Plug-in methods in classification

Evgenii Chzhen

UNIVERSITÉ PARIS-EST

UNIVERSITÉ PARIS-EST
ÉCOLE DOCTORALE MSTIC

Thèse de doctorat
Spécialité : Mathématiques Appliquées

Evgenii Chzhen

Plug-in methods in classification

Thèse dirigée par: Mohamed Hebiri
Florence Merlevède
Joseph Salmon

Soutenue publiquement le 25 Septembre 2019, devant le jury composé de

Cristina	**Butucea**	ENSAE CREST	Examinatrice
Arnak	**Dalalyan**	ENSAE CREST	Examinateur
Christophe	**Giraud**	Université Paris-Sud	Rapporteur
Mohamed	**Hebiri**	Université Paris-Est	Directeur
Clément	**Marteau**	Université Lyon I	Rapporteur
Florence	**Merlevède**	Université Paris-Est	Directrice
Joseph	**Salmon**	Université de Montpellier	Directeur
Alexandre	**Tsybakov**	ENSAE CREST	Président

Acknowledgements

I would like to express my sincere gratitude to all my supervisor. Thank you Florence for agreeing to be my main supervisor, I would not be where I am now without your support. Joseph and Mohamed – supervisors, mentors, colleagues, friends. It is really difficult to overestimate your impact on me. Both of you were always there when I needed it the most. Maybe, what is more important, you were there for me when I did not even understand that I need it! Looking back I realize that I did not appreciate enough all the scientific and life lessons that you gave me. As we say in Russia "better late than never". Thank you Joseph! Thank you Mohamed!

Thank you Christophe Giraud and Clément Marteau for taking your time to review this manuscript. To all of my jury Cristina Butucea, Arnak Dalalyan, and Alexander Tsybakov to have accepted to be on my defence. I am extremely thankful to all of you! Thank you Cristina, it was you who introduced me to the modern statistics with you course in Marne. Thank you Arnak, for giving me the opportunity to teach with you, for writing me the recommendations, and for your continuous support! Thank you Sasha, your research served me as the main inspiration throughout these years, I am looking forward to teaching with you. Thank you Christophe Giraud for agreeing to have me in your group as a postdoc.

Thank you Christophe Denis with whom we have shared a lot of scientific moments together. I am grateful to Zaid Harchaoui and his group at UW for teaching me optimization and giving me a taste of machine learning. Thank you Luca Oneto and Massimialiano Pontil, it is a big pleasure to work with you! I would like to thank everyone from LAMA! Thank you Audrey and Christiane, the best and the most efficient duo in the lab! I would like to thank all the professors and teacher at MIPT. Especially to my teachers of calculus, quantum mechanics, and functional analysis, these courses were amazing. Thank you Alexander Komech – my first mentor. You introduced me to research, Hartree-Fock equations, distribution theory, Sobolev spaces, and France.

To all my friends (даже четвертому) who were constantly teaching and inspiring me!

I would like to thank my parents for their support and wisdom.

Mónika, would all of this be even possible without you? Спасибо, szeretlek!

Contents

1 Introduction **4**
- 1.0.1 Examples of settings . . . 8
- 1.0.2 Empirical Risk Minimization . . . 9
- 1.0.3 Plug-in classifiers . . . 12

1.1 Binary classification . . . 16
- 1.1.1 Standard setup . . . 17
- 1.1.2 F-score setup (Section 2.1) . . . 18
- 1.1.3 Constrained classification: general framework . . . 20
- 1.1.4 Constrained classification: fairness (Section 2.2) . . . 23

1.2 Multi-class classification . . . 26
- 1.2.1 Standard setup . . . 26
- 1.2.2 Confidence set setup (Chapter 3) . . . 27

1.3 Multi-label classification (Chapter 4) . . . 27

1.4 Organization of the manuscript . . . 29

1.5 Resumé en français . . . 30
- 1.5.1 F-score . . . 30
- 1.5.2 Classification équitable . . . 31
- 1.5.3 Ensembles des confiances . . . 32
- 1.5.4 Multi-label . . . 32

2 Binary classification **34**

2.1 F-score . . . 34
- 2.1.1 Introduction . . . 34
- 2.1.2 The problem formulation . . . 35
- 2.1.3 Related works and contributions . . . 38
- 2.1.4 Main results . . . 39
- 2.1.5 Conclusion . . . 44
- 2.1.6 Proofs . . . 44

2.2 Fair binary classification . . . 55
- 2.2.1 Introduction . . . 55
- 2.2.2 Optimal Equal Opportunity classifier . . . 57
- 2.2.3 Proposed procedure . . . 59
- 2.2.4 Consistency . . . 61
- 2.2.5 Experimental results . . . 63
- 2.2.6 Optimal classifier independent of sensitive feature . . . 65
- 2.2.7 Conclusion . . . 67
- 2.2.8 Proofs . . . 67

3 Multi-class classification **89**
3.1 Confidence set approach . . . 89
3.1.1 Introduction . . . 89
3.1.2 Main contributions . . . 96
3.1.3 Class of confidence sets . . . 97
3.1.4 Lower bounds . . . 98
3.1.5 Upper bounds . . . 102
3.1.6 Discussion . . . 109
3.1.7 Conclusion . . . 110
3.1.8 Proofs . . . 111

4 Multi-label classification **136**
4.1 Constrained approach . . . 136
4.1.1 Introduction . . . 136
4.1.2 Framework and notation . . . 138
4.1.3 Control over sparsity . . . 139
4.1.4 Control over false positives . . . 141
4.1.5 Discussion . . . 145
4.1.6 Conclusion . . . 146
4.1.7 Proofs . . . 146

5 Concluding remarks **153**

6 Appendix **154**

Chapter 1

Introduction

The problem of classification is one of the most popular and the most classical problems of statistics and pattern recognition [Devroye et al., 1996, Vapnik, 1998]. In this manuscript, it is assumed that two data samples are provided – a labeled one $\mathcal{D}_n^{\mathrm{L}} = \{(X_i, Y_i)\}_{i=1}^{n}$ and an unlabeled one $\mathcal{D}_N^{\mathrm{U}} = \{X_i\}_{i=n+1}^{n+N}$ with some $n \in \mathbb{N}$ and $N \in \mathbb{N} \cup \{0\}$. In classification framework, for each $i \in \{1, \ldots, n+N\}$ the element X_i belongs to some feature space $\mathcal{X}$ and for each $i \in \{1, \ldots, n\}$ the element Y_i belongs to some finite space $\mathcal{Y}$ of labels or classes. Typically, each pair $(X_i, Y_i) \in \mathcal{D}_n^{\mathrm{L}}$ is called a labeled (supervised) observation and each vector $X_i \in \mathcal{D}_N^{\mathrm{U}}$ is called an unlabeled (unsupervised) observation. The high level objective is to construct an algorithm $\hat{g}$, using the data samples $\mathcal{D}_n^{\mathrm{L}}, \mathcal{D}_N^{\mathrm{U}}$, which for a new observation $X \in \mathcal{X}$ outputs its label Y. At this point several questions can be asked about this formulation:

- What is the nature of our data $\mathcal{D}_n^{\mathrm{L}}, \mathcal{D}_N^{\mathrm{U}}$?
- What is a new observation X and what does it mean to predict its label Y? In particular, what does it mean that X has a label Y?
- What is an algorithm $\hat{g}$ and what do we mean by saying that it is based on the data? How do we know that $\hat{g}$ is "good"?

A possible way to address all of the above questions is provided by the theory of statistics, which poses certain assumptions on the data generating process, furthermore, it gives a precise way to define the notion of an algorithm $\hat{g}$ and its goodness. Concerning the nature of the data, it is assumed that the data generating process is probabilistic. That is, we assume that there exists a random pair $(X, Y) \sim \mathbb{P}$ such that $X \sim \mathbb{P}_X$ and $Y|X \sim \mathbb{P}_{Y|X}$ and the observed samples $\mathcal{D}_n^{\mathrm{L}}, \mathcal{D}_N^{\mathrm{U}}$ are *i.i.d.* from $\mathbb{P}$ (*i.e.*, $\mathcal{D}_n^{\mathrm{L}} \overset{i.i.d.}{\sim} \mathbb{P}$) and from $\mathbb{P}_X$ (*i.e.*, $\mathcal{D}_N^{\mathrm{U}} \overset{i.i.d.}{\sim} \mathbb{P}_X$) respectively.

Remark. *In this context there are two types of methods that can be constructed* – supervised *ones and* semi-supervised *ones. While the former methods are based only on the labeled dataset $\mathcal{D}_n^{\mathrm{L}}$, the later can also leverage the information provided by the unlabeled dataset $\mathcal{D}_N^{\mathrm{U}}$. Let us point out that most of the classical settings do not assume the availability of the unlabeled dataset $\mathcal{D}_N^{\mathrm{U}}$. However, in some setups one can rigorously justify the introduction of the* unlabeled *set and in several contributions of the present manuscript this dataset is used to build* semi-supervised *algorithms.*

Under the above probabilistic assumptions, a generic classification problem can be described by a tuple $(\mathcal{X}, \mathcal{Y}, \mathcal{A}, \mathcal{R}, \mathcal{G}_\theta, \mathcal{P})$ where[1]

- $\mathcal{X}$ is a measurable space of features (typically $\mathbb{R}^d$ with Borel sigma-algebra);
- $\mathcal{Y}$ is the label space, such that $|\mathcal{Y}| < +\infty$, where $|\mathcal{Y}|$ denotes the cardinal of the set $\mathcal{Y}$;
- $\mathcal{A}$ is a set of possible predicted values;
- $\mathcal{G} := \mathcal{G}(\mathcal{X}, \mathcal{A})$ is the set of all measurable functions from $\mathcal{X}$ to $\mathcal{A}$, which are called prediction rules or classifiers;
- $\mathcal{R} : \mathcal{G} \to \mathbb{R}_+$ is a risk function, that describes the performance of any prediction rule;
- $\mathcal{G}_\theta \subset \mathcal{G}(\mathcal{X}, \mathcal{A})$ is a subset of measurable functions (predictions with certain properties), parametrized by some *known* parameter θ belonging to some topological space Θ;
- $\mathcal{P}$ is a family of possible joint distributions on $\mathcal{X} \times \mathcal{Y}$.

In several standard formulations of classification problems $\mathcal{A} = \mathcal{Y}$ and $\mathcal{G}_\theta = \mathcal{G}(\mathcal{X}, \mathcal{Y})$. For instance, in standard binary classification settings we have $\mathcal{A} = \mathcal{Y} = \{0, 1\}$ and $\mathcal{G}_\theta = \mathcal{G}(\mathcal{X}, \mathcal{Y})$ is the set of all binary valued functions from $\mathcal{X}$. However, such a formulation does not allow to model more complex situations that emerge from applications. For example, in binary classification with reject option [Chow, 1957, 1970, Herbei and Wegkamp, 2006], the set of predicted outcomes is given by $\mathcal{A} = \{0, 1, ®\}$ where the symbol ® is interpreted as reject, that is, a definitive prediction is not provided.

The set $\mathcal{G}_\theta$ carries all the properties that are expected by a statistician from an optimal classifier. Such properties might actually depend on the unknown distribution $\mathbb{P}$ and thus the main difficulty is arising from the fact that the class $\mathcal{G}_\theta$ is unknown beforehand. For instance, returning to the binary classification setup with $\mathcal{Y} = \{0, 1\}$ and $\mathcal{G} = \{g : \mathcal{X} \to \{0, 1\}\}$ we can define $\mathcal{G}_\theta$ as

$$\mathcal{G}_\theta = \{g \in \mathcal{G} \, : \, \mathbb{P}_X(g(X) = 1) = \theta\} \quad \text{(toy example)} \ , \tag{1.1}$$

for some $\theta \in \Theta = [0, 1]$. For each fixed marginal distribution $\mathbb{P}_X$, this set consists of those predictions g whose probability to predict 1 is equal to some predefined parameter θ. Clearly, different marginal distributions $\mathbb{P}_X$ yield different sets $\mathcal{G}_\theta$ (sometimes $\mathcal{G}_\theta$ can be empty and we discuss this issue later in the introduction). Moreover, since the marginal distribution $\mathbb{P}_X$ is unknown beforehand, the whole set $\mathcal{G}_\theta$ of predictions with prescribed properties is unknown. This toy example reveals the main difficulty connected with the introduction of the set $\mathcal{G}_\theta$. We will further use this example to explain several notions that are arising in the constrained estimation framework.

