression vector  $p(\cdot)$ . In this part we make use of Pinsker's inequality recalled in Lemma A.2.

In order to discriminate the supervised and the semi-supervised procedures we make use of Definition 1.4. Notice that every supervised procedure thanks to Definition 1.4 is not 'sensitive' to the expectation taken *w.r.t.* the unlabeled dataset  $\mathcal{D}_N$ , that is, randomness is only induced by the labeled dataset  $\mathcal{D}_n$ . This strategy allows to eliminate the dependence of the lower bound on the size of the unlabeled dataset  $\mathcal{D}_N$  for supervised procedures. Indeed, let  $\hat{\Gamma}$  be any supervised estimator in the sense of Definition 1.4, then for any real valued function of confidence sets  $Z$  we have

$$\mathbb{E}_{(\mathcal{D}_n, \mathcal{D}_N)}[\mathbb{E}_{\mathbb{P}_X} Z(\hat{\Gamma}(X; \mathcal{D}_n, \mathcal{D}_N))] = \mathbb{E}_{\mathcal{D}_n}[\mathbb{E}_{\mathbb{P}_X} Z(\hat{\Gamma}(X; \mathcal{D}_n, \mathcal{D}'_N))] ,$$

with  $\mathcal{D}'_N$  being an arbitrary set of  $N$  points in  $\mathbb{R}^d$ .

### C.1. Part I: $(N + n)^{-1/2}$

Here we prove that the rate  $(N + n)^{-1/2}$  is optimal for semi-supervised methods, as already mentioned the rate for the supervised methods can be obtained by formally setting  $N = 0$ . The constant  $C', C, c$  are always assumed to be independent of  $N, n$  and can differ from line to line. Let us fix  $\beta \in \{1, \dots, \lfloor K/2 \rfloor - 1\}$  and  $K \geq 5$ . For a positive constant  $C < 1/2$  we define the following sequence

$$\kappa_{N,n} = C(N + n)^{-\frac{1}{2}} < 0.5 .$$

To prove the lower bound we construct two distribution  $\mathbb{P}_0$  and  $\mathbb{P}_1$  on  $\mathbb{R}^d$  sharing the same regression function  $p(\cdot) = (p_1(\cdot), \dots, p_K(\cdot))$  and with different marginals admitting densities  $\mu_0, \mu_1$ . First, for a fixed parameter  $0 < \rho < 1$  and fixed constants  $0 < r_0 < r_1 < r_2 < r_3 < r_4$  to be specified we define the

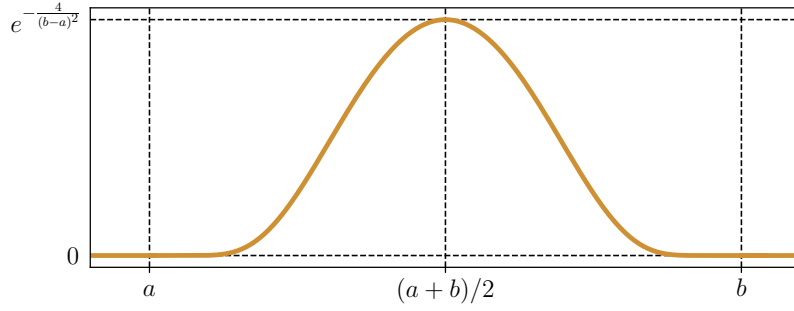


Fig 1: Bump function:  $x \mapsto \psi_{a,b}(x)$ . Importantly, this function is supported on  $(a, b)$  and is infinitely smooth.

following sets

$$\begin{aligned}
 \mathcal{X}_0 &= \{x \in \mathbb{R}^d : \|x\| \leq r_0\} \quad , \\
 \mathcal{X}_1 &= \left\{ x \in \mathbb{R}^d : \left\| x - \underbrace{(r_1 + \rho, 0, \dots, 0)^\top}_{\in \mathbb{R}^d} \right\| \leq \rho/2 \right\} \quad , \\
 \mathcal{X}_2 &= \left\{ x \in \mathbb{R}^d : \left\| x - \underbrace{(r_2 + \rho, 0, \dots, 0)^\top}_{\in \mathbb{R}^d} \right\| \leq \rho \right\} \quad , \\
 \mathcal{X}_3 &= \left\{ x \in \mathbb{R}^d : \left\| x - \underbrace{(r_3 + \rho, 0, \dots, 0)^\top}_{\in \mathbb{R}^d} \right\| \leq \rho/2 \right\} \quad , \\
 \mathcal{X}_4 &= \{x \in \mathbb{R}^d : r_4 \leq \|x\| \leq 2r_4\} \quad .
 \end{aligned}$$

Let us denote by  $o_i = (r_i + \rho, 0, \dots, 0)^\top$  for  $i = 1, 2, 3$  the centers of  $\mathcal{X}_1$ ,  $\mathcal{X}_2$  and  $\mathcal{X}_3$ . Using these sets we define the regression vector as

$$\begin{aligned}
p_1(x) = \dots = p_{2\beta}(x) &= \begin{cases} \frac{1}{2\beta} - \frac{\varphi_0(x)}{2\beta}, & x \in \mathcal{X}_0 \\ \frac{3K+2\beta}{8K\beta} - \frac{\varphi_1(x)}{2\beta}, & x \in \mathcal{X}_1 \\ \frac{2K+4\beta}{8K\beta} - \frac{\varphi_2(x)}{2\beta}, & x \in \mathcal{X}_2 \\ \frac{K+6\beta}{8K\beta} - \frac{\varphi_3(x)}{2\beta}, & x \in \mathcal{X}_3 \\ \frac{1}{K} - \frac{\varphi_4(x)}{2\beta}, & x \in \mathcal{X}_4 \end{cases} \\
p_{2\beta+1}(x) = \dots = p_K(x) &= \begin{cases} \frac{\varphi_0(x)}{K-2\beta}, & x \in \mathcal{X}_0 \\ \frac{1}{4K} + \frac{\varphi_1(x)}{K-2\beta}, & x \in \mathcal{X}_1 \\ \frac{2K-4\beta}{4K(K-2\beta)} + \frac{\varphi_2(x)}{K-2\beta}, & x \in \mathcal{X}_2 \\ \frac{3K-6\beta}{4K(K-2\beta)} + \frac{\varphi_3(x)}{K-2\beta}, & x \in \mathcal{X}_3 \\ \frac{1}{K} + \frac{\varphi_4(x)}{K-2\beta}, & x \in \mathcal{X}_4 \end{cases} ,
\end{aligned}$$

In order to define the functions  $\varphi_i$  for  $i = 0, \dots, 4$  we first define a one dimensional function of two real-valued parameters  $a < b$

$$\psi_{a,b}(x) = \begin{cases} \exp\left(-\frac{1}{(b-x)(x-a)}\right), & x \in (a, b) \\ 0, & \text{otherwise} \end{cases} .$$

Figure 1 illustrates the behavior of  $\psi_{a,b}$  function in one dimension. Note that for every  $a, b \in \mathbb{R}$  the function above is infinitely smooth. Using the definition of  $\psi_{a,b}$  we define the functions  $\varphi_i$  for  $i = 0, \dots, 4$  as

$$\begin{aligned}
\varphi_0(x) &= \frac{C'}{2} \left( \frac{K-2\beta}{8K\beta} \wedge \frac{1}{4K} \right) \psi_{-1, r_0}(\|x\|) , \\
\varphi_i(x) &= \frac{C'}{2} \rho^\gamma \left( \frac{K-2\beta}{8K\beta} \wedge \frac{1}{4K} \right) \left( \frac{\|x - o_i\|}{\rho} \right)^{2\lceil \frac{\gamma}{2} \rceil} \psi_{-1, 1} \left( \frac{\|x - o_i\|}{\rho} \right), \quad i = 1, 3 , \\
\varphi_2(x) &= \frac{C'}{2} \rho^\gamma \left( \frac{K-2\beta}{8K\beta} \wedge \frac{1}{4K} \right) \psi_{-1, 1} \left( \frac{\|x - o_2\|}{\rho} \right) , \\
\varphi_4(x) &= \frac{C'}{2} \left( \frac{K-2\beta}{8K\beta} \wedge \frac{1}{4K} \right) \psi_{r_4, 2r_4}(\|x\|) ,
\end{aligned}$$

and the constant  $C' \leq 1$  is chosen small enough so that each function  $\varphi_i$  for  $i = 0, \dots, 4$  is  $(\gamma, L)$ -Hölder. Let us point out that such value  $C'$  exists and is independent of  $n, N$ , indeed, the mapping

$$x \mapsto C' \|x\|^{2\lceil \frac{\gamma}{2} \rceil} \psi_{-1, 1}(\|x\|) ,$$

is infinitely smooth, thus it is  $(\gamma, L)$ -Hölder for a properly chosen  $C'$ . Figure 2 demonstrates the behavior of the considered construction in one dimension. Note that  $\varphi_i(x)$  for  $i = 1, 3$  are obtained from the previous mapping by rescaling which preserves the Hölder constant  $L$ . Same reasoning applies to  $\varphi_i$  for  $i = 0, 2, 4$ .



