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

EVGENII CHZHEN, CHRISTOPHE DENIS* and MOHAMED HEBIRI

Université Paris-Saclay
CNRS, Inria, Laboratoire de mathématiques d’Orsay
91405 Orsay, France
E-mail: evgenii.chzhen@universite-paris-saclay.fr

Université Paris-Est – Marne-la-Vallée
5 boulevard Descartes
77454 Marne-la-Vallée cedex 2
E-mail: christophe.denis@u-pem.fr; mohamed.hebiri@u-pem.fr

We study supervised and semi-supervised algorithms in the set-valued classification framework with controlled expected size. While the former methods can use only \(n\) labeled samples, the latter are able to make use of \(N\) additional unlabeled data. We obtain semi-supervised minimax rates of convergence under the \(\alpha\)-margin assumption and a \(\beta\)-Hölder condition on the conditional distribution of labels. Our analysis implies that if no further assumption is made, there is no supervised method that outperforms the semi-supervised estimator proposed in this work – the best achievable rate for any supervised method is \(O(n^{-1/2})\), even if the margin assumption is extremely favorable; on the contrary, the developed semi-supervised estimator can achieve faster \(O((n/\log n)^{-\frac{1+\alpha}{2(\beta+d)}})\) rate of convergence provided that sufficiently many unlabeled samples are available. We also show that under additional smoothness assumption, supervised methods are able to achieve faster rates and the unlabeled sample cannot improve the rate of convergence. Finally, a numerical study supports our theory and emphasizes the relevance of the assumptions we required from an empirical perspective.

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

Keywords: multi-class classification, set-valued classification, minimax optimality, semi-supervised classification.

1. Introduction

Let \(K \geq 2\) and \((X,Y) \in \mathbb{R}^d \times [K] := \{1,\ldots,K\}\) be a random couple following a distribution \(\mathbb{P}\) on \(\mathbb{R}^d \times [K]\). The vector \(X \in \mathbb{R}^d\) is seen as the vector of features with the marginal distribution \(\mathbb{P}_X\) and \(Y \in [K]\) is the corresponding class. In the framework of the set-valued classification, the decision rule is defined as a mapping \(\Gamma: \mathbb{R}^d \rightarrow 2^{[K]}\), and then belongs to the set \(\Upsilon\) of all measurable functions from \(\mathbb{R}^d\) to \(2^{[K]}\) (Here \(2^{[K]}\) stands for the set of all subsets of \([K]\)). That is, instead of a single label prediction, a set-valued classifier \(\Gamma\) is able to output a set of possible label candidates. Two parameters naturally
arise in this framework: the error and the size of a set-valued classifier $\Gamma$ defined as

$$P(\Gamma) = P(Y \notin \Gamma(X)),$$

$$S(\Gamma) = \mathbb{E} |\Gamma(X)|,$$

where $|\cdot|$ stands for the cardinality of a finite set. It is clear that the best set-valued classifier in terms of the error outputs all classes simultaneously while the best set-valued classifier in terms of the size does not output anything. Thus, a trade-off between the two quantities ought to be found by a good set-valued classifier. In this work, we focus on outputs with moderate size to gain the interpretability of the resulting set-valued classifier, this idea is explained in details in Section 2, where we motivate the considered framework. We fix $s \in (0, K)$, which is chosen by a practitioner, and the best set-valued classifier $\Gamma^*_s$ whose expected size is controlled by $s$ is defined as

$$s\text{-Oracle: } \Gamma^*_s = \arg \min \{P(\Gamma) : \Gamma \in \Upsilon \text{ s.t. } S(\Gamma) \leq s\}.$$  

(1.1)

Let us point out that the feasible set $\{\Gamma \in \Upsilon : S(\Gamma) \leq s\}$ of the above problem is distribution dependent. It implies that a priori we cannot decide whether a given set-valued classifier is feasible. However, this set only depends on the marginal distribution $\mathbb{P}_X$ of the features, which motivates us to introduce unlabeled sample in the observational model. Hence, we are interested in the semi-supervised setup of this problem. That is, in what follows it is assumed that two independent samples are provided – labeled $D^n_L = \{(X_1, Y_1), \ldots, (X_n, Y_n)\} \overset{i.i.d.}{\sim} \mathbb{P}$ and unlabeled $D^u_N = \{X_{n+1}, \ldots, X_{n+N}\} \overset{i.i.d.}{\sim} \mathbb{P}_X$ both being independent from $(X, Y)$. The statistical goal is to construct an empirical rule (a set-valued classifier) $\hat{\Gamma} : (\mathbb{R}^d \times [K])^n \times (\mathbb{R}^d)^N \to \Upsilon$, which mimics the behavior of $\Gamma^*_s$. Our work addresses the following questions:

Q1. What is a minimax setup in this problem and what are the minimax rates of convergence?

Q2. Can we statistically justify the introduction of the unlabeled data $D^n_u$ from the minimax perspective? To be more precise, we would like to understand whether the rates of convergence are affected by $N$ – the size of the unlabeled sample.

Neither of these natural questions have been considered and answered in the previous literature.

The paper is organized as follows. Our exposition starts with Section 2, where we motivate and explain the benefits of the set-valued classification with controlled expected size. Section 3 is devoted to the review of previous works and a more detailed discussion of our contributions. We introduce additional notation and introduce the minimax framework in Section 4. The next two sections are devoted to the exposition of the main results. In Sections 5 and 6 we derive minimax lower and upper bounds respectively and discuss the implications of our results. In particular, Section 6.3 develops a finer analysis of the impact of unlabeled data. A numerical study is provided in Section 7. There we illustrate the performance of set-valued classifiers and focus on the assumptions from an

---

$^1$By agreement $N = 0$ means that $D^u_N = \emptyset$
empirical perspective. Finally, in Section 8 we summarize our contributions and draw future research directions.

2. Motivation and comparison to other set-valued frameworks

Multi-class classification problems emerge in numerous applications ranging from image recognition to medical diagnosis (Russakovsky et al., 2015; LeCun and Cortes, 2010; You et al., 2011; Zhu et al., 2013). The conventional goal of the multi-class classification is to predict a single class $Y$ for the feature vector $X$. However, some multi-class classification problems might be very complex due to the possible ambiguity that may occur between classes. For instance, this is the case in the celebrated ImageNet problem (Russakovsky et al., 2015). These ambiguous cases tend to become more and more common in modern applications due to their complexity. For such problems, the conventional single-output classifier might lead to an unsatisfactory performance and it is better to provide a set of possible candidates instead of a single one. Actually, even if we are faced with a relatively unambiguous problem (e.g., MNIST problem LeCun and Cortes (2010)), it might be still beneficial to allow to predict a set of candidates if the accuracy can be significantly improved.

As soon as we start considering sets as possible predictions, a natural question is related to the interpretability and the informativeness of the outputs – larger sets are less informative but more likely to contain the underlying truth, while smaller behave oppositely. As an initial attempt to build a set-valued classifier, it is tempting to look for the one that is defined as

$$\Gamma^*_{\text{top}}(s) \in \arg \min \{ \mathbb{P}(Y \notin \Gamma(X)) : \forall x \in \mathbb{R}^d |\Gamma(x)| = s \} ,$$

with some $s \in [K]$. Setting $s = 1$ the corresponding $\Gamma^*_{\text{top}(1)}$ reduces to the Bayes optimal classifier in the classical single-output prediction. We refer to (Oh, 2017; Lapin, Hein and Schiele, 2015) and references therein for developments in this direction. However, such an approach does not take into account inhomogeneous structure of the problem. That is, in the regions where the classification is easy, a good set-valued classifier should output less candidates and it should output more candidates for difficult regions. The constraint on the expected size allows to bypass this drawback. It is important to emphasize that unlike $\Gamma^*_{\text{top}(1)}$, the $s$-Oracle defined in Eq. (1.1) with $s = 1$ does not correspond to the Bayes optimal classifier in the classical single-output prediction. Indeed, the constraint imposed on $\Gamma^*_1$ is in expectation, while $\Gamma^*_{\text{top}(1)}$ satisfies almost sure constraint. Of course, since $|\Gamma^*_{\text{top}(1)}(x)| = 1$, then $S(\Gamma^*_{\text{top}(1)}) = 1$ and by optimality of $\Gamma^*_1$ we have $\mathbb{P}(\Gamma^*_1) \leq \mathbb{P}(\Gamma^*_{\text{top}(1)})$. Thus, the 1-Oracle $\Gamma^*_1$ is always at least as good as the Bayes classifier and can potentially be superior to it (see Section 7 for an empirical illustration).

Yet another way to define a set-valued classification framework is proposed by Vovk (2002a,b); Vovk, Gammerman and Shafer (2005) and statistically addressed by Sadinle,
Lei and Wasserman (2018), where for a fixed \( \alpha \in (0, 1) \), they define \( \Gamma^*_\alpha \) as

\[
\text{Controlled error: } \quad \Gamma^*_\alpha \in \arg \min \{ \mathbb{E}[\Gamma(X)] : \mathbb{P}(Y \neq \Gamma(X)) \leq \alpha \} ,
\]

that is, \( \Gamma^*_\alpha \) is the “smallest” set-valued classifier with controlled probability of error. Even though this framework is intuitive, in certain situations it suffers from the lack of interpretability and the lack of stability w.r.t. the parameter \( \alpha \).

Let us construct a simple example to illustrate this phenomena. Assume that \( K \geq 10 \) is an even integer and let us define the following distribution \( \mathbb{P} \) of the pair \((X, Y)\):

- The marginal distribution of the features \( \mathbb{P}_X \) is uniform on \([0, 1]^d \cup [1, 2]^d\).
- The conditional distribution of the label \( p_k(x) := \mathbb{P}(Y = k|X = x) \) for \( k \in [K] \) satisfies

\[
\begin{cases}
    p_1(x) := 0.75 & \forall k \in \left[0, \frac{K}{2}\right], \quad p_k(x) := \frac{0.3}{K-2} & \forall k > \frac{K}{2}, \\
    p_1(x) := p_2(x) := p_3(x) = 0.25 & \forall k \in [4, \frac{K}{2}], \quad p_k(x) := \frac{0.3}{K-6} & \forall k > \frac{K}{2}, \\
    p_1(x) := 0.2 & \forall k \in [1, 2], \quad p_k(x) := \frac{0.2}{K} & \forall k \in [3, K].
\end{cases}
\]

\( \forall x \in [0, 1]^d \)

\( \forall x \in (1, 2]^d \)

The above example is tailored to reflect the inhomogeneous structure of the classification problem. Indeed, on the set \([0, 1]^d\) there is one dominant class and the classification is easy, whereas on the set \((1, 2]^d\) the classification is more difficult as three labels can equally explain the observation. Moreover, if we decide to rely on the previous framework and encounter this type of distributions, then for all values \( \alpha \leq 0.1 \) the set-valued classifier \( \Gamma^*_\alpha \) satisfies \( |\Gamma^*_\alpha(X)| \geq \frac{K}{2} \) almost surely \( \mathbb{P}_X \). That is, for the most intuitive range of levels\(^2\) \( \alpha \leq 0.1 \), the optimal set-valued classifier with controlled error \( \Gamma^*_\alpha \) is too large and its use is limited. This situation occurs when the probability of almost each individual class is very low; yet, their total probability is relatively high. Returning to our example, for large values of \( K \) and all \( x \in [0, 1]^d \) the values of \( p_2, \ldots, p_K \) are negligible. However, their total impact, that is \( p_2 + \ldots + p_K \), adds up to 0.25. In this situation the optimal set-valued classifier with controlled error \( \Gamma^*_\alpha \) is deemed to output almost all classes if the goal is to control the error. Nevertheless, this framework can be well suited for applications where the control on the error is crucial and even massive outputs are acceptable.

The situation is different for the framework considered in this paper, for \( s = 2 \), the set-valued classifier (s-Oracle) \( \Gamma^*_s \) outputs only one label on \([0, 1]^d\) and three labels on \((1, 2]^d\), while preserving its essential feature of having the controlled size in expectation. Consequently, the considered framework gains in both interpretability and stability of the outcome and diminish the impact of unlikely labels. This setup was considered first by Denis and Hebiri (2017), where they proposed an algorithm based on convex empirical risk minimization. They did not prove that their rate is optimal and did not study the fundamental limits of semi-supervised learning. Moreover, their result does not imply consistency of their algorithm in the absence of the margin assumption. In this work we fill these gaps by addressing the optimality questions and describing those regimes where

\(^2\)The level 0.1 is selected only for the sake of simplicity and other values can be considered.
unlabeled data strictly improves the rates of convergence. Our contribution also reveals that the previous rate, though under another set of assumptions, is suboptimal.

3. Related work

In what follows we would like to discuss the relation of the present work with the previous contributions and highlight the core differences. The scope of our work is two-fold – we study a non-standard setup of classification problem and theoretically treat the benefit of unlabeled data. Consequently, our literature review consists of two parts related to these subjects.

3.1. Classification literature

The set-valued classification is naturally related to the standard classification settings. It has been widely studied from the minimax point of view in the binary classification framework, see for instance (Yang, 1999; Mammen and Tsybakov, 1999; Györfi et al., 2002; Massart and Nédélec, 2006; Audibert and Tsybakov, 2007). In particular, 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. They derive fast rates of convergence (faster than $n^{-1/2}$) 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.

Minimax setup of the set-valued framework can also be related to the level set estimation problem (Hartigan, 1987; Polonik, 1995; Tsybakov, 1997; Rigollet and Vert, 2009). To draw this relation, we prove in Proposition 4.2, that under a mild assumption, for every $s \in (0, K)$, the set-valued classifier ($s$-Oracle) $\Gamma^*_s$ is given by

$$\Gamma^*_s(x) = \{k \in [K] : P(Y = k | X = x) \geq \theta_s\} ,$$

where $\theta_s$ is an unknown threshold, which depends on the distribution $P$ of $(X, Y)$. Similarly, the problem of level set estimation, focuses on the estimation of

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

where $f$ is the density of the observations and $\lambda > 0$ is some fixed value. Rigollet and Vert (2009) study plug-in density level set estimators through the measure of symmetric differences and the excess mass. In particular, they derive fast rates of convergence, that is faster than $n^{-1/2}$, for the excess mass. An important difference is that, in the level set estimation problem, the threshold $\lambda$ is chosen beforehand; whereas in our work, the threshold $\theta_s$ depends on the distribution of the data which makes the statistical analysis more involved.
3.2. Semi-supervised learning

In a range of classical problems of statistics, the inference is solely governed by the behavior of the conditional distribution $P_{Y|X}$ (for instance regression or binary classification). Unlike those situations, the optimal set-valued classifier $\Gamma^*_s$, in this work, depends on the whole distribution $P$. Precisely due to this reason, we allow to observe extra unlabeled data $D_U$ to better estimate the marginal distribution $P_X$ of $X$. From the practical point of view, whenever unlabeled data are available, it is always reasonable to assume that $N \geq n$, since the labeling process is much more expensive than feature gathering process. Yet, we do not explicitly require this relation between $N$ and $n$ and we distinguish two global practical regimes: semi-supervised ($N > 0$) and supervised ($N = 0$). In the former the statistician has access to an additional unlabeled sample and can potentially exploit it for better prediction, while in the latter, the statistician only observes the labeled sample.

Semi-supervised methods are studied in several papers (Vapnik, 1998; Rigollet, 2007; Singh, Nowak and Zhu, 2009; Bellec et al., 2018) and references therein. Some contributions aim at improving a given supervised estimator with the help of unlabeled data and demonstrate this improvement empirically. In contrast, our paper aims at understanding whether the semi-supervised methods should be considered superior to their supervised counterparts from minimax point of view. Our findings reveal that the semi-supervised approach might or might 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).

3.3. Main contributions

Below we summarize our contributions.

- Our results focus on the case where the conditional distribution of $Y$ belongs to a Hölder class and the joint distribution $P$ satisfies the margin condition. Under these assumptions, we establish lower bounds on the minimax rates for the set-valued classification framework with controlled expected size.
- Furthermore, we describe the situations when the semi-supervised ($N > 0$) estimation should be considered statistically superior to its supervised ($N = 0$) 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. Also, we show that in our set-valued classification setup, supervised procedures cannot achieve fast rates, that is, their rates of convergence cannot be faster than $n^{-1/2}$ (in contrast to some other classical settings (Audibert and Tsybakov, 2007; Rigollet and Vert, 2009; Herbei and Wegkamp, 2006) that allow to achieve faster rates for supervised methods).
- We provide a set-valued classification procedure which is optimal or optimal up to a 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.

• Our final contributions further explore the limits of semi-supervised estimation. We establish that under extra assumption, which are rather restrictive, the unlabeled data cannot improve the rate of convergence and only labeled data suffices.

4. Framework and notation

First let us introduce some generic notation that is used throughout this work. For two numbers \(a, a' \in \mathbb{R}\) we denote by \(a \lor a'\) (resp. \(a \land a'\)) 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 \(a\). For a positive integer \(m\) we denote by \([m]\), the set of the first positive integers up to \(m\). 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 \(B(x, r)\). 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 \(p : \mathbb{R}^d \to \mathbb{R}^K\) and a Borel measure \(\mu\) on \(\mathbb{R}^d\) we define the infinity norm of \(p\) as \(\|p\|_{\infty, \mu} := \inf\{C \geq 0 : \max_{k \in [K]} |p_k(x)| \leq C, \text{ a.e. } x \in \mathbb{R}^d \text{ w.r.t. } \mu\}\). In this work \(C\) or its lower-cased versions always refers to some constants which might be different 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.

We denote by \(\mathbb{P}_X\) the marginal distribution of \(X \in \mathbb{R}^d\) and by \(p(\cdot) := (p_1(\cdot), \ldots, p_K(\cdot))^\top\) the regression function defined for all \(k \in [K]\) and all \(x \in \mathbb{R}^d\) as \(p_k(x) := \mathbb{P}(Y = k|X = x)\). For any sets \(A, A' \subset [K]\) we denote by \(A \triangle A'\) their symmetric difference.

4.1. Properties of s-Oracle

The complete set of assumptions that is used for minimax analysis is postponed to Section 4.3, where we describe the family of joint distributions of \((X, Y)\). However, let us start by the following mild continuity assumption.