In this probabilistic framework, there is at least one classifier g^* which is seen superior to others. We call this classifier g^* as Bayes optimal predictor: it minimizes the risk $\mathcal{R}(\cdot)$ over the set of classifiers with prescribed properties. Formally, a Bayes optimal predictor g^* satisfies

$$g^* \in \arg\min \{\mathcal{R}(g) \, : \, g \in \mathcal{G}_\theta\} \quad \text{(Bayes rule)} \ .$$

[1]Here and later it is assumed that all spaces are (linear) topological and are equipped with their Borel sigma-algebra. All measures are assumed to be Borel measures.

This Bayes rule is seen as a theoretical benchmark in the sense that there is no other classifier $g \in \mathcal{G}_\theta$ with lower risk. If we consider, the standard setup of binary classification, the most common choice of the risk $\mathcal{R}(\cdot)$ is given by the probability of misclassification as

$$\mathcal{R}(g) = \mathbb{P}(Y \neq g(X)) \ .$$

Importantly, in the context of the present manuscript, the set $\mathcal{G}_\theta$ should not be confused with the family of concepts typically considered in (agnostic) PAC-learning literature. Unlike (agnostic) PAC settings, we shall consider classes $\mathcal{G}_\theta$ which are potentially much larger than standard classes considered in the learning community [Vapnik, 1998]. For instance, typical classes studied in the learning literature are half-spaces (linear classifiers) or some Boolean forms of classes with finite VC-dimension [Vapnik and Chervonenkis, 1971, Blumer et al., 1989].

One of the goals of a practitioner is to mimic the Bayes rule. To this end, we need to build an algorithm or an estimator $\hat{g}$ which is a measurable mapping defined as

$$\hat{g} : \bigcup_{n,N \in \mathbb{N}} (\mathcal{X} \times \mathcal{Y})^n \times \mathcal{X}^N \to \mathcal{G}(\mathcal{X}, \mathcal{A}) \quad \text{(estimator)} \ .$$

In other words, an estimator maps each possible combination of the input data to a prediction rule. A proper way of writing $\hat{g}$ evaluated at some point $x \in \mathcal{X}$ would be $(\hat{g}(\mathcal{D}_n^{\mathrm{L}}, \mathcal{D}_N^{\mathrm{U}}))(x)$, however, in this manuscript we omit the dependence on $\mathcal{D}_n^{\mathrm{L}}, \mathcal{D}_N^{\mathrm{U}}$ and write $\hat{g}(x)$ instead.

Once the notion of an estimator is defined, it remains to understand how to compare different algorithms, that is, how to evaluate their performances. One of the possible ways to assert the performance of an algorithm $\hat{g}$ is to introduce the notion of excess risk defined as

$$\mathbb{E}_{(\mathcal{D}_n^{\mathrm{L}}, \mathcal{D}_N^{\mathrm{U}})}[\mathcal{E}(\hat{g})] := \mathbb{E}_{(\mathcal{D}_n^{\mathrm{L}}, \mathcal{D}_N^{\mathrm{U}})} \left| \mathcal{R}(\hat{g}) - \inf_{g \in \mathcal{G}_\theta} \mathcal{R}(g) \right| \quad \text{(excess risk)} \ ,$$

where the expectation is taken *w.r.t.* the distribution of data $\mathcal{D}_n^{\mathrm{L}}, \mathcal{D}_N^{\mathrm{U}}$. For a good algorithm $\hat{g}$, the sequence $\mathbb{E}_{(\mathcal{D}_n^{\mathrm{L}}, \mathcal{D}_N^{\mathrm{U}})}[\mathcal{E}(\hat{g})]$ should decrease with the growth of n, N as fast as possible. Besides, since our original Bayes rule lies in some space $\mathcal{G}_\theta$, a constructed estimator $\hat{g}$ should not deviate from this set too much. In other words, for a good algorithm $\hat{g}$ there exists $g \in \mathcal{G}_\theta$ such that

$$\hat{g} \longrightarrow g \in \mathcal{G}_\theta \text{ as } n, N \to \infty \quad \text{(prescribed property)} \ ,$$

where the convergence will be specified depending on the considered context.

Note that the notion of the excess risk has the absolute value in its definition, this is due to the fact that in some scenarios one can not guarantee the inclusion of the algorithm $\hat{g}$ in the set $\mathcal{G}_\theta$. Indeed, recall the toy example of binary classification with $\mathcal{G}_\theta$ defined in Eq. (1.1) as

$$\mathcal{G}_\theta = \{g \in \mathcal{G} \ : \ \mathbb{P}_X(g(X) = 1) = \theta\} \ .$$

Since $\mathcal{G}_\theta$ is distribution dependent in this setup, our goal is to build an algorithm $\hat{g}$ which converges to some member of $\mathcal{G}_\theta$. Note that in this case, as well as in various other settings, the set $\mathcal{G}_\theta$ is described by a system of constraints (inequalities or equalities). Therefore, a possible strategy is to construct an algorithm which does not violate these constraints "too much" in finite sample regime or satisfies these constraints asymptotically as n and N grow.

For example, returning to the above set $\mathcal{G}_\theta$ (introduced in Eq. (1.1)), a good algorithm $\hat{g}$ would satisfy

$$\mathbb{E}_{(\mathcal{D}_n^{\mathrm{L}},\mathcal{D}_N^{\mathrm{U}})} \left|\mathbb{P}_X(\hat{g}(X)=1)-\theta\right| \to 0 \ .$$

This thesis is mostly concerned with minimax settings of the above problem, that is, given some family of joint distributions $\mathcal{P}$ on $\mathcal{X}\times\mathcal{Y}$ we study the following minimax risks

$$\inf_{\hat{g}} \max_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{(\mathcal{D}_n^{\mathrm{L}},\mathcal{D}_N^{\mathrm{U}})} \left[\mathcal{E}(\hat{g})\right] \quad \text{(minimax excess risk)} \ ,$$

where the infimum is taken over all possible estimators and the excess risk is based on $\mathcal{R}(\cdot)$. As we shall see, in some situations we can show that the Bayes optimal rule g^* is a minimizer of some *other* θ-risk $\mathcal{R}_\theta(\cdot)$ over *all* possible predictions. Formally, for some problems we can show that the Bayes optimal prediction satisfies

$$g^* \in (\arg\min\{\mathcal{R}_\theta(g) \ : \ g\in\mathcal{G}\}) \bigcap (\arg\min\{\mathcal{R}(g) \ : \ g\in\mathcal{G}_\theta\}) \neq \emptyset \ .$$

In particular, under certain assumptions on the distribution $\mathbb{P}$ and for specific choices of $\mathcal{G}_\theta$, we can show that

- the Bayes rule g^* satisfies the property prescribed by $\mathcal{G}_\theta$, that is, $g^*\in\mathcal{G}_\theta$;
- if g satisfies the property prescribed by $\mathcal{G}_\theta$ (*i.e.*, $g\in\mathcal{G}_\theta$), then we have $\mathcal{R}(g^*)\leq\mathcal{R}(g)$;
- for *all* predictions $g\in\mathcal{G}$, we have $\mathcal{R}_\theta(g^*)\leq\mathcal{R}_\theta(g)$.

Let us provide a simple example when such $\mathcal{R}_\theta$ is available. For some $\theta\in(0,1)$, consider the set $\mathcal{G}_\theta$ defined in Eq. (1.1) and the following Bayes rule

$$g^* \in \arg\min\{\mathcal{R}(g) \ : \ \mathbb{P}(g(X)=1)=\theta\} \ ,$$

with the misclassification risk $\mathcal{R}(g) := \mathbb{P}(Y\neq g(X))$. We also assume that the Cumulative Distribution Function (CDF) of $\eta(X) := \mathbb{E}[Y|X]$ defined as $G_{\eta(X)}(t) := \mathbb{P}(\eta(X)\leq t)$ is continuous on $(0,1)$ and we denote by $\bar{G}_{\eta(X)}(\cdot) = 1-G_{\eta(X)}(\cdot)$ its complement. Then, we can show that the following statement hold true

Statement. *Under the above assumptions we have*

- *the set $\mathcal{G}_\theta$ is not empty and there exists a Bayes rule $g^*\in\mathcal{G}_\theta$;*
- *a Bayes rule g^* can be written for all $x\in\mathbb{R}^d$ as*

$$g^*(x) = \mathbb{1}_{\left\{\eta(x)\geq\frac{1+\lambda^*}{2}\right\}} \ , \tag{1.2}$$

  where $\lambda^* = 2\bar{G}^{-1}_{\eta(X)}(\theta)-1$ with $\bar{G}^{-1}_{\eta(X)}(\cdot)$ being the (generalized) inverse of $\bar{G}_{\eta(X)}(\cdot)$;
- *the classifier g^* in Eq. (1.2) satisfies*

$$g^* \in \arg\min\left\{\underbrace{\mathbb{P}(Y\neq g(X)) + \lambda^*\mathbb{P}(g(X)=1)}_{\theta\text{-risk } \mathcal{R}_\theta(\cdot)} \ : \ g\in\mathcal{G}\right\} \ ,$$

$$g^* \in \arg\min\left\{\underbrace{\mathbb{P}(Y\neq g(X))}_{\text{risk } \mathcal{R}(\cdot)} \ : \ \mathbb{P}(g(X)=1)=\theta\right\} \ .$$

Proof. The proof goes by application of the general constrained classification framework described in Section 1.1.3. □

Notice that in this case the θ-risk $\mathcal{R}_\theta$ is given as

$$\mathcal{R}_\theta(g) = \mathbb{P}(Y \neq g(X)) + \lambda^* \mathbb{P}(g(X) = 1) \ .$$

Thus, we have seen that the problem of constrained classification can be reduced to a study of the cost-sensitive risk $\mathcal{R}_\theta$. However, it is important to note that, unlike classical cost-sensitive risks, here the value of λ^* is *unknown* beforehand.

If we can show that g^* minimizes some $\mathcal{R}_\theta$ over all classifiers, then we rather focus on another excess risk given as

$$\mathbb{E}_{(\mathcal{D}_n^{\mathrm{L}}, \mathcal{D}_N^{\mathrm{U}})}[\mathcal{E}_\theta(\hat{g})] := \mathbb{E}_{(\mathcal{D}_n^{\mathrm{L}}, \mathcal{D}_N^{\mathrm{U}})}[\mathcal{R}_\theta(\hat{g})] - \inf_{g \in \mathcal{G}} \mathcal{R}_\theta(g) \quad \text{(excess } \theta\text{-risk)} \ ,$$

and study the minimax version of $\mathbb{E}_{(\mathcal{D}_n^{\mathrm{L}}, \mathcal{D}_N^{\mathrm{U}})}[\mathcal{E}_\theta(\hat{g})]$. This notion of the excess risk reflects the common wisdom that the Bayes optimal classifier is minimizing the risk (θ-risk in this case) over *all* possible classifiers in $\mathcal{G}$. Finally, we can incorporate both constraints and risk errors into a one excess risk combining both $\mathcal{E}(\cdot)$ and the constraint violations. Returning again to our example $\mathcal{G}_\theta$ from Eq. (1.1) we can consider the following discrepancy of an algorithm $\hat{g}$

$$\mathbb{E}_{(\mathcal{D}_n^{\mathrm{L}}, \mathcal{D}_N^{\mathrm{U}})}[\mathcal{E}^{\mathrm{D}}(\hat{g})] = \mathbb{E}_{(\mathcal{D}_n^{\mathrm{L}}, \mathcal{D}_N^{\mathrm{U}})}[\mathcal{E}(\hat{g})] + \mathbb{E}_{(\mathcal{D}_n^{\mathrm{L}}, \mathcal{D}_N^{\mathrm{U}})} |\mathbb{P}_X(\hat{g}(X) = 1) - \theta| \quad \text{(discrepancy)} \ .$$

The first term of the above discrepancy controls the deviation of $\hat{g}$ from the Bayes rule g^* in terms of the risk $\mathcal{R}(\cdot)$, while the second term controls the violation of the constraints.

1.0.1 Examples of settings

In this part we mainly focus on possible properties that might be required from the Bayes optimal classifier. Most of the examples below will be discussed in details later in this chapter. What follows is a brief description of popular (constrained) classification frameworks. Yet, by no means this list should be regarded as exhaustive since more and more (constrained) classification frameworks are emerging with time. The list below is separated into three parts: Binary classification; Multi-class classification; Multi-label classification.