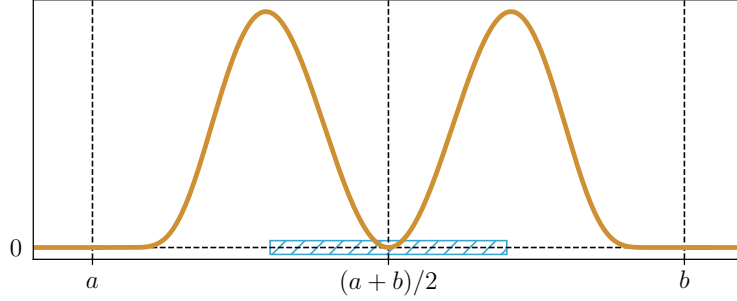


Fig 2: Dumped bump function:  $x \mapsto (x - \frac{a+b}{2})^{2\lceil \frac{\gamma}{2} \rceil} \psi_{a,b}(x)$ . Importantly, this function behaves as polynomial of even degree  $2\lceil \frac{\gamma}{2} \rceil$  in the affinity of  $\frac{a+b}{2}$ , while being infinitely smooth and supported on  $(a, b)$ . It means that if we select a measure which is supported in the affinity of  $\frac{a+b}{2}$  (light-blue hatched region) the function on the plot is essentially polynomial *w.r.t.* such a measure.

Since  $\beta < K/2$  one can check that the following relations hold true

$$\begin{aligned} 0 < \frac{1}{4K} &< \frac{2K - 4\beta}{4K(K - 2\beta)} < \frac{3K - 4\beta}{4K(K - 2\beta)} < \frac{3K - 6\beta}{4K(K - 2\beta)} < \frac{1}{K} \\ &< \frac{K + 6\beta}{8K\beta} < \frac{2K + 4\beta}{8K\beta} < \frac{3K + 2\beta}{8K\beta} < \frac{1}{2\beta} , \end{aligned}$$

which will help us to ensure that the thresholds under  $\mathbb{P}_0, \mathbb{P}_1$  are  $\frac{3K+2\beta}{8K\beta}$  and  $\frac{K+6\beta}{8K\beta}$  respectively. Now, we define two marginal distributions  $\mu_0, \mu_1$  by their densities as

$$\mu_0(x) = \begin{cases} \frac{1/2}{\text{Leb}(\mathcal{X}_0)}, & x \in \mathcal{X}_0 \\ \frac{\kappa_{N,n}}{\text{Leb}(\mathcal{X}_1)}, & x \in \mathcal{X}_1 \\ \frac{\kappa_{N,n}}{\text{Leb}(\mathcal{X}_2)}, & x \in \mathcal{X}_2 \\ \frac{\kappa_{N,n}}{\text{Leb}(\mathcal{X}_3)}, & x \in \mathcal{X}_3 \\ \frac{1/2 - 3\kappa_{N,n}}{\text{Leb}(\mathcal{X}_4)}, & x \in \mathcal{X}_4 \end{cases}, \quad \mu_1(x) = \begin{cases} \frac{1/2 - 3\kappa_{N,n}}{\text{Leb}(\mathcal{X}_0)}, & x \in \mathcal{X}_0 \\ \frac{\kappa_{N,n}}{\text{Leb}(\mathcal{X}_1)}, & x \in \mathcal{X}_1 \\ \frac{\kappa_{N,n}}{\text{Leb}(\mathcal{X}_2)}, & x \in \mathcal{X}_2 \\ \frac{\kappa_{N,n}}{\text{Leb}(\mathcal{X}_3)}, & x \in \mathcal{X}_3 \\ \frac{1/2}{\text{Leb}(\mathcal{X}_4)}, & x \in \mathcal{X}_4 \end{cases},$$

and both  $\mu_0, \mu_1$  are equal to zero in unspecified regions. Clearly, the strong density assumption is satisfied on  $\mathcal{X}_0$  and  $\mathcal{X}_4$  since the density is lower and upper-bounded by a constant independent of both  $N, n$ . The parameter  $\rho$  is chosen such that the strong density assumption on  $\mathcal{X}_i$  for  $i = 1, 2, 3$  is satisfied. Notice that

$$\text{Leb}(\mathcal{X}_i) = c\rho^d ,$$

for some constant  $c > 0$  independent of  $N, n$ , thus we set  $\rho = C(N + n)^{-1/2d}$ . For these hypotheses one can easily check that the thresholds  $G_0^{-1}(\beta), G_1^{-1}(\beta)$

and the optimal  $\beta$ -sets  $\Gamma_0^*, \Gamma_1^*$  are given as

$$\begin{aligned} G_0^{-1}(\beta) &= \frac{3K+2\beta}{8K\beta}, \quad G_1^{-1}(\beta) = \frac{K+6\beta}{8K\beta}, \\ \Gamma_0^*(x) &= \begin{cases} \{1, \dots, 2\beta\}, & x \in \mathcal{X}_0 \\ \emptyset, & \text{otherwise} \end{cases}, \\ \Gamma_1^*(x) &= \begin{cases} \{1, \dots, 2\beta\}, & x \in \mathcal{X}_0 \cup \mathcal{X}_1 \cup \mathcal{X}_2, \\ \emptyset, & \text{otherwise} \end{cases}. \end{aligned}$$

**The margin assumption:** we are in position to check the margin Assumption 2.1. Let  $t_0 = \frac{1}{2} \left( \frac{K-2\beta}{8K\beta} \wedge \frac{1}{4K} \right)$ , thus for every  $k \in \{2\beta+1, \dots, K\}$  and every  $t \leq t_0$  we have

$$\mathbb{P}_0(|p_k(X) - G_0^{-1}(\beta)| \leq t) = 0, \quad \mathbb{P}_1(|p_k(X) - G_1^{-1}(\beta)| \leq t) = 0,$$

moreover for every  $k \in \{1, \dots, 2\beta\}$  and every  $t \leq t_0$  we can write

$$\begin{aligned} &\mathbb{P}_0\left(|p_k(X) - G_0^{-1}(\beta)| \leq t\right) \\ &= \mathbb{P}_0\left(\frac{C'}{2}\rho^\gamma \left(\frac{K-2\beta}{8K\beta} \wedge \frac{1}{4K}\right) \left(\frac{\|x - o_1\|}{\rho}\right)^{2\lceil \frac{\gamma}{2} \rceil} \psi_{-1,1}\left(\frac{\|x - o_1\|}{\rho}\right) \leq 2\beta t, X \in \mathcal{X}_1\right), \\ &\mathbb{P}_1\left(|p_k(X) - G_1^{-1}(\beta)| \leq t\right) \\ &= \mathbb{P}_1\left(\frac{C'}{2}\rho^\gamma \left(\frac{K-2\beta}{8K\beta} \wedge \frac{1}{4K}\right) \left(\frac{\|x - o_3\|}{\rho}\right)^{2\lceil \frac{\gamma}{2} \rceil} \psi_{-1,1}\left(\frac{\|x - o_3\|}{\rho}\right) \leq 2\beta t, X \in \mathcal{X}_3\right). \end{aligned}$$

Hence, for the 0 hypothesis there exists  $c$  independent of  $N, n$  such that

$$\mathbb{P}_0(|p_k(X) - G_0^{-1}(\beta)| \leq t) \leq \mathbb{P}_0\left(\left(\frac{\|x - o_1\|}{\rho}\right)^{2\lceil \frac{\gamma}{2} \rceil} \psi_{-1,1}\left(\frac{\|x - o_1\|}{\rho}\right) \leq c^2 \rho^{-\gamma} t, X \in \mathcal{X}_1\right)$$

Therefore we can write using the strong density assumption

$$\begin{aligned} \mathbb{P}_0(|p_k(X) - G_0^{-1}(\beta)| \leq t) &\leq \int_{\|x - o_1\| \leq \rho/2} \mathbf{1}_{\left\{\left(\frac{\|x - o_1\|}{\rho}\right)^{2\lceil \frac{\gamma}{2} \rceil} \psi_{-1,1}\left(\frac{\|x - o_1\|}{\rho}\right) \leq c^2 \rho^{-\gamma} t\right\}} d\mu_0(x) \\ &\leq C \int_{\|x - o_1\| \leq \rho/2} \mathbf{1}_{\left\{\left(\frac{\|x - o_1\|}{\rho}\right)^{2\lceil \frac{\gamma}{2} \rceil} \psi_{-1,1}\left(\frac{\|x - o_1\|}{\rho}\right) \leq c^2 \rho^{-\gamma} t\right\}} dx \\ &= C \int_{\|x\| \leq \rho/2} \mathbf{1}_{\left\{\left(\frac{\|x\|}{\rho}\right)^{2\lceil \frac{\gamma}{2} \rceil} \psi_{-1,1}\left(\frac{\|x\|}{\rho}\right) \leq c^2 \rho^{-\gamma} t\right\}} dx \\ &= C \rho^d \int_{\|x\| \leq 1/2} \mathbf{1}_{\left\{\|x\|^{2\lceil \frac{\gamma}{2} \rceil} \psi_{-1,1}(\|x\|) \leq c^2 \rho^{-\gamma} t\right\}} dx, \end{aligned}$$

Finally notice that for every  $x \in \mathbb{R}^d$  such that  $\|x\| \leq 1/2$  we have for some  $C > 0$

$$\psi_{-1,1}(\|x\|) \geq \psi_{-1,1}(1/2) \geq C,$$

which implies that for some positive  $C, C'$  independent of  $N, n$  we can write

$$\begin{aligned} \mathbb{P}_0(|p_k(X) - G_0^{-1}(\beta)| \leq t) &\leq C\rho^d \int_{\|x\| \leq 1/2} \mathbf{1}_{\{\|x\|^{2\lceil \frac{\gamma}{2} \rceil} \leq C'\rho^{-\gamma}t\}} dx \\ &= C\rho^d \int_{\|x\| \leq 1/2} \mathbf{1}_{\{\|x\| \leq C'^{1/(2\lceil \frac{\gamma}{2} \rceil)} \rho^{-\gamma/(2\lceil \frac{\gamma}{2} \rceil)} t^{1/(2\lceil \frac{\gamma}{2} \rceil)}\}} dx \\ &\leq C\rho^{d(1-\gamma/(2\lceil \frac{\gamma}{2} \rceil))} t^{d/(2\lceil \frac{\gamma}{2} \rceil)}. \end{aligned}$$

This implies that for as long as  $\alpha \leq d/(2\lceil \frac{\gamma}{2} \rceil)$  (and since we have  $\gamma \leq 2\lceil \frac{\gamma}{2} \rceil$ ) the margin assumption is satisfied. Moreover, these conditions imply that  $\alpha\gamma \leq d$ , which we will also require while proving the supervised part of the rate. Same reasoning can be carried out for the case of the first hypothesis  $\mathbb{P}_1$  on the set  $\mathcal{X}_3$ .