**Assumption 4.1 (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 4.2 (Denis and Hebiri (2017)).** Fix $s \in (0, K)$, and let the function $G : [0, 1] \to [0, K]$ 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 4.1 an $s$-Oracle set-valued classifier $\Gamma_s^*$ can be obtained for all $x \in \mathbb{R}^d$ as

$$\Gamma_s^*(x) = \{ k \in [K] : p_k(x) \geq G^{-1}(s) \} ,$$

(4.1)

where we denote by $G^{-1}$ the generalized inverse of $G$ defined for all $s \in (0, K)$ as $G^{-1}(s) := \inf \{ t \in [0, 1] : G(t) \leq s \}$.

Let us draw the attention to the fact that the threshold $G^{-1}(s)$ depends on the joint distribution $\mathbb{P}$ and thus, is unknown beforehand. Under the continuity assumption, the considered framework is well posed in the sense that the $s$-Oracle set-valued classifier $\Gamma_s^*$ is unique up to changes on sets of $\mathbb{P}_X$ zero measure.

**Theorem 4.3.** For every $s \in (0, K)$, under Assumption 4.1 the $s$-Oracle set-valued classifier $\Gamma_s^*$ defined in Proposition 4.2 is unique up to changes on $\mathbb{P}_X$ zero measure. That is, for all $\Gamma : \mathbb{R}^d \to 2^{[K]}$ with $S(\Gamma) \leq s$ either of the following conditions hold

- $\mathbb{P}(\Gamma) > \mathbb{P}(\Gamma_s^*)$,
- $\Gamma(x) = \Gamma_s^*(x)$ for almost all $x \in \mathbb{R}^d$ w.r.t. $\mathbb{P}_X$.

Proof of this result can be found in Appendix A. The continuity assumption ensures another description of the optimal set-valued classifier. Namely the next proposition establishes that $s$-Oracle can be obtained via an unconstrained minimization, which trades-off the error and the size.

**Proposition 4.4 (Denis and Hebiri (2017)).** Assume that Assumption 4.1 is fulfilled, then the $s$-Oracle defined in Eq. (4.1) is a minimizer of the following risk

$$R_s(\Gamma) = \mathbb{P}(\Gamma) + G^{-1}(s) S(\Gamma) .$$

(4.2)

Consequently, the accuracy of a set-valued classifier $\Gamma$ can be for instance quantified according to its excess risk

$$R_s(\Gamma) - R_s(\Gamma_s^*) = \sum_{k=1}^{K} \mathbb{E}_{\mathbb{P}_X} \left[ |p_k(X) - G^{-1}(s)| 1_{\{k \in \Gamma(X) \Delta \Gamma_s^*(X)\}} \right].$$
4.2. Minimax setup

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

$$
\Psi^H_{n,N}(\hat{\Gamma}; \mathcal{P}) := \sup_{\mathcal{P}} \mathbb{E}_{(D^L_n, D^U_N)} \mathbb{E}_{\mathcal{F}_X} \left| \hat{\Gamma}(X) \triangle \Gamma^*_s(X) \right| \quad \text{(Hamming risk)},
$$

$$
\Psi^E_{n,N}(\hat{\Gamma}; \mathcal{P}) := \sup_{\mathcal{P}} \mathbb{E}_{(D^L_n, D^U_N)} R_s(\hat{\Gamma}) - R_s(\Gamma^*_s) \quad \text{(Excess risk)},
$$

$$
\Psi^D_{n,N}(\hat{\Gamma}; \mathcal{P}) := \sup_{\mathcal{P}} \mathbb{E}_{(D^L_n, D^U_N)} \left[ \left| P(\hat{\Gamma}) - P(\Gamma^*_s) \right| + \left| s - S(\hat{\Gamma}) \right| \right] \quad \text{(Discrepancy)}
$$

where $\mathbb{E}_{(D^L_n, D^U_N)}$ denotes the expectation $w.r.t.$ the joint distribution of data $D^L_n$ and $D^U_N$. These maximal risks are arising in a natural way in the context of the set-valued classification with controlled expected size. The risk $\Psi^H_{n,N}(\hat{\Gamma}; \mathcal{P})$ corresponds to the estimation of the $s$-Oracle through the Hamming distance. The second risk is directly connected with Proposition 4.4, which gives a description of the $s$-Oracle as a minimizer of $R_s(\cdot)$. The discrepancy risk reflects the ultimate goal of the considered framework: we would like to construct a procedure $\hat{\Gamma}$ that exhibits a low error $P(\hat{\Gamma})$ and low cardinal discrepancy $|s - S(\hat{\Gamma})|$.

In this paper we study the minimax rate of convergence (Tsybakov, 2008) which arises naturally in the context of set-valued classification.

**Definition 4.5** (Minimax rate of convergence). For a given family $\mathcal{P}$ of joint distributions on $\mathbb{R}^d \times [K]$ the minimax rates are defined as

$$
\Psi^\Box_{n,N}(\mathcal{P}) := \inf_{\hat{\Gamma}} \Psi^\Box_{n,N}(\hat{\Gamma}; \mathcal{P}),
$$

where $\Box$ is $H$, $E$ or $D$ and the infimum is take over set-valued classifiers based on $D^L_n$ and $D^U_N$. The sequence $\Psi^\Box_{n,0}(\mathcal{P})$ corresponds to the supervised regime, while the sequence $\Psi^\Box_{n,N}(\mathcal{P})$ for $N \geq 1$ corresponds to the semi-supervised regime.

The distinction between supervised ($N = 0$) and semi-supervised ($N > 0$) setups is practically intuitive. However, the actual theoretical frontier between these setups is a bit more complicated. Note that, as long as $\Psi^\Box_{n,N}(\mathcal{P})$ decays polynomially with the growth of $n, N$ one has $\Psi^\Box_{n,N}(\mathcal{P}) \gg \Psi^\Box_{n,N+1}(\mathcal{P})$. More generally, $\Psi^\Box_{n,O(n)}(\mathcal{P}) \gg \Psi^\Box_{n,0}(\mathcal{P})$, that is, if $N$ is $O(n)$, the impact of unlabeled data on the rate of convergence is only marginal. Consequently, we can appeal to three regimes: i) purely supervised ($N = 0$); ii) essentially supervised ($N > 0$ and $N = O(n)$); iii) purely semi-supervised ($n = o(N)$).

From the theoretical standpoint i) and ii) are not distinguishable and only in iii) there is a hope to establish faster rates of convergence. Finally, we remark that from the practical standpoint it is reasonable to think that either $N = 0$ (no unlabeled data is given); or $N \gg n$ (a lot of unlabeled data is given). While we do not explicitly assume any relations on $n$ and $N$, our results have the most natural interpretations when $N = 0$ or $N \gg n$. 


4.3. Assumptions

In this part we state all the assumptions used in this work and state the family of distributions \( P \) which drives the minimax rates. We start by adapting the already classical margin assumption to our framework.

**Assumption 4.6** \((\alpha\text{-margin assumption})\). We say that the distribution \( P \) of the pair \((X, Y) \in \mathbb{R}^d \times [K]\) satisfies \( \alpha \text{-margin assumption} \) if there exists \( C_1 > 0 \) and \( t_0 \in (0,1) \) such that for every positive \( t \leq t_0 \)

\[
P_X \left( 0 < \left| p_k(X) - G^{-1}(s) \right| \leq t \right) \leq C_1 t^\alpha .
\]

The exponent \( \alpha \) will directly influence the rates of convergence and the classification problem gets easier with the growth of this parameter. Due to the continuity Assumption 4.1 it holds that \( \lim_{t \to +0} P_X \left( 0 < \left| p_k(X) - G^{-1}(s) \right| \leq t \right) = 0 \). Thus, the \( \alpha \text{-margin} \) Assumption 4.6 specifies the rate of this convergence. Finally, the restriction of the range of \( t \) to \([0, t_0]\) in \( \alpha \text{-margin} \) Assumption 4.6 does not affect its global behavior as for all \( t \in [0,1] \) we have

\[
P_X \left( 0 < \left| p_k(X) - G^{-1}(s) \right| \leq t \right) \leq c_1 t^\alpha ,
\]

with \( c_1 = C_1 \vee t_0^{-\alpha} \).

The second assumption restricts the set of possible marginal distributions of the feature vectors. Following (Audibert and Tsybakov, 2007), we first introduce the notion of regular set. 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

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

**Definition 4.7** (Strong density). We say that the probability measure \( P_X \) on \( \mathbb{R}^d \) satisfies the \((\mu_{\text{min}}, \mu_{\text{max}}, c_0, r_0)\)-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 w.r.t. the Lebesgue measure such that \( \mu(x) = 0 \) for all \( x \in \mathbb{R}^d \setminus A \) and \( 0 < \mu_{\text{min}} \leq \mu(x) \leq \mu_{\text{max}} < \infty \), \( \forall x \in A \).

Let us mention, that there are various ways to relax this assumption. For instance, it is possible to get rid of the lower bound on the density (Audibert and Tsybakov, 2007; Kpotufe and Martinet, 2018). Besides, the compactness of the support can also be relaxed and replaced by a proper tail condition (Gadat, Klein and Marteau, 2016). This type of relaxations are not altering our conclusions about the effect of unlabeled data and thus, for simplicity, we provide the analysis under the strong density assumption.

The next assumption is standard in non-parametric statistics, and states that the conditional distribution of \( Y \) is smooth.

**Definition 4.8** (Hölder class, Tsybakov (2008)). We say that a function \( h : \mathbb{R}^d \to \mathbb{R} \) is \((\beta, L)\)-Hölder for \( \beta > 0 \) and \( L > 0 \) if \( h \) is \( \lfloor \beta \rfloor \) times continuously differentiable and \( \forall x, x' \in \mathbb{R}^d \) we have

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

Finally, we are in position to define the family of distributions \( \mathcal{P} \) that governs the rates of convergence.

**Definition 4.9.** We denote by \( \mathcal{P}(L, \beta, \alpha) \) the set of joint distributions on \( \mathbb{R}^d \times [K] \) which satisfy the following conditions

- the marginal \( \mathbb{P}_X \) satisfies the \((\mu_{\text{min}}, \mu_{\text{max}}, c_0, r_0)\)-strong density,
- for all \( k \in [K] \) the \( k \)th regression function \( p_k(\cdot) = \mathbb{P}(Y = k|X = \cdot) \) belongs to the \((\beta, L, \mathbb{R}^d)\)-Hölder class, that is, \( p_k \in \mathcal{H}(\beta, L, \mathbb{R}^d) \) for all \( k \in [K] \),
- for all \( k \in [K] \) the regression function \( p_k \) satisfy the \((C_1, t_0, \alpha)\)-Margin assumption,
- for all \( k \in [K] \), the cumulative distribution function \( F_{p_k} \) of \( p_k(X) \) is continuous.

The family of distributions \( \mathcal{P}(L, \beta, \alpha) \) resembles the one considered in (Audibert and Tsybakov, 2007) in the context of binary classification. The major difference is the continuity Assumption 4.1, which poses certain restrictions and does not allow to re-use in a straightforward way their construction for lower bounds.

5. Lower bounds

In this section we establish minimax lower bounds on the introduced risk measures. Our rates highlight the benefit of the semi-supervised approaches in the context of the set-valued classification with controlled expected size.

**Theorem 5.1.** Let \( K \geq 3, s \in [\lfloor K/2 \rfloor - 1] \). If \( 2\alpha \left[ \frac{d}{2} \right] \leq d \), then there exist constants \( c, c', c'' > 0 \) such that for all \( n, N \in \mathbb{N} \) it holds that

\[
\Psi_{n,N}^H(\mathcal{P}(L, \beta, \alpha)) \geq c \left( n^{-\frac{\alpha}{2\beta+\delta}} \sqrt{(n+N)^{-1/2}} \right) \quad \text{(Hamming risk)}
\]

\[
\Psi_{n,N}^E(\mathcal{P}(L, \beta, \alpha)) \geq c' \left( n^{-\frac{(1+\alpha)d}{2\beta+\delta}} \sqrt{(n+N)^{-1/2}} \right) \quad \text{(Excess risk)}
\]

\[
\Psi_{n,N}^D(\mathcal{P}(L, \beta, \alpha)) \geq c'' \left( n^{-\frac{(1+\alpha)d}{2\beta+\delta}} \sqrt{(n+N)^{-1/2}} \right) \quad \text{(Discrepancy)}
\]

First of all, based on these results, we observe that the lower bound for the Hamming risk \( \Psi_{n,N}^H \) is slower than those for the other risks, which is explained by the structure of the Hamming risk. We remark that the second part of the rate in all three cases is \((n+N)^{-1/2}\) instead of \(N^{-1/2}\). Actually, if \( n \gg N \), from purely minimax perspective it is impossible to obtain a lower bound with \(N^{-1/2}\) instead of \((n+N)^{-1/2}\). Indeed, one can always split the labeled sample erasing labels from one of them. Such splitting artificially augments the size of unlabeled sample \( N \) by some fraction of \( n \). Of course, the regime \( n \gg N \) is not particularly interesting, since, despite the fact that \( N \neq 0 \), still...
corresponds to the essentially supervised setup. Secondly, the above lower bounds imply that the best rate in the purely or essentially supervised regimes is $n^{-1/2}$ across all the risk. Therefore, even if the margin assumption is very strong ($\alpha \gg 1$) supervised methods ($N = 0$) cannot achieve fast rates. This fact is the major difference with classical setups where the value of threshold is known (such as classification and level set estimation). Indeed, under the same assumptions on the family of distributions, without the continuity Assumption 4.1, the minimax rate in those frameworks is $n^{-(1+\alpha)\beta/(2\beta+d)}$ as proved for instance in (Audibert and Tsybakov, 2007; Rigollet and Vert, 2009) and unlabeled data cannot improve it. In contrast, this limitation can be neglected in the purely semi-supervised regime ($N \gg n$). Indeed, for sufficiently large unlabeled sample, the dominant term in the lower bound is of order $n^{-(1+\alpha)\beta/(2\beta+d)}$, which can be faster than $n^{-1/2}$. To be more precise, when we consider $\Psi_{n,N}^E$ or $\Psi_{n,N}^D$ the following relations are necessary to get fast rates of convergence

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

The condition on the left hand side ensures that we have enough unlabeled data to eliminate the impact of not knowing threshold $G^{-1}(s)$ in Eq. (4.1). Whereas, the condition on the right hand side ensures that the classification problem with “known” threshold admits fast rates. The above discussion suggests that the lack of knowledge of the threshold $G^{-1}(s)$ is significant, and the considered framework is more difficult from the statistical perspective than its more classical counterparts.

Final remarks we could make before going further concerns the assumption on the parameters $\alpha$, $\beta$, and $s$. The condition $2\alpha^\left\lfloor \frac{\beta}{2} \right\rfloor \leq d$ in the lower bounds is slightly more restrictive than the conditions given in (Audibert and Tsybakov, 2007) (they have $\alpha\beta \leq d$). 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 4.1. In contrast, the construction of Rigollet and Vert (2009) satisfies $^3$ Assumption 4.1 but their lower bound is limited by the condition $\alpha\beta \leq 1$, that is, it does not cover the fast rates as long as the dimension $d > 2$. Moreover, in Theorem 5.1, it is assumed that $s$ is an integer bounded from above by $\lfloor K/2 \rfloor − 1$. The assumption that $s$ is integer is made to simplify the proofs and can be relax by a straightforward, but heavy, modification of our construction. Meanwhile, the assumption that $s \leq \lfloor K/2 \rfloor − 1$ can be only relaxed to $s \leq \lfloor \alpha K \rfloor − 1$ with $\alpha < (K−1)/K$. The choice of $\alpha = 1/2$ is made from the consideration that in most of the practical situations it is unrealistic to ask for more than a half candidates in average.

### 5.1. Proof sketch

Before sketching the proof of the lower bounds, let us first display connections between the different minimax rates. These links are used to prove Theorem 5.1.

$^3$Modified properly to fit the classification framework.
Set-valued classification

Proposition 5.2. Let \( \Gamma \) be a measurable function from \( \mathbb{R}^d \) to \( 2^{[K]} \), \( s \in (0,K) \) and assume that Assumption 4.1 is fulfilled, then

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

Furthermore, if additionally Assumption 4.6 is satisfied with \( \alpha > 0 \), then there exist \( C > 0 \) which depends only on \( K, \alpha, C_1 \) such that for any pair of set-valued classifiers \( \Gamma, \Gamma' \) it holds that

\[
\mathbb{E}_{\mathbb{P}_X} \left| \Gamma(X) \Delta \Gamma'(X) \right| \leq C \left( R_s(\Gamma) - R_s(\Gamma') \right)^{\alpha/(\alpha+1)}.
\]  

Proposition 5.3. For any \( K \geq 2, s \in [K] \) and \( n,N \in \mathbb{N} \) the following relation between minimax rates holds:

\[
\Psi_{n,N}^H(\mathbb{P}) \geq \Psi_{n,N}^D(\mathbb{P}) \geq \Psi_{n,N}^E(\mathbb{P}).
\]

Proposition 5.2, and in particular Eq. (5.1) gives an easy way to establish a lower bound on the excess risk \( \Psi_{n,N}^E(\mathbb{P}) \) via a lower bound on the Hamming distance \( \Psi_{n,N}^H(\mathbb{P}) \). However, this approach does not allow to get \( (N+n)^{-1/2} \) part of the rate in the lower bound for \( \Psi_{n,N}^E(\mathbb{P}) \) and other methods should be exploited. Moreover, Proposition 5.3 allows to prove a lower bound on the discrepancy \( \Psi_{n,N}^D(\mathbb{P}) \) with the correct rate via the lower bound on the excess risk \( \Psi_{n,N}^E(\mathbb{P}) \).

Finally, let us briefly explain the main ideas behind the proof of the lower bounds in Theorem 5.1. In order to get the correct rate, we separately prove two lower bounds on the minimax rates. 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 \( G^{-1}(s) \). Therefore, the first lower bound is independent of the size of unlabeled data \( N \) and the second lower bound is of order \( (n+N)^{-1/2} \). To be more specific, the first lower bound is closely related to the one provided by Audibert and Tsybakov (2007); Rigollet and Vert (2009), however, the continuity Assumption 4.1 makes the proof more involved and results in a final construction of hypotheses that differs significantly. This part of our lower bound relies on Fano’s inequality in the form provided by Birgè (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, while the regression function \( p \) remains the same. Crucially, these marginal distributions admit two different values of threshold \( G^{-1}(s) \) and thus two different s-Oracle classifiers.