- Binary classification, $\mathcal{Y} = \{0, 1\}$:
  - Standard setup:
    Classifier: $g : \mathbb{R}^d \to \{0, 1\}$
    Risk: $\mathbb{P}(Y \neq g(X))$
    Property: no properties required
  - F-score setup:
    Classifier: $g : \mathbb{R}^d \to \{0, 1\}$
    Risk: $1 - \frac{2\mathbb{P}(Y=1, g(X)=1)}{\mathbb{P}(Y=1)+\mathbb{P}(g(X)=1)}$
    Property: no properties required
  - Reject option [Chow, 1957]: ® stands for reject
    Classifier: $g : \mathbb{R}^d \to \{0, 1, ®\}$
    Risk: $\mathbb{P}\left(Y \neq g(X) \,|\, g(X) \neq ®\right)$
    Property: $\mathbb{P}(g(X) = ®) \leq \alpha$

- Fairness with equal opportunity [Hardt et al., 2016]: $S \in \{0,1\}$ is a sensitive attribute
  Classifier: $g : \mathbb{R}^d \times \{0,1\} \to \{0,1\}$
  Risk: $\mathbb{P}\left(Y \neq g(X,S)\right)$
  Property: $\mathbb{P}\left(g(X,S)=1 \,|\, Y=1, S=1\right) = \mathbb{P}\left(g(X,S)=1 \,|\, Y=1, S=0\right)$

- Multi-class classification: $\mathcal{Y} = [K] := \{1, \ldots, K\}$
  - Standard setup:
    Classifier: $g : \mathbb{R}^d \to [K]$
    Risk: $\mathbb{P}(Y \neq g(X))$
    Property: no properties required
  - Confidence set setup: $2^{[K]}$ is the set of all subsets of $[K]$
    Classifier: $\Gamma : \mathbb{R}^d \to 2^{[K]}$
    Risk: $\mathbb{P}(Y \notin \Gamma(X))$
    Property: $\mathbb{E}\,|\Gamma(X)| \leq \beta$

- Multi-label classification: $\mathcal{Y} = \{0,1\}^L$
  - Standard Hamming setup:
    Classifier: $g : \mathbb{R}^d \to \{0,1\}^L$
    Risk: $\sum_{l=1}^L \mathbb{P}(Y^l \neq g^l(X))$
    Property: no properties required
  - Standard 0/1 setup:
    Classifier: $g : \mathbb{R}^d \to \{0,1\}^L$
    Risk: $\mathbb{P}(Y \neq g(X))$
    Property: no properties required
  - Controlled false positive:
    Classifier: $g : \mathbb{R}^d \to \{0,1\}^L$
    Risk: $\sum_{l=1}^L \mathbb{P}(Y^l = 1, g^l(X) = 0)$
    Property: $\sum_{l=1}^L \mathbb{P}\left(Y^l = 0, g^l(X) = 1 \,|\, X\right) \leq \beta$, a.s.

The rest of the chapter aims at providing a brief overview of these examples with an emphasis on the contributions of the author.

1.0.2 Empirical Risk Minimization

For this discussion let us disregard the constraints, that is, in this section we shall talk only about unconstrained classification problems. In other words, using notation of Chapter 1 we have $\mathcal{G}_\theta = \mathcal{G}$. Moreover, we limit this discussion to standard settings of binary classification, yet, the same arguments can be extended to other classification problems.

Clearly, our ultimate goal as statisticians and practitioners is to build a classification procedure which enjoys strong theoretical guarantees and demonstrates superior practical performance. However, achieving both goals is a notoriously difficult task from both theoretical and applied perspectives. Indeed, on the one hand, some modern state-of-the-art methods are extremely involved and their analysis is notably difficult when possible; on the

other hand, theoretically well understood methods might perform inferior to their more modern versions. Nevertheless, there are at least two well established approaches to construct classifiers: Empirical Risk Minimization (ERM) methods and plug-in methods.

Both approaches enjoy strong theoretical guarantees and have been studied in various settings. It was pointed out by Audibert and Tsybakov [2007] and later by Rigollet and Vert [2009] that ERM and plug-in methods should not be directly compared in theory, since the analysis of both have been carried under different sets of assumptions. The author of the thesis does not advocate for either of those methods. This section should be seen as a short overview of theoretical results available for ERM algorithms. Apart from this section, this thesis is mainly addressing plug-in approaches.

Let us first put some context for concreteness. Consider the setting of *standard* binary classification $(X, Y) \in \mathbb{R}^d \times \{0, 1\}$ with $(X, Y) \sim \mathbb{P}$ and assume that we observe n *i.i.d.* points from $\mathbb{P}$ denoted by $\mathcal{D}_n^{\mathrm{L}} = \{(X_i, Y_i)\}_{i=1}^n$. Using $\mathcal{D}_n^{\mathrm{L}}$, the goal is to construct $\hat{g}$ which approximates the following rule

$$g^* \in \arg\min_{g \in \mathcal{G}} \mathcal{R}(g) \ ,$$

where for all classifiers $g : \mathbb{R}^d \to \{0, 1\}$ its risk is defined as $\mathcal{R}(g) := \mathbb{P}(Y \neq g(X))$. We additionally would like this rule to satisfy

$$\lim_{n \to \infty} \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\mathcal{D}_n^{\mathrm{L}}}[\mathcal{R}(\hat{g}) - \mathcal{R}(g^*)] = 0 \ , \tag{1.3}$$

where $\mathcal{P}$ is some family of joint distributions on $\mathbb{R}^d \times \{0, 1\}$. There is a deep reason to restrict the family $\mathcal{P}$ of distributions, due to the following negative result.

Theorem 1 ([Devroye, 1982, Audibert, 2009])**.** *Let $(X, Y) \in \mathbb{R}^d \times \{0, 1\}$ and $\mathcal{P}$ be the set of all joint distributions on $\mathbb{R}^d \times \{0, 1\}$, then there is no $\hat{g}$ which satisfies Equation (1.3).*

After this negative result it becomes apparent that if we want to further develop theory for classification we need to restrict the possible family of distributions $\mathcal{P}$ or find other workarounds.

One can always propose an algorithm $\hat{g}$ which seems intuitive (thanks to the law of large numbers): as we do not have an access to the real distribution $\mathbb{P}$ but only to some realization $\mathcal{D}_n^{\mathrm{L}}$ we replace the risk by its empirical version

$$\hat{g} \in \arg\min_{g \in \mathcal{G}} \hat{\mathcal{R}}(g) \ , \tag{1.4}$$

where the empirical risk is defined as $\hat{\mathcal{R}}(g) := \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{Y_i \neq g(X_i)\}}$. Clearly, for each *fixed* classifier g thanks to the strong law of large numbers we have

$$\hat{\mathcal{R}}(g) \to \mathcal{R}(g) \ ,$$

where the converges holds almost surely. Besides, the central limit theorem gives us an idea of the rate of this convergence. Yet, the strong law of large numbers works only for a *fixed* classifier g and a more desirable result would be its uniform variant which would state that

$$\overset{\text{a.s.}}{\lim_{n \to \infty}} \sup_{g \in \mathcal{G}} \left| \hat{\mathcal{R}}(g) - \mathcal{R}(g) \right| = 0 \ , \tag{1.5}$$

where $\lim_{n\to\infty}^{\text{a.s.}}$ means that the convergence is almost sure. This type of results would at least allow us to obtain for the algorithm $\hat{g}$ from Eq. (1.4)

$$\lim_{n\to\infty}^{\text{a.s.}} \left|\hat{\mathcal{R}}(\hat{g}) - \mathcal{R}(\hat{g})\right| = 0 \ .$$

Non-asymptotic versions of such result, often called generalization bounds, suggests that whenever the empirical risk $\hat{\mathcal{R}}(\hat{g})$ is small, the real risk $\mathcal{R}(\hat{g})$ is also small. Importantly, the value of the empirical risk $\hat{\mathcal{R}}(\hat{g})$ is known in principle as it is based on the data. It is apparent that the result in Eq. (1.5) is impossible for $\mathcal{G}$ being defined as all binary classifiers from $\mathbb{R}^d$. A possible workaround is to study the following decomposition

$$\mathcal{R}(\hat{g}_{\mathcal{F}}) - \mathcal{R}(g^*) = \underbrace{(\mathcal{R}(\hat{g}_{\mathcal{F}}) - \mathcal{R}(g^*_{\mathcal{F}}))}_{\text{Stochastic error}} + \underbrace{(\mathcal{R}(g^*_{\mathcal{F}}) - \mathcal{R}(g^*))}_{\text{Systematic error}} \ ,$$

where $\hat{g}_{\mathcal{F}}, g^*_{\mathcal{F}}$ are defined for some $\mathcal{F} \subset \mathcal{G}$ as

$$\hat{g}_{\mathcal{F}} \in \arg\min_{g\in\mathcal{F}} \hat{\mathcal{R}}(g) \,, \qquad\qquad g^*_{\mathcal{F}} \in \arg\min_{g\in\mathcal{F}} \mathcal{R}(g) \ ,$$

respectively and the infimum is taken over some *fixed* subset $\mathcal{F}$ of $\mathcal{G}$. The size of the subset $\mathcal{F}$ gives the bias-variance trade-off – a central notion of statistics.

The systematic error is purely deterministic and can be studied with the means of approximation theory, meanwhile statistical theory can control the stochastic part of the decomposition. Additionally notice that if we restrict our attention to the distributions $\mathbb{P}$ for which the Bayes optimal classifier g^* belongs to the class $\mathcal{F}$, the systematic term is zero and our sole goal is to control the stochastic term. This road is typically, yet not always, taken by researchers and the focus is skewed towards the stochastic term.

For this term, a lot of results are available in the literature, the most famous being due to Vapnik and Chervonenkis [1971] who showed that if $\mathcal{F}$ is not "too big" then the stochastic error converges to zero with rate $\mathcal{O}(\sqrt{\log n/n})$. The key point of this result relies on the description of the complexity of $\mathcal{F}$ using *purely* combinatorial notion, which is typically called the VC-dimension (VC-dim) of $\mathcal{F}$. The work by Vapnik and Chervonenkis was highly influential in statistics and in discrete and computational geometry [Matoušek, 2002]. An interesting point of their result is that there is no curse of dimensionality in the rate of convergence of the stochastic term – a typical phenomenon in non-parametric statistics. Yet, it should be noted that the VC-dim is hidden in $\mathcal{O}(\sqrt{\log n/n})$ and in most of the situations it does depend on the dimension of the feature space $\mathcal{X}$, but only polynomially or even linearly for half spaces or its Boolean forms [Blumer et al., 1989, Eisenstat and Angluin, 2007].

Later, this rate was sharpened to $\mathcal{O}(\sqrt{1/n})$ with means of chaining [Sudakov, 1976, Dudley, 1967, Ledoux and Talagrand, 1991] and the famous packing lemma of Haussler for VC-classes [Haussler, 1995]. Further development showed that under extra low noise assumption, Tsybakov's margin assumption or Bernstein's type condition, this rate can be even further improved to $\mathcal{O}(1/n)$ [Mammen and Tsybakov, 1999, Tsybakov, 2004, Bartlett et al., 2005, Tsybakov and van de Geer, 2005, Massart and Nédélec, 2006].

Practically, however, these results have little impact – the minimization problem in Eq. (1.4) is non-convex and non-differentiable, thus leads to algorithms very difficult to implement in practice. A possible direction to alleviate this issue is to convexify each indicator in the sum of the empirical risk and to convexify the class of predictions. Both can be achieved if we consider classifiers g of the form $g(\cdot) = \mathbb{1}_{\{f(\cdot)\geq 0\}}$ for some real-valued function

f. That is, we associate each classifier g with some real valued score function f and in this case the misclassification can be written[2] as $\mathbb{1}_{\{Y \neq g(X)\}} = \mathbb{1}_{\{(2Y-1)f(X) \leq 0\}}$. Since the mapping $x \mapsto \mathbb{1}_{\{x \leq 0\}}$ is univariate, one can build its convexification in several ways. Various type of convexifications lead to different algorithms, such as logistic regression, support vector machines (SVM) or boosting among others. This approach was theoretically justified in the work[3] of Zhang [2004] and later generalized by Bartlett et al. [2006]. In the later, the authors showed that for appropriately chosen convexification, the excess risk of any classifier g can be upper-bounded by the "convexified" excess risk of the corresponding score function f. Thus, they reduced the study of ERM to the study of convexified ERM, to which the theory of Vapnik and Chervonenkis can also be applied.