Finally, the parameters  $r_0, r_1, r_2, r_3$  are chosen as constants independent of  $n, N$  such that there exists a smooth connection between the parts of the regression functions  $p_k(\cdot)$  which are defined on  $\mathcal{X}_0, \mathcal{X}_1, \mathcal{X}_2, \mathcal{X}_3, \mathcal{X}_4$ . Notice that such a choice is possible since by the construction of functions  $\varphi_i$  for  $i = 0, 1, 2, 3, 4$  they are zeroed-out on the boundaries of  $\mathcal{X}_0, \mathcal{X}_1, \mathcal{X}_2, \mathcal{X}_3, \mathcal{X}_4$ . Thus in the region  $\mathbb{R}^d \setminus \bigcup_{i=0}^4 \mathcal{X}_i$  it is sufficient to construct a function which connects four different constants smoothly. We avoid this over complication on this part and hope that the guidelines provided above are sufficient for the understanding.

Notice that the constructed distributions are satisfying Assumption 1.1 since the measures are only defined on  $\mathcal{X}_0, \mathcal{X}_1, \mathcal{X}_2, \mathcal{X}_3, \mathcal{X}_4$  and the regression functions on these sets are not concentrated around any constant.

Before proceeding to the final stage of the proof let us mention that in what follows we use the de Finetti (de Finetti, 1972, 1974) notation which is common in probability. That is, given a probability measure  $\mathbb{P}$  on some measurable space  $(\Omega_0, \mathcal{A}_0)$  and a measurable function  $X : (\Omega_0, \mathcal{A}_0) \rightarrow (\mathbb{R}, \text{Borel}(\mathbb{R}))$  we write

$$\mathbb{P}[X] := \mathbb{E}[X].$$

**Bound on the KL-divergence:** we start by computing the KL-divergence between  $\mu_0$  and  $\mu_0$

$$\begin{aligned} \text{KL}(\mu_0, \mu_1) &:= \int_{\mathbb{R}^d} \mu_0(x) \log \left( \frac{\mu_0}{\mu_1} \right) dx = \sum_{i=0}^4 \int_{x \in \mathcal{X}_i} \mu_0(x) \log \left( \frac{\mu_0(x)}{\mu_1(x)} \right) dx \\ &= \frac{1}{\text{Leb}(\mathcal{X}_0)} \int_{x \in \mathcal{X}_0} \frac{1}{2} \log \left( \frac{1/2}{1/2 - 3\kappa_{N,n}} \right) dx \\ &\quad + \frac{1}{\text{Leb}(\mathcal{X}_4)} \int_{x \in \mathcal{X}_4} \left( \frac{1}{2} - 3\kappa_{N,n} \right) \log \left( \frac{1/2 - 3\kappa_{N,n}}{1/2} \right) dx \\ &= \frac{1}{2} \log \left( \frac{1/2}{1/2 - 3\kappa_{N,n}} \right) \\ &\quad + \left( \frac{1}{2} - 3\kappa_{N,n} \right) \log \left( \frac{1/2 - 3\kappa_{N,n}}{1/2} \right) \\ &= -3\kappa_{N,n} \log(1 - 6\kappa_{N,n}) \leq 36\kappa_{N,n}^2. \end{aligned}$$

**Lower bound for the Hamming risk:** first of all let us introduce the following notation for  $i = 0, 1$

$$H(\hat{\Gamma}, \Gamma_i^*) := \mu_i \left| \hat{\Gamma}(X) \Delta \Gamma_i^*(X) \right| .$$

Recall that we are interested in the following quantity

$$\inf_{\hat{\Gamma}} \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{(\mathcal{D}_n, \mathcal{D}_N)} \mathbb{E}_{X \sim \mathbb{P}_X} \left| \hat{\Gamma}(X) \Delta \Gamma_\beta^*(X) \right| ,$$

since the hypotheses  $\mathbb{P}_0, \mathbb{P}_1 \in \mathcal{P}$  we can write

$$2 \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{(\mathcal{D}_n, \mathcal{D}_N)} \mathbb{E}_{X \sim \mathbb{P}_X} \left| \hat{\Gamma}(X) \Delta \Gamma_\beta^*(X) \right| \geq (*) ,$$

where  $(*)$  is defined as

$$(*) = \mu_0^{\otimes(n+N)} \otimes \mathbb{P}_{Y|X}^{\otimes n} H(\hat{\Gamma}, \Gamma_0^*) + \mu_1^{\otimes(n+N)} \otimes \mathbb{P}_{Y|X}^{\otimes n} H(\hat{\Gamma}, \Gamma_1^*) ,$$

thus, for the Hamming risk we can write

$$(*) \geq \mu_0^{\otimes(n+N)} \otimes \mathbb{P}_{Y|X}^{\otimes n} \left( \frac{d\mu_1^{\otimes(n+N)} \otimes \mathbb{P}_{Y|X}^{\otimes n}}{d\mu_0^{\otimes(n+N)} \otimes \mathbb{P}_{Y|X}^{\otimes n}} \wedge 1 \right) \left( H(\hat{\Gamma}, \Gamma_0^*) + H(\hat{\Gamma}, \Gamma_1^*) \right) .$$

Now we focus our attention to the sum of two Hamming differences which appearing on the right hand side of the above inequality

$$\begin{aligned} H(\hat{\Gamma}, \Gamma_0^*) + H(\hat{\Gamma}, \Gamma_1^*) &= \mu_0 \sum_{k=1}^K \mathbf{1}_{k \in \hat{\Gamma}(X) \Delta \Gamma_0^*(X)} + \mu_1 \sum_{k=1}^K \mathbf{1}_{k \in \hat{\Gamma}(X) \Delta \Gamma_1^*(X)} \\ &\geq \mu_0 \left( \frac{d\mu_1}{d\mu_0} \wedge 1 \right) \sum_{k=1}^K \mathbf{1}_{k \in \hat{\Gamma}(X) \Delta \Gamma_0^*(X)} \\ &\quad + \mu_0 \left( \frac{d\mu_1}{d\mu_0} \wedge 1 \right) \sum_{k=1}^K \mathbf{1}_{k \in \hat{\Gamma}(X) \Delta \Gamma_1^*(X)} \\ &\geq \mu_0 \left( \frac{d\mu_1}{d\mu_0} \wedge 1 \right) \sum_{k=1}^K \mathbf{1}_{k \in \Gamma_1^*(X) \Delta \Gamma_0^*(X)} \quad (\text{Triangle inequality}) \\ &= 2\beta \mu_0 \left( \frac{d\mu_1}{d\mu_0} \wedge 1 \right) (\mathbf{1}_{\mathcal{X}_1} + \mathbf{1}_{\mathcal{X}_2}) \\ &= 2\beta \int_{\mathbb{R}^d} \left( \frac{\mu_1(x)}{\mu_0(x)} \wedge 1 \right) (\mathbf{1}_{\mathcal{X}_1} + \mathbf{1}_{\mathcal{X}_2}) d\mu_0(x) \\ &= 2\beta \int_{\mathcal{X}_1} \left( \frac{\mu_1(x)}{\mu_0(x)} \wedge 1 \right) d\mu_0(x) + 2\beta \int_{\mathcal{X}_2} \left( \frac{\mu_1(x)}{\mu_0(x)} \wedge 1 \right) d\mu_0(x) \\ &= 2\beta \mathbb{P}_0(\mathcal{X}_1 \cup \mathcal{X}_2) \geq 2\beta \kappa_{n,N} . \end{aligned}$$

Substituting this lower bound into the initial inequality we arrive at

$$\begin{aligned}
(*) &\geq 2\beta\kappa_{n,N}\mu_0^{\otimes(n+N)} \otimes \mathbb{P}_{Y|X}^{\otimes n} \left( \frac{d\mu_1^{\otimes(n+N)} \otimes \mathbb{P}_{Y|X}^{\otimes n}}{d\mu_0^{\otimes(n+N)} \otimes \mathbb{P}_{Y|X}^{\otimes n}} \bigwedge 1 \right) \\
&= 2\beta\kappa_{n,N} \left( 1 - \text{TV} \left( \mu_0^{\otimes(n+N)} \otimes \mathbb{P}_{Y|X}^{\otimes n}, \mu_1^{\otimes(n+N)} \otimes \mathbb{P}_{Y|X}^{\otimes n} \right) \right) \\
&= 2\beta\kappa_{n,N} \left( 1 - \text{TV} \left( \mu_0^{\otimes(n+N)}, \mu_1^{\otimes(n+N)} \right) \right) \\
&\geq 2\beta\kappa_{n,N} \left( 1 - \sqrt{\frac{1}{2} \text{KL} \left( \mu_0^{\otimes(n+N)}, \mu_1^{\otimes(n+N)} \right)} \right) \quad (\text{Pinsker's inequality}) \\
&\geq 2\beta\kappa_{n,N} \left( 1 - \kappa_{n,N} \sqrt{n+N} \right),
\end{aligned}$$

which implies the desired lower bound on the Hamming risk.