6. Upper bounds

In this section, we show that we can build a set-valued classifier that achieves, up to a logarithmic factor, the lower bounds stated in Theorem 5.1. In other words, our estimator is nearly optimal in the minimax sense. To come straight to the point, we delay the construction of the estimator to Section 6.1 and focus right now on the upper bounds.
Theorem 6.1. Let $K \in \mathbb{N}$, $s \in (0, K)$, then there exists an estimator $\hat{\Gamma}$ and constants $C, C', C'' > 0$ such that for all $n, N \in \mathbb{N}$ we have

$$
\Psi_{n,N}^H(\hat{\Gamma}; \mathcal{P}(L, \beta, \alpha)) \leq C \left( n^{-\frac{s}{2s+2}} \sqrt{(n + N)^{-1/2}} \right) \quad \text{(Hamming risk)},
$$

$$
\Psi_{n,N}^E(\hat{\Gamma}; \mathcal{P}(L, \beta, \alpha)) \leq C' \left( \frac{n}{\log n} \right)^{-\frac{(1+\alpha)\beta}{2s+2}} \sqrt{(n + N)^{-1/2}} \quad \text{(Excess risk)},
$$

$$
\Psi_{n,N}^D(\hat{\Gamma}; \mathcal{P}(L, \beta, \alpha)) \leq C'' \left( \frac{n}{\log n} \right)^{-\frac{(1+\alpha)\beta}{2s+2}} \sqrt{(n + N)^{-1/2}} \quad \text{(Discrepancy)}.
$$

An immediate conclusion from the above result is that the lower bounds of Theorems 5.1 are achievable. In particular, in the case of Hamming risk, the rates are optimal; whereas for the Excess risk and the Discrepancy, the upper bounds match the lower bounds up to a logarithmic factor. Thus, the necessary conditions for the fast rates of semi-supervised methods are also sufficient.

Let us finally mention that the presence of the logarithmic factor in these upper bounds is due to $\ell_\infty$-norm estimation and will be discussed in details later (see Lemma 6.7). It is also instructive to consider the case $N = +\infty$, which formally corresponds to classification with known marginal distribution $\mathbb{P}_X$. In this case, the rate of convergence depends only on the size of the labeled sample and the obtained rates resemble classical results in classification and level set estimation (Audibert and Tsybakov, 2007; Rigollet and Vert, 2009). Intuitively, when $N = +\infty$ the set of set-valued classifiers which satisfy the expected size constraint is known and we do not need to pay a price for its estimation.

6.1. Construction of the estimator

The construction of $\hat{\Gamma}$ that reaches the rates in the former upper bounds involves a preliminary estimators $\hat{p}_k$ of the regression functions $p_k$, $k \in [K]$. We recall that we have two datasets at hand, labeled one $D^L_n$ and unlabeled one $D^U_N$. Depending on the relation between $N$ and $n$ there are two theoretically different regimes: i) the purely or essentially supervised regime when $N = 0$ or $N = O(n)$. In this case we split the labeled sample – one sub-sample is used for estimating the conditional probabilities and the other sub-sample (along with the unlabeled data) is used to estimate $G^{-1}(s)$. In this case, splitting is essential from both theoretical and practical points of views (see Figure 2); ii) the purely semi-supervised regime when $N \gg n$. In contrast, here splitting the labeled sample is useless. One can use the whole labeled sample for the estimation of the conditional probabilities and only the unlabeled sample to estimate the threshold. However, in order to unify our statements and avoid introduction of two estimators depending on $n$ and $N$ we consider the strategy that always splits the labeled dataset $D^L_n$. Importantly, this splitting does not inflate the resulting rate of convergence and only affects leading constants.
Let us point out that Assumption 6.2 induces that there exists a constant $C > 0$ such that for all $n \geq 2$ and all $\alpha > 0$

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

Assumption 6.2 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) with $\delta_0 = 0$ and for a choice of the bandwidth parameter equal to $n^{-1/(2\beta + d)}$, see (Audibert and Tsybakov, 2007, Theorem 3.2) and (Lei, 2014, Proposition 1). At the same time, Assumption 6.3 can always be satisfied by slightly processing any estimator $\hat{p}$. Indeed, assume Assumption 6.3 fails to be satisfied by some estimator $\hat{p}$. It means that there exists a measurable 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 to $\hat{p}$ such regions can no longer exist. To make the above argument formal, we state the following result whose proof can be found in Appendix B.

**Lemma 6.4.** Assume that $\tilde{p} \in \mathcal{P}([n/2])$, $\bar{p}$ is a measurable function of $\mathcal{D}_{[n/2]}$, $\max_{k \in [K]} \|\bar{p}_k - \tilde{p}_k\|_{\infty, \mathbb{P}_X} \leq n^{-\beta/(2\beta + d)}$ almost surely w.r.t. the distribution of $\mathcal{D}_{[n/2]}$, $\tilde{p}$ satisfies Assumption 6.2 with $\delta_0 = 1$, $\tilde{p}$ satisfies Assumption 6.3.

4One can take any constant fraction of $n$. 

---

**Assumption 6.2 (Exponential concentration).** There exist estimators $\hat{p}_k$ for all $k \in [K]$ based on $\mathcal{D}_{[n/2]}$ and positive constants $C_1, C_2$ and $\delta_0 \geq 0$ such that for all $k \in [K]$ and all $n \geq 2$ we have for all $\delta > \delta_0 n^{-\beta/(2\beta + d)}$

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

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

**Assumption 6.3 (Continuity of CDF).** For all $k \in [K]$, conditionally on the data, 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 $\mathbb{P}^{\otimes [n/2]}$ continuous on $(0, 1)$.
Considering the above, neither Assumption 6.2 nor Assumption 6.3 are restrictive, as both of them are satisfied by either classical estimators or by their slightly modified versions.

**Remark 6.5.** The proof of Lemma 6.4 is constructive – it introduces a deterministic perturbation to \( \bar{p} \) without damaging its statistical properties. However, the construction of this lemma is impractical – it is not computationally tractable. To alleviate this issue, a simpler strategy can be proposed. Instead of the deterministic modification, one can make a randomized one. That is, given estimators \( \bar{p}_1, \ldots, \bar{p}_K \), we add a small bounded continuous noise \( \xi_1, \ldots, \xi_K \) independent from \( X \) and the training sample which yields \( \hat{p}_k(X) = \bar{p}_k(X) + \xi_k \). As a consequence, the resulting set-valued classifier is randomized and \( t \mapsto \mathbb{P}_{X,\xi_k}(\hat{p}_k(X) \leq t) \) is continuous almost surely. We illustrate this strategy in the numerical Section 7 as it is more practical.

Recall that according to Proposition 4.2 the optimal set-valued classifier \( \Gamma^*_s \) involves the threshold \( G^{-1}(s) \) in its expression. Moreover, \( G^{-1}(s) \) is distribution dependent, and thus ought to be estimated using data. A straightforward estimation procedure can be constructed using the unlabeled dataset \( \mathcal{D}^U_N \). To make our presentation mathematically correct we introduce the following notation \( \mathcal{D}^L_n = \mathcal{D}^U_n \cup \mathcal{D}_{[n/2]} \), where \( \mathcal{D}_{[n/2]} \) is the dataset used to build the estimators \( \hat{p}_k \) for \( k \in [K] \). Now, all the labels are removed from \( \mathcal{D}_{[n/2]} \). That is, \( \mathcal{D}_{[n/2]} \) consists of \( [n/2] \) i.i.d. samples from \( \mathbb{P}_X \). Consequently, the semi-supervised estimator of \( G(\cdot) \) is defined as

\[
\hat{G}(\cdot) = \frac{1}{[n/2] + N} \sum_{X \in \mathcal{D}^L_n} \sum_{k=1}^K \mathbf{1}_{\{\hat{p}_k(X) > \cdot\}} 
\]

Finally, the set-valued classification procedure \( \hat{\Gamma} \) is defined point-wise as

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

An important property of the introduced estimator \( \hat{\Gamma} \) is obtained via Assumption 6.3, which describes the deviation of the size of \( \hat{\Gamma} \) from the desired level \( s \).

**Proposition 6.6.** Let \( \hat{p}_k \) for all \( k \in [K] \) be arbitrary estimators of the regression functions constructed using \( \mathcal{D}_{[n/2]} \) that satisfies Assumption 6.3 and \( \mathbb{P}_X \) admits a density, then there exists a positive constant \( C \) such that for all \( n, N \in \mathbb{N} \) it holds that

\[
\mathbb{E}_{(\mathcal{D}^L_n, \mathcal{D}^U_N)} \left| \mathbf{S}(\hat{\Gamma}) - s \right| \leq C(N + n)^{-1/2}.
\]

The obtained rate is parametric and does not depend on the quality of \( \hat{p}_k \) as it holds for any estimator which satisfies Assumption 6.3. In addition, the condition of Proposition 6.6

\[ 5 \text{We split the data to ensure that the unlabeled sample is larger than the labeled one. From practical perspective this splitting is unnecessary as long as } N \geq n. \]
only asks for the density of $P_X$, which makes the construction of set-valued classifier $\hat{\Gamma}$ practically relevant even in the absence of non-parametric assumptions in consideration. Assumption 6.3 is crucial for this theoretical result and, as it is illustrated in Section 7, it is also crucial in practice. Moreover, the size of labeled data $n$ enters this bound only because we split data $\mathcal{D}_n^L = \mathcal{D}_{[n/2]} \cup \mathcal{D}_{[n/2]}$. Without this step the rate would be $O(N^{-1/2})$, which is of the same order as long as $N \geq n$, that is, we are in the purely semi-supervised regime.

6.2. Proof sketch of Theorem 6.1

To show that the estimators introduced in this section satisfy the statement of Theorem 6.1 we significantly refine the proof technique used in (Denis and Hebiri, 2017), since their approach gives only suboptimal rates. That is, we introduce an intermediate quantity $\tilde{G}(\cdot) := \sum_{k=1}^K P_X(\hat{p}_k(X) > \cdot)$, and the associated set-valued classifier, which we refer to as the pseudo Oracle classifier given for all $x \in \mathbb{R}^d$ by

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

The classifier $\tilde{\Gamma}$ assumes knowledge of the marginal distribution $P_X$, it is seen as an idealized version of $\hat{\Gamma}$, and formally corresponds to the case $N = +\infty$. This is why the pseudo Oracle $\tilde{\Gamma}$ is not an estimator and only serves as an intermediate quantity in the proof. Besides, thanks to Assumption 6.3, this pseudo Oracle satisfies the size constraints, that is $S(\tilde{\Gamma}) = s$ almost surely. This simple fact is the first step in the proof of Proposition 6.6.

A crucial step of our analysis is the following lemma, that bounds the difference between $\tilde{G}^{-1}(s)$ and $G^{-1}(s)$ in terms of the difference between $\hat{p}_k$’s and $p_k$’s.

**Lemma 6.7** (Upper bound on the thresholds). Let Assumption 4.1 be satisfied, then for all $s \in (0, K)$

$$|G^{-1}(s) - \tilde{G}^{-1}(s)| \leq \|p - \hat{p}\|_{\infty, P_X}, \quad \text{almost surely} \quad P^\otimes n \otimes P_X^\otimes N.$$ 

The proof of Lemma 6.7 uses elementary properties of the generalized inverse functions which are provided in Appendix. Besides, let us mention, that the difference $|G^{-1}(s) - \tilde{G}^{-1}(s)|$ resembles the Wasserstein infinity distance which gives an alternative approach to prove Lemma 6.7, see (Bobkov and Ledoux, 2016). Lemma 6.7 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).

It is intuitively clear that if, on top of Lemma 6.7, we manage to control the difference $|\tilde{G}^{-1}(s) - G^{-1}(s)|$ then the proof of the upper bound would simply follow the arguments of Audibert and Tsybakov (2007). Yet, such a control is not feasible under our assumptions. To see this, notice that conditionally on $\mathcal{D}_n$ the quantity $|\tilde{G}^{-1}(s) - G^{-1}(s)|$
resembles the deviation of quantile from its empirical version. However, classical result\textsuperscript{6} on asymptotic normality of sample quantiles (Ma and Robinson, 1998, Theorem 2) tells that in order to have a central limit theorem with \((n+N)^{-1/2}\) rate it is necessary and sufficient to require \(\tilde{G}'(\tilde{G}^{-1}(s)) > 0\). From the minimax perspective, this condition cannot be satisfied since we do not require any lower bound on the derivative of \(G(\cdot)\).

To alleviate this drawback, we construct the estimator \(\hat{p}\) which satisfies continuity Assumption 6.3. This construction allows to avoid the direct control of \(|\tilde{G}^{-1}(s) - \hat{G}^{-1}(s)|\) and leverage standard properties of generalized inverse of a continuous and monotone function. In Section 6.3 we demonstrate that the upper bound can be improved if we assume that the derivative of \(G(\cdot)\) is uniformly lower bounded, that is, \(G^{-1}(\cdot)\) is Lipschitz continuous.

6.3. Upper bound under extra assumptions

In previous sections we developed minimax analysis of set-valued approach for multi-class classification problem. Under standard non-parametric assumptions we have shown that no fully supervised estimator can achieve fast rates and that this issue can be alleviated by augmenting the data with unlabeled samples. In this section our goal is to understand whether the above phenomena occurs under even stronger assumptions on the joint distribution of \((X,Y)\). As already pointed out, the rate of convergence of \(\hat{\Gamma}\) can be improved if we consider those distributions from \(\mathcal{P}(L,\beta,\alpha)\) for which the function \(G^{-1}(\cdot)\) is Lipschitz.

Assumption 6.8. There exists a constant \(M > 0\) such that for all \(t,t' \in [0,1]\) it holds that \(|G^{-1}(t) - G^{-1}(t')| \leq M |t-t'|\).

From now on we consider the family of joint distributions \(\mathcal{P}(L,\beta,\alpha,M)\) of \((X,Y)\) which satisfies all assumptions in Definition 4.9 and satisfies Assumption 6.8 with some constant \(M > 0\). Let us point out that the family \(\mathcal{P}(L,\beta,\alpha,M)\) is much smaller than \(\mathcal{P}(L,\beta,\alpha)\). Indeed, \(\mathcal{P}(L,\beta,\alpha,M)\) excludes all those \(G\) functions which are locally constants. Moreover, as \(M\) is assumed to be independent of the sample size, there are no \(G\) functions whose slope is locally decreasing with the data. It actually means that if we take all functions \(G\) which are generated by the distributions from \(\mathcal{P}(L,\beta,\alpha,M)\) then these functions are not dense (in infinity norm) in the set of all functions \(G\) generated by the distributions from \(\mathcal{P}(L,\beta,\alpha)\). Thus, the family \(\mathcal{P}(L,\beta,\alpha,M)\) is considerably smaller than the one without the Lipschitz assumption \(\mathcal{P}(L,\beta,\alpha)\).

The main statement in this section is the following uniform over \(\mathcal{P}(L,\beta,\alpha,M)\) upper bound on the excess risk of the estimator constructed in Section 6.1.

\textsuperscript{6}We can arrive to a similar conclusion from (Bobkov and Ledoux, 2016, Theorem 5.11)
Theorem 6.9. Let $K \in \mathbb{N}, s \in (0, K)$, then there exists a semi-supervised estimator $\hat{\Gamma}$ and a constant $C > 0$ such that for all $n, N \in \mathbb{N}$ we have

$$\Psi_{n,N}(\hat{\Gamma}; \mathcal{P}(L, \beta, \alpha, M)) \leq C \left( \frac{n}{\log n} \right)^{-\frac{1+\alpha}{2}},$$

Let us discuss this result before presenting its proof. First of all notice that we have an obvious relation $\mathcal{P}(L, \beta, \alpha, M) \subset \mathcal{P}(L, \beta, \alpha)$ for all $M > 0$, therefore it is expected that the upper bound in this case can only be faster than its counterpart in Theorem 6.1. Moreover, we would like to emphasize that the upper bound in Theorem 6.9 is independent from the size of the unlabeled dataset $N$. It demonstrates that in the family $\mathcal{P}(L, \beta, \alpha)$ there is a subfamily with additional restrictive assumptions which neglects the effect of unlabeled sample.

In order to leverage the Lipschitzness, we rely on the results from (Bobkov and Ledoux, 2016) on the Wasserstein infinity distance between empirical and real distributions. Let us first define the Wasserstein infinity distance between two probability measures.

Definition 6.10. Let $\mu$ and $\nu$ be two Borel probability measures on $[0, 1]$. Denote by $F_\mu$ and $F_\nu$ the CDFs of $\mu$ and $\nu$ respectively. Then, the Wasserstein infinity distance is defined as

$$W_\infty(\mu, \nu) = \sup_{0 < \theta < 1} \left| F_\mu^{-1}(\theta) - F_\nu^{-1}(\theta) \right|,$$

where $F_\mu^{-1}$ and $F_\nu^{-1}$ are the generalized inverses of $F_\mu$ and $F_\nu$ respectively.

The next result is not present in (Bobkov and Ledoux, 2016), yet can be easily recovered from their Theorem 5.11.

Theorem 6.11. Let $Z, Z_1, \ldots, Z_n$ be i.i.d. $[0, 1]$-valued random variables distributed according to $\mu$. Denote by $\hat{\mu}_n$ the empirical distribution of $Z_1, \ldots, Z_n$, then for all $\alpha \geq 0$ there exists $c > 0$ such that for all $n \in \mathbb{N}$ it holds that

$$\mathbb{E} W_\infty^{1+\alpha}(\mu, \hat{\mu}_n) \leq c \left( \left\| F_\mu^{-1} \right\|_{\text{Lip}} \right)^{1+\alpha} n^{-\frac{1+\alpha}{2}},$$

where $\left\| F_\mu^{-1} \right\|_{\text{Lip}}$ is the Lipschitz constant of $F_\mu^{-1}$.

For the sake of completeness we provide the proof of this result in Appendix F. In fact, the only difference between Theorem 6.11 and (Bobkov and Ledoux, 2016, Theorem 5.11) is the control of moments of the Wasserstein infinity distance. That is, the result of (Bobkov and Ledoux, 2016, Theorem 5.11) is stated with $\alpha = 0$ but for our purposes we need to control all moments starting from the first one. Theorem 6.11 provides a way to control $|G^{-1}(s) - G^{-1}_{n,N}(s)|$, where $G^{-1}_{n,N}(s)$ is the generalized inverse of

$$G_{n,N}(\cdot) = \frac{1}{\lceil n/2 \rceil + N} \sum_{X \in \mathcal{D}_N \cup \mathcal{D}_{\lceil n/2 \rceil}} \sum_{k=1}^{K} 1_{\{p_k(X)>\cdot\}}.$$
Unlike \( \hat{G} \), which is used to construct our set-valued classifier, the function \( G_{n,N} \) is based on the real function \( p_k \)'s and not their estimates. In order to clarify the relation of Theorem 6.11 with the difference \( |G^{-1}(s) - G_{n,N}^{-1}(s)| \) we require the following lemma.