Based on the convex risks, a related way to construct an algorithm $\hat{g}$ is based on Penalized Empirical Risk Minimization (PERM), where we add a convex (typically Tikhonov) penalization term to the empirical risk. Several learning guarantees can be proven for such type of estimators using the notion of stability [Bousquet and Elisseeff, 2002, Shalev-Shwartz et al., 2010]. In a nutshell, an algorithm $\hat{g}$ is uniformly stable[4] if an arbitrary perturbation of *one* data point cannot change this algorithm drastically. This approach bears some similarities with the bounded differences concentration inequality of McDiarmid [1989], which is widely used to demonstrate theoretical properties in the ERM context.

1.0.3 Plug-in classifiers

Consider the general settings of the beginning of Chapter 1 with $\mathcal{Y} = \{0, 1\}$. A central object in the analysis of binary classification problems is the regression function defined as $\eta(\cdot) = \mathbb{E}[Y|X = \cdot]$ since it describes how likely $Y = 1$. Another important object is the marginal distribution $\mathbb{P}_X$ of the feature vector $X \in \mathbb{R}^d$, which carries the information about the most typical observations $X \in \mathbb{R}^d$. In a large variety of settings we can show that the optimal classifier g^* is given for all $x \in \mathbb{R}^d$ by

$$g^*(x) = \begin{cases} 1, & \text{if } A\left(\eta(x), \mathbb{P}_X, \tau(\eta, \mathbb{P}_X)\right) \geq 0 \\ 0, & \text{otherwise} \end{cases},$$

for some *known* real valued function A and some *known* real-valued function τ. In this general description of the Bayes rule, we point out that the function A depends on η pointwise through the first argument and on the marginal distribution $\mathbb{P}_X$ through the second argument. Whereas, the function τ might depend on the *whole* regression function η through the first argument and likewise on the marginal distribution $\mathbb{P}_X$ through the second argument. Let us provide several classical examples where the description of the optimal classifier above is available.

Example 1.0.1 (Standard case)**.** *In the standard settings of binary classification without constraints, the Bayes optimal classifier [Devroye et al., 1996] is given as*

$$g^*(x) = \begin{cases} 1, & \textit{if } \eta(x) - \frac{1}{2} \geq 0 \\ 0, & \textit{otherwise} \end{cases}. \tag{1.6}$$

Notice that in this case the marginal distribution of the feature vector $X \in \mathbb{R}^d$ does not affect the Bayes optimal classifier

[2] One additionally should take care of the event $\mathbb{1}_{\{(2Y-1)f(X)=0\}}$.

[3] This result is typically called Zhang's Lemma.

[4] At this moment there are a lot of different ways to define stability of an algorithm.

Example 1.0.2 (F-score case). *In this settings we define a Bayes optimal classifier as*

$$g^* \in \arg\min_{g \in \mathcal{G}} \left\{ 1 - \frac{2\mathbb{P}(Y=1, g(X)=1)}{\mathbb{P}(Y=1) + \mathbb{P}(g(X)=1)} \right\} \ .$$

And one can establish [Zhao et al., 2013] that

$$g^*(x) = \begin{cases} 1, & \text{if } \eta(x) - \theta^* \geq 0 \\ 0, & \text{otherwise} \end{cases} \ ,$$

with θ^* being a solution in θ of

$$\theta \mathbb{E}_{\mathbb{P}_X}[\eta(X)] = \mathbb{E}_{\mathbb{P}_X} \max\{\eta(X) - \theta, 0\} \ .$$

Notice that in this case θ^* plays a role of the function $\tau(\cdot, \cdot)$, since it depends on the whole *regression function η and on the marginal distribution $\mathbb{P}_X$. For more details on classification with F-score see Section 1.1.2 of this introduction or Section 2.1 for the results that can be obtained in this setup.*

Example 1.0.3 (Classification with reject option). *Another setup is given by the problem of binary classification with reject option. In this setting our classifiers are such that $g : \mathbb{R}^d \to \{0, 1, ®\}$ and we define a Bayes optimal classifier for some $\beta \in [0, 1]$ as*

$$g^* \in \arg\min \{\mathbb{P}(Y \neq g(X) \,|\, g(X) \neq ®) \ : \ \mathbb{P}(g(X) = ®) \leq \beta\} \ .$$

In words, the Bayes optimal classifier has controlled rate of rejection. Under continuity assumptions on the cumulative distribution function (CDF) of $\eta(X)$ one can demonstrate [Chow, 1957, 1970] that

$$g^*(x) = \begin{cases} \mathbb{1}_{\{\eta(x) \geq \frac{1}{2}\}}, & \text{if } \max\{\eta(x), 1 - \eta(x)\} - \theta^* \geq 0 \\ ®, & \text{otherwise} \end{cases} \ .$$

In this case to define θ^* we need to introduce the following CDF

$$G(t) = \mathbb{P}(\max\{\eta(X), 1 - \eta(X)\} \leq t) \ ,$$

and θ^* is given by $G^{-1}(1 - \beta)$. Again, the threshold θ^* plays the role of the function $\tau(\cdot, \cdot)$. This expression for the optimal classifier can be traced back to the work of Chow [1957] in the context of Information Retrieval. We refer an interested reader to the work of Denis and Hebiri [2015b], who analyzed plug-in approach to this problem.

In all these cases a simple procedure to mimic g^* is based on the plug-in approach which uses two samples $\mathcal{D}_n^{\mathrm{L}}$ (labeled) and $\mathcal{D}_N^{\mathrm{U}}$ (unlabeled). Using the labeled part we can solve the regression problem by constructing $\hat{\eta}$"$\approx$"η and using the unlabeled data we can approximate the marginal distribution of $\mathbb{P}_X$ by its empirical version $\hat{\mathbb{P}}_X$"$\approx$"$\mathbb{P}_X$. Thus, the plug-in approach boils down to an algorithm $\hat{g}$ which is defined for all $x \in \mathbb{R}^d$ by

$$\hat{g}(x) = \begin{cases} 1, & \text{if } A\left(\hat{\eta}(x), \hat{\mathbb{P}}_X, \tau(\hat{\eta}, \hat{\mathbb{P}}_X)\right) \geq 0 \\ 0, & \text{otherwise} \end{cases} \ , \tag{1.7}$$

where we replaced all the unknown quantities by their approximations based on data. For this plug-in rule we build $\hat{\eta}$ using $\mathcal{D}_n^{\mathrm{L}}$ and use the empirical marginal distribution $\hat{\mathbb{P}}_X$ defined as $\hat{\mathbb{P}}_X = \frac{1}{N}\sum_{X \in \mathcal{D}_N^{\mathrm{U}}} \delta_X$. Recall that the function A and τ are assumed to be known, thus we do not need to estimate them. The knowledge of A and τ can hardly be called an assumption, it should rather be viewed as a result which establishes the form of a Bayes optimal classifier. In Section 6 we provide this form of the Bayes rule for a rather general framework of constrained classification.

One can immediately notice when we *could* expect an improvement introducing the unlabeled data. On the one hand, the Bayes classifier in Example 1.0.1 does not depend on the marginal distribution $\mathbb{P}_X$, thus, without extra structural assumptions (*e.g.,* cluster assumption [Rigollet, 2007]) on the regression function η we should not expect that the unlabeled sample can help. On the other hand, Bayes classifiers in Examples 1.0.2 and 1.0.3 do depend on the marginal distribution $\mathbb{P}_X$ and the unlabeled sample might improve the approximation accuracy of $\mathbb{P}_X$. Yet, the dependence of the Bayes classifier on the marginal distribution does not guarantee that any improvement of semi-supervised approaches over supervised ones can be shown[5]. For instance, in case of F-scores, it is shown in Section 2.1 that semi-supervised techniques cannot outperform supervised ones.

Besides, it is clear that the performance of the plug-in algorithm $\hat{g}$ is *at least* governed by the goodness of $\hat{\eta}$. Without going deeper into the details on the approximation of the marginal distribution and as a starting point, it is interesting to understand what kind of guarantees we can obtain for $\hat{\eta}$ and under what kind of assumptions.

From classification to regression

Non-parametric regression gives one of the possible approaches to address the question of estimation of the regression function η. Typically, results of this field of statistics are used as some sort of black box which guarantees existence of "optimal" estimators. Importantly, these existence results are constructive and can be realized in polynomial time. This section serves as a compact collection of some results which were extensively used by the author in various classification frameworks. For much more profound development and introduction we refer to [Giné and Nickl, 2015] and [Tsybakov, 2009].

Non-parametric statistics studies problems in which the unknown parameter belongs to an infinite dimensional space. In our case, the unknown parameter is the regression function η on which we have restrictions due to the nature of the problem, namely for all $x \in \mathbb{R}^d$ we have

$$0 \leq \eta(x) \leq 1 \ .$$

Unfortunately, even if we put ourselves in a space of all measurable bounded functions, this is not enough to guarantee existence of a uniformly good estimator $\hat{\eta}$. This is because, the space of all measurable bounded functions on $\mathbb{R}^d$ is too large[6] and further assumptions are required. At least, we would like to work in a totally bounded space in order to be able to construct ϵ-nets and execute standard arguments of the theory of empirical processes and make use of concentration inequalities.

A natural notion that allows to restrict ourselves to a totally bounded space is to assume that the regression function η is smooth. Smoothness can be described in different ways,

[5] The notion of semi-supervised and supervised estimator is introduced rigorously in Chapter 3

[6] For instance, this space is not totally bounded with respect to the sup norm.

such as Hölder, Sobolev, Nikolskii or Besov smoothness among others [Adams, 1975, Rudin, 1987, Simon, 1990, Sobolev, 1991, Besov et al., 1996]. As it is not the main concern of the present manuscript and to ease the presentation, we stick to the most basic notion of smooth functions. Namely, we are interested in Hölder smooth functions [Rudin, 1987], which says that $\eta : \mathbb{R}^d \to [0,1]$ is (L,β)-Hölder smooth for some $\beta \in (0,1]$ and $L > 0$ if and only if for all $x, x' \in \mathbb{R}^d$ we have

$$|\eta(x) - \eta(x')| \leq L \|x - x'\|_2^\beta \quad .$$

Now denote by $\Sigma(L,\beta,\mathbb{R}^d)$ the set of all (L,β)-Hölder smooth functions on $\mathbb{R}^d$ which are valued in $[0,1]$. Note that this definition gives something non-trivial[7] only in the case of $0 < \beta \leq 1$. The set $\Sigma(L,\beta,\mathbb{R}^d)$ is in some sense simultaneously small and large. Indeed, smooth functions seem to be rather common which is an argument for the largeness of this space; yet, this class admits a finite ϵ-net [Kolmogorov and Tikhomirov, 1961] which significantly simplifies theoretical analysis in this space. Besides, it appears that the notion of smoothness perfectly fits in the idea of non-parametric statistics and various results can be derived in this context [Ibragimov and Khasminskii, 1981, Tsybakov, 2009, Giné and Nickl, 2015].

In our settings, the feature vector $X \in \mathbb{R}^d$ is random and follows some marginal distribution $\mathbb{P}_X$. For simplicity we only consider those distributions which admit uniformly lower- and upper-bounded densities μ *w.r.t.* the Lebesgue measure and are supported on the unit cube $[0,1]^d$. This is referred to as the "very strong density" assumption.