**Lower bound for the  $\beta$  excess risk:** this part is analogues to the case of the Hamming distance. Let us recall that for every  $\hat{\Gamma}$  we have the following expression for  $i = 0, 1$

$$D(\hat{\Gamma}, \Gamma_i^*) := R_\beta(\hat{\Gamma}) - R_\beta(\Gamma_i^*) = \mu_i \sum_{k=1}^K |p_k(X) - G^{-1}(\beta)| \mathbf{1}_{\{k \in \hat{\Gamma}(X) \Delta \Gamma_i^*(X)\}}.$$

Again, recall that we are interested in

$$\inf_{\hat{\Gamma}} \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{(\mathcal{D}_n, \mathcal{D}_N)} [R_\beta(\hat{\Gamma})] - R(\Gamma_\beta^*)$$

similarly to the previous case, since the hypotheses  $\mathbb{P}_0, \mathbb{P}_1 \in \mathcal{P}$  we can write

$$2 \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{(\mathcal{D}_n, \mathcal{D}_N)} [R_\beta(\hat{\Gamma})] - R(\Gamma_\beta^*) \geq (**),$$

where  $(**)$  is defined as

$$(**) = \mu_0^{\otimes(n+N)} \otimes \mathbb{P}_{Y|X}^{\otimes n} D(\hat{\Gamma}, \Gamma_0^*) + \mu_1^{\otimes(n+N)} \otimes \mathbb{P}_{Y|X}^{\otimes n} D(\hat{\Gamma}, \Gamma_1^*),$$

we can write

$$(**) \geq \mu_0^{\otimes(n+N)} \otimes \mathbb{P}_{Y|X}^{\otimes n} \left( \frac{d\mu_1^{\otimes(n+N)} \otimes \mathbb{P}_{Y|X}^{\otimes n}}{d\mu_0^{\otimes(n+N)} \otimes \mathbb{P}_{Y|X}^{\otimes n}} \bigwedge 1 \right) (D(\hat{\Gamma}, \Gamma_0^*) + D(\hat{\Gamma}, \Gamma_1^*))$$

and we continue in a similar fashion

$$\begin{aligned}
D(\hat{\Gamma}, \Gamma_0^*) + D(\hat{\Gamma}, \Gamma_1^*) &= \mu_0 \sum_{k=1}^K \left| p_k(X) - \frac{3K+2\beta}{8K\beta} \right| \mathbf{1}_{k \in \hat{\Gamma}(X) \Delta \Gamma_0^*(X)} \\
&\quad + \mu_1 \sum_{k=1}^K \left| p_k(X) - \frac{K+6\beta}{8K\beta} \right| \mathbf{1}_{k \in \hat{\Gamma}(X) \Delta \Gamma_1^*(X)} \\
&\geq \mu_0 \sum_{k=1}^{2\beta} \left| p_k(X) - \frac{3K+2\beta}{8K\beta} \right| \mathbf{1}_{k \in \hat{\Gamma}(X) \Delta \Gamma_0^*(X)} \mathbf{1}_{X \in \mathcal{X}_2} \\
&\quad + \mu_1 \sum_{k=1}^{2\beta} \left| p_k(X) - \frac{K+6\beta}{8K\beta} \right| \mathbf{1}_{k \in \hat{\Gamma}(X) \Delta \Gamma_1^*(X)} \mathbf{1}_{X \in \mathcal{X}_2},
\end{aligned}$$

since  $\mu_0(x) = \mu_1(x)$  for all  $x \in \mathcal{X}_2$  we obtain

$$\begin{aligned}
& D(\hat{\Gamma}, \Gamma_0^*) + D(\hat{\Gamma}, \Gamma_1^*) \\
& \geq \mu_0 \left( \sum_{k=1}^{2\beta} \left| p_k(X) - \frac{3K+2\beta}{8K\beta} \right| \mathbf{1}_{k \in \hat{\Gamma}(X) \Delta \Gamma_0^*(X)} \mathbf{1}_{X \in \mathcal{X}_2} \right. \\
& \quad \left. + \sum_{k=1}^{2\beta} \left| p_k(X) - \frac{K+6\beta}{8K\beta} \right| \mathbf{1}_{k \in \hat{\Gamma}(X) \Delta \Gamma_1^*(X)} \mathbf{1}_{X \in \mathcal{X}_2} \right) \\
& \geq \mu_0 \left( \sum_{k=1}^{2\beta} \left( \left| p_k(X) - \frac{3K+2\beta}{8K\beta} \right| \wedge \left| p_k(X) - \frac{K+6\beta}{8K\beta} \right| \right) \mathbf{1}_{k \in \Gamma_1^*(X) \Delta \Gamma_0^*(X)} \mathbf{1}_{X \in \mathcal{X}_2} \right) \\
& = \mu_0 \left( \sum_{k=1}^{2\beta} \left( \left| p_k(X) - \frac{3K+2\beta}{8K\beta} \right| \wedge \left| p_k(X) - \frac{K+6\beta}{8K\beta} \right| \right) \mathbf{1}_{X \in \mathcal{X}_2} \right) \\
& = \mu_0 \left( 2\beta \left( \left| \frac{2K+4\beta}{8K\beta} - \frac{\varphi_2(X)}{2\beta} - \frac{3K+2\beta}{8K\beta} \right| \wedge \left| \frac{2K+4\beta}{8K\beta} - \frac{\varphi_2(X)}{2\beta} - \frac{K+6\beta}{8K\beta} \right| \right) \mathbf{1}_{X \in \mathcal{X}_2} \right) \\
& = \mu_0 \left( 2\beta \left( \left| \frac{K-2\beta}{8K\beta} + \frac{\varphi_2(X)}{2\beta} \right| \wedge \left| \frac{K-2\beta}{8K\beta} - \frac{\varphi_2(X)}{2\beta} \right| \right) \mathbf{1}_{X \in \mathcal{X}_2} \right) \\
& = \mu_0 \left( 2\beta \left| \frac{K-2\beta}{8K\beta} - \frac{\varphi_2(X)}{2\beta} \right| \mathbf{1}_{X \in \mathcal{X}_2} \right),
\end{aligned}$$

then, since  $\frac{\varphi_2(x)}{\beta} \leq \frac{K-2\beta}{8K\beta}$  for all  $x \in \mathcal{X}_2$ , we have

$$D(\hat{\Gamma}, \Gamma_0^*) + D(\hat{\Gamma}, \Gamma_1^*) \geq \frac{2\beta(K-2\beta)}{16K\beta} \mu_0(\mathcal{X}_2) = \frac{K-2\beta}{8K} \kappa_{n,N}.$$

Thus,

$$\begin{aligned}
(**) & \geq \frac{K-2\beta}{8K} \kappa_{n,N} \left( 1 - \text{TV} \left( \mu_0^{\otimes(n+N)} \otimes \mathbb{P}_{Y|X}^{\otimes n}, \mu_1^{\otimes(n+N)} \otimes \mathbb{P}_{Y|X}^{\otimes n} \right) \right) \\
& \geq \frac{K-2\beta}{8K} \kappa_{n,N} \left( 1 - \sqrt{\frac{1}{2} \text{KL} \left( \mu_0^{\otimes(n+N)} \otimes \mathbb{P}_{Y|X}^{\otimes n}, \mu_1^{\otimes(n+N)} \otimes \mathbb{P}_{Y|X}^{\otimes n} \right)} \right) \\
& \geq \frac{K-2\beta}{8K} \kappa_{n,N} \left( 1 - \kappa_{n,N} \sqrt{n+N} \right)
\end{aligned}$$

Which concludes the first part of the lower bounds.

### C.2. Part II: $n^{-\alpha\gamma/(2\gamma+d)}$

In this section we prove that in case of the Hamming risk  $\Psi^H$  the rate  $n^{-\alpha\gamma/(2\gamma+d)}$  is minimax optimal. Notice, that thanks to Proposition 3.1 a lower bound of order  $n^{-\alpha\gamma/(2\gamma+d)}$  on the Hamming risk  $\Psi^H$  immediately implies a lower bound of order  $n^{-(\alpha+1)\gamma/(2\gamma+d)}$  on both  $\Psi^E$  and  $\Psi^D$ .

The proof is based on the reduction of the Hamming risk to a multiple hypotheses testing problem and an application of Fano's inequality provided by Birgé (2005) recalled in Lemma A.1.

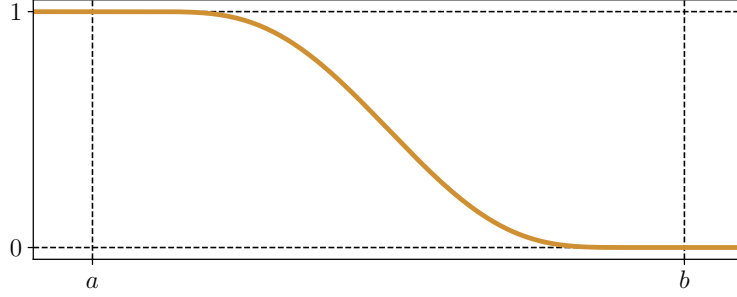


Fig 3: Integrated bump:  $x \mapsto \frac{\int_x^\infty \psi_{a,b}(t)dt}{\int_a^b \psi_{a,b}(t)dt}$ . Importantly, this function is infinitely smooth and is equal to one or zero only outside of the interval  $(a, b)$ .