**Lemma 6.12.** Let \( Z = \sum_{k=1}^K p_k(X)1_{(\epsilon=\epsilon_k)} \), where \( \epsilon \) is distributed uniformly on \([K]\) and is independent from \( X \). Let \( F \) be the cumulative distribution function of \( Z \), then

\[
F(\cdot) = 1 - \frac{1}{K} G(\cdot), \quad G^{-1}(\cdot) = F^{-1}\left(1 - \frac{\cdot}{K}\right), \quad \|G^{-1}\|_{\text{Lip}} = \frac{\|F^{-1}\|_{\text{Lip}}}{K}.
\]

Notice that we cannot yet apply Theorem 6.11 to our problem, since \( 1 - G_{n,N}/K \) is not an empirical distribution of \( Z_1, \ldots, Z_{N+[n/2]} \). This issue can be alleviated thanks to the convexity of the Wasserstein distances.

**Theorem 6.13.** Under Assumption 6.8 there exists \( C > 0 \) that depends only on \( \alpha \) such that

\[
\mathbb{E}_{(\mathbb{D}^n, \mathbb{P}^n)}[|G^{-1}(s) - G_{n,N}^{-1}(s)|^{1+\alpha}] \leq C (KM)^{1+\alpha} (n + N)^{-\frac{\alpha+\alpha}{2}}.
\]

**Proof.** We first notice that the distribution of \( Z \) from Lemma 6.12 is a pushforward measure. Indeed, consider \( \mathbb{R}^d \times [K] \) the product space equipped with its Borel sigma algebra and measure \( \nu = \mathbb{P}_X \otimes U([K]) \), where \( U([K]) \) is the uniform distribution on \([K]\). Let \( f : \mathbb{R}^d \times [K] \to [0,1] \) be defined for all \((x,\epsilon) \in \mathbb{R}^d \times [K] \) as \( f(x,\epsilon) := p_\epsilon(x) \), then \( Z \) is distributed according to \( \mu = f\#\nu \), where \( f\# \) denotes the pushforward of measure \( \cdot \) by \( f \), and \( G \) is the CDF of \( \mu \). It is not difficult to see that \( G_{n,N} \) is a CDF of \( \tilde{\mu} \), where \( \tilde{\mu} = f\#\tilde{\nu} \) with \( \tilde{\nu} = \tilde{\mathbb{P}}_X \otimes U([K]) \). Let \( \epsilon_1, \ldots, \epsilon_{N+[n/2]} \) be i.i.d. realizations of \( \epsilon \), and \( \tilde{\mu}_{N+[n/2]} \) be the empirical measure of \( Z_i = p_{\epsilon_i}(X_i) \) for each \( i \in [N+[n/2]] \), then \( \tilde{\mu} = \mathbb{E}(\epsilon_1, \ldots, \epsilon_{N+[n/2]})[\tilde{\mu}_{N+[n/2]}] \), where \( \mathbb{E}(\epsilon_1, \ldots, \epsilon_{N+[n/2]}) \) denotes the expectation taken w.r.t. \( \epsilon_1, \ldots, \epsilon_{N+[n/2]} \). Using convexity (we refer the reader to Bobkov and Ledoux (2016) for more details) of Wasserstein distance and of \( x \mapsto x^{1+\alpha} \) for \( x \geq 0 \) we can write for all \( s \)

\[
|G^{-1}(s) - G_{n,N}^{-1}(s)|^{1+\alpha} \leq \mathcal{W}_\infty^{1+\alpha}(\mu, \tilde{\mu}) = \mathcal{W}_\infty^{1+\alpha}(\mu, \mathbb{E}(\epsilon_1, \ldots, \epsilon_{N+[n/2]})[\tilde{\mu}_{N+[n/2]}])
\leq \mathbb{E}(\epsilon_1, \ldots, \epsilon_{N+[n/2]})[\mathcal{W}_\infty^{1+\alpha}(\mu, \tilde{\mu}_{N+[n/2]})],
\]

where for the first inequality we used Lemma 6.12. We conclude by applying Theorem 6.11 of Bobkov and Ledoux (2016) to the expectation of \( \mathcal{W}_\infty(\mu, \tilde{\mu}_{N+[n/2]}) \) and Lemma 6.12 to determine the Lipschitz constant of \( F_{\mu}^{-1} \).

The proof Theorem 6.9 combines Theorem 6.13 and the argument of Audibert and Tsybakov (2007). Complete proof can be found in Appendix F. Let us point out the difference between proofs of Theorems 6.1 and 6.9. The proof of Theorem 6.1 introduces the pseudo-oracle \( \hat{\Gamma} \) and partly relies on the continuity Assumption 6.3 posed on the estimator \( \hat{p} \). In contrast, the proof of Theorem 6.9 does not require this assumption, yet, this bound is valid for a much smaller family of distributions, since the function \( G^{-1} \) is
Set-valued classification

assumed to be Lipschitz. Instead of the pseudo-oracle \( \tilde{\Gamma} \), which “knows” the marginal distribution of the features, the proof of Theorem 6.9 is based on \( G_{n,N} \), introduced above, which can be seen as a pseudo-oracle that “knows” the conditional distribution of the labels instead. The reason of such a discrepancy is dictated by the difficulty of estimating the conditional distribution of \( Y \) given \( X \). Indeed, on \( P(L,\beta,\alpha) \) the dominant term of the upper bound is connected with the estimation of the marginal distribution of the features. Meanwhile, on \( P(L,\beta,\alpha,M) \) the dominant term is connected with the estimation of the conditional distribution of labels. We do not provide the minimax lower bound on the family \( P(L,\beta,\alpha,M) \), however, it can be recovered from the proof of Theorem 5.1 (only the first part of the proof) by straightforward but cumbersome modification.

Lastly, let us provide a simple intuition describing the role of the Lipschitzness condition on \( G^{-1} \). This type of conditions is well known in the analysis of order statistics and sample quantiles. For example, consider a random variable \( X \sim \frac{1}{2} U([-2,-1]) + \frac{1}{2} U([1,2]) \) and denote by \( F : \mathbb{R} \to [0,1] \) the CDF of \( X \) and let \( F_n \) be the empirical CDF of \( (X_1,\ldots,X_n) \) i.i.d. copies of \( X \). Our goal is to understand the statistical properties of \( F_n^{-1}(1/2) = X(\frac{1}{2}) \) namely, how to control its deviation from \( F^{-1}(1/2) = -1 \). However, the quantile function of \( X \) is not Lipschitz around \( 1/2 \), thus this control is impossible. Indeed, notice that if there are at least \( \frac{n}{2} + 1 \) realizations of \( (X_1,\ldots,X_n) \) that end up in \([1,2]\), then we have \( |F_n^{-1}(1/2) - F^{-1}(1/2)| \geq 1/2 \). Therefore, we have \( \mathbb{P}(|F_n^{-1}(1/2) - F^{-1}(1/2)| \geq 1/2) \geq \mathbb{P}(V > \frac{n}{2}) \longrightarrow 1/2 \), where \( V \) is the binomial random variable with parameters \((n,1/2)\).

7. Numerical illustrations

In this section we empirically support our theoretical results. One of our primary objectives is to provide experimental evidences that highlight the importance of the assumptions involved in our framework. In particular, we focus on Assumption 6.3 that is required from \( \hat{p}_1,\ldots,\hat{p}_K \) and moreover, in case \( N = 0 \), we highlight the importance of data splitting for construction of \( \hat{p}_1,\ldots,\hat{p}_K \) and estimation of \( G \). To this end, we compare the performance of the set-valued classifier \( \hat{\Gamma} \) against several natural alternatives:

a) classifier that violates the continuity Assumption 6.3: exploits an estimator \( \hat{p}_k \) of the regression functions \( p_k \) so that Assumption 6.3 is violated;

b) classifier that violates the independence between samples \( D^L_N \) and \( D^U_N \): uses the \( X_i \)’s used for training the regression functions \( \hat{p}_k \) to also estimate \( G \) function.

We conduct all the experiments using the MNIST dataset (LeCun and Cortes, 2010) which is composed of images of handwritten digits from 0 to 9. The goal is to predict which digit is present on the image.

Recall the definition (6.1), of the proposed set-valued classifier \( \hat{\Gamma} \). This predictor relies on a base estimator \( \hat{p}_k \) of \( p_k(X) = \mathbb{P}(Y = k \mid X) \) and, generally, can be applied on top of any such estimator. For the base estimator \( \hat{p}_k \) we select the random forest (Breiman, 2001).

\footnote{For simplicity \( n \) is assumed to be even.}
Chzhen, Denis and Hebiri

provided by the scikit-learn python library (Pedregosa et al., 2011). The random forest is chosen for the following reasons: i) certain theoretical studies consider the random forest (or its modifications) in non-parametric setup (see e.g., (Arlot and Genuer, 2014; Mourtada, Gaïffas and Scornet, 2020) and references therein) ii) it is known to achieve nearly state-of-the-art accuracy in practice (Denil, Matheson and De Freitas, 2014) iii) it does not satisfy the continuity Assumption 6.3 (it provides constant predictions on non-empty parts of the feature space), which is necessary for our guarantees. In view of Remark 6.5, we randomize the set-valued classifier by adding a small continuous bounded noise to the predicted regression function when we want Assumption 6.3 to be satisfied.

Experimental setup. For our experiments we collect three independent sets: a labeled one $D^n_l = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ of size $n$, an unlabeled one $D^n_u = \{X_{n+1}, \ldots, X_{n+N}\}$ of size $N$, and a test set $T_M = \{(X_{n+N+1}, Y_{n+N+1}), \ldots, (X_{n+N+M}, Y_{n+N+M})\}$ of size $M$. These datasets are sampled randomly from the whole MNIST dataset without replacement. We fix $n = 1000$, $M = 1000$, and $N \in \{0, 10, 20, \ldots, 290, 300\}$.

Let $\hat{p}_k(\cdot; D)$, $k \in [K]$ be the output of the random forest classifier trained on some labeled data $D$. To build the set-valued classifiers from a) and b) in the beginning of this section, we define four estimators of the function $G$ using $\hat{p}_k$ – two for each point.

For a) We want to gather empirical evidences for the importance of the continuity Assumption 6.3. The set-valued classifier is built in a semi-supervised manner and we use the data in $D^n_u$ for the estimation of $G$ function.

- **Continuous:** Since random forest is used to estimate the $p_k$’s, Assumption 6.3 does not hold. We add the smoothing random variables $\xi_{ik}$’s so that Assumption 6.3 holds. We define

$$\hat{G}^{+c}(t) = \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} 1\{\hat{p}_k(X_{n+i}; D^n_l) + \xi_{ik} > t\};$$

- **Not Continuous:** Since random forest is used to estimate the $p_k$’s, Assumption 6.3 does not hold. We define

$$\hat{G}^{-c}(t) = \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} 1\{\hat{p}_k(X_{n+i}; D^n_l) > t\};$$

For b) The conclusion from the point a) will clearly demonstrate that continuity is crucial. We consider here two set-valued classifiers that satisfy Assumption 6.3 and analyze whether data splitting degrades performance.

- **With splitting:** We split the data to build the set-valued classifier. We use $2n/3$ samples from $D^n_l$ to train the $\hat{p}_k$’s (that is, $D = (X_{i'}, Y_{i'})_{i' = n/3 + 1}^n$) and

---

8Except when splitting is performed, we use all the data in $D^n_l$ to train the $\hat{p}_k$’s, i.e., $D = D^n_l$.

9In our experiments we sample $\xi_{ik}$ independently from the uniform distribution on $(0, 10^{-6})$.

10If $n/3$ is not an integer, consider its rounding.
drop labels from the remaining $n/3$ to build

$$
\hat{G}^{+sp}(t) = \frac{1}{n/3} \sum_{i=1}^{n/3} \sum_{k=1}^{K} \mathbf{1}\{\hat{p}_k(X_i; (X_{i'}, Y_i'))_{i'=n/3+1}^{n} + \xi_{ik} > t\}. 
$$

Without splitting: We set $D = D_n^{L_i}$ and we reuse the $X_i$’s from $D_n^{L_i}$ (pulling labels out) to build

$$
\hat{G}^{-sp}(t) = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} \mathbf{1}\{\hat{p}_k(X_i; D_k) + \xi_{ik} > t\}. 
$$

Consequently, the corresponding set-valued classifiers are defined as

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

where $\square$ is $+sp$, $-sp$, $+c$, or $-c$. Note that according to the developed theory, we expect that 1) the continuity Assumption 6.3 is important 2) the splitting is necessary for good performance. That is, we expect that the estimators $\hat{\Gamma}_s^{+sp}$ (With splitting) and $\hat{\Gamma}_s^{+c}$
(Continuous) have superior performance when compared to $\hat{\Gamma}^{-sp}$ (Without splitting) and $\hat{\Gamma}^{-c}$ (Not continuous) respectively. Note also that both $\hat{\Gamma}^{+sp}$ and $\hat{\Gamma}^{+c}$ satisfy our assumptions and then are supported by our theory. However, $\hat{\Gamma}^{+sp}$ is a supervised method since it uses only the dataset $D_n$ to construct the set-valued classifier. In contrast, $\hat{\Gamma}^{+c}$ is semi-supervised since it uses an additional unlabeled sample.

At last, for any set-valued classifier $\hat{\Gamma}$ we compute empirical versions of the size and of the error as

$$P(\hat{\Gamma}) := \frac{1}{M} \sum_{i=1}^{M} \mathbb{1}\{y_{n+N+i} \notin \hat{\Gamma}(X_{n+N+i})\}, \quad S(\hat{\Gamma}) := \frac{1}{M} \sum_{i=1}^{M} |\hat{\Gamma}(X_{n+N+i})|.$$  \hspace{1cm} (7.1)

The whole scheme is repeated 30 times and we report averages and standard deviations of the error and size respectively.

**General observations.** Let us focus on the set-valued $\hat{\Gamma}^{+c}$ (Continuous) which is the semi-supervised method presented in the previous sections. It satisfies all the required assumptions and hence Proposition 6.6 is applicable to this classifier. Figure 1-right-bottom displays the expected size of $\hat{\Gamma}^{+c}$ for $s=2$. It highlights that even with a moderate unlabeled sample size $N$, the set-valued classifier has the prescribed expected size. In addition, when we compare the values of the errors $P(\hat{\Gamma}^{+c})$ in all the boxes in Figure 3, we observe that this error is indeed decreasing w.r.t. $s$. It is important to notice that the error of the corresponding single-output classifier (pure random forest without the second step) is 0.140 and then the use of the set-valued classifier is relevant already for moderate values of the size as $s = 2$. In this case, $P(\hat{\Gamma}^{+c}_2)$ equals 0.014 – the error is 10 times lower. Even with $s = 1$, the error of $\hat{\Gamma}^{+c}_1$ is lower than that of the corresponding single-output classifier and is equal to 0.105.

**Continuity of the estimator.** As it was mentioned in the beginning of this section, the random forest classifier does not satisfy Assumption 6.3. Our goal here is to understand the importance of this assumption. Figure 1-left demonstrates that in the absence of the continuity Assumption 6.3, the set-valued classifier $\hat{\Gamma}^{-c}$ (Not continuous) which does not modify $\hat{p}_1, \ldots, \hat{p}_K$ has a systematic bias in terms of the size across a wide range of $N$. Meanwhile, $\hat{\Gamma}^{+c}$ (Continuous) successfully captures the prescribed size in average, see Figure 1-right. We also note that in both cases the variance of the outcome reduces with the growth of the unlabeled data $N$. Finally, we highlight that the error of $\hat{\Gamma}^{-c}$ is slightly lower than that of $\hat{\Gamma}^{+c}$. However, this is attributed to the larger output size and not its superior performance.

**Data splitting.** The second important conclusion we report deals with the relevance of the independence condition between the dataset used to estimate the regression functions and the dataset used to estimate the function $G$. Figure 2-right displays that the set-valued classifier $\hat{\Gamma}^{-sp}$ (Without splitting) consistently over-estimates the size. By time, the expected size can even be twice as large as the desired size. In contrast, the size of the set-valued classifier $\hat{\Gamma}^{+sp}$ (With splitting) follows the diagonal line illustrating again that the proposed construction succeeds to satisfy the size constraint. According to the error, we can see from Figure 2-left that the set-valued classifier without splitting outperforms
Figure 2: Importance of splitting in case $D^U_N = \emptyset$. Set-valued classifier $\hat{\Gamma}^{-sp}_s$ uses the same labeled data twice. Set-valued classifier $\hat{\Gamma}^{+sp}_s$ splits labeled data to force independence.

slightly the set-valued classifier with splitting. However, this should be tempered by the fact that the size of the former is larger. In addition this could be related to the fact that we used more data to estimate the $p_k$’s for the method without splitting.

**Which sizes do we get?** We end this section by a thinner description of the size of the set-valued classifier $\hat{\Gamma}^{+c}_s$ (*Continuous*). For a single outcome of the experiment we report on Figure 3 the distribution of the size of the set-valued classifier. One draw the following conclusion. First of all we note that setting $s = 1$ is not equivalent to the set-up of the single-output classification. Indeed, the plot on top-left of Figure 3 shows that even though most of the times the corresponding set-valued classifier outputs only one candidates, there are situations where no labels or two predicted label candidates are provided. Moreover, the error of $\hat{\Gamma}^{+c}_1$ (with $s = 1$) is 0.105, while the error of the corresponding single-output classifier (pure random forest without the second step) is 0.14. Hence, set-valued classifiers can improve the performance even in the case of $s = 1$. Besides, note that for values of $s = 2, 3$ the corresponding set-valued classifier *significantly improves* the error, while having small size in average. Finally, we highlight how the set-valued classifier with the controlled expected size is different from a more naive procedure, which outputs $s$ most likely classes, that is, top-$s$ procedure (see also Section 2).