We define a family of joint distributions $\mathcal{P}(L,\beta)$ on $\mathbb{R}^d \times \{0,1\}$ such that for all $\mathbb{P} \in \mathcal{P}(L,\beta)$ the regression function $\eta \in \Sigma(L,\beta,\mathbb{R}^d)$ and $\mathbb{P}_X$ satisfies the very strong density assumption. Finally, once the family of distributions is defined we can start asking questions about guarantees; assume that we have an estimator $\hat{\eta}$ based on $\mathcal{D}_n^{\mathrm{L}} \overset{i.i.d.}{\sim} \mathbb{P}$ and some $0 < q < \infty$, $1 \leq p < \infty$ then its[8] maximal ℓ_p risk is given as

$$\Psi_n^{p,q}(\hat{\eta}, \mathcal{P}(L,\beta)) := \sup_{\mathbb{P} \in \mathcal{P}(L,\beta)} \mathbb{E}_{\mathcal{D}_n^{\mathrm{L}}} \left(\int_{\mathbb{R}^d} |\eta(x) - \hat{\eta}(x)|^p \, d\mu(x) \right)^{\frac{q}{p}} \quad ,$$

and we want this sequence to be as small as possible. For the case $p = \infty$ we define this risk as

$$\Psi_n^{\infty,q}(\hat{\eta}, \mathcal{P}(L,\beta)) := \sup_{\mathbb{P} \in \mathcal{P}(L,\beta)} \mathbb{E}_{\mathcal{D}_n^{\mathrm{L}}} \|\eta - \hat{\eta}\|_\infty^q \quad ,$$

where the infinity norm is interpreted as an essential supremum with respect to the distribution $\mathbb{P}_X$ of the vector $X \in \mathbb{R}^d$. In this context for all $q \in (0,\infty), p \in [1,\infty]$ the minimax rate of convergence over the class $\mathcal{P}(L,\beta)$ is a sequence $\psi_n^{p,q}$ for which there exist two positive constants $0 < c \leq C < \infty$ such that for all $n \in \mathbb{N}$ it holds that

$$c\psi_n^{p,q} \leq \inf_{\hat{\eta}} \Psi_n^{p,q}(\hat{\eta}, \mathcal{P}(L,\beta)) \leq C\psi_n^{p,q} \quad .$$

Here the infimum is taken over all estimator $\hat{\eta}$, that is, over all measurable functions of $\mathcal{D}_n^{\mathrm{L}}$.

The next result is a major one in non-parametric statistics. It is also available for other notions of smoothness and in a lot of situations it formally boils down to the replacement of β by some other effective smoothness β'.

[7] It is possible to extend this notion to $\beta > 1$, see [Tsybakov, 2009].

[8] Or its q^{th} moment.

Theorem 2 (Stone [1982], Tsybakov [1986], Korostelëv and Tsybakov [1993]). *There exists an estimator $\hat{\eta}$ based on $\mathcal{D}_n^{\mathrm{L}}$ such that for all $0 < q < \infty$ and $1 \leq p < \infty$ we have*

$$\sup_{\mathbb{P} \in \mathcal{P}(L,\beta)} \mathbb{E}_{\mathcal{D}_n^{\mathrm{L}}} \left(\int_{\mathbb{R}^d} |\eta(x) - \hat{\eta}(x)|^p \, d\mu(x) \right)^{\frac{q}{p}} \lesssim n^{-\frac{q\beta}{2\beta+d}} \quad ,$$

and if $p = \infty$, then it holds that for all $0 < q < \infty$

$$\sup_{\mathbb{P} \in \mathcal{P}(L,\beta)} \mathbb{E}_{\mathcal{D}_n^{\mathrm{L}}} \|\eta - \hat{\eta}\|_\infty^q \lesssim \left(\frac{n}{\log n} \right)^{-\frac{q\beta}{2\beta+d}} \quad .$$

Moreover, these rates are minimax, that is, the best possible estimator $\hat{\eta}$ also achieves these rates.

Note how the dimension d is included in the rate of convergence, that is, the larger the dimension the worse the rate is. This phenomena is known as the curse of dimensionality and is intrinsic to the non-parametric regression. More restrictive models such as single/multi-index, composite functions [Hristache et al., 2001b,a, Juditsky et al., 2009], or recent work on neural nets [Schmidt-Hieber, 2017] allow to improve the dependence on the dimension.

In any case, Theorem 2 already allows to obtain several non-trivial results concerning the rates of convergence of plug-in method in binary classification [Yang, 1999]. However, it might give an over-pessimistic convergence rate as discovered by [Audibert and Tsybakov, 2007]. Namely, these authors showed that the rate can be specified and the curse of dimensionality might be alleviated in classification settings; Section 1.1 gives a review on this line of work. The core argument of Audibert and Tsybakov [2007] is based on the use of another type of guarantees, namely, instead of the bound in expectation, they showed that an exponential deviation inequality can be obtained.

Theorem 3 (Audibert and Tsybakov [2007]). *There exists an estimator $\hat{\eta}$ based on $\mathcal{D}_n^{\mathrm{L}}$ and constants C_1, C_2 such that for any $\delta > 0$ and $n \geq 1$ it holds that*

$$\sup_{\mathbb{P} \in \mathcal{P}} \mathbb{P}^{\otimes n} \left(|\hat{\eta}(x) - \eta(x)| \geq \delta \right) \leq C_1 \exp\left(-C_2 n^{-\frac{2\beta}{2\beta+d}} \delta^2 \right) \quad ,$$

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

Similar results are available in the context of density estimation [Rigollet and Vert, 2009, Jiang, 2017]. Theorem 3 allows to perform very sharp analysis for the binary classification using the peeling technique popularized by Audibert and Tsybakov [2007], and obtain minimax optimal rates of convergence. In this thesis this inequality will be used many times in various contexts.

1.1 Binary classification

In binary classification each instance $X \in \mathbb{R}^d$ is associated with a binary label $Y \in \{0, 1\}$. Additionally, it is assumed that (X, Y) follows some distribution $\mathbb{P}$ on $\mathbb{R}^d \times \{0, 1\}$. This framework has been successfully used in various applied scenarios such as medicine [Güvenir et al., 1998, Gil et al., 2012], spam detection [Sahami et al., 1998], and credit scoring [Louzada et al., 2016] among others.

1.1.1 Standard setup

The most basic and well-studied setting is the one we call *standard binary* classification. Recall, that the risk in this setup is given by the probability of misclassification

$$\mathcal{R}(g) = \mathbb{P}(Y \neq g(X)) \quad \text{(misclassification)} \ .$$

In this case we do not want to assign any properties to our estimator and our only goal is to classify a point X as accurately as possible in the sense of minimizing the probability of misclassification.

Example 1.0.1 states that the Bayes rule in this setup is tightly related to the regression function $\eta(x) := \mathbb{P}(Y = 1|X = x) = \mathbb{E}[Y|X = x]$ [Devroye et al., 1996], that is, for all $x \in \mathbb{R}^d$

$$g^*(x) = \mathbb{1}_{\{\eta(x) \geq 1/2\}} \quad \text{(Bayes rule)} \ .$$

Hence, the Bayes classifier compares the regression function to $1/2$ and makes the prediction accordingly. As discussed in the previous section, a natural attempt to construct an algorithm $\hat{g}$ is to estimate the regression function η by some estimator $\hat{\eta}$ and use the following plug-in strategy for all $x \in \mathbb{R}^d$

$$\hat{g}(x) = \mathbb{1}_{\{\hat{\eta}(x) \geq 1/2\}} \quad \text{(plug-in estimator)} \ ,$$

which tries to mimic the Bayes rule by thresholding some estimate $\hat{\eta}$ of the regression function η. The theoretical justification of this approach can be obtained from the following upper bound on the excess risk [Devroye et al., 1996]

$$\mathbb{E}_{\mathcal{D}_n^{\mathrm{L}}}[\mathcal{E}(\hat{g})] := \mathbb{E}_{\mathcal{D}_n^{\mathrm{L}}}[\mathcal{R}(\hat{g})] - \mathcal{R}(g^*) \leq 2\mathbb{E}_{\mathcal{D}_n^{\mathrm{L}}} \int_{\mathcal{X}} |\hat{\eta}(x) - \eta(x)| \, d\mathbb{P}_X(x) \ ,$$

which suggests that the problem of classification can be linked to the regression problem. A drawback of such an approach is the resulting rate of convergence, which is directly linked to the rate of estimation of η. Recall Theorem 2, which states that if η is a β-Hölder smooth functions, then the rate of convergence [Stone, 1982, Tsybakov, 2009] is given by

$$\mathbb{E}_{\mathcal{D}_n^{\mathrm{L}}}[\mathcal{E}(\hat{g})] \leq 2\mathbb{E}_{\mathcal{D}_n^{\mathrm{L}}} \int_{\mathcal{X}} |\hat{\eta}(x) - \eta(x)| \, d\mathbb{P}_X(x) \lesssim n^{-\frac{\beta}{2\beta+d}} = n^{-\frac{1}{2+d/\beta}} \ ,$$

uniformly over this class. Again notice, that this rate is always slower than $n^{-1/2}$ and the bound degrades drastically with the growth of the dimension d. One might hope that the upper bound on the excess risk via the regression risk is loose and a better rate could be obtained with a more sophisticated technique. Unfortunately, it was shown in [Yang, 1999] that these rates are minimax optimal which resulted in a criticism of the plug-in approach and a more favorable opinion on Empirical Risk Minimization (ERM) type methods.

The situation has changed after the work [Audibert and Tsybakov, 2007], where the authors showed that the results of Yang [1999] are too pessimistic, or in some sense "too minimax" in a lot of situations. The key property that allowed to improve the rate of convergence is a precise description of how the regression function η behaves around the decision threshold $1/2$. That is, Audibert and Tsybakov [2007] discriminated the regression functions η not only by their smoothness but also by their concentration around $1/2$. This strategy was originally explored by Polonik [1995], Tsybakov [1997] in the context of density level set estimation and later in the context of classification by [Mammen and Tsybakov, 1999]. In the case of the binary classification, this assumption is known as the margin, low noise, or Mammen-Tsybakov assumption

Assumption 1 (Margin assumption)**.** *The regression function η is such that there exist constants $\alpha \geq 0$ and $c > 0$ such that for all $\delta > 0$*

$$\mathbb{P}_X\left(0 < \left|\eta(X) - \frac{1}{2}\right| \leq \delta\right) \leq c\delta^\alpha \ .$$

The value of $\alpha = 0$ corresponds to the case of no assumption, in this case the regression function $\eta(x)$ can be concentrated around $1/2$ and the result of [Yang, 1999] is rather focused on this scenario. In contrast, the larger value of α we have, the simpler the classification problem is. The "simplest" case being "$\alpha = \infty$", it describes a situation when there is a "corridor" between $\eta(X)$ and the threshold $1/2$. In other words a regression function η satisfies the margin assumption with $\alpha = \infty$ if there exists $h > 0$ such that

$$\left|\eta(X) - \frac{1}{2}\right| \geq h \quad \text{almost surely} \ .$$

This case is often referred as Massart's low noise condition due to the work [Massart and Nédélec, 2006], where the authors provided an extensive analysis of ERM algorithm over VC-classes under this assumption.

Importantly, Audibert and Tsybakov [2007] have shown that there exists a plug-in algorithm $\hat{g}$ whose rate of convergence under the margin assumption satisfies

$$\mathbb{E}_{\mathcal{D}_n^{\mathrm{L}}}[\mathcal{E}(\hat{g})] \lesssim n^{-\frac{(1+\alpha)\beta}{2\beta+d}} \ ,$$

and this rate is minimax optimal over a typical non-parametric family of distributions. Clearly, unlike the previous rate $n^{-\frac{\beta}{2\beta+d}}$ provided by Yang [1999], under the margin assumption the rate can be significantly improved. In the same work Audibert and Tsybakov [2007] showed that depending on the interplay between α, β, d the rate can be slow (slower than $n^{-1/2}$), fast (in between $n^{-1/2}$ and n^{-1}) and *even* super-fast (faster than n^{-1})[9]. These results also suggest that the plug-in methods should not be considered inferior to the ERM methods. It also emphasizes the intrinsic role of the margin type assumption in the analysis of classification methods.

Remark. *Let us emphasize that essentially the study of non-parametric settings of standard binary setup is reduced to a study of an efficient estimation of the regression function η. Indeed, as the Bayes optimal classifier is given by thresholding of the regression function on the level $1/2$, which is known beforehand, it is expected that a good approximation of η thresholded by $1/2$ would allow to obtain a good approximation in classification framework. Yet, most of the contributions of this manuscript are concerned with situations when this threshold is* distribution dependent*, and it ought to be estimated using data.*

1.1.2 F-score setup (Section 2.1)

The probability of misclassification is still widely used in practice to evaluate performance of an algorithm. Practically, this risk is suitable in the situation of class-balanced distributions, that is, in the case when $\mathbb{P}(Y = 1) \approx \mathbb{P}(Y = 0)$. Such a distribution would typically yield a well-balanced dataset, thanks to the law of large numbers, a situation where algorithms tailored to optimize the misclassification risk shine the most. Once the condition $\mathbb{P}(Y =$

[9] In the same work these authors showed that the family of distributions for which the rate can be super-fast is very poor.