Assume that  $K \geq 5$  and fix some  $\beta \in \{2, \dots, (K-2) \wedge \lfloor K/2 \rfloor\}$ , define the regular grid on  $[0, 1]^d$  as

$$G_q := \left\{ \left( \frac{2k_1+1}{2q}, \dots, \frac{2k_d+1}{2q} \right)^\top : k_i \in \{0, \dots, q-1\}, i = 1, \dots, d \right\},$$

and denote by  $n_q(x) \in G_q$  as the closest point to of the grid  $G_q$  to the point  $x \in \mathbb{R}^d$ . Such a grid defines a partition of the unit cube  $[0, 1]^d \subset \mathbb{R}^d$  denoted by  $\mathcal{X}'_1, \dots, \mathcal{X}'_{q^d}$ . Besides, denote by  $\mathcal{X}'_{-j} := \{x \in \mathbb{R}^d : -x \in \mathcal{X}'_j\}$  for all  $j = 1, \dots, q^d$ . For a fixed integer  $m \leq q^d$  and for any  $j \in \{1, \dots, m\}$  define  $\mathcal{X}_i := \mathcal{X}'_i$ ,  $\mathcal{X}_{-i} := \mathcal{X}'_{-i}$ . Additionally we introduce the following set  $\mathcal{X}_0 = \mathcal{B}(0, (4q)^{-1})$ . For every  $w \in W := \{-1, 1\}^m$  we build the distribution  $\mathbb{P}_w \in \mathcal{P}_W$ , such that, the marginal distribution  $\mathbb{P}_{w,X}$  is independent of  $w \in \{-1, 1\}^m$  and the regression vector  $(p_1^w(x), \dots, p_K^w(x))$  is constructed as

$$\begin{aligned}
p_1^w(x) &= \dots = p_{\beta-1}^w(x) = v + \frac{c'}{\beta-1} + \frac{g(x)}{\beta-1}, \\
p_\beta^w(x) &= \begin{cases} v + \phi(x), & \text{if } x \in \mathcal{X}_0 \\ v + w_i \varphi(x - n_q(x)), & \text{if } x \in \mathcal{X}_i \\ v - w_i \varphi(x - n_q(x)), & \text{if } x \in \mathcal{X}_{-i} \\ \frac{1}{K}, & \text{if } x \in \mathcal{B}(0, \sqrt{d}) \setminus \left( \bigcup_{i=-m, i \neq 0}^m \mathcal{X}_i \right) \\ \frac{3v}{2} + g(x), & \text{if } x \in \mathbb{R}^d \setminus \mathcal{B}(0, \sqrt{d} + \rho) \\ v + \xi(x), & \text{if } x \in \mathcal{B}(0, \sqrt{d} + \rho) \setminus \mathcal{B}(0, \sqrt{d}) \end{cases}, \\
p_{\beta+1}^w(x) &= \begin{cases} v - \phi(x), & \text{if } x \in \mathcal{X}_0 \\ v - w_i \varphi(x - n_q(x)), & \text{if } x \in \mathcal{X}_i \\ v + w_i \varphi(x - n_q(x)), & \text{if } x \in \mathcal{X}_{-i} \\ v, & \text{if } x \in \mathcal{B}(0, \sqrt{d}) \setminus \left( \bigcup_{i=-m, i \neq 0}^m \mathcal{X}_i \right) \\ \frac{v}{2} - g(x), & \text{if } x \in \mathbb{R}^d \setminus \mathcal{B}(0, \sqrt{d} + \rho) \\ v - \xi(x), & \text{if } x \in \mathcal{B}(0, \sqrt{d} + \rho) \setminus \mathcal{B}(0, \sqrt{d}) \end{cases}, \\
p_{\beta+2}^w(x) &= \dots = p_K^w(x) = v - \frac{c'}{K-\beta-1} - \frac{g(x)}{K-\beta-1},
\end{aligned}$$

where  $v \in [0, 1]$ ,  $\varphi : \mathbb{R}^d \mapsto \mathbb{R}_+$ , and  $\xi : \mathbb{R}^d \mapsto \mathbb{R}_+$  are to be specified. The constants  $v, c'$  are set as

$$v = \frac{1}{K}, \quad c' = \frac{(\beta-1)(K-\beta-1)}{K^2}$$

The function  $\xi$  is constructed as

$$\xi(x) = \frac{v}{2} \bar{u} \left( \frac{\|x\|_2 - \sqrt{d}}{\rho} \right), \quad \bar{u}(x) = 1 - \frac{\int_x^\infty \psi_{0,1}(t) dt}{\int_0^1 \psi_{0,1}(t) dt},$$

the function  $\bar{u}$  is infinitely many times differentialble, is equal to zero on  $(-\infty, 0]$  and to one on  $[1, +\infty)$ . Figure 3 shows the behavior of  $1 - \bar{u}$ . Taking the constant  $\rho > 0$  big enough independently of  $N, n$  we can ensure that the function  $\xi$  is  $(\gamma, L)$ -Hölder.

The function  $\phi$  is constructed similarly to the previous part of the rate, that is, for  $\phi$  we choose

$$\phi(x) = C_\phi (2q)^{-\gamma} \left( \frac{\|x\|}{(2q)^{-1}} \right)^{2 \lceil \frac{\gamma}{2} \rceil} \psi_{-1,1} \left( \frac{\|x\|}{(2q)^{-1}} \right),$$

with  $C_\phi$  being sufficiently small such that  $\phi(\cdot)$  is  $(\gamma, L)$ -Hölder and upper-bounded by  $c'/2 \wedge v/4$ . For the function  $\varphi$  we consider the following construction

$$\varphi(x) = C_\varphi q^{-\gamma} \left( u_2 \left( \frac{\|x\|}{q^{-1}} \right) + \psi_{-\frac{1}{4}, \frac{1}{4}} \left( \frac{\|x\|}{q^{-1}} \right) \right),$$



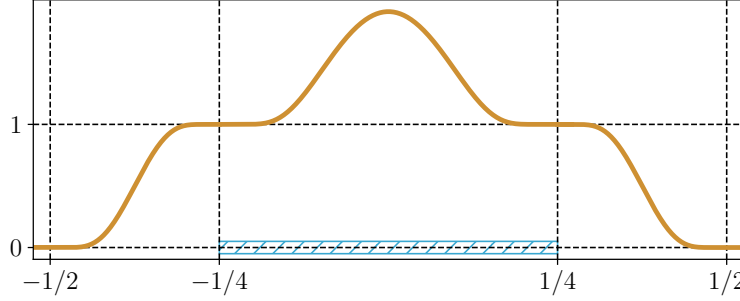


Fig 4: The function  $x \mapsto u_2(|x|) + \psi_{-1/4, 1/4}(x)$ . Importantly, this function is infinitely smooth nowhere concentrates at any constant on  $(-1/4, 1/4)$ . If the marginal measure is supported on the light-blue hatched region the function is lower- and upper-bounded almost surely.

where  $u_2(\cdot)$  is defined as

$$u_2(x) = \frac{\int_x^\infty \psi_{\frac{1}{4}, \frac{1}{2}}(t) dt}{\int_{1/4}^{1/2} \psi_{\frac{1}{4}, \frac{1}{2}}(t) dt}.$$

Figure 4 explains the behavior of this function and helps for better understanding of our results. The constant  $C_\varphi$  is chosen in such a way that the constructed function  $\varphi(\cdot)$  is  $(\gamma, L)$ -Hölder and and upper-bounded by  $c'/2 \wedge v/4$ . Notice that the function  $\varphi(x)$  for all  $x \in \mathcal{B}(0, (4q)^{-1})$  satisfies

$$C_\varphi q^{-\gamma} \leq \varphi(x) \leq C_\varphi q^{-\gamma} \left(1 + \psi_{-1/4, 1/4}(0)\right) \leq 2C_\varphi q^{-\gamma}.$$

Finally, the function  $g$  is any  $(\gamma, L)$ -Hölder function with sufficiently bounded variation which is not concentrated around any constant, for example

$$g(x) = C_g \bar{u} \left( \|x\| - \sqrt{d} - \rho \right) \cos \left( \|x\| - \sqrt{d} - \rho \right),$$

For  $C_g$  chosen small enough to ensure that it is  $(\gamma, L)$ -Hölder and has a bounded by  $c'/2 \wedge v/4$  variation.

It remains to define the marginal distribution of the vector  $X \in \mathbb{R}^d$ . We select a Euclidean ball in  $\mathbb{R}^d$  denoted by  $A_0$  that has an empty intersection with  $\mathcal{B}(0, \sqrt{d} + \rho)$  and whose Lebesgue measure is  $\text{Leb}(A_0) = 1 - mq^{-d}$ . The density  $\mu$  of the marginal distribution of  $X \in \mathbb{R}^d$  is constructed as

- $\mu(x) = \frac{\tau}{\text{Leb}(\mathcal{B}(0, (4q)^{-1}))}$  for every  $z \in G_q \cup \{0\}$  and every  $x \in \mathcal{B}(z, (4q)^{-1})$  or  $x \in \mathcal{B}(-z, (4q)^{-1})$ ,
- $\mu(x) = \frac{1-2m\tau}{\text{Leb}(A_0)}$  for every  $x \in A_0$ ,
- $\mu(x) = 0$  for every other  $x \in \mathbb{R}^d$ ,

for some  $\tau$  to be specified. Now, we check that the distributions constructed above belong to the set  $\mathcal{P}$  for every  $w \in W$ . Namely, we check the following list of assumption

- The functions  $p_1^w, \dots, p_K^w$  are defining some regression function for every  $w \in W$ . That is, for each  $x \in \mathbb{R}^d$  we have  $\sum_{k=1}^K p_k^w(x) = 1$  and  $0 \leq p_k^w(x) \leq 1$ ,
- the functions  $p_1^w, \dots, p_K^w$  are  $(\gamma, L)$ -Hölder,
- the function  $G_w(t) := \sum_{k=1}^K \int_{\mathbb{R}^d} \mathbf{1}_{\{p_k^w(x) \geq t\}} \mu(x) dx$  is continuous,
- the threshold  $G^{-1}(\beta)$  is equal to  $v$  for every  $w \in W$ ,
- the marginal distribution satisfies the strong density assumption,
- the regression function satisfies  $\alpha$ -margin assumption.