8. **Conclusion**

In this work we motivated and analyzed the set-valued framework of multi-class classification with controlled expected size from the minimax point of view. We established minimax lower bounds and provided a procedure which is optimal or optimal up to extra
logarithmic factor depending on the measure of performance. Our analysis implies that a sufficiently large unlabeled sample can strictly improve the rate of convergence, which advocates that the semi-supervised techniques ought be favored in this framework. To complement our analysis, we established that on a much smaller family of distributions which satisfy Lipschitzness assumption, the role of unlabeled data is negligible from the minimax point of view. Empirical study showed practical relevance of the proposed construction and highlighted our theoretical results. In future it would be valuable to analyze the considered framework relaxing the strong density assumption. Extensions to non-compact supports are of particular interest.

References


Appendix

In this part we provide missing proofs for our results. Let us first briefly describe the structure of this supplementary material.

- In Appendix A we provide the proof of Theorem 4.3, which establishes uniqueness of the s-Oracle classifier.
- In Appendix B we describe how to modify any preliminary estimator to satisfy Assumption 6.3. That is, we prove Lemma 6.4.
- We introduce some technical results that we use in the proofs of our main upper and lower bounds in Appendix C.
- In Appendix D and Appendix E are devoted to the proofs of main results of Sections 6 and 5 respectively.
- Finally, in Appendix F we prove the upper bound when $G^{-1}(\cdot)$ is assumed to be Lipschitz.

Appendix A: Uniqueness of the Oracle set-valued classifier

Proof of Theorem 4.3. The proof of this result goes by contradiction. Assume that there exists $\Gamma$ such that $S(\Gamma) \leq s$, $P(\Gamma) = P(\Gamma^*_s)$, and

$$
\sum_{k=1}^{K} \mathbb{P}_X(k \in \Gamma(X) \triangle \Gamma^*_s(X)) > 0 .
$$

Thanks to Proposition 5.2 and the assumption $P(\Gamma) = P(\Gamma^*_s)$, we have

$$
0 = P(\Gamma) - P(\Gamma^*_s) = \sum_{k=1}^{K} \mathbb{E}_P[X|p_k(X) - G^{-1}(s)\mathbb{1}_{\{k \in \Gamma(X) \triangle \Gamma^*_s(X)\}}] + G^{-1}(s)(s - S(\Gamma)) .
$$

Since $S(\Gamma) \leq s$, the r.h.s. of the above equation involves two positive terms whose sum is 0. Therefore, they are both null and we have $G^{-1}(s)(s - S(\Gamma)) = 0$ and

$$
\sum_{k=1}^{K} \mathbb{E}_P[X|p_k(X) - G^{-1}(s)\mathbb{1}_{\{k \in \Gamma(X) \triangle \Gamma^*_s(X)\}}] = 0 .
$$

Due to the fact that the random variable $|p_k(X) - G^{-1}(s)\mathbb{1}_{\{k \in \Gamma(X) \triangle \Gamma^*_s(X)\}}$ is almost surely non-negative, we must have almost surely $|p_k(X) - G^{-1}(s)\mathbb{1}_{\{k \in \Gamma(X) \triangle \Gamma^*_s(X)\}} = 0$ for each $k \in [K]$. Finally, thank to the continuity assumption 4.1 we have $\mathbb{P}_X(p_k(X) = G^{-1}(s)) = 0$ for every $k \in [K]$, which implies that

$$
\sum_{k=1}^{K} \mathbb{P}_X(k \in \Gamma(X) \triangle \Gamma^*_s(X)) = 0 .
$$

This contradicts our assumption and the proof is finished. \qed
Appendix B: Modification to satisfy continuity

Proof of Lemma 6.4. The proof of this lemma goes in two steps.

Step 1. Let Leb be the Lebesgue measure and for all $k \in [K]$ consider for some estimator $\hat{p}_k$

$$F_{\hat{p}_k}^{\text{Leb}}(t) = \text{Leb}\{x \in \mathbb{R}^d : \hat{p}_k(x) \leq t\}.$$ 

Lemma B.1. Assume that $\mathbb{P}_X$ admits a density with respect to the Lebesgue measure. Then, the continuity of $F_{\hat{p}_k}^{\text{Leb}}$ implies the continuity of $F_{\hat{p}_k}$

Proof of this result can be found in the end of the present section.

Step 2. Equipped with the previous result it is sufficient to construct $\hat{p}_k$ whose $F_{\hat{p}_k}^{\text{Leb}}$ is continuous. Consider $\bar{p}_k$ form the statement of Lemma 6.4. Denote by $T$ the set of all discontinuity points of $F_{\bar{p}_k}^{\text{Leb}}$. For each $t \in T$ consider $X_k^t = \{x \in \mathbb{R}^d : \bar{p}_k(x) = t\}$. Notice that since $T$ is countable, and for each $t \in [0, 1]$ the set $X_k^t$ is measurable, we have that $X_k := \bigcup_{t \in T} X_k^t$ is measurable as well. Consider an arbitrary measurable function $\varphi : \mathbb{R}^d \to [0, 1]$ such that for all $t \in \mathbb{R}$ it holds that $\text{Leb}\{x \in \mathbb{R}^d : \varphi(x) = t\} = 0$. For instance it can be $\varphi(x) = |\sin(\|x\|_2)|$. Define $\hat{p}_k$ for all $x \in \mathbb{R}^d$ as $\hat{p}_k(x) = \bar{p}_k(x) + a_n \varphi(x) 1_{\{x \in X_k^t\}}$, where $a_n = n^{-\beta/(2\beta + d)}$. Since $\varphi$ and $X_k$ are measurable, the function $\hat{p}_k$ is measurable by construction. Moreover, for all $x \in \mathbb{R}^d$ it holds that $|\hat{p}_k(x) - \bar{p}_k(x)| \leq a_n$. Thus, the first three entries of our statement are satisfied. It remains to check that $F_{\hat{p}_k}^{\text{Leb}}$ is discontinuous and use Lemma B.1. The proof goes by contradiction. Assume that there exists $t_0 \in [0, 1]$ such that $F_{\hat{p}_k}^{\text{Leb}}$ is discontinuous in $t_0$. Therefore,

$$0 < A := \text{Leb}\{x \in \mathbb{R}^d : \bar{p}_k(x) + a_n \varphi(x) 1_{\{x \in X_k^t\}} = t_0\}.$$ 

Notice that we can write

$$A = \text{Leb}\{x \in X_k^t : \bar{p}_k(x) + a_n \varphi(x) = t_0\} + \text{Leb}\{x \in \mathbb{R}^d \setminus X_k^t : \bar{p}_k(x) = t_0\}.$$ 

By construction of $X_k$ we have for all $t \in [0, 1]$ that

$$\text{Leb}\{x \in \mathbb{R}^d \setminus X_k^t : \bar{p}_k(x) = t\} = 0.$$ 

Moreover, we have

$$\text{Leb}\{x \in X_k^t : \bar{p}_k(x) + a_n \varphi(x) = t_0\} \leq \sum_{t \in T} \text{Leb}\{x \in \mathbb{R}^d : a_n \varphi(x) = t_0 - t\} = 0.$$
Set-valued classification

where the last equality holds due to the fact that all level sets of $\varphi$ have zero Lebesgue measure. We conclude that $A = 0$ and therefore arrive to a contradiction. The above implies that for each $k \in [K]$ the function $F_{\hat{p}_k}^{\text{Leb}}$ is continuous, thus $F_{\hat{p}_k}$ is continuous by Lemma B.1. Importantly, note that the construction of $\hat{p}$ is data driven and does not rely on any unknown quantity.

Proof of Lemma B.1. The proof of this result goes by contradiction. Assume that there exists $t_0 \in [0, 1]$ such that $F_{\hat{p}_k}$ is discontinuous in $t_0$. Therefore, we have

$$P_X(\hat{p}_k(X) = t_0) = F_{\hat{p}_k}(t_0) - \lim_{\epsilon \to 0^-} F_{\hat{p}_k}(t_0 + \epsilon) > 0.$$  

Since $P_X$ admits a density with respect to the Lebesgue measure it holds that

$$\text{Leb}\{x \in \mathbb{R}^d : \hat{p}_k(x) = t_0\} > 0,$$

implying that $F_{\hat{p}_k}^{\text{Leb}}$ is discontinuous at $t_0$.

Appendix C: Technical results

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 $P_1, P_2$ on some space measurable space $(X, A)$ the Kullback-Leibler divergence between $P_1$ and $P_2$ is defined as

$$KL(P_1, P_2) := \begin{cases} \int_X \log \left( \frac{dP_1}{dP_2} \right) dP_1, & \text{supp}(P_1) \subset \text{supp}(P_2), \\ +\infty, & \text{otherwise} \end{cases},$$

and the total variation distance is defined as

$$TV(P_1, P_2) := \sup_{A \in \mathcal{A}} |P_1(A) - P_2(A)|.$$

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

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

$$\min_{i \in \{0, 1, \ldots, m\}} P_i(A_i) \leq \left( 0.71 \sqrt{\frac{1}{m} \sum_{i=1}^m KL(P_i, P_0)} \right).$$

Lemma C.2 (Pinsker’s inequality). Given any two probability measures $P_1, P_2$ on some measurable space $(X, A)$ we have

$$TV(P_1, P_2) \leq \sqrt{\frac{1}{2} KL(P_1, P_2)}.$$
Lemma C.3 (Hoeffding’s inequality (Hoeffding, 1963)). Let $b > 0$ be a real number, and $N$ be a positive integer. Let $X_1, \ldots, X_N$ be $N$ random variables having values in $[0, b]$, then

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

Proposition C.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 $p : \mathbb{R}^d \to [0, 1]^K$ be a vector function, we define for all $t \in [0, 1]$ and all $s \in (0, K)$

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

Then,

- for all $t \in (0, 1)$ and $s \in (0, K)$ we have $G^{-1}(s) \leq t \iff G(t) \leq s$.
- 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 $s \in (0, K)$ we have $G(G^{-1}(s)) = s$.

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 C.5. Let $\varepsilon$ distributed from a uniform distribution on $[K]$ and $Z_1, \ldots, Z_K$, $K$ 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

$$
H(Z) \overset{d}{=} U \quad \text{and} \quad H^{-1}(U) \overset{d}{=} Z ,
$$

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

Proof. First we note that for every $t \in [0, 1]$, $\Pr (H(Z) \leq t) = \Pr (Z \leq H^{-1}(t))$. Moreover,
we have
\[ P(H(Z) \leq t) = \sum_{k=1}^{K} P(Z \leq H^{-1}(t), \varepsilon = k) \]
\[ = \frac{1}{K} \sum_{k=1}^{K} P(Z_k \leq H^{-1}(t)) \quad \text{(with } \varepsilon \text{ independent of } Z) \]
\[ = H(H^{-1}(t)) \]
\[ = t \quad \text{(with } H \text{ continuous)}. \]

To conclude the proof, we observe that
\[ P(H^{-1}(U) \leq t) = P(U \geq H(t)) = \frac{1}{K} \sum_{k=1}^{K} P(Z_k \leq t) \]
\[ = \sum_{k=1}^{K} P(Z_k \leq t, \varepsilon = k) = P(Z \leq t). \]

\[ \square \]

Appendix D: Upper bounds

In this section we prove Theorem 6.1. Thus, in this section for simplicity we omit the subscript SSE from \( \hat{\Gamma} \). Recall that our dataset consists of three parts \( D_{\lfloor n/2 \rfloor}, D_{\lceil n/2 \rceil}, D_U \).

The set \( D_{\lfloor n/2 \rfloor} \) is used to construct an estimator \( \hat{p} \) of the regression function \( p \), that is, \( \hat{p} \) is independent from both \( D_{\lceil n/2 \rceil}, D_U \). The other two sets \( D_{\lfloor n/2 \rfloor}, D_U \) are used in a semi-supervised manner to estimate the threshold, that is, we erase the labels from \( D_{\lfloor n/2 \rfloor} \). Let \( s \in [K-1] \), and also recall the definition of the proposed semi-supervised estimator for a given \( x \in \mathbb{R}^d \)
\[ \hat{\Gamma}(x) = \{ k \in [K] : \hat{p}_k(x) \geq \hat{G}^{-1}(s) \} , \]
with \( \hat{p}_k(x) \) satisfying Assumptions 6.3, 6.2 for all \( k \in [K] \). Moreover, \( \hat{G}^{-1}(s) \) defined as the generalized inverse of
\[ \hat{G}(t) = \frac{1}{|n/2| + N} \sum_{x \in D_{\lfloor n/2 \rfloor} \cup D_{\lceil n/2 \rceil}} \sum_{k=1}^{K} 1_{\hat{p}_k(x) > t} , \]
where \( t \in [0,1] \). Additionally, recall that the s-Oracle is given as
\[ \Gamma^*_s(x) = \{ k \in [K] : p_k(x) \geq G^{-1}(s) \} , \quad \text{(D.1)} \]
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

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

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 $\tilde{G}^{-1}(s)$ of $\tilde{\Gamma}$ to the true value of the threshold $G^{-1}(s)$.

**Lemma D.1** (Upper bound on the thresholds). Let $\mathbb{Q}_X$ be a Borel probability measure on $\mathbb{R}^d$. For two vector functions $p, \hat{p} : \mathbb{R}^d \to [0, 1]^K$, we define

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

Then for every $s \in (0, K)$

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

Before deriving this result let us mention that there are no assumptions on $\mathbb{Q}_X$, that is, this result can be applied to any $\mathbb{Q}_X$. This fact is important for both Theorems 6.1 and 6.9 and is used with the real distribution of the features ($\mathbb{Q}_X = \mathbb{P}_X$) in Theorem 6.1 and its empirical counterpart ($\mathbb{Q}_X$ is based on the unlabeled sample) in Theorem 6.9.

**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 C.4 to the first inequality we get for all $t \in [0, 1]$

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

thus, for $t = G^{-1}(s)$ with $s \in (0, K)$ we have thanks to the fact that $G(G^{-1}(s)) \leq s$ and the fact that $\tilde{G}^{-1}$ is deceasing, we have $\tilde{G}^{-1}(s) \leq \tilde{G}^{-1}(G(G^{-1}(s)))$. This yields

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

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

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

$$
\sum_{k=1}^{K} Q_X \left( \hat{p}_k(X) > t + \|\hat{p} - p\|_{\infty, Q_X} \right) \leq \sum_{k=1}^{K} Q_X (p_k(X) > t) \leq \sum_{k=1}^{K} Q_X \left( \hat{p}_k(X) > t - \|\hat{p} - p\|_{\infty, Q_X} \right) ,
$$

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

$$
Q_X (\hat{p}_k(X) > t + |\hat{p}_k(X) - p_k(X)|) \leq Q_X (p_k(X) > t) \leq Q_X (\hat{p}_k(X) > t - |\hat{p}_k(X) - p_k(X)|) ,
$$

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

We are in position to prove Theorem 6.1, let us point out that the most difficult part in Theorem 6.1 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

$$
R_{s}(\hat{\Gamma}) - R_{s}(\Gamma^*_s) = \left( R_{s}(\hat{\Gamma}) - R_{s}(\Gamma^*_s) \right) + \left( R_{s}(\hat{\Gamma}) - R_{s}(\hat{\Gamma}) \right) .
$$

Recall that thanks to Proposition 5.2 we have

$$
R_{1} = \sum_{k=1}^{K} E \left[ |p_k(X) - G^{-1}(s)| \mathbb{1}_{\{k \in \hat{\Gamma}(X) \triangle \Gamma^*_s(X)\}} \right] .
$$

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

$$
\left\{ \begin{array}{c}
p_k(X) - G^{-1}(s) \geq 0 \\
\hat{p}_k(X) - \tilde{G}^{-1}(s) < 0
\end{array} \right. \quad \text{or} \quad \left\{ \begin{array}{c}
p_k(X) - G^{-1}(s) < 0 \\
\hat{p}_k(X) - \tilde{G}^{-1}(s) \geq 0
\end{array} \right. ,
$$
holds. Thus on the event \( k \in \tilde{\Gamma}(X) \triangle \Gamma_s^*(X) \) we have

\[
\left| p_k(X) - G^{-1}(s) \right| \leq \left| \hat{p}_k(X) - p_k(X) + G^{-1}(s) - \tilde{G}^{-1}(s) \right|
\]

\[
\leq \left| \hat{p}_k(X) - p_k(X) \right| + \left| G^{-1}(s) - \tilde{G}^{-1}(s) \right|.
\]

Therefore, for \( R_1 \) using Lemma D.1 and the observations above we can write

\[
R_1 \leq \sum_{k=1}^{K} \mathbb{E} \left[ \left| p_k(X) - G^{-1}(s) \right| \mathbb{1}_{\left\{ \left| p_k(X) - G^{-1}(s) \right| \leq \left| \hat{p}_k(X) - p_k(X) \right| + \left| G^{-1}(s) - \tilde{G}^{-1}(s) \right| \right\}} \right]
\]

\[
\leq \sum_{k=1}^{K} \mathbb{E} \left[ \left| p_k(X) - G^{-1}(s) \right| \mathbb{1}_{\left\{ \left| p_k(X) - G^{-1}(s) \right| \leq 2 \| \hat{p} - p \|_{\infty, \mathbb{P}_X} \right\}} \right]
\]

\[
\leq \sum_{k=1}^{K} \mathbb{E} \left[ 2 \| \hat{p} - p \|_{\infty, \mathbb{P}_X} \mathbb{1}_{\left\{ \left| p_k(X) - G^{-1}(s) \right| \leq 2 \| \hat{p} - p \|_{\infty, \mathbb{P}_X} \right\}} \right]
\]

\[
= 2 \| \hat{p} - p \|_{\infty, \mathbb{P}_X} \sum_{k=1}^{K} \mathbb{P}_X \left( \left| p_k(X) - G^{-1}(s) \right| \leq 2 \| \hat{p} - p \|_{\infty, \mathbb{P}_X} \right),
\]

finally, using the margin Assumption 4.6 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 6.2 we get

\[
\mathbb{E}_{(\mathcal{D}_n, \mathcal{D}_n')} R_1 \leq c_1 C_2 2^{1+\alpha} K \left( \frac{n}{\log n} \right) \left( \frac{\| \hat{p} - p \|_{\infty, \mathbb{P}_X}^{1+\alpha}}{2^{1+\alpha}} \right).
\]