$1) \approx \mathbb{P}(Y = 0)$ fails to be satisfied, the resulting dataset might be highly unbalanced and a method which optimizes misclassification risk yields an unsatisfying performance. In practical applications, a popular way to assess the performance of an algorithm in the unbalanced setup is to use the F-score, whose roots can be traced back to the Information Retrieval (IR) literature [van Rijsbergen, 1974, Lewis, 1995]. The F-score of a classifier g is defined as

$$F_1(g) := \frac{2\mathbb{P}(Y = 1, g(X) = 1)}{\mathbb{P}(Y = 1) + \mathbb{P}(g(X) = 1)} \qquad \text{(F-score)} \ . \tag{1.8}$$

Intuitively, the numerator, in the definition of the F-score, is big for those classifiers maximizing true positives (*i.e.,* $\mathbb{P}(Y = 1, g(X) = 1)$). Clearly, a naive classifier $g \equiv 1$ does maximize this value, at this moment the denominator in the F-score kicks in. Indeed, the F-score, defined in Eq. (1.8), decreases with the growth of the denominator and thus, an optimal classifier for the F-score gives a trade-off between the probability of true positive and the probability of predicting one.

Besides, the F-score can be linked to Precision and Recall – two basic notions in the IR community. Formally, Precision and Recall of a binary classifier $g : \mathbb{R}^d \to \{0, 1\}$ are defined as

$$\text{Precision}(g) := \frac{\mathbb{P}(Y = 1, g(X) = 1)}{\mathbb{P}(g(X) = 1)}, \quad \text{Recall}(g) := \frac{\mathbb{P}(Y = 1, g(X) = 1)}{\mathbb{P}(Y = 1)} \ .$$

In classification framework high Precision of g means that instances $X \in \mathbb{R}^d$ classified as $g(X) = 1$ are likely to have the real labels being $Y = 1$. Whereas, high Recall of g means that instances $X \in \mathbb{R}^d$ with real label $Y = 1$ are likely to be classified correctly by g. In contrast, the high Precision of g says nothing about instances $X \in \mathbb{R}^d$ with $Y = 1$ that were not classified correctly and, likewise, the high Recall of g says nothing about the instances $X \in \mathbb{R}^d$ which were classified as $g(X) = 0$. Typically, neither Precision nor Recall are considered separately. Instead, classifiers are compared in terms of one measure with a fixed budget for the other or both measures are blended into a single one. One of the most popular examples of such a measure, which combines both Precision and Recall, is the aforementioned F-score which is seen as their harmonic average

$$F_1(g) = \left(\frac{\text{Precision}^{-1}(g) + \text{Recall}^{-1}(g)}{2} \right)^{-1} \ .$$

Let us mention that the F-score is not a risk measure but rather a score measure, that is, our goal is to *maximize* it.

In this setup natural statistical questions emerge.

Q.1: What is the form of the Bayes optimal classifier g^* in this context?

Q.2: How to construct a consistent algorithm for this problem and under which assumptions?

Q.3: What is the minimax rate for the problem of classification with the F-score?

The first question was already answered in [Zhao et al., 2013] who showed that the Bayes classifier g^* can be defined for all $x \in \mathbb{R}^d$ as

$$g^*(x) = \mathbb{1}_{\{\eta(x) > \theta^*\}} \ , \tag{1.9}$$

with $\theta^* \in [0,1]$ being a threshold which satisfies

$$\theta^* \mathbb{P}(Y=1) = \mathbb{E}(\eta(X) - \theta^*)_+ \ .$$

This results will be extended to a more general definition of the F-score in Section 2.1 (see Theorem 4). Here for all $a \in \mathbb{R}$ we denote by $(a)_+$ its positive part, that is, $(a)_+ = \max\{a, 0\}$. The result of Zhao et al. [2013] demonstrates that in order to transition from the standard setting of binary classification with misclassification risk to the F-score setup it is sufficient to have an access to the threshold θ^* or to its estimate. Unfortunately, these authors did not study the statistical side of this problem and did not provide any procedure to estimate neither g^* nor θ^*.

In Section 2.1 of Chapter 2 we address these statistical questions in details. In particular, we propose a *semi-supervised* algorithm $\hat{g}$, which under assumptions similar to the ones used by Audibert and Tsybakov [2007] satisfies

$$F_1(g^*) - \mathbb{E}_{(\mathcal{D}_n^{\mathrm{L}}, \mathcal{D}_N^{\mathrm{U}})}\left[F_1(\hat{g})\right] \lesssim n^{-\frac{(1+\alpha)\beta}{2\beta+d}} \ .$$

It is additionally shown that the above rate is minimax optimal. The rate above is obtained under a modified version of the margin assumption, which in this setup reads for all $\delta > 0$ as

$$\mathbb{P}_X(0 < |\eta(X) - \theta^*| \leq \delta) \leq c\delta^\alpha \ ,$$

for some $c, \alpha > 0$. On the first glance it might be surprising to consider semi-supervised procedures, since the algorithm of Audibert and Tsybakov [2007] is supervised. However, it becomes more evident if we rewrite the condition on θ^* in the following form

$$\theta^* \mathbb{E}_{X \sim \mathbb{P}_X}[\eta(X)] = \mathbb{E}_{X \sim \mathbb{P}_X}(\eta(X) - \theta^*)_+ \ .$$

On the one hand, would the regression function η be available for a statistician, a natural approach to estimate θ^* is to replace $\mathbb{E}_{X \sim \mathbb{P}_X}$ by its empirical version on some unlabeled dataset $\mathcal{D}_N^{\mathrm{U}}$, which leads to a fully unsupervised "algorithm". On the other hand, if we had an access to the distribution $\mathbb{P}_X$ of the feature vector $X \in \mathbb{R}$, then a logical estimator of g^* is thresholding of $\hat{\eta}$ on a level $\tilde{\theta}$ satisfying

$$\tilde{\theta} \mathbb{E}_{X \sim \mathbb{P}_X}[\hat{\eta}(X)] = \mathbb{E}_{X \sim \mathbb{P}_X}(\hat{\eta}(X) - \tilde{\theta})_+ \ ,$$

which leads to a fully supervised "algorithm". Obviously, neither η nor $\mathbb{P}_X$ are available in reality, yet, having a labeled sample $\mathcal{D}_n^{\mathrm{L}}$ and an unlabeled sample $\mathcal{D}_N^{\mathrm{U}}$ would allow to estimate these quantities efficiently. Section 2.1 of Chapter 2 describes this classification algorithm and establishes its optimality in the minimax sense.

1.1.3 Constrained classification: general framework

This section is an attempt to put forward a class of constrained binary classification problems in which the plug-in approach is expected to work well. The goal here is to define a family of problems with easily accessible Bayes optimal classifier and also to provide a general machinery that may be useful in similar contexts

We consider the following settings of binary classification: given $(Z, Y) \in \mathcal{Z} \times \{0, 1\}$ distributed according to $\mathbb{P}$, a classifier is a measurable mapping $g : \mathcal{Z} \to \{0, 1\}$. Let us denote the marginal distribution of Z by $\mathbb{P}_Z$. In this part, we use the notation $Z \in \mathcal{Z}$

instead of $X \in \mathbb{R}^d$ to account for more complex setups such as fair binary classification of equal opportunity (see Section 1.1.4), where the variable $Z \in \mathcal{Z}$ is represented by a tuple $(X, S) \in \mathbb{R}^d \times \{0, 1\}$. This fairness setup is discussed in details in Section 1.1.4 and in Section 2.2 where we provide theoretical and experimental analyses of this problem.

For a given classifier g we assign its risk $\mathcal{R}(g)$ and we assume that $\mathcal{R}(g)$ can be expressed as

$$\mathcal{R}(g) = \mathbb{E}_{Z\sim\mathbb{P}_Z}\left[A_{\mathbb{P}}(Z) + B_{\mathbb{P}}(Z)\mathbb{1}_{\{g(Z)=1\}}\right] \quad \text{(risk)} \ , \tag{1.10}$$

for some bounded measurable functions $A_{\mathbb{P}} : \mathcal{Z} \to \mathbb{R}$ and $B_{\mathbb{P}} : \mathcal{Z} \to \mathbb{R}$ which are allowed to depend on the unknown distribution $\mathbb{P}$. Additionally, we assume that the constraint on the classifier can be expressed by one equality of the form[10]:

$$\mathbb{E}_{Z\sim\mathbb{P}_Z}[\bar{A}_{\mathbb{P}}(Z) - \bar{B}_{\mathbb{P}}(Z)\mathbb{1}_{\{g(Z)=1\}}] = 0 \quad \text{(constraints)} \ , \tag{1.11}$$

for some bounded measurable functions $\bar{A}_{\mathbb{P}} : \mathcal{Z} \to \mathbb{R}$ and $\bar{B}_{\mathbb{P}} : \mathcal{Z} \to \mathbb{R}$. We can now define a Bayes optimal classifier in this context as

$$g^* \in \arg\min\left\{\mathcal{R}(g) \ : \ \mathbb{E}_{Z\sim\mathbb{P}_Z}[\bar{A}_{\mathbb{P}}(Z) - \bar{B}_{\mathbb{P}}(Z)\mathbb{1}_{\{g(Z)=1\}}] = 0\right\} \ .$$

The next proposition shows that the proposed framework admits a large family of classification problems and can be used in various settings.

Proposition 1. *Let $L_{\mathbb{P}} : \{0, 1\} \times \{0, 1\} \to \mathbb{R}$ be a measurable loss function and $C_{\mathbb{P}} : \{0, 1\} \times \{0, 1\} \to \mathbb{R}$ a measurable constraint function, and consider the following problem*

$$\min\left\{\mathbb{E}_{(Z,Y)\sim\mathbb{P}}[L_{\mathbb{P}}(g(Z), Y)] \ : \ \mathbb{E}_{(Z,Y)\sim\mathbb{P}}[C_{\mathbb{P}}(g(Z), Y)] = 0\right\} \ ,$$

then this constrained binary classification formulation admits the representation from Equations (1.10), (1.11), with

$$\begin{aligned}
A_{\mathbb{P}}(Z) &= L_{\mathbb{P}}(0,1)\mathbb{P}(Y=1|Z) + L_{\mathbb{P}}(0,0)\mathbb{P}(Y=0|Z) \ , \\
B_{\mathbb{P}}(Z) &= (L_{\mathbb{P}}(1,1) - L_{\mathbb{P}}(0,1))\mathbb{P}(Y=1|Z) + (L_{\mathbb{P}}(1,0) - L_{\mathbb{P}}(0,0))\mathbb{P}(Y=0|Z) \ , \\
\bar{A}_{\mathbb{P}}(Z) &= C_{\mathbb{P}}(0,1)\mathbb{P}(Y=1|Z) + C_{\mathbb{P}}(0,0)\mathbb{P}(Y=0|Z) \ , \\
\bar{B}_{\mathbb{P}}(Z) &= (C_{\mathbb{P}}(0,1) - C_{\mathbb{P}}(1,1))\mathbb{P}(Y=1|Z) + (C_{\mathbb{P}}(0,0) - C_{\mathbb{P}}(1,0))\mathbb{P}(Y=0|Z) \ .
\end{aligned}$$

In general the set $\left\{g \ : \ \mathbb{E}_{Z\sim\mathbb{P}_Z}[\bar{A}_{\mathbb{P}}(Z) - \bar{B}_{\mathbb{P}}(Z)\mathbb{1}_{\{g(Z)=1\}}] = 0\right\}$ might be empty and thus the Bayes classifier is ill-defined in such a case. The next assumption gives a sufficient condition under which this issue can be bypassed and this assumption is at the core of this section. Let us also mention that in some scenarios one can get rid of it, which we shall discuss later in this section.

Assumption 2. *The random variable $\bar{B}_{\mathbb{P}}(Z) - B_{\mathbb{P}}(Z)$ is bounded $\mathbb{P}_Z$-almost surely. Moreover, the mapping*

$$t \mapsto \mathbb{P}_Z\left(\lambda\bar{B}_{\mathbb{P}}(Z) - B_{\mathbb{P}}(Z) \leq t\right) \ ,$$

is continuous for all $\lambda \in \mathbb{R}$. Which is the same as to say that the random variable $\bar{B}_{\mathbb{P}}(Z) - B_{\mathbb{P}}(Z)$ does not have atoms.