**The regression function is well defined:** to see this, notice that for every  $w \in W$  and every  $x \in \mathbb{R}^d$  we have by construction

$$\begin{aligned} p_{\beta+1}^w(x) + p_{\beta}^w(x) &= 2v, \\ \sum_{k=1}^{\beta-1} p_k^w(x) + \sum_{k=\beta+2}^K p_k^w(x) &= (K-2)v, \end{aligned}$$

and the combination of both with  $v = 1/K$  implies that  $\sum_{k=1}^K p_k^w(x) = 1$ . Moreover, as long as  $\sup_{x \in \mathcal{X}_i} \varphi(x) \leq v/2$  for every  $i = -m, \dots, -1, 1, \dots, m$  we have for every  $x \in \mathbb{R}^d$

$$0 < v/2 \leq p_{\beta+1}^w(x) \leq 3v/2 \leq 1, \quad 0 < v/2 \leq p_{\beta}^w(x) \leq 3v/2 \leq 1,$$

and by construction of the function  $g$  we have for every  $k = 1, \dots, \beta-1$ , every  $x \in \mathbb{R}^d$  and every  $w \in W$

$$0 \leq p_k^w(x) \leq v + \frac{3c'}{2(\beta-1)},$$

due to the choice of  $c', v$  we have

$$v + \frac{3c'}{2(\beta-1)} = \frac{1}{K} + \frac{3(K-\beta-2)}{2K^2} \leq \frac{2}{K} \leq 1.$$

Similarly, for every  $k = \beta+2, \dots, K$ , every  $x \in \mathbb{R}^d$  and every  $w \in W$

$$v - \frac{3c'}{2(K-\beta-1) \wedge (\beta-1)} \leq p_k^w(x) \leq 1,$$

and with the choice of  $v, c'$  specified above and the constraint  $\beta \leq \lfloor K/2 \rfloor$  we have

$$v - \frac{3c'}{2(K-\beta-1)} = \frac{1}{K} - \frac{3(\beta-1)}{2K^2} \geq \frac{1}{K} - \frac{3(K/2-1)}{2K^2} = \frac{1}{4K} + \frac{3}{2K^2} \geq 0.$$

Thus, the construction above defines some regression function for every  $w \in W$ .

**The regression function is  $(\gamma, L)$ -Hölder:** this implication follows immediately from the construction of  $\varphi, \xi, g$ .

**Continuity of  $G(t)$ :** first let us show that  $\int_{\mathbb{R}^d} \mathbf{1}_{p_k^w(x) \geq t} \mu(x) dx$  is continuous for every  $k \in [K]$ . For  $k = 1, \dots, \beta - 1, \beta + 2, \dots, K$  the continuity follows from the fact that  $g$  is not concentrated around any constant. For  $k = \beta, \beta + 1$  we first write

$$\begin{aligned} \int_{\mathbb{R}^d} \mathbf{1}_{p_k^w(x) \geq t} \mu(x) dx &= \sum_{c \in G_q \cup -G_q}^m \frac{\tau}{\text{Leb}(\mathcal{B}(c, (4q)^{-1}))} \int_{\mathcal{B}(c, (4q)^{-1})} \mathbf{1}_{p_k^w(x) \geq t} dx \\ &+ \frac{1 - 2m\tau}{\text{Leb}(A_0)} \int_{A_0} \mathbf{1}_{p_k^w(x) \geq t} dx, \end{aligned}$$

thus for this choice of  $k$  the continuity follows from the fact that  $\varphi$  and  $g$  are not concentrated around any constant.

**Threshold  $G^{-1}(\beta) = v$ :** to see this notice that for every  $w \in W$ ,

$$\sum_{k=1}^K \mathbf{1}_{p_k^w(x) \geq v} = \beta, \quad \text{a.e. } \mu,$$

and the condition on the threshold follows from the continuity of  $G(\cdot)$ . Besides, the corresponding  $\beta$ -Oracle sets  $\Gamma_w^*$  are given for every  $w \in W$  as

$$\Gamma_w^*(x) = \begin{cases} \{1, \dots, \beta - 1, \beta\}, & x \in \mathcal{X}_i, w_i = 1, \\ \{1, \dots, \beta - 1, \beta + 1\}, & x \in \mathcal{X}_i, w_i = -1, \\ \{1, \dots, \beta - 1, \beta\}, & x \in \mathcal{X}_{-i}, w_i = -1, \\ \{1, \dots, \beta - 1, \beta + 1\}, & x \in \mathcal{X}_{-i}, w_i = 1, \\ \{1, \dots, \beta - 1, \beta\}, & x \in \mathbb{R}^d \setminus (\bigcup_{i=-m}^m \mathcal{X}_i) \end{cases}$$

**The strong density assumption:** the strong density assumption can be checked following the proof of (Audibert and Tsybakov, 2007, Theorem 3.5) where an analogous construction of the marginal distribution was considered.

**$\alpha$ -margin assumption:** for all  $t \leq t_0 := v/4$ , all  $k \in [K] \setminus \{\beta, \beta + 1\}$  and all  $w \in W$  we have

$$\mu(|p_k^w(X) - v| \leq t) = 0,$$

thus for  $k \in [K] \setminus \{\beta, \beta + 1\}$  the margin assumption is satisfied. It remains to check that the margin assumption is satisfied for  $k \in \{\beta, \beta + 1\}$ . Fix an arbitrary  $w \in W$  and  $k = \beta$ , then for all  $t \leq t_0$  we can write

$$\begin{aligned} \mu\left(|p_k^w(X) - v| \leq t\right) &= \sum_{i=-m}^m \mu(|p_k^w(X) - v| \leq t, X \in \mathcal{X}_i) \\ &= \sum_{i=-m, i \neq 0}^m \mu(\varphi(X - n_q(X)) \leq t, X \in \mathcal{X}_i) + \mu(\phi(X) \leq t, X \in \mathcal{X}_0). \end{aligned}$$

We separately upper-bound both terms which appear on the right hand side of the equality.

$$\begin{aligned}\mu(\phi(X) \leq t, X \in \mathcal{X}_0) &= \frac{\tau}{\text{Leb}(\mathcal{B}(0, (4q)^{-1}))} \int_{\mathcal{B}(0, (4q)^{-1})} \mathbf{1}_{\{\phi(X) \leq t\}} dx \\ &= \frac{\tau}{\text{Leb}(\mathcal{B}(0, (4q)^{-1}))} \int_{\mathcal{B}(0, (4q)^{-1})} \mathbf{1}_{\left\{C_\phi(2q)^{-\gamma} \left(\frac{\|x\|}{(2q)^{-1}}\right)^{2\lceil \frac{\gamma}{2} \rceil} \psi_{-1,1}\left(\frac{\|x\|}{(2q)^{-1}}\right) \leq t\right\}} dx \\ &= \frac{C\tau q^{-d}}{\text{Leb}(\mathcal{B}(0, (4q)^{-1}))} \int_{\mathcal{B}(0, 1/2)} \mathbf{1}_{\{(\|x\|)^{2\lceil \frac{\gamma}{2} \rceil} \psi_{-1,1}(\|x\|) \leq C_\phi^{-1}(2q)^\gamma t\}} dx ,\end{aligned}$$

clearly there exists a constant  $C$  such that for all  $x \in \mathcal{B}(0, 1/2)$  we have

$$\psi_{-1,1}(\|x\|) \geq C ,$$

Therefore for some constant  $C > 0$  we can write

$$\begin{aligned}\mu(\phi(X) \leq t, X \in \mathcal{X}_0) &\leq \frac{C\tau q^{-d}}{\text{Leb}(\mathcal{B}(0, (4q)^{-1}))} \int_{\mathcal{B}(0, 1/2)} \mathbf{1}_{\{\|x\| \leq C(q)^{\gamma/2\lceil \frac{\gamma}{2} \rceil} t^{1/2\lceil \frac{\gamma}{2} \rceil}\}} dx \\ &\leq \frac{C\tau q^{-d(1-\gamma/2\lceil \frac{\gamma}{2} \rceil)}}{\text{Leb}(\mathcal{B}(0, (4q)^{-1}))} t^{d/2\lceil \frac{\gamma}{2} \rceil} ,\end{aligned}$$

thanks to the strong density assumption we can write for some  $C > 0$

$$\mu(\phi(X) \leq t, X \in \mathcal{X}_0) \leq Cq^{-d(1-\gamma/2\lceil \frac{\gamma}{2} \rceil)} t^{d/2\lceil \frac{\gamma}{2} \rceil} .$$