For \( R_2 \) the following trivial upper bound holds

\[
R_2 = \left( \mathbb{P}(\hat{\Gamma}) - \mathbb{P}(\tilde{\Gamma}) \right) + G^{-1}(s) \left( S(\hat{\Gamma}) - S(\tilde{\Gamma}) \right)
\]

\[
= \sum_{k=1}^{K} \mathbb{E} \left( p_k(X) - G^{-1}(s) \right) \mathbb{1}_{\{ k \in \hat{\Gamma}(X) \}} - \mathbb{1}_{\{ k \in \tilde{\Gamma}(X) \}}
\]

\[
\leq \sum_{k=1}^{K} \mathbb{E} \left| \mathbb{1}_{\{ k \in \hat{\Gamma}(X) \}} - \mathbb{1}_{\{ k \in \tilde{\Gamma}(X) \}} \right| \mathbb{E}[\hat{\Gamma}(X) \triangle \tilde{\Gamma}(X)]
\]

\[
= \sum_{k=1}^{K} \mathbb{E} \left| \hat{p}_k(X) \geq G^{-1}(s) \right| - \left| \tilde{p}_k(X) \geq G^{-1}(s) \right|.
\]
now, thanks to the first property of Proposition C.4 we can write

\[
R_2 \leq \sum_{k=1}^{K} \mathbb{E} \left| \mathbf{1}\{\hat{G}(\hat{p}_k(X)) \leq s\} - \mathbf{1}\{\hat{G}(\hat{p}_k(X)) \leq s\} \right|
\]

\[
\leq \sum_{k=1}^{K} \mathbb{P}_X \left( \left| \hat{G}(\hat{p}_k(X)) - \tilde{G}(\hat{p}_k(X)) \right| \geq \left| \hat{G}(\hat{p}_k(X)) - s \right| \right)
\]

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 \( k \in [K] \)

\[
A_0^k = \{ \left| \tilde{G}(\hat{p}_k(X)) - s \right| \leq \delta \}
\]

\[
A_j^k = \{ 2^{j-1} \delta < \left| \tilde{G}(\hat{p}_k(X)) - s \right| \leq 2^j \delta \}, \quad j \geq 1.
\]

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

\[
\sum_{k=1}^{K} \mathbb{P}_X \left( \left| \hat{G}(\hat{p}_k(X)) - \tilde{G}(\hat{p}_k(X)) \right| \geq \left| \tilde{G}(\hat{p}_k(X)) - s \right| \right) = (D.3)
\]

\[
\sum_{k=1}^{K} \sum_{j \geq 0} \mathbb{P}_X \left( \left| \hat{G}(\hat{p}_k(X)) - \tilde{G}(\hat{p}_k(X)) \right| \geq \left| \tilde{G}(\hat{p}_k(X)) - s \right|, A_j^k \right).
\]

Now, we consider \( \varepsilon \) uniformly distributed on \([K]\) independent of the data and \( X \). Conditional on the data and under Assumption 6.3, we apply Lemma C.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

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

Hence, for all \( j \geq 0 \), we obtain

\[
\sum_{k=1}^{K} \mathbb{P}_X(A_j^k) \leq \sum_{k=1}^{K} \mathbb{P}_X \left( \left| \tilde{G}(\hat{p}_k(X)) - s \right| \leq 2^j \delta \right) \leq 2^{j+1} \delta .
\]  

(D.4)

Next, we observe that for all \( j \geq 1 \)

\[
\sum_{k=1}^{K} \mathbb{P}_X \left( \left| \tilde{G}(\hat{p}_k(X)) - \hat{G}(\hat{p}_k(X)) \right| \geq \left| \hat{G}(\hat{p}_k(X)) - s \right|, A_j^k \right) \leq (D.5)
\]

\[
\sum_{k=1}^{K} \mathbb{P}_X \left( \left| \hat{G}(\hat{p}_k(X)) - \tilde{G}(\hat{p}_k(X)) \right| \geq 2^{j-1} \delta, A_j^k \right).
\]
Thus, we obtain that

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

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

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

$$= \sum_{k=1}^{K} \sum_{j \geq 0} \mathbb{E}_{(\mathcal{D}_{n/2}, \mathcal{D}_{n/2}, \mathcal{D}_N^+, X \sim \mathbb{P}_X)} \mathbf{1}_{\{|\hat{G}(\hat{p}_k(X)) - \hat{G}(\hat{p}_k(X))| \geq 2j^{-1}\delta\}} \mathbf{1}_{\{A_j^k\}}.$$

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

$$\mathbb{E}_{(\mathcal{D}_{n/2}, \mathcal{D}_{n/2}, \mathcal{D}_N^+, X \sim \mathbb{P}_X)} \mathbf{1}_{\{|\hat{G}(\hat{p}_k(X)) - \hat{G}(\hat{p}_k(X))| \geq 2j^{-1}\delta\}} \mathbf{1}_{\{A_j^k\}} =$$

$$\mathbb{E}_{(\mathcal{D}_{n/2}, X \sim \mathbb{P}_X)} \mathbb{E}_{(\mathcal{D}_{n/2}, \mathcal{D}_N^+)} \left[ \mathbf{1}_{\{|\hat{G}(\hat{p}_k(X)) - \hat{G}(\hat{p}_k(X))| \geq 2j^{-1}\delta\}} \mathbf{1}_{\{A_j^k\}} \right].$$

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

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

Therefore, treating $A_j^k$ separately, we get from inequalities of Eqs. (D.3), (D.4), and (D.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 $s$

$$\mathbb{E}_{(\mathcal{D}_n^+, \mathcal{D}_N^+)} \mathbb{E}_{X \sim \mathbb{P}_X} \left| \hat{\Gamma}(X) \Delta \Gamma_s^*(X) \right| \leq \mathbb{E}_{(\mathcal{D}_n^+, \mathcal{D}_N^+)} \mathbb{E}_{X \sim \mathbb{P}_X} \left| \hat{\Gamma}(X) \Delta \Gamma_s^*(X) \right| + \mathbb{E}_{(\mathcal{D}_n^+, \mathcal{D}_N^+)} \mathbb{E}_{X \sim \mathbb{P}_X} \left| \hat{\Gamma}(X) \Delta \hat{\Gamma}(X) \right|.$$

Notice that for the term $\mathbb{E}_{(\mathcal{D}_n^+, \mathcal{D}_N^+)} \mathbb{E}_{X \sim \mathbb{P}_X} \left| \hat{\Gamma}(X) \Delta \hat{\Gamma}(X) \right|$ we can re-use the proof technique used for the term $R_2$ in Eq. (D.2). Thus, it remain to upper-bound the term
\[ \mathbb{E}_{(P^*_k, \mathbb{P}^*_k)} \mathbb{E}_{X \sim \mathbb{P}_X} \left| \hat{\Gamma}(X) \triangle \Gamma^*_s(X) \right|. \] 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 \( \Gamma^*_s \) and \( \hat{\Gamma} \) we can write for \( (*) = \mathbb{E}_{X \sim \mathbb{P}_X} \left| \hat{\Gamma}(X) \triangle \Gamma^*_s(X) \right| \)

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

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

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

Similar conditions are satisfied if \( \hat{p}_k(X) < \tilde{G}^{-1}(s) \) and \( p_k(X) \geq G^{-1}(s) \). Using the above arguments we can upper-bound \( (*) \) as

\[
(*) \leq \sum_{k=1}^{K} \mathbb{P}_X \left( |p_k(X) - G^{-1}(s)| \leq |\hat{p}_k(X) - p_k(X)| \right) \\
+ 1_{\{\tilde{G}^{-1}(s) \leq G^{-1}(s)\}} \sum_{k=1}^{K} \mathbb{P}_X \left( \tilde{G}^{-1}(s) < \hat{p}_k(X) < G^{-1}(s) \right) \\
+ 1_{\{G^{-1}(s) \leq \tilde{G}^{-1}(s)\}} \sum_{k=1}^{K} \mathbb{P}_X \left( G^{-1}(s) < \hat{p}_k(X) < \tilde{G}^{-1}(s) \right) \\
= \sum_{k=1}^{K} \mathbb{P}_X \left( |p_k(X) - G^{-1}(s)| \leq |\hat{p}_k(X) - p_k(X)| \right) \\
+ \left| \hat{G} \left( \tilde{G}^{-1}(s) \right) - \hat{G} \left( G^{-1}(s) \right) \right| .
\]

Thanks to the continuity Assumption 6.3 on the estimator and the continuity Assumption 4.1 on the distribution we clearly have \( \hat{G} \left( \tilde{G}^{-1}(s) \right) = s = G \left( G^{-1}(s) \right) \). Moreover, we can write

\[
\left| \hat{G} \left( \tilde{G}^{-1}(s) \right) - \hat{G} \left( G^{-1}(s) \right) \right| = \left| G \left( G^{-1}(s) \right) - \hat{G} \left( G^{-1}(s) \right) \right| \\
\leq \sum_{k=1}^{K} \mathbb{E}_{X \sim \mathbb{P}_X} \left| 1_{\{\hat{p}_k(X) \geq G^{-1}(s)\}} - 1_{\{p_k(X) \geq G^{-1}(s)\}} \right| \\
\leq \sum_{k=1}^{K} \mathbb{P}_X \left( |p_k(X) - G^{-1}(s)| \leq |\hat{p}_k(X) - p_k(X)| \right) .
\]
Thus, our bound reads as

\[(\ast) \leq 2 \sum_{k=1}^{K} \mathbb{P}_X (|p_k(X) - G^{-1}(s)| \leq |\hat{p}_k(X) - p_k(X)|) .\]

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 6.2. 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 4.1, 6.3 on the distribution and the estimator respectively.

Appendix E: Proof of the lower bounds

This section is devoted to the proof of the lower bounds provided by Theorem 5.1. 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 Theorem 5.1 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 \(G^{-1}(s)\).

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 4.1 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 C.2.

E.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 \(s \in \{1, \ldots, \lfloor K/2 \rfloor - 1\}\) 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 \(P_0\) and \(P_1\) on \(\mathbb{R}^d\) sharing the same regression function \(p(\cdot) = (p_1(\cdot), \ldots, p_K(\cdot))\) 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 following sets
\begin{align*}
X_0 &= \{ x \in \mathbb{R}^d : \|x\| \leq r_0 \} , \\
X_1 &= \left\{ x \in \mathbb{R}^d : \left\| x - \left( r_1 + \rho, 0, \ldots, 0 \right) \right\|_{\mathbb{R}^d} \leq \rho/2 \right\} , \\
X_2 &= \left\{ x \in \mathbb{R}^d : \left\| x - \left( r_2 + \rho, 0, \ldots, 0 \right) \right\|_{\mathbb{R}^d} \leq \rho \right\} , \\
X_3 &= \left\{ x \in \mathbb{R}^d : \left\| x - \left( r_3 + \rho, 0, \ldots, 0 \right) \right\|_{\mathbb{R}^d} \leq \rho/2 \right\} , \\
X_4 &= \{ x \in \mathbb{R}^d : r_4 \leq \|x\| \leq 2r_4 \} .
\end{align*}

Let us denote by $o_i = (r_i + \rho, 0, \ldots, 0)^T$ for $i = 1, 2, 3$ the centers of $X_1, X_2$ and $X_3$. Using these sets we define the regression vector as
\[ p_1(x) = \ldots = p_{2s}(x) = \begin{cases} 
\frac{1}{2s} - \frac{\varphi_0(x)}{2s}, & x \in X_0 \\
\frac{3K+2s}{8Ks} - \frac{\varphi_1(x)}{2s}, & x \in X_1 \\
\frac{3K+4s}{8Ks} - \frac{\varphi_2(x)}{2s}, & x \in X_2 \\
\frac{K+2s}{8Ks} - \frac{\varphi_3(x)}{2s}, & x \in X_3 \\
\frac{1}{K} - \frac{\varphi_4(x)}{2s}, & x \in X_4 
\end{cases} \]

\[ p_{2s+1}(x) = \ldots = p_K(x) = \begin{cases} 
\frac{\varphi_0(x)}{K-2s}, & x \in X_0 \\
\frac{1}{K} + \frac{\varphi_1(x)}{K-2s}, & x \in X_1 \\
\frac{3K-4s}{8(K-2s)} + \frac{\varphi_2(x)}{K-2s}, & x \in X_2 \\
\frac{3K-6s}{8(K-2s)} + \frac{\varphi_3(x)}{K-2s}, & x \in X_3 \\
\frac{1}{K} + \frac{\varphi_4(x)}{K-2s}, & x \in X_4 
\end{cases} \]

In order to define the functions \( \varphi_i \) for \( i = 0, \ldots, 4 \) 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 4 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 \( i = 0, \ldots, 4 \) as

\[ \varphi_0(x) = \frac{C'}{2} \left( K - 2s \right) \left( \frac{1}{8Ks} \vee \frac{1}{4K} \right) \psi_{-1,0}(||x||), \]

\[ \varphi_i(x) = \frac{C'}{2} \rho^\beta \left( K - 2s \right) \left( \frac{1}{8Ks} \vee \frac{1}{4K} \right) \psi_{-1,1} \left( \frac{||x - \alpha_i||}{\rho} \right), \quad i = 1, 3, \]

\[ \varphi_2(x) = \frac{C'}{2} \rho^\beta \left( K - 2s \right) \left( \frac{1}{8Ks} \vee \frac{1}{4K} \right) \psi_{-1,1} \left( \frac{||x - \alpha_3||}{\rho} \right), \]

\[ \varphi_4(x) = \frac{C'}{2} \left( K - 2s \right) \left( \frac{1}{8Ks} \vee \frac{1}{4K} \right) \psi_{-1,2r_4}(||x||), \]

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

\[ x \mapsto C' \left( ||x||^{2\left(\frac{\beta}{2}\right)} \psi_{-1,1}(||x||) \right), \]

is infinitely smooth, thus it is \((\beta, L)\)-Hölder for a properly chosen \( C' \). Figure 5 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 re-scaling which preserves the Hölder constant \( L \). Same reasoning applies to \( \varphi_i \) for \( i = 0, 2, 4 \).
Set-valued classification

\[ \left( a + b \right) / 2 \]

Figure 5: Dumped bump function: \( x \mapsto \left( x - \frac{a+b}{2} \right)^2 \psi_{a,b}(x) \). Importantly, this function behaves as polynomial of even degree \( 2 \left\lceil \beta \right\rceil \) 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 \( s < K/2 \) one can check that the following relations hold true

\[
0 < \frac{1}{4K} < \frac{2K - 4s}{4K(K - 2s)} < \frac{3K - 4s}{4K(K - 2s)} < \frac{3K - 6s}{4K(K - 2s)} < \frac{1}{K},
\]

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

\[
\mu_0(x) = \begin{cases} 
\frac{1}{2} \frac{\text{Leb}(X_0)}{\text{Leb}(X_0)}, & x \in X_0 \\
\frac{\kappa_{N,n}}{\text{Leb}(X_1)}, & x \in X_1 \\
\frac{\kappa_{N,n}}{\text{Leb}(X_2)}, & x \in X_2, \\
\frac{1/2-3\kappa_{N,n}}{\text{Leb}(X_3)}, & x \in X_3 \\
\frac{1/2-3\kappa_{N,n}}{\text{Leb}(X_4)}, & x \in X_4 
\end{cases}
\]

and both \( \mu_0, \mu_1 \) are equal to zero in unspecified regions. Clearly, the strong density assumption is satisfied on \( X_0 \) and \( X_4 \) 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 \( X_i \) for \( i = 1, 2, 3 \) is satisfied. Notice that

\[
\text{Leb}(X_i) = c \rho^d,
\]

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 \( G_0^{-1}(s), G_1^{-1}(s) \) and the optimal s-sets
$\Gamma_0^\ast, \Gamma_1^\ast$ are given as

$$G_0^{-1}(s) = \frac{3K + 2s}{8Ks}, \quad G_1^{-1}(s) = \frac{K + 6s}{8Ks},$$

$$\Gamma_0^\ast(x) = \begin{cases} \{1, \ldots, 2s\}, & x \in \mathcal{X}_0 \\ \emptyset, & \text{otherwise} \end{cases},$$

$$\Gamma_1^\ast(x) = \begin{cases} \{1, \ldots, 2s\}, & x \in \mathcal{X}_0 \cup \mathcal{X}_1 \cup \mathcal{X}_2, \\ \emptyset, & \text{otherwise} \end{cases}.$$  

**The margin assumption:** we are in position to check the margin Assumption 4.6. Let $t_0 = \frac{1}{2} \left( \frac{K - 2s}{8Ks} \land \frac{1}{4K} \right)$, thus for every $k \in \{2s + 1, \ldots, K\}$ and every $t \leq t_0$ we have

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

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

$$\mathbb{P}_0 \left( |p_k(X) - G_0^{-1}(s)| \leq t \right) = \mathbb{P}_0 \left( \frac{\|x - \alpha_1\|}{\rho} \leq \frac{2}{\rho} \psi_{-1,1} \left( \frac{\|x - \alpha_1\|}{\rho} \right) \leq 2st, X \in \mathcal{X}_1 \right),$$

$$\mathbb{P}_1 \left( |p_k(X) - G_1^{-1}(s)| \leq t \right) = \mathbb{P}_1 \left( \frac{\|x - \alpha_1\|}{\rho} \leq \frac{2}{\rho} \psi_{-1,1} \left( \frac{\|x - \alpha_1\|}{\rho} \right) \leq 2st, X \in \mathcal{X}_3 \right).$$

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

$$\mathbb{P}_0 \left( |p_k(X) - G_0^{-1}(s)| \leq t \right) \leq \mathbb{P}_0 \left( \left( \frac{\|x - \alpha_1\|}{\rho} \right)^{2/\beta} \psi_{-1,1} \left( \frac{\|x - \alpha_1\|}{\rho} \right) \leq c^2 \rho^{-\beta} t, X \in \mathcal{X}_1 \right).$$

Therefore we can write using the strong density assumption

$$\mathbb{P}_0 \left( |p_k(X) - G_0^{-1}(s)| \leq t \right) \leq \int_{\|x - \alpha_1\| \leq \rho/2} \frac{1}{\left( \left( \frac{\|x - \alpha_1\|}{\rho} \right)^{2/\beta} \psi_{-1,1} \left( \frac{\|x - \alpha_1\|}{\rho} \right) \leq c^2 \rho^{-\beta} t \right)} d\mu_0(x)$$

$$\leq C \int_{\|x - \alpha_1\| \leq \rho/2} \frac{1}{\left( \left( \frac{\|x - \alpha_1\|}{\rho} \right)^{2/\beta} \psi_{-1,1} \left( \frac{\|x - \alpha_1\|}{\rho} \right) \leq c^2 \rho^{-\beta} t \right)} dx$$

$$= C \int_{\|x\| \leq \rho/2} \frac{1}{\left( \left( \frac{\|x\|}{\rho} \right)^{2/\beta} \psi_{-1,1} \left( \frac{\|x\|}{\rho} \right) \leq c^2 \rho^{-\beta} t \right)} dx = C \rho^\beta \int_{\|x\| \leq 1/2} \frac{1}{\left( \|x\|^{2/\beta} \psi_{-1,1}(\|x\|) \leq c^2 \rho^{-\beta} t \right)} dx.$$
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

$$P_0(|p_k(X) - G^{-1}_0(s)| \leq t) \leq C\rho^d \int_{\|x\| \leq 1/2} 1\{\|x\|^2 \leq C\rho^{-\beta} t\} dx \leq C\rho^d \int_{\|x\| \leq 1/2} 1\{\|x\| \leq C\rho^{-\beta/t}(2\beta/2)\} dx \leq C\rho^d(1-\beta/t)\rho^{d/(2\beta/2)}.$$

This implies that as long as $\alpha \leq d/(2\beta/2)$ (and since we have $\beta \leq 2\beta$) the margin assumption is satisfied. Moreover, these conditions imply that $\alpha\beta \leq d$, 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 $P_1$ on the set $X_3$.