[10] The minus sign in the expression for the constraints is simply for convenience.

One possible way to relax this assumption is to consider randomized classifiers, which instead of a definitive prediction 0 or 1 can output a distribution on $\{0,1\}$. We leave the extension to randomized classifiers for future work.

Lemma 1 (Bayes rule)**.** *Under Assumption 2 a Bayes optimal classifier g^* can be obtained for all $z \in \mathcal{Z}$ as*

$$g_{\lambda^*}(z) = \mathbb{1}_{\left\{\lambda^* \bar{B}_{\mathbb{P}}(z) - B_{\mathbb{P}}(z) > 0\right\}} \ ,$$

where λ^* is determined as a root of

$$\lambda \mapsto \mathbb{E}_{Z\sim\mathbb{P}_Z}\left[\bar{B}_{\mathbb{P}}(Z)\mathbb{1}_{\left\{\lambda\bar{B}_{\mathbb{P}}(Z) - B_{\mathbb{P}}(Z) > 0\right\}}\right] - \mathbb{E}_{Z\sim\mathbb{P}_Z}[\bar{A}_{\mathbb{P}}(Z)] \ .$$

Moreover, for every classifier g we can write

$$\mathcal{R}(g) - \mathcal{R}(g^*) = \mathbb{E}_{Z\sim\mathbb{P}_Z}\left|B_{\mathbb{P}}(Z) - \lambda^*\bar{B}_{\mathbb{P}}(Z)\right| \mathbb{1}_{\{g(Z)\neq g^*(Z)\}} - \lambda^*(\mathbb{E}_{Z\sim\mathbb{P}_Z}[\bar{A}_{\mathbb{P}}(Z) - \bar{B}_{\mathbb{P}}(Z)\mathbb{1}_{\{g(Z)=1\}}]) \ .$$

The proof of this result is given in Appendix and it relies on weak duality, which combined with the continuity Assumption 2 can be turned into strong duality.

Remark 1. *Assumption 2 can be relaxed significantly if one can guarantee the existence of a root of*

$$\lambda \mapsto \mathbb{E}_{Z\sim\mathbb{P}_Z}\left[\bar{B}_{\mathbb{P}}(Z)\mathbb{1}_{\left\{\lambda^*\bar{B}_{\mathbb{P}}(Z) - B_{\mathbb{P}}(Z) > 0\right\}}\right] - \mathbb{E}_{Z\sim\mathbb{P}_Z}[\bar{A}_{\mathbb{P}}(Z)] \ .$$

This existence is ensured due to the fact that the equation

$$\mathbb{E}_{Z\sim\mathbb{P}_Z}\left[\bar{B}_{\mathbb{P}}(Z)\mathbb{1}_{\left\{\lambda\bar{B}_{\mathbb{P}}(Z) - B_{\mathbb{P}}(Z) > 0\right\}}\right] - \mathbb{E}_{Z\sim\mathbb{P}_Z}[\bar{A}_{\mathbb{P}}(Z)] = 0 \ ,$$

is the first order optimality condition[11] of a concave maximization problem under Assumption 2. Besides, notice that if a classifier g satisfies constraints of Equation (1.11) we have

$$\mathcal{R}(g) - \mathcal{R}(g^*) = \mathbb{E}_{Z\sim\mathbb{P}_Z}\left|B_{\mathbb{P}}(Z) - \lambda^*\bar{B}_{\mathbb{P}}(Z)\right| \mathbb{1}_{\{g(Z)\neq g^*(Z)\}} \ ,$$

which resembles classical expression for the excess risk in the standard settings of Binary classification. However, sometimes it is difficult or even impossible to guarantee that g satisfies constraints of Equation (1.11).

The most straightforward application of this framework is the standard setup of binary classification. In this settings, typically the space $\mathcal{Z} = \mathbb{R}^d$ and the elements of this space are seen as feature vector and denoted by X. Recall that in the context of standard binary classification the standard choice of the risk is the probability of misclassification defined for every classifier $g : \mathbb{R}^d \to \{0,1\}$ as

$$\mathcal{R}(g) = \mathbb{P}(Y \neq g(X)) \ .$$

The constraints in this case can be formally written as $0 = 0$, that is, no-constraints are forced on the classifiers and $\bar{A}_{\mathbb{P}}(X) \equiv \bar{B}_{\mathbb{P}}(X) \equiv 0$. Let us demonstrate, that the classical results are easily recovered with the introduced framework. Using Proposition 1 with $L_{\mathbb{P}}(\cdot,\cdot) = \mathbb{1}_{\{\cdot\neq\cdot\}}$

[11] Condition on λ^*

we get , $A_{\mathbb{P}}(X) = \eta(X)$ and $B_{\mathbb{P}}(X) = 1 - 2\eta(X)$, where $\eta(X) = \mathbb{E}[Y|X]$ is the regression function. Finally, thanks to Lemma 1, the Bayes optimal classifier is given for all $x \in \mathbb{R}^d$ by

$$g_{\lambda^*}(x) = \mathbb{1}_{\left\{\lambda^* \bar{B}_{\mathbb{P}}(x) - B_{\mathbb{P}}(x) > 0\right\}} = \mathbb{1}_{\{-(1-2\eta(x))>0\}} = \mathbb{1}_{\{\eta(x)>1/2\}} \ ,$$

which recovers the expression of the Bayes classifier in Example 1.0.1. Finally, for any binary classifier g we have in this context the following familiar expression for the excess risk

$$\mathcal{R}(g) - \mathcal{R}(g^*) = \mathbb{E}_{X\sim\mathbb{P}_X} \left|B_{\mathbb{P}}(X)\right| \mathbb{1}_{\{g(X)\neq g^*(X)\}} = \mathbb{E}_{X\sim\mathbb{P}_X} \left|1 - 2\eta(X)\right| \mathbb{1}_{\{g(X)\neq g^*(X)\}} \ .$$

This is a starting point of the theoretical assessment in case of the binary classification. Let us mention that the arguments above do not require Assumption 2 to be satisfied, which is posed in the proposed framework. Though, following Remark 1 it is obvious that in this case any λ satisfies

$$\mathbb{E}_{X\sim\mathbb{P}_X}\left[\underbrace{\bar{B}_{\mathbb{P}}(X)}_{\equiv 0} \mathbb{1}_{\left\{\lambda \bar{B}_{\mathbb{P}}(X) - B_{\mathbb{P}}(X) > 0\right\}}\right] - \mathbb{E}_{X\sim\mathbb{P}_X}[\underbrace{\bar{A}_{\mathbb{P}}(X)}_{\equiv 0}] = 0 \ ,$$

thus Assumption 2 can be alleviated. Clearly, this reasoning can be effortlessly generalized to cost-sensitive settings. More involved examples are possible, one of them being the fair binary classification discussed in details in Section 2.2.

A nice thing about this framework is that the derivation of the Bayes rule is very robust with respect to the classification problem. In other words, it can be easily extended to multi-class and multi-label framework as well as to other type of constraints. Moreover, it immediately suggests to use the plug-in approach described in Section 1.0.3 as the Bayes classifier obtained in Lemma 1 is of the form given in Eq. (1.6). The main restriction in this framework is the fact that the risk and the constraints are required to be "*linear*" functionals of classifier g. In Chapter 4 we shall see an example of constraints that cannot be written as a linear functional of the classifier. One of these examples is concerned with multi-label classification with an almost sure control over false positive errors (see Chapter 4 for a precise definition).

Once the expression for the excess risk is derived it is tempting to understand whether the analysis of Audibert and Tsybakov [2007] can be extended to this general setting.
Open question: are the fast rates of convergence achievable under the following margin assumption?

Assumption 3. *The distribution $\mathbb{P}_Z$ of Z is such that there exists $\alpha > 0$ and $c > 0$ for which we have for all $\delta > 0$*

$$\mathbb{P}_Z\left(0 < \left|B_{\mathbb{P}}(Z) - \lambda^* \bar{B}_{\mathbb{P}}(Z)\right| \leq \delta\right) \leq c\delta^{\alpha} \ .$$

We conjecture that the answer to this question is negative in general, yet, in some cases the fast rates of convergence are possible. We leave this line of research for the future.

1.1.4 Constrained classification: fairness (Section 2.2)

Another prominent setup of binary classification, where plug-in approaches can lead to state-of-the-art performance is the so called fair binary classification of equal opportunity [Hardt et al., 2016]. In this setup we slightly modify the observed data. Instead of plain $(X, Y) \in$

$\mathbb{R}^d \times \{0,1\}$ we additionally observe one more binary *feature* $S \in \{0,1\}$. This feature S is often referred as a sensitive feature or a protected attribute [Hardt et al., 2016, Barocas et al., 2018], with gender or race being typical interpretations. This observation model is motivated by the idea that historically the data that we collect are actually biased towards a more favorable decisions in either $S = 0$ or $S = 1$. In this manuscript we prefer to avoid extra discussion on legal, ethical, and sociological points of view and stick to the statistical side of the question, an interested reader can learn more from the excellent book [Barocas and Selbst, 2016] and references therein.

Our goal as statisticians has been formally stated by Hardt et al. [2016], where the authors proposed a way to measure the level of fairness of any given classifier $g : \mathbb{R}^d \times \{0,1\} \to \{0,1\}$. Namely, Hardt et al. [2016] introduce the following definition of equal opportunity

Definition 1 (Equal Opportunity [Hardt et al., 2016])**.** *A classifier $g : \mathbb{R}^d \times \{0,1\} \to \{0,1\}$ is called fair in terms of the equal opportunity if*

$$\mathbb{P}\left(g(X,S) = 1 \,|\, S = 1, Y = 1\right) = \mathbb{P}\left(g(X,S) = 1 \,|\, S = 0, Y = 1\right) \quad ,$$

where $\mathbb{P}$ is the distribution of the random tuple (X, S, Y).

A nice intuitive explanation of this definition is given in Hardt et al. [2016]: "... people who pay back $[Y = 1]$[12] their loan, have an equal opportunity of getting $[g(X,S) = 1]$[13] the loan in the first place (without specifying any requirement for those that will ultimately default)". Luckily, this setup perfectly fits the idea of constrained binary classification.

Using this definition of fair classifiers our goal is to approximate the following fair optimal (fair Bayes) classifier

$$g^* \in \arg\min_{g \in \mathcal{G}} \left\{\mathcal{R}(g) \,:\, \mathbb{P}\left(g(X,S) = 1 \,|\, Y = 1, S = 1\right) = \mathbb{P}\left(g(X,S) = 1 \,|\, Y = 1, S = 0\right)\right\} \quad ,$$

where the risk is given by $\mathcal{R}(g) = \mathbb{P}(Y \neq g(X,S))$. This is another example of a problem, where the constraints on the classifier are distribution dependent. Dependency on the *unknown* distribution does not allow in principle to construct an estimator $\hat{g}$ which would satisfy

$$\mathbb{P}\left(\hat{g}(X,S) = 1 \,|\, Y = 1, S = 1\right) = \mathbb{P}\left(\hat{g}(X,S) = 1 \,|\, Y = 1, S = 0\right) \quad ,$$

almost surely. Thus, apart from the classical excess risk, we additionally would like to control the magnitude of how this constraint is violated, that is, our goal is to control the unfairness of $\hat{g}$ defined as

$$\mathbb{E}\left|\mathbb{P}\left(\hat{g}(X,S) = 1 \,|\, Y = 1, S = 1\right) - \mathbb{P}\left(\hat{g}(X,S) = 1 \,|\, Y = 1, S = 0\right)\right| \quad .$$

This question is discussed in details in Section 2.2, where a semi-supervised plug-in procedure is proposed and its consistency is established. Besides, numerical results of Section 2.2 suggest that with properly chosen estimator of $\mathbb{P}\left(Y = 1 \,|\, X, s\right)$ for $s \in \{0,1\}$, the proposed procedure achieves state-of-the-art performance.

[12] Inserted by the author.

[13] Inserted by the author.

Other measures of fairness

Equal Opportunity is not the only possible way to define the set of fair classifiers and other notions have been recently introduced in the literature. It seems that at this moment[14] the agreement on which notion of fairness to use and in which domains is still missing. Consequently, in this part we want to provide other possible ways to define the notion of fairness. All the following examples are formulated as particular instances of the constrained classification framework with different sets of desired properties $\mathcal{G}_\theta$.