Thus since  $1 - \gamma/2\lceil \frac{\gamma}{2} \rceil \geq 0$  and  $d/2\lceil \frac{\gamma}{2} \rceil \geq \alpha$  we can write for some  $C > 0$

$$\mu(\phi(X) \leq t, X \in \mathcal{X}_0) \leq Ct^\alpha .$$

To finish this part it remains to upper-bound the other term in the margin assumption

$$\sum_{i=-m, i \neq 0}^m \mu(\varphi(X - n_q(X)) \leq t, X \in \mathcal{X}_i) = \frac{2m\tau}{\text{Leb}(\mathcal{B}(0, (4q)^{-1}))} \int_{\mathcal{B}(0, (4q)^{-1})} \mathbf{1}_{\{\varphi(X) \leq t\}} dx ,$$

using the fact that the function  $\varphi(x)$  for all  $x \in \mathcal{B}(0, (4q)^{-1})$  satisfies

$$C_\varphi q^{-\gamma} \leq \varphi(x) \leq C_\varphi q^{-\gamma} \left(1 + \psi_{-\frac{1}{4}, \frac{1}{4}}(0)\right) \leq 2C_\varphi q^{-\gamma} ,$$

we can write for all  $t \leq C_\varphi q^{-\gamma}$

$$\sum_{i=-m, i \neq 0}^m \mu(\varphi(X - n_q(X)) \leq t, X \in \mathcal{X}_i) = 0 ,$$

moreover, for all  $t \geq 2C_\varphi q^{-\gamma}$  we can write

$$\sum_{i=-m, i \neq 0}^m \mu(\varphi(X - n_q(X)) \leq t, X \in \mathcal{X}_i) \leq 2m\tau ,$$

and finally for  $t \in (C_\varphi q^{-\gamma}, 2C_\varphi q^{-\gamma})$  we can write

$$\begin{aligned} \sum_{i=-m, i \neq 0}^m \mu(\varphi(X - n_q(X)) \leq t, X \in \mathcal{X}_i) &= \frac{2m\tau}{\text{Leb}(\mathcal{B}(0, (4q)^{-1}))} \int_{\mathcal{B}(0, (4q)^{-1})} \mathbf{1}_{\{\varphi(X) \leq t\}} dx \\ &\leq \frac{2m\tau}{\text{Leb}(\mathcal{B}(0, (4q)^{-1}))} \int_{\mathcal{B}(0, (4q)^{-1})} \mathbf{1}_{\{C_\varphi q^{-\gamma} \leq t\}} dx \\ &= 2m\tau . \end{aligned}$$

The above implies that for some constant  $C > 0$  we have for all  $t \leq t_0$

$$\begin{aligned} \sum_{i=-m, i \neq 0}^m \mu(\varphi(X - n_q(X)) \leq t, X \in \mathcal{X}_i) &\leq 2\tau m \mathbf{1}_{\{t \leq 2C_\varphi q^{-\gamma}\}} \\ &\leq C\tau m q^{\gamma\alpha} t^\alpha . \end{aligned}$$

Thus the margin assumption is satisfied as long as

- $\tau m = \mathcal{O}(q^{-\gamma\alpha})$ ;
- $2\lceil \frac{\gamma}{2} \rceil \alpha \leq d$ .

Similarly one can check that the margin assumption is satisfied for  $k = \beta + 1$

**Bound on the KL-divergence:** we are in position to upper-bound the KL divergence between any two hypotheses. Fix some  $w, w' \in W$ , then using the upper bound on  $\varphi(\cdot)$  we can write for some  $C > 0$

$$\begin{aligned} \text{KL}(\mathbb{P}_w, \mathbb{P}_{w'}) &\leq 2 \sum_{i=-m, i \neq 0}^m \mu \left( \varphi(X - n_q(X)) \log \left( \frac{1 + \varphi(X - n_q(X))}{1 - \varphi(X - n_q(X))} \right), X \in \mathcal{X}_i \right) \\ &\leq C m \tau q^{-2\gamma} \end{aligned}$$

**How many hypotheses to take:** let us recall the following result which is a version of Varshamov-Gilbert bound ([Gilbert, 1952](#); [Varshamov, 1957](#)).

**Lemma C.1.** *Let  $\delta(w, w')$  denote the Hamming distance between  $w, w' \in W$  given by*

$$\delta(w, w') := \sum_{i=1}^m \mathbf{1}_{\{w_i \neq w'_i\}} .$$

*There exists  $\mathcal{W} \subset W$  such that for all  $w \neq w' \in \mathcal{W}$  we have*

$$\delta(w, w') \geq \frac{m}{4} ,$$

*and  $\log |\mathcal{W}| \geq \frac{m}{8}$ .*

Denote  $\mathcal{W} \subset W$  the set provided by Lemma C.1 and by  $\mathcal{P}_{\mathcal{W}}$  the set of distributions  $\mathbb{P}^w$  with  $w \in \mathcal{W}$ . Taking into account all the above we conclude that  $\mathcal{P}_{\mathcal{W}}$  satisfies the assumptions of our result.

**Lower bound on the Hamming risk (applying Birgé's Lemma A.1):** finally, we are in position to lower bound the hamming risk. Recall that we are interested in the following quantity

$$\inf_{\hat{\Gamma}} \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{(\mathcal{D}_n, \mathcal{D}_N)} \mathbb{E}_{\mathbb{P}_X} \left| \hat{\Gamma}(X) \Delta \Gamma_{\beta}^*(X) \right| .$$

The rest of the proof follows standard arguments, which again using the de Finetti notation read as

$$\inf_{\hat{\Gamma}} \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{(\mathcal{D}_n, \mathcal{D}_N)} \mathbb{E}_{\mathbb{P}_X} \left| \hat{\Gamma}(X) \Delta \Gamma_{\beta}^*(X) \right| \geq \inf_{\hat{\Gamma}} \sup_{w \in \mathcal{W}} \mu^{\otimes N} \otimes \mathbb{P}_w^{\otimes n} \mu \left( \left| \hat{\Gamma}(X) \Delta \Gamma_w^*(X) \right| \right) .$$

Denote by  $\hat{w}$  the following minimizer

$$\hat{w} \in \arg \min_{w \in \mathcal{W}} \mu \left( \left| \hat{\Gamma}(X) \Delta \Gamma_w^*(X) \right| \right) ,$$

thus if  $w \neq \hat{w}$  we can write using the definition of  $\hat{w}$  and the triangle inequality

$$\begin{aligned} 2\mu \left( \left| \hat{\Gamma}(X) \Delta \Gamma_w^*(X) \right| \right) &\geq \mu \left( \left| \hat{\Gamma}(X) \Delta \Gamma_w^*(X) \right| \right) + \mu \left( \left| \hat{\Gamma}(X) \Delta \Gamma_{\hat{w}}^*(X) \right| \right) \\ &\geq \mu \left( \left| \Gamma_{\hat{w}}^*(X) \Delta \Gamma_w^*(X) \right| \right) \geq 2\delta(w, \hat{w})\mu(\mathcal{X}_0) \\ &= 2\delta(w, \hat{w})\tau \geq \frac{m\tau}{2} . \end{aligned}$$

These arguments and Birgé's lemma A.1 imply that

$$\begin{aligned} \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{(\mathcal{D}_n, \mathcal{D}_N)} \mathbb{E}_{\mathbb{P}_X} \left| \hat{\Gamma}(X) \Delta \Gamma_{\beta}^*(X) \right| &\geq \frac{m\tau}{4} \sup_{w \in \mathcal{W}} \mu^{\otimes N} \otimes \mathbb{P}_w^{\otimes n} (w \neq \hat{w}) \\ &\geq \frac{m\tau}{4} \left( 0.29 \vee 1 - \frac{\sum_{w \in \mathcal{W} \setminus \{\hat{w}\}} \text{KL}(\mu^{\otimes N} \otimes \mathbb{P}_w^{\otimes n}, \mu^{\otimes N} \otimes \mathbb{P}_{\hat{w}}^{\otimes n})}{|\mathcal{W} - 1| \log |\mathcal{W}|} \right) . \end{aligned}$$

Since the marginal distribution of the vector  $X \in \mathbb{R}^d$  is shared among the hypotheses, using the upper-bound on the KL-divergence and the conditions on  $\mathcal{W}$  we get for some  $C > 0$

$$\sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{(\mathcal{D}_n, \mathcal{D}_N)} \mathbb{E}_{\mathbb{P}_X} \left| \hat{\Gamma}(X) \Delta \Gamma_{\beta}^*(X) \right| \geq \frac{m\tau}{4} (1 - Cn\tau q^{-2\gamma}) .$$

Finally, let  $q = \lfloor \bar{C}n^{1/(2\gamma+d)} \rfloor$ ,  $\tau = \lfloor C'q^{-d} \rfloor$  and  $m = \lfloor C''q^{d-\alpha\gamma} \rfloor$  for some  $\bar{C}, C', C'' > 0$  small enough we get for some  $C > 0$  and  $c < 1$

$$\sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{(\mathcal{D}_n, \mathcal{D}_N)} \mathbb{E}_{\mathbb{P}_X} \left| \hat{\Gamma}(X) \Delta \Gamma_{\beta}^*(X) \right| \geq Cn^{-\alpha\gamma/(2\gamma+d)} (1 - c) .$$

One can easily verify that this choice of parameters  $\tau, m, q$  is possible as long as  $2\lceil \frac{\gamma}{2} \rceil \alpha \leq d$  and clearly with our choice we have  $\tau m = \mathcal{O}(q^{-\alpha\gamma})$ . As already mentioned the lower bound for the excess risk and the discrepancy follows from Propositions 3.1 and 3.2.