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 $X_0, X_1, X_2, X_3, X_4$. Notice that such a choice is possible since by the construction of functions $\phi_i$ for $i = 0, 1, 2, 3, 4$ they are zeroed-out on the boundaries of $X_0, X_1, X_2, X_3, X_4$. Thus in the region $\mathbb{R}^d \setminus \bigcup_{i=0}^4 X_i$ 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 4.1 since the measures are only defined on $X_0, X_1, X_2, X_3, X_4$ 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 $P$ on some measurable space $(\Omega_0, \mathcal{A}_0)$ and a measurable function $X : (\Omega_0, \mathcal{A}_0) \to (\mathbb{R}, \text{Borel}(\mathbb{R}))$ we write

$$P[X] := E[X].$$

**Bound on the KL-divergence:** we start by computing the KL-divergence between
\[ KL(\mu_0, \mu_1) := \int_{\mathbb{R}^d} \mu_0(x) \log \left( \frac{\mu_0(x)}{\mu_1(x)} \right) \, dx = \sum_{i=0}^{4} \int_{x \in X_i} \mu_0(x) \log \left( \frac{\mu_0(x)}{\mu_1(x)} \right) \, dx \]

\[ = \frac{1}{\text{Leb}(X_0)} \int_{x \in X_0} \frac{1}{2} \log \left( \frac{1/2}{1/2 - 3\kappa_{N,n}} \right) \, dx \]

\[ + \frac{1}{\text{Leb}(X_4)} \int_{x \in X_4} (1/2 - 3\kappa_{N,n}) \log \left( \frac{1/2}{1/2} \right) \, dx \]

\[ = \frac{1}{2} \log \left( \frac{1/2}{1/2 - 3\kappa_{N,n}} \right) \]

\[ + \frac{1}{2} (1/2 - 3\kappa_{N,n}) \log \left( \frac{1/2}{1/2} \right) \]

\[ = -3\kappa_{N,n} \log (1 - 6\kappa_{N,n}) \leq 36\kappa_{N,n}^2. \]

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

\[ H(\hat{\Gamma}, \Gamma^*_i) := \mu_i \left| \hat{\Gamma}(X) \triangle \Gamma^*_i(X) \right|. \]

Recall that we are interested in the following quantity

\[ \inf_{\hat{\Gamma}} \sup_{P \in \mathcal{P}} \mathbb{E}_{X \sim P_X} \left| \hat{\Gamma}(X) \triangle \Gamma^*_s(X) \right|, \]

since the hypotheses \( P_0, P_1 \in \mathcal{P} \) we can write

\[ 2 \sup_{P \in \mathcal{P}} \mathbb{E}_{(P^*_h, P^*_N)_X \sim P_X} \left| \hat{\Gamma}(X) \triangle \Gamma^*_s(X) \right| \geq (\ast), \]

where \((\ast)\) is defined as

\[ (\ast) = \mu_0^{\otimes (n+N)} \otimes \mathbb{P}^{\otimes n}_{X \mid X} H(\hat{\Gamma}, \Gamma^*_0) + \mu_1^{\otimes (n+N)} \otimes \mathbb{P}^{\otimes n}_{X \mid X} H(\hat{\Gamma}, \Gamma^*_1), \]

thus, for the Hamming risk we can write

\[ (\ast) \geq \mu_0^{\otimes (n+N)} \otimes \mathbb{P}^{\otimes n}_{X \mid X} \left( \frac{d\mu_1^{\otimes (n+N)} \otimes \mathbb{P}^{\otimes n}_{X \mid X}}{d\mu_0^{\otimes (n+N)} \otimes \mathbb{P}^{\otimes n}_{X \mid X} \wedge 1} \right) \left( H(\hat{\Gamma}, \Gamma^*_0) + H(\hat{\Gamma}, \Gamma^*_1) \right). \]

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

\[
H(\hat{\Gamma}, \Gamma_0^*) + H(\hat{\Gamma}, \Gamma_1^*) = \mu_0 \sum_{k=1}^{K} 1_{k \in \hat{\Gamma}(X) \triangle \Gamma_0^*(X)} + \mu_1 \sum_{k=1}^{K} 1_{k \in \hat{\Gamma}(X) \triangle \Gamma_1^*(X)}
\]

\[
\geq \mu_0 \left( \frac{d\mu_1}{d\mu_0} \wedge 1 \right) \sum_{k=1}^{K} 1_{k \in \hat{\Gamma}(X) \triangle \Gamma_0^*(X)}
\]

\[
+ \mu_0 \left( \frac{d\mu_1}{d\mu_0} \wedge 1 \right) \sum_{k=1}^{K} 1_{k \in \hat{\Gamma}(X) \triangle \Gamma_1^*(X)}
\]

\[
\geq \mu_0 \left( \frac{d\mu_1}{d\mu_0} \wedge 1 \right) \sum_{k=1}^{K} 1_{k \in \hat{\Gamma}_i^*(X) \triangle \Gamma_0^*(X)} \quad \text{(Triangle inequality)}
\]

\[
= 2s\mu_0 \left( \frac{d\mu_1}{d\mu_0} \wedge 1 \right) (1_{X_1} + 1_{X_2})
\]

\[
= 2s \int_{\mathbb{R}^d} \left( \frac{\mu_1(x)}{\mu_0(x)} \wedge 1 \right) (1_{X_1} + 1_{X_2}) d\mu_0(x)
\]

\[
= 2s \int_{X_1} \left( \frac{\mu_1(x)}{\mu_0(x)} \wedge 1 \right) d\mu_0(x) + 2s \int_{X_2} \left( \frac{\mu_1(x)}{\mu_0(x)} \wedge 1 \right) d\mu_0(x)
\]

\[
= 2s P_0(X_1 \cup X_2) \geq 2s \kappa_{n,N}.
\]

Substituting this lower bound into the initial inequality we arrive at

\[
(*) \geq 2s \kappa_{n,N} \mu_0^{\otimes (n+N)} \otimes P_0^{\otimes n} \left( \frac{d\mu_1^{\otimes (n+N)} \otimes P_0^{\otimes n} \otimes P_0|Y|X}{d\mu_0^{\otimes (n+N)} \otimes P_0^{\otimes n} \otimes P_0|Y|X} \wedge 1 \right)
\]

\[
= 2s \kappa_{n,N} \left( 1 - TV \left( \mu_0^{\otimes (n+N)} \otimes P_0|Y|X, \mu_1^{\otimes (n+N)} \otimes P_0|Y|X \right) \right)
\]

\[
= 2s \kappa_{n,N} \left( 1 - TV \left( \mu_0^{\otimes (n+N)}, \mu_1^{\otimes (n+N)} \right) \right)
\]

\[
\geq 2s \kappa_{n,N} \left( 1 - \sqrt{\frac{1}{2} KL \left( \mu_0^{\otimes (n+N)} \otimes (n+N) \right)} \right) \quad \text{( Pinsker’s inequality)}
\]

\[
\geq 2s \kappa_{n,N} \left( 1 - \kappa_{n,N} \sqrt{n+N} \right),
\]

which implies the desired lower bound on the Hamming risk.

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

\[
D(\hat{\Gamma}, \Gamma_i^*):= R_s(\hat{\Gamma}) - R_s(\Gamma_i^*) = \mu_i \sum_{k=1}^{K} \left| p_k(X) - G^{-1}(s) \right| 1_{\{k \in \hat{\Gamma}(X) \triangle \Gamma_i^*(X)\}}
\]
Again, recall that we are interested in
\[
\inf_{\Gamma} \sup_{P \in \mathcal{P}} \mathbb{E}_{(P, D_L, D_U)}[R_a(\hat{\Gamma})] - R(\Gamma^*_s)
\]
similarly to the previous case, since the hypotheses \(P_0, P_1 \in \mathcal{P}\) we can write
\[
2 \sup_{P \in \mathcal{P}} \mathbb{E}_{(P, D_L, D_U)}[R_a(\hat{\Gamma})] - R(\Gamma^*_s) \geq (**),
\]
where (**) is defined as
\[
(**) = \mu_0^{\otimes (n+N)} \otimes P_{X|Y}^{\otimes n} D(\hat{\Gamma}, \Gamma^*_0) + \mu_1^{\otimes (n+N)} \otimes P_{X|Y}^{\otimes n} D(\hat{\Gamma}, \Gamma^*_1),
\]
we can write
\[
(**) \geq \mu_0^{\otimes (n+N)} \otimes P_{X|Y}^{\otimes n} \left( \frac{d\mu_1^{\otimes (n+N)} \otimes P_{X|Y}^{\otimes n}}{d\mu_0^{\otimes (n+N)} \otimes P_{X|Y}^{\otimes n}} \wedge 1 \right) \left( D(\hat{\Gamma}, \Gamma^*_0) + D(\hat{\Gamma}, \Gamma^*_1) \right)
\]
and we continue in a similar fashion
\[
D(\hat{\Gamma}, \Gamma^*_0) + D(\hat{\Gamma}, \Gamma^*_1) = \mu_0 \sum_{k=1}^{K} p_k(X) - \frac{3K + 2s}{8Ks} 1_{k \in \hat{\Gamma}(X) \Delta \Gamma^*_0(X)} + \mu_1 \sum_{k=1}^{K} p_k(X) - \frac{K + 6s}{8Ks} 1_{k \in \hat{\Gamma}(X) \Delta \Gamma^*_1(X)} \geq \mu_0 \sum_{k=1}^{2s} p_k(X) - \frac{3K + 2s}{8Ks} 1_{k \in \hat{\Gamma}(X) \Delta \Gamma^*_0(X)} 1_{X \in X_2} + \mu_1 \sum_{k=1}^{2s} p_k(X) - \frac{K + 6s}{8Ks} 1_{k \in \hat{\Gamma}(X) \Delta \Gamma^*_1(X)} 1_{X \in X_2}.
\]
Set-valued classification

since \( \mu_0(x) = \mu_1(x) \) for all \( x \in X_2 \) we obtain

\[
D(\hat{\Gamma}, \Gamma_0^\ast) + D(\hat{\Gamma}, \Gamma_1^\ast) \\
\geq \mu_0 \left( \sum_{k=1}^{2s} p_k(X) - \frac{3K + 2s}{8Ks} \right) \mathbb{1}_{k\in\Gamma(X)\triangle \Gamma_1^\ast(X)} 1_{X\in X_2} \\
+ \sum_{k=1}^{2s} \left( p_k(X) - \frac{K + 6s}{8Ks} \right) \mathbb{1}_{k\in\Gamma(X)\triangle \Gamma_1^\ast(X)} 1_{X\in X_2} \\
\geq \mu_0 \left( \sum_{k=1}^{2s} \left( p_k(X) - \frac{3K + 2s}{8Ks} \right) \bigwedge \left( p_k(X) - \frac{K + 6s}{8Ks} \right) \mathbb{1}_{X\in X_2} \right) \\
= \mu_0 \left( \sum_{k=1}^{2s} \left( \frac{2s}{K + 4s} - \frac{K - 2s}{8Ks} \right) \bigwedge \left( \frac{3K + 2s}{8Ks} - \frac{K - 2s}{8Ks} \right) \mathbb{1}_{X\in X_2} \right) \\
= \mu_0 \left( \frac{2s}{K + 4s} - \frac{K - 2s}{8Ks} \bigwedge \left( \frac{3K + 2s}{8Ks} - \frac{K - 2s}{8Ks} \right) \mathbb{1}_{X\in X_2} \right) \\
= \mu_0 \left( \frac{2s}{K + 4s} - \frac{K - 2s}{8Ks} \bigwedge \left( \frac{3K + 2s}{8Ks} - \frac{K - 2s}{8Ks} \right) \mathbb{1}_{X\in X_2} \right)
\]

then, since \( \frac{\varphi(x)}{s} \leq \frac{K + 2s}{8Ks} \) for all \( x \in X_2 \), we have

\[
D(\hat{\Gamma}, \Gamma_0^\ast) + D(\hat{\Gamma}, \Gamma_1^\ast) \geq \frac{2s(K - 2s)}{16Ks} \mu_0(X_2) = \frac{K - 2s}{8K} \kappa_{n,N}.
\]

Thus,

\[
(\ast\ast) \geq \frac{K - 2s}{8K} \kappa_{n,N} \left( 1 - \text{TV} \left( \mu_0^{(n+N)} \otimes P_{Y|X}^{\otimes n}, \mu_1^{(n+N)} \otimes P_{Y|X}^{\otimes n} \right) \right) \\
\geq \frac{K - 2s}{8K} \kappa_{n,N} \left( 1 - \sqrt{\frac{1}{2} \text{KL} \left( \mu_0^{(n+N)} \otimes P_{Y|X}^{\otimes n}, \mu_1^{(n+N)} \otimes P_{Y|X}^{\otimes n} \right) } \right) \\
\geq \frac{K - 2s}{8K} \kappa_{n,N} \left( 1 - \kappa_{n,N} \sqrt{n + N} \right)
\]

Which concludes the first part of the lower bounds.

**E.2. Part II: \( n^{-\alpha\beta/(2\beta+d)} \)**

In this section we prove that in case of the Hamming risk \( \Psi^H \) the rate \( n^{-\alpha\beta/(2\beta+d)} \) is minimax optimal. Notice, that thanks to Proposition 5.2 a lower bound of order \( n^{-\alpha\beta/(2\beta+d)} \) on the Hamming risk \( \Psi^H \) immediately implies a lower bound of order \( n^{-(\alpha+1)\beta/(2\beta+d)} \) on both \( \Psi^B \) 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 C.1.
Figure 6: 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 $s \in \{2, \ldots, (K - 2) \wedge \lfloor K/2 \rfloor\}$, define the regular grid on $[0, 1]^d$ as

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

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 $X'_1, \ldots, X'_q$. Besides, denote by $X'_{-j} := \{x \in \mathbb{R}^d : -x \in X'_j\}$ for all $j = 1, \ldots, q^d$. For a fixed integer $m \leq q^d$ and for any $j \in \{1, \ldots, m\}$ define $X'_i := X'_{-i}, X'_i := X'_{-i}$. Additionally we introduce the following set $X_0 = B(0, (4q)^{-1})$. For every $w \in W := \{-1, 1\}^m$ we build the distribution $P_w \in \mathcal{P}_W$, such that, the marginal distribution $P_{w,X}$ is independent of $w \in \{-1, 1\}^m$ and the regression vector $(p_1^w(x), \ldots, p_K^w(x))$ is constructed as
Set-valued classification

\[ p^w_1(x) = \ldots = p^w_{s-1}(x) = v + \frac{c'}{s-1} + \frac{g(x)}{s-1}, \]

\[ p^w_s(x) = \begin{cases} 
  v + \phi(x), & \text{if } x \in X_0 \\
  v + w_i \varphi(x - n_q(x)), & \text{if } x \in X_i \\
  v - w_i \varphi(x - n_q(x)), & \text{if } x \in X_{-i} \\
 \end{cases} \]

\[ \frac{1}{K}, \quad \text{if } x \in B(0, \sqrt{d}) \setminus \left( \bigcup_{i=1}^{m} \left( \bigcup_{i \neq 0} X_i \right) \right), \]

\[ \frac{3v}{2} + g(x), \quad \text{if } x \in \mathbb{R}^d \setminus B(0, \sqrt{d} + \rho) \]

\[ v + \xi(x), \quad \text{if } x \in B(0, \sqrt{d} + \rho) \setminus B(0, \sqrt{d}) \]

\[ \frac{v}{2} - g(x), \quad \text{if } x \in \mathbb{R}^d \setminus B(0, \sqrt{d} + \rho) \]

\[ v - \xi(x), \quad \text{if } x \in B(0, \sqrt{d} + \rho) \setminus B(0, \sqrt{d}) \]

\[ p^w_{s+1}(x) = \ldots = p^w_K(x) = v - \frac{c'}{K - s - 1} - \frac{g(x)}{K - s - 1}, \]

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

\[ v = \frac{1}{K}, \quad c' = \frac{(s-1)(K-s-1)}{K^2} \]

The function \( \xi \) is constructed as

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

the function \( \bar{u} \) is infinitely many times differentiable, is equal to zero on \( (-\infty, 0] \) and to one on \([1, +\infty)\). Figure 6 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 \((\beta, L)\)-Hölder.