Definition 2 (Demographic Parity [Calders et al., 2009]). *A classifier $g : \mathbb{R}^d \times \{0,1\} \to \{0,1\}$ is called fair in terms of the demographic parity if*

$$\mathbb{P}\left(g(X,S)=1 \,|\, S=1\right) = \mathbb{P}\left(g(X,S)=1 \,|\, S=0\right) \quad ,$$

where $\mathbb{P}$ is the distribution of the random tuple (X,S,Y).

The notion of demographic parity implies that the probability of assigning an instance $X \in \mathbb{R}^d$ to the positive class (*i.e.,* $g(X,S)=1$) is the same for both values of the protected attribute $S \in \{0,1\}$.

Definition 3 (Equal Odds [Hardt et al., 2016]). *A classifier $g : \mathbb{R}^d \times \{0,1\} \to \{0,1\}$ is called fair in terms of the equal odds if*

$$\mathbb{P}\left(g(X,S)=1 \,|\, S=1, Y=y\right) = \mathbb{P}\left(g(X,S)=1 \,|\, S=0, Y=y\right) \quad \textit{for all} \quad y \in \{0,1\} \quad ,$$

where $\mathbb{P}$ is the distribution of the random tuple (X,S,Y).

The notion of equal odds implies that the prediction $g(X,S)$ and the protected attribute $S \in \{0,1\}$ are independent conditional on the true label $Y \in \{0,1\}$.

Definition 4 (Disparate Treatment [Zafar et al., 2017]). *A classifier $g : \mathbb{R}^d \times \{0,1\} \to \{0,1\}$ is called fair in terms of the disparate treatment if*

$$\mathbb{P}\left(g(X,S)=y \,|\, X=x, S=s\right) = \mathbb{P}\left(g(X,S)=y \,|\, X=x\right) \quad \textit{for all} \quad y,s \in \{0,1\}, x \in \mathbb{R}^d \quad ,$$

where $\mathbb{P}$ is the distribution of the random tuple (X,S,Y).

The notion of disparate treatment implies that the probability of prediction $g(X,S)$ to assign a feature $x \in \mathbb{R}^d$ to the positive class (*i.e.,* $g(X,S)=1$), is the same for both values of the protected attribute $S \in \{0,1\}$.

Again, we would like to avoid extra discussion on the sociological and ethical impact of these definitions as this is not the subject of the present manuscript. However, these examples are interesting from the point of view of the constrained classification framework and, in particular, in view of Section 1.1.3.

Let us also mention that sometimes in applications of fairness, the dependency of a classifier g on the protected attribute is forbidden. In other words, the set of classifiers is defined as $x \mapsto g(x)$, that is, these classifiers do not take into account the protected feature $S \in \{0,1\}$ for prediction. For example, in the case of the equal opportunity, the Bayes classifier g^* would be given as a solution of the following optimization problem

$$\min\left\{\mathbb{P}(Y \neq g(X)) \,:\, \mathbb{P}\left(g(X)=1 \,|\, Y=1, S=1\right) = \mathbb{P}\left(g(X)=1 \,|\, Y=1, S=0\right)\right\} \quad .$$

[14] At the moment of writing this manuscript.

In Section 2.2 we also derive the expression of the Bayes classifier in this case and propose a semi-supervised procedure to mimic this Bayes rule. Importantly, for this setup we assume that the constructed algorithm $\hat{g}$ has an access to labeled data $\mathcal{D}_n^{\mathrm{L}} = \{(X_i, S_i, Y_i)\}_{i=1}^n$ which includes the protected attribute. This does not contradict the requirement of independence from $S \in \{0, 1\}$ for the algorithm $\hat{g}$. Indeed, the proposed algorithm $\hat{g}$ uses the protected attribute *only* at the training phase but not at the prediction phase.

1.2 Multi-class classification

In multi-class classification each instance $X \in \mathbb{R}^d$ is associated with a class Y that takes values in $[K] := \{1, \dots, K\}$. Additionally, it is assumed that (X, Y) follows some distribution $\mathbb{P}$ on $\mathbb{R}^d \times [K]$. Typical examples of applications of this framework include image classification [LeCun et al., 1998], web advertising [Beygelzimer et al., 2009], and document categorization [Dekel and Shamir, 2010] to name few.

1.2.1 Standard setup

Similarly, to the binary case described in Section 1.1.1, for the standard settings of multi-class classification we would like to study the misclassification risk defined for all classifiers g as

$$\mathcal{R}(g) = \mathbb{P}(Y \neq g(X)) \ .$$

By analogy with the binary case we can define K different regression functions $\eta_k(x) = \mathbb{P}(Y = k | X = x)$ for all $k \in [K]$ and all $x \in \mathbb{R}^d$. Besides, one can easily show that a Bayes optimal classifier g^* can be defined for all $x \in \mathbb{R}^d$ as

$$g^*(x) = \arg\max_{k \in [K]} \eta_k(x) \ ,$$

which in case $K = 2$ reduces to the well-known binary rule. Logically, one would expect that the minimax analysis of Audibert and Tsybakov [2007] can be carried out to these settings. However, in order to extend their analysis to the multi-class case one also needs to understand how to adapt the notion of margin assumption in this setting. A possible extension was given by Dinh et al. [2015], where the authors assumed that the regression functions are such that

$$\mathbb{P}(0 < \eta_{(1)}(X) - \eta_{(2)}(X) \leq \delta) \leq c\delta^\alpha \ ,$$

where $\eta_{(1)}(\cdot)$ and $\eta_{(2)}(\cdot)$ are the largest and the second largest regression functions. Dinh et al. [2015] showed that there exists an algorithm $\hat{g}$ which achieves

$$\mathbb{E}[\mathcal{R}(\hat{g})] - \mathcal{R}(g^*) \lesssim n^{-\frac{(1+\alpha)\beta}{2\beta+d}} \ ,$$

where β is the common smoothness parameter of the η_k's. Even though they did not provide a lower bound on the excess risk, it is expected that their rate is optimal as one can always construct $\eta_1, \dots, \eta_K$ such that $\eta_3 \equiv \dots \equiv \eta_K \equiv 0$ and execute the classical arguments using only η_1, η_2 to obtain the lower-bound.

1.2.2 Confidence set setup (Chapter 3)

Another prominent approach to the problem of multi-class classification is through confidence sets. The goal in this setup is very different from its standard counterpart. In the confidence set approach our intention is to find a measurable mapping $\Gamma : \mathbb{R}^d \to 2^{[K]}$, where $2^{[K]}$ is the set of all subsets of $[K]$, which is in some sense optimal. Any measurable mapping Γ of this type is called a confidence set and for each $x \in \mathbb{R}^d$ we denote by $|\Gamma(x)|$ the cardinal of the set $\Gamma(x)$.

A logical way to generalize the misclassification risk considered in the standard setup is to define the following quantity for each Γ

$$\mathrm{P}(\Gamma) := \mathbb{P}(Y \notin \Gamma(X)) \ ,$$

which is viewed as an error of the confidence set Γ. It becomes apparent that minimization of $\mathrm{P}(\cdot)$ is useless without any constraints. Indeed, the set $\Gamma^{\mathrm{all}} \equiv [K]$ would minimize this quantity for any possible distributions $\mathbb{P}$. The main reason why the confidence set Γ^{all} is unsuitable is its size, that is, the number of classes that it outputs. This discussion gives a motivation to introduce another property of a set Γ which is related to its size. It can be done in at least two straightforward ways – for each confidence set Γ we can define its size as

$$\begin{aligned} \mathrm{I}_{\sup}(\Gamma) &= \sup_{x \in \mathbb{R}^d} |\Gamma(x)| \, , && \text{(largest size)} \ , \\ \mathrm{I}_{\mathbb{E}}(\Gamma) &= \mathbb{E} \, |\Gamma(X)| \, , && \text{(expected size)} \ . \end{aligned}$$

Using any of these two notions for the size we can define a Bayes confidence set as

$$\Gamma^*_\beta \in \arg\min \{\mathrm{P}(\Gamma) \ : \ \mathrm{I}_\square(\Gamma) \leq \beta\} \ \ ,$$

where $\beta \in [K]$ is a parameter that controls the size of the optimal set and $\square$ stands for sup or $\mathbb{E}$. Interestingly, under some mild assumptions which are discussed in details in Section 3.1, one can demonstrate that the Bayes confidence set actually satisfies $\mathrm{I}_\square(\Gamma^*_\beta) = \beta$.

This problem with *expected* size was first considered by Denis and Hebiri [2017] who provided an ERM type algorithm to mimic Γ^*_β. Later, some non-trivial consequences were made in [Chzhen et al., 2019a] concerning the performance of semi-supervised algorithms in this context. Namely, it is shown that properly defined *supervised* methods cannot achieve rates which are faster than $1/\sqrt{n}$ even under a suitable margin assumption. Whereas, some *semi-supervised* approaches allow to bypass this issue and allow to recover typical minimax rates of convergence, provided that the size of unlabeled set is sufficiently large.

In Section 3.1 the problem of confidence set classification with controlled *expected* size is considered and semi-supervised algorithms are studied from the minimax point of view.

1.3 Multi-label classification (Chapter 4)

In the setup of multi-label classification, unlike binary or multi-class case, each instance $X \in \mathbb{R}^d$ is associated with a binary vector $Y \in \{0, 1\}^L$. This framework encompasses a number of applications such as text categorization [Gao et al., 2004, Tsoumakas et al., 2009, Partalas et al., 2015], functional genomics [Barutcuoglu et al., 2006], image classification [Li et al., 2014], and recommendation systems [Agrawal et al., 2013, Prabhu et al., 2018] among others.

Statistical framework of this problem assumes that the couple (X, Y) is random and it follows some distribution $\mathbb{P}$ on $\mathbb{R}^d \times \{0,1\}^L$. A standard notion of a classifier g in this context is then given as $g : \mathbb{R}^d \to \{0,1\}^L$, that is, a classifier is a vector valued binary function defined on $\mathbb{R}^d$. In this context there are two straightforward ways to generalize the misclassification risk, both of which lead to different Bayes optimal classifiers, see [Dembczyński et al., 2012] and references therein. The first one is the 0/1-risk, which is defined for all classifiers g as

$$\mathcal{R}^{0/1}(g) = \mathbb{P}(Y \neq g(X)) \qquad (0/1 \text{ risk}) \ .$$

Note that this risk compares the whole vector Y to the prediction $g(X)$. The 0/1-risk does not take into account the idea that $g(X)$ can be "approximately" correct in the sense that it makes mistakes only on few coordinates of $Y \in \{0,1\}^L$. This observation motivates to consider a modified version of this risk, which is often referred as the Hamming risk, defined for all classifiers g as

$$\mathcal{R}^{\mathrm{H}}(g) = \frac{1}{L} \sum_{l=1}^{L} \mathbb{P}(Y^l \neq g^l(X)) \qquad (\text{Hamming risk}) \ .$$

Contrary to the 0/1-risk the later takes into account mistakes on each coordinate of the vector $Y \in \{0,1\}^L$ separately.

It is interesting to point out that modern applications of multi-label classification are typically asymmetric in their treatment of the labels, see [Prabhu and Varma, 2014] and references therein. To illustrate this, consider the following example connected to the image recognition – for some $l \in [L]$ the outcome $Y^l = 1$ means that the object assigned to l^{th} coordinate is present on a picture $X \in \mathbb{R}^d$. In this setup it is unreasonable to expect that there are pictures with *all* possible objects, that is, the label vector $Y \in \{0,1\}^L$ is actually sparse. Apart from that, in many practical applications, it is much more beneficial to correctly predict the occurrence of $Y^l = 1$, than $Y^l = 0$. This idea motivated the community to consider the following risk

$$\mathcal{R}^{\mathrm{fn}}(g) = \frac{1}{L} \sum_{l=1}^{L} \mathbb{P}(Y^l = 1, g^l(X) = 0) \ ,$$

which only considers the false negative errors of a classifier g.

Clearly, minimization of $\mathcal{R}^{\mathrm{fn}}$ without any constraints does not make any sense, since $g(X) \equiv (1, \ldots, 1)^\top$ is optimal in this case. Thus, some restrictions are necessary on the set of possible classifiers