### Appendix D: Inconsistency of top- $\beta$ approach

In this section we prove Proposition 3.3. The proof builds an explicit construction of a distribution  $\mathbb{P}$  whose  $\beta$ -Oracle satisfies  $|\Gamma_\beta^*(x)| > \beta$  for all  $x$  in some  $A \subset \mathbb{R}^d$  with  $\mathbb{P}_X(A) > 0$ . Clearly, if such a distribution exists then there is no estimator in  $\hat{\Upsilon}_\beta$  that would consistently estimate this  $\beta$ -Oracle. Let  $\beta \in [0, \dots, \lfloor K/2 \rfloor - 1]$  be a fixed integer and  $K \geq 3$ . For the proof of the theorem we shall construct one distribution  $\mathbb{P}$  for which none of the estimators with a fixed information can perform well. We start by specifying the marginal distribution of  $X \in \mathbb{R}^d$ . We start the construction by specifying the density  $\mu$  of the marginal distribution  $\mathbb{P}_X$ . Define a disk in  $\mathbb{R}^d$  for some positive  $r \leq r'$  as  $\mathcal{D}(r, r') = \{x \in \mathbb{R}^d : r \leq \|x\| \leq r'\}$ . First of all fix some parameters  $r_1 < r_2 < 2r_2 < r_3 < 2r_3$  which are independent from  $n, N$ . The density  $\mu$  is supported on  $\mathcal{B}(0, r_1) \cup \mathcal{D}(r_2, 2r_2) \cup \mathcal{D}(r_3, 2r_3)$ .

Moreover,

- $\mu(x) = \frac{\frac{\beta}{\beta+1} - \text{Leb}(\mathcal{B}(0, r_1))}{\text{Leb}(\mathcal{D}(r_2, 2r_2))}$  for all  $x \in \mathcal{D}(r_2, 2r_2)$ ,
- $\mu(x) = \frac{1}{(\beta+1)\text{Leb}(\mathcal{D}(r_3, 2r_3))}$  for all  $x \in \mathcal{D}(r_3, 2r_3)$ ,
- $\mu(x) = 1$ , for all  $x \in \mathcal{B}(0, r_1)$ ,
- $\mu(x) = 0$  otherwise,

where  $r_1 > 0$  is chosen small enough to ensure that  $\frac{\beta}{\beta+1} - \text{Leb}(\mathcal{B}(0, r_1)) > 0$ .

The regression function  $p(\cdot) = (p_1(\cdot), \dots, p_K(\cdot))^\top$  are defined as

$$p_1(x) = \dots = p_{\beta+1}(x) = \begin{cases} \frac{1}{2(\beta+1)} + C_L \frac{1 - \cos\left(\frac{2\pi}{r_1}\|x\|\right)}{\beta+1}, & x \in \mathcal{B}(0, r_1) \\ \frac{1}{2(\beta+1)} + \frac{g(x)}{\beta+1}, & x \in \mathcal{D}(r_1, r_2) \\ \frac{1}{\beta+1} - C_L \frac{1 - \cos\left(\frac{2\pi}{r_2}\|x\|\right)}{\beta+1}, & x \in \mathcal{D}(r_2, 2r_2) \\ \frac{1}{\beta+1} - \frac{\xi(x)}{\beta+1}, & x \in \mathcal{D}(2r_2, r_3) \\ \frac{1}{4(\beta+1)} - C_L \frac{1 - \cos\left(\frac{2\pi}{r_3}\|x\|\right)}{\beta+1}, & x \in \mathbb{R}^d \setminus \mathcal{B}(0, r_3) \end{cases},$$

$$p_{\beta+2}(x) = \dots = p_K(x) = \begin{cases} \frac{1}{2(K-\beta-1)} - C_L \frac{1 - \cos\left(\frac{2\pi}{r_1}\|x\|\right)}{\beta+1}, & x \in \mathcal{B}(0, r_1) \\ \frac{1}{2(K-\beta-1)} - \frac{g(x)}{K-\beta-1}, & x \in \mathcal{D}(r_1, r_2) \\ C_L \frac{1 - \cos\left(\frac{2\pi}{r_2}\|x\|\right)}{K-\beta-1}, & x \in \mathcal{D}(r_2, 2r_2) \\ \frac{\xi(x)}{K-\beta-1}, & x \in \mathcal{D}(2r_2, r_3) \\ \frac{3}{4(K-\beta-1)} + C_L \frac{1 - \cos\left(\frac{2\pi}{r_3}\|x\|\right)}{\beta+1}, & x \in \mathbb{R}^d \setminus \mathcal{B}(0, r_3) \end{cases},$$

where the constant  $C_L$  is chosen small enough to ensure that these functions are  $(\gamma, L)$ -Hölder and have sufficiently small variation. Consider an arbitrary infinitely many times differentiable function  $v : \mathbb{R} \mapsto [0, 1]$  which satisfies  $v(x) = 0$  for all  $x \leq 0$  and  $v(x) = 1$  for all  $x \geq 1$ . Then, the functions  $g(\cdot)$  and  $\xi(\cdot)$  are defined as  $g(x) = \frac{1}{2}v\left(\frac{\|x\| - r_1}{r_2 - r_1}\right)$ ,  $\xi(x) = \frac{3}{4}v\left(\frac{\|x\| - 2r_2}{r_3 - 2r_2}\right)$ . The above construction

defines a distribution  $\mathbb{P}$  for which we have

$$G^{-1}(\beta) = \frac{1}{2(\beta+1)}$$

$$\Gamma_{\beta}^*(x) = \begin{cases} \{1, \dots, \beta+1\}, & x \in \mathcal{B}(0, r_1) \cup \mathcal{D}(r_2, 2r_2) \\ \emptyset, & \text{otherwise} \end{cases}.$$

Indeed, let us evaluate the following quantity under the assumption that  $\beta \leq \lfloor K/2 \rfloor - 1$

$$\begin{aligned} \sum_{k=1}^K \int \mathbf{1}_{p_k(x) \geq G^{-1}(\beta)} \mu(x) dx &= (\beta+1) \left( \int_{\mathcal{B}(0, r_1)} \mu(x) dx + \int_{\mathcal{D}(r_2, 2r_2)} \mu(x) dx \right) \\ &= (\beta+1) \left( \text{Leb}(\mathcal{B}(0, r_1)) + \left( \frac{\beta}{\beta+1} - \text{Leb}(\mathcal{B}(0, r_1)) \right) \right) \\ &= \beta. \end{aligned}$$

Thus, using this distribution we can write for any classifier  $\hat{\Gamma} \in \hat{\Upsilon}_{\beta}$  with fixed cardinal

$$\begin{aligned} \mathbb{P}(\hat{\Gamma}) - \mathbb{P}(\Gamma_{\beta}^*) &= \int_{\mathbb{R}^d} \sum_{k=1}^K |p_k(x) - G^{-1}(\beta)| \mathbf{1}_{k \in \hat{\Gamma}(x) \triangle \Gamma^*(x)} \mu(x) dx \\ &\geq \int_{\mathcal{D}(r_2, 2r_2)} \left| \frac{1}{(\beta+1)} - C_L \frac{1 - \cos\left(\frac{2\pi}{r_2} \|x\|\right)}{\beta+1} - \frac{1}{2(\beta+1)} \right| \mu(x) dx \\ &= \int_{\mathcal{D}(r_2, 2r_2)} \left| \frac{1}{2(\beta+1)} - C_L \frac{1 - \cos\left(\frac{2\pi}{r_2} \|x\|\right)}{\beta+1} \right| \frac{\frac{\beta}{\beta+1} - \text{Leb}(\mathcal{B}(0, r_1))}{\text{Leb}(\mathcal{D}(r_2, 2r_2))} dx, \end{aligned}$$

where the first inequality follows from the observation that for  $x \in \mathcal{D}(r_2, 2r_2)$  there is always at least one label  $k$  such that  $k \in \hat{\Gamma}(x) \triangle \Gamma^*(x)$ . Thus, since the constant  $C_L$  is chosen to satisfy  $2C_L/(\beta+1) \leq 1/4(\beta+1)$  we have for any  $\hat{\Gamma} \in \hat{\Upsilon}_{\beta}$

$$\mathbb{P}(\hat{\Gamma}) - \mathbb{P}(\Gamma_{\beta}^*) \geq \frac{\frac{\beta}{\beta+1} - \text{Leb}(\mathcal{B}(0, r_1))}{4(\beta+1)},$$

If  $r_1$  is such that  $\text{Leb}(\mathcal{B}(0, r_1)) \leq \frac{\beta}{2(\beta+1)}$  we get

$$\mathbb{P}(\hat{\Gamma}) - \mathbb{P}(\Gamma_{\beta}^*) \geq \frac{\beta}{8(\beta+1)^2}, \quad \text{almost surely}.$$

By construction, the regression vector is  $(\gamma, L)$ -Hölder and the density is lower- and upper-bounded by some positive constants. Hence, it remains to check that the constructed distribution satisfies the  $\alpha$ -margin assumption. This can be achieved by an appropriate choice of  $r_1$ . Indeed, on the sets  $\mathcal{D}(r_2, 2r_2) \cup$



$\mathcal{D}(r_3, 2r_3)$  there is a “corridor” of constant size between the regression functions and the threshold  $G^{-1}(\beta)$ . The threshold  $G^{-1}(\beta)$  is only approached by the regression function on the set  $\mathcal{B}(0, r_1)$ . As all the parameters in our construction are independent from  $n, N \in \mathbb{N}$  we can find a value  $r_1$  being small enough so that the  $\alpha$ -margin assumption is verified for a fixed  $\alpha > 0$ .