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

\[ \phi(x) = C_\phi (2q)^{-\beta} \left( \frac{\|x\|}{(2q)^{-1}} \right)^{2[\frac{\gamma}{q}]} \psi_{1,1} \left( \frac{\|x\|}{(2q)^{-1}} \right), \]

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

\[ \varphi(x) = C_\varphi q^{-\beta} \left( u_2 \left( \frac{\|x\|}{q^{-1}} \right) + \psi_{1,1} \left( \frac{\|x\|}{q^{-1}} \right) \right), \]
Figure 7 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 $(\beta, L)$- Hölder and and upper-bounded by $c'/2 \wedge v/4$. Notice that the function $\varphi(x)$ for all $x \in B(0,(4q)^{-1})$ satisfies

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

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

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

For $C_g$ chosen small enough to ensure that it is $(\beta, L)$- 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 $B(0,\sqrt{d} + \rho)$ 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

- $\mu(x) = \frac{\tau}{\text{Leb}(B(0,(4q)^{-1})))}$ for every $z \in G_q \cup \{0\}$ and every $x \in B(z,(4q)^{-1}))$ or $x \in B(-z,(4q)^{-1}))$, for every $\tau \in \mathbb{R}$,
• \( \mu(x) = \frac{1-2m}{\text{card}(A_0)} \) 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 \( p_{w1}^w, \ldots, p_{wK}^w \) are defining some regression function for every \( w \in W \). That is, for each \( x \in \mathbb{R}^d \) we have \( \sum_{k=1}^{K} p_{wk}^w(x) = 1 \) and \( 0 \leq p_{wk}^w(x) \leq 1 \),
• the functions \( p_{w1}^w, \ldots, p_{wK}^w \) are \((\beta, L)\)-Hölder,
• the function \( G_w(t) := \sum_{k=1}^{K} \int_{\mathbb{R}^d} 1_{\{p_{wk}^w(x) \geq t\}} \mu(x) dx \) is continuous,
• the threshold \( G_w^{-1}(s) \) 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

\[
p_{w+1}^w(x) + p_s^w(x) = 2v , \]
\[
\sum_{k=1}^{s-1} p_k^w(x) + \sum_{k=s+2}^{K} p_k^w(x) = (K-2)v ,
\]

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

\[
0 < v/2 \leq p_{s+1}^w(x) \leq 3v/2 \leq 1 , \quad 0 < v/2 \leq p_s^w(x) \leq 3v/2 \leq 1 ,
\]

and by construction of the function \( g \) we have for every \( k = 1, \ldots, s-1 \), every \( x \in \mathbb{R}^d \) and every \( w \in W \)

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

due to the choice of \( c', v \) we have

\[
v + \frac{3c'}{2(s-1)} = \frac{1}{K} + \frac{3(K-s-2)}{2K^2} \leq \frac{2}{K} \leq 1 .
\]

Similarly, for every \( k = s+2, \ldots, K \), every \( x \in \mathbb{R}^d \) and every \( w \in W \)

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

and with the choice of \( v, c' \) specified above and the constraint \( s \leq |K/2| \) we have

\[
v - \frac{3c'}{2(K-s-1)} = \frac{1}{K} - \frac{3(s-1)}{2K^2} \geq \frac{1}{K} - \frac{3(K/2-1)}{2K^2} = \frac{1}{4K} + \frac{3}{2K^2} \geq 0 .
\]
Thus, the construction above defines some regression function for every \( w \in W \).

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

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

\[
\int_{\mathbb{R}^d} 1_{p_k^w(x) \geq t} \mu(x) dx = \sum_{c \in G_s \cup G_q} \int_{B(c, (4q)^{-1})} 1_{p_k^w(x) \geq t} \mu(x) dx + \frac{1 - 2m \tau}{\text{Leb(A}_0)} \int_{A_0} 1_{p_k^w(x) \geq t} dx ,
\]

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

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

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

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

\[
\Gamma^s_w(x) = \begin{cases} 
\{1, \ldots, s-1, s\}, & x \in X_i, w_i = 1, \\
\{1, \ldots, s-1, s+1\}, & x \in X_i, w_i = -1, \\
\{1, \ldots, s-1, s\}, & x \in X_{-i}, w_i = -1, \\
\{1, \ldots, s-1, s+1\}, & x \in X_{-i}, w_i = 1, \\
\{1, \ldots, s-1, s\}, & x \in \mathbb{R}^d \setminus (\bigcup_{k=-m}^{m} X_i) 
\end{cases}
\]

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 \( k \in [K] \setminus \{s, s+1\} \) and all \( w \in W \) we have

\[
\mu \left( |p_k^w(X) - v| \leq t \right) = 0 ,
\]

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

\[
\mu \left( |p_k^w(X) - v| \leq t \right) = \sum_{i=-m}^{m} \mu \left( |p_k^w(X) - v| \leq t, X \in X_i \right) = \sum_{i=-m, i \neq 0}^{m} \mu \left( \varphi(X - n_q(X)) \leq t, X \in X_i \right) + \mu \left( \phi(X) \leq t, X \in X_0 \right) .
\]
Set-valued classification

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

\[
\mu(\phi(X) \leq t, X \in X_0) = \frac{\tau}{\text{Leb}(B(0, (4q)^{-1}))} \int_{B(0,(4q)^{-1})} 1_{\{\phi(X) \leq t\}} dx
\]

\[
= \frac{\tau}{\text{Leb}(B(0, (4q)^{-1}))} \int_{B(0,(4q)^{-1})} 1_{\{C_\phi(2q)^{-\beta} \left( \frac{x_1}{(2q)^{1/2}} \right)^{\beta/2} \leq t \}} dx
\]

\[
= \frac{C \tau q^{-d}}{\text{Leb}(B(0, (4q)^{-1}))} \int_{B(0, (4q)^{-1})} 1_{\{\|x\| \leq C \phi q^{-\beta} \}} dx ,
\]

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

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

Therefore for some constant \( C > 0 \) we can write

\[
\mu(\phi(X) \leq t, X \in X_0) \leq \frac{C \tau q^{-d}}{\text{Leb}(B(0, (4q)^{-1}))} \int_{B(0, 1/2)} 1_{\{\|x\| \leq C \phi^{-\beta/2} t^{1/2} \}} dx
\]

\[
\leq \frac{C \tau q^{-d(1-\beta/2)[\beta/2]}}{\text{Leb}(B(0, (4q)^{-1}))} t^{d/2}[\beta/2] ,
\]
	hanks{To the strong density assumption we can write for some \( C > 0 \)

\[
\mu(\phi(X) \leq t, X \in X_0) \leq C t^\alpha .
\]

Thus since \( 1 - \beta/2[\beta/2] \geq 0 \) and \( d/2[\beta/2] \geq \alpha \) we can write for some \( C > 0 \)

\[
\mu(\phi(X) \leq t, X \in X_0) \leq C t^\alpha .
\]

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

\[
\sum_{i=-m,i\neq0}^m \mu(\varphi(X-nq(X)) \leq t, X \in X_i) = \frac{2m \tau}{\text{Leb}(B(0, (4q)^{-1}))} \int_{B(0,(4q)^{-1})} 1_{\{\varphi(X) \leq t\}} dx ,
\]

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

\[
C \varphi q^{-\beta} \leq \varphi(x) \leq C \varphi q^{-\beta} \left( 1 + \psi_{-\frac{1}{4},1}(0) \right) \leq 2C \varphi q^{-\beta} ,
\]

we can write for all \( t \leq C \varphi q^{-\beta} \)

\[
\sum_{i=-m,i\neq0}^m \mu(\varphi(X-nq(X)) \leq t, X \in X_i) = 0 ,
\]
moreover, for all $t \geq 2Cq^{-\beta}$ 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 $t \in (Cq^{-\beta}, 2Cq^{-\beta})$ we can write

$$\sum_{i=-m, i \neq 0}^{m} \mu (\varphi(X - n_q(X)) \leq t, X \in \mathcal{X}_i) = 2m\tau \leq \text{Leb}(B(0, (4q-1)^{-1})) \int_{B(0, (4q-1)^{-1})} 1_{\{Cq^{-\beta} \leq t\}} dx \leq Cmq^{3\alpha}t^\alpha.$$

Thus the margin assumption is satisfied as long as

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

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

**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 $C > 0$

$$\text{KL}(\mathbb{P}_w, \mathbb{P}_{w'}) \leq 2 \sum_{i=-m, i \neq 0}^{m} \mu (\varphi(X - n_q(X)) \leq t, X \in \mathcal{X}_i) \leq 2\tau m 1_{\{t \leq 2Cq^{-\beta}\}} \leq C\tau mq^{3\alpha}t^\alpha.$$

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

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

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

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 $\log |\mathcal{W}| \geq \frac{m}{n}.$
Denote \( W \subset W \) the set provided by Lemma E.1 and by \( \mathcal{P}_W \) the set of distributions \( \mathbb{P}^w \) with \( w \in W \). Taking into account all the above we conclude that \( \mathcal{P}_W \) satisfies the assumptions of our result.

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

\[
\inf \sup_{\mathcal{P}} \mathbb{E}_{(\mathcal{D}_{\alpha}^L, \mathcal{D}_{\alpha}^U)} \mathbb{E}_{P_X} \left| \tilde{\Gamma}(X) \Delta \Gamma_s(X) \right|.
\]

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

\[
\inf \sup_{\mathcal{P}} \mathbb{E}_{(\mathcal{D}_{\alpha}^L, \mathcal{D}_{\alpha}^U)} \mathbb{E}_{P_X} \left| \tilde{\Gamma}(X) \Delta \Gamma_s(X) \right| \geq \inf \sup_{\mathcal{P}} \mathbb{E}_{P_X} \left| \tilde{\Gamma}(X) \Delta \Gamma_s(X) \right|.
\]

Denote by \( \hat{w} \) the following minimizer

\[
\hat{w} \in \arg \min_{w \in W} \left( \left| \tilde{\Gamma}(X) \Delta \Gamma_s(X) \right| \right),
\]

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

\[
2\mu \left( \left| \tilde{\Gamma}(X) \Delta \Gamma_s(X) \right| \right) \geq \mu \left( \left| \tilde{\Gamma}(X) \Delta \Gamma_s(X) \right| \right) + \mu \left( \left| \tilde{\Gamma}(X) \Delta \Gamma_s(X) \right| \right)
\]

\[
\geq \mu \left( \left| \tilde{\Gamma}(X) \Delta \Gamma_s(X) \right| \right) \geq 2\delta(w, \hat{w}) \mu(X_0)
\]

\[
= 2\delta(w, \hat{w}) \tau \geq \frac{m\tau}{2}.
\]

These arguments and Birgé’s lemma C.1 imply that

\[
\sup_{P \in \mathcal{P}} \mathbb{E}_{(\mathcal{D}_{\alpha}^L, \mathcal{D}_{\alpha}^U)} \mathbb{E}_{P_X} \left| \tilde{\Gamma}(X) \Delta \Gamma_s(X) \right| \geq \frac{m\tau}{4} \sup_{w \in W} \mu^\otimes N \otimes P_w^\otimes n \left( w \neq \hat{w} \right)
\]

\[
\geq \frac{m\tau}{4} \left( 0.29 \sqrt{1 - \frac{\sum_{w \in W \setminus \{w\}} \text{KL}(\mu^\otimes N \otimes P_w^\otimes n, \mu^\otimes N \otimes P_{\hat{w}}^\otimes n)}{|W - 1| \log |W|} } \right).
\]

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 \( W \) we get for some \( C > 0 \)

\[
\sup_{P \in \mathcal{P}} \mathbb{E}_{(\mathcal{D}_{\alpha}^L, \mathcal{D}_{\alpha}^U)} \mathbb{E}_{P_X} \left| \tilde{\Gamma}(X) \Delta \Gamma_s(X) \right| \geq \frac{m\tau}{4} \left( 1 - Cn\tau q^{-2\beta} \right).
\]

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

\[
\sup_{P \in \mathcal{P}} \mathbb{E}_{(\mathcal{D}_{\alpha}^L, \mathcal{D}_{\alpha}^U)} \mathbb{E}_{P_X} \left| \tilde{\Gamma}(X) \Delta \Gamma_s(X) \right| \geq Cn^{-\alpha\beta/(2\beta + d)} (1 - c).
\]

One can easily verify that this choice of parameters \( \tau, m, q \) is possible as long as \( 2\lfloor \frac{\beta}{2} \rfloor \alpha d \leq d \) and clearly with our choice we have \( \tau m = O(q^{-\alpha\beta}) \). As already mentioned the lower bound for the excess risk and the discrepancy follows from Propositions 5.2 and 5.3.
Appendix F: Upper bound under extra assumption

Let us first state the result describing the deviation of uniform distribution from its empirical counterpart in Wasserstein infinity distance.

**Lemma F.1.** Let $\mu$ be the uniform distribution on $(0,1)$, and $\mu_n$ be its empirical counterpart based on $n$ i.i.d. samples from $\mu$. Then, for any $\alpha > 0$ there exists $C_\alpha > 0$ such that for all $n > 0$ it holds that

$$
E[W^{1+\alpha}_\infty(\mu, \mu_n)] \leq C_\alpha n^{-\frac{1+\alpha}{2}}.
$$

**Proof.** It is easy to verify that if $F$ and $F_n$ are CDF and empirical CDF of $\mu$, then

$$
W_\infty(\nu, \nu_n) = \|F - F_n\| := \sup_{t \in (0,1)}|F(t) - F_n(t)|.
$$

Moreover, thanks to DKW inequality we have for all $r > 0$

$$
P(\|F - F_n\| \geq r) \leq 2 \exp\left(-2nr^2\right).
$$

Thus, we can write

$$
E[W^{1+\alpha}_\infty(\nu, \nu_n)] = \int_0^\infty P(\|F - F_n\| \geq r \frac{1}{\sqrt{n}}) \, dr \leq 2 \int_0^\infty \exp\left(-2 \left(n^{\frac{1+\alpha}{2}} r\right)^\frac{2}{1+\alpha}\right) \, dr = 2n^{-\frac{1+\alpha}{2}} \int_0^\infty \exp\left(-2r^{\frac{2}{1+\alpha}}\right) \, dr,
$$

since $\int_0^\infty \exp\left(-2r^{\frac{2}{1+\alpha}}\right) \, dr = c_\alpha < \infty$ we conclude.

**Proof of Theorem 6.11.** Proof of this result closely follows the one presented in (Bobkov and Ledoux, 2016). We observe that thanks to the convexity of Wasserstein infinity distance it holds that

$$
W^{1+\alpha}_\infty(\mu, \mu_n) \leq E[W^{1+\alpha}_\infty(\mu, \mu_n)],
$$

where $\mu_n'$ is an independent copy of $\mu_n$, and the expectation is taken w.r.t. the distribution of $\mu_n'$. Finally, Theorem 5.11 in (Bobkov and Ledoux, 2016) states that

$$
W^{1+\alpha}_\infty(\mu_n', \mu_n) \leq \|F_{\mu}^{-1}\|_{\text{Lip}} W^{1+\alpha}_\infty(\nu_n', \nu_n),
$$

where $\nu_n, \nu_n'$ are independent empirical CDFs of uniform distribution from Lemma F.1. Thus, using the triangle inequality and denoting by $\nu$ the uniform measure on $(0,1)$ we can write

$$
W^{1+\alpha}_\infty(\nu_n', \nu_n) \leq 2^{1+\alpha} W^{1+\alpha}_\infty(\nu, \nu_n).
$$
Finally, thanks to Lemma F.1 we have
\[ E[W_\infty^{1+\alpha}(\mu, \mu_n)] \leq 2^{1+\alpha} \| F_\mu^{-1} \|_{\text{Lip}}^{1+\alpha} E[W_\infty^{1+\alpha}(\nu, \nu_n)] \leq C_\alpha 2^{1+\alpha} \| F_\mu^{-1} \|_{\text{Lip}}^{1+\alpha} n^{-\frac{1+\alpha}{2}}, \]
and we conclude by setting \( c = C_\alpha 2^{1+\alpha}. \)

**Proof of Theorem 6.9.** Recall that according to Proposition 5.2 the excess risk of any \( \Gamma \) is expressed as
\[ R_n(\Gamma) - R_n(\Gamma^*_n) = \sum_{k=1}^K E_{P_X} \left[ |p_k(X) - G^{-1}(s)| \mathbf{1}_{(k \in \Gamma(X) \cap \Gamma^*_n(X))} \right] . \]
Thus, for any \( P \in \mathcal{P}(L, \beta, \alpha, M) \) we can write for \( \hat{\Gamma} \) defined in Section 6.1 that
\[ R_n(\Gamma) - R_n(\Gamma^*_n) \leq \sum_{k=1}^K E_{P_X} \left[ |p_k(X) - G^{-1}(s)| \mathbf{1}_{\{|p_k(X) - G^{-1}(s)| \leq 2|p_k(X) - \hat{p}_k(X)|\}} \right] . \]

The first term \( R_1 \) on the right hand side of the above inequality can be handled similarly to (Audibert and Tsybakov, 2007). To handle the second term \( R_2 \) we first define
\[ G_{n,N}(\cdot) = \frac{1}{\lfloor n/2 \rfloor} + N \sum_{X \in D^\alpha_n \cup D^\alpha_{[n/2]}} \sum_{k=1}^K \mathbf{1}_{\{|p_k(X) - G^{-1}(s)| \leq 2|G^{-1}(s) - \hat{G}^{-1}(s)|\}} , \]
which is seen as an empirical version of \( G(\cdot) \) without knowledge of the marginal distribution \( P_X \). Therefore we continue as
\[ R_2 \leq \sum_{k=1}^K E_{P_X} \left[ |p_k(X) - G^{-1}(s)| \mathbf{1}_{\{|p_k(X) - G^{-1}(s)| \leq 4|G^{-1}(s) - G_{n,N}(s)|\}} \right] . \]

For \( R_2^2 \) we use Lemma D.1 proven in Appendix and obtain the bound
\[ |G^{-1}(s) - G_{n,N}(s)| \leq \| p - \hat{p} \|_\infty . \]
Thus $R_2^2$ can be handled using similar machinery as in the proof of Theorem 6.1. Therefore, it remain to bound $R_1^1$. For this term we note that thanks to the margin assumption we have

$$R_1^1 \leq KC_14^{1+\alpha}\left|G^{-1}(s) - G_{n,N}^{-1}(s)\right|^{1+\alpha}.$$

Using Theorem 6.13 we get for some $C > 0$

$$E_{(D_n,D_N)}[R_1^1] \leq CC_1K^{2+\alpha}(n + N)^{-\frac{1+\alpha}{2}}.$$

Finally, recall that $R_1$ and $R_2^2$ can be bounded for some $C' > 0$ as

$$E_{(D_n,D_N)}[R_2 + R_2^2] \leq C' \left(\frac{n}{\log n}\right)^{-\frac{(1+\alpha)\beta}{2\beta + d}} + C(n + N)^{-\frac{1+\alpha}{2}}.$$

which implies that for some $C > 0$ that is independent from both $n, N \in \mathbb{N}$ we have

$$\Psi_{n,N}(\hat{\Gamma}; \mathcal{P}(L, \beta, \alpha, M)) \leq C \left(\frac{n}{\log n}\right)^{-\frac{(1+\alpha)\beta}{2\beta + d}} + C(n + N)^{-\frac{1+\alpha}{2}}.$$

To conclude we notice that since $\beta/(2\beta + d) < 1/2$, then for all $N \in \mathbb{N}$ it holds that

$$(n + N)^{-\frac{1+\alpha}{2}} < \left(\frac{n}{\log n}\right)^{-\frac{(1+\alpha)\beta}{2\beta + d}}.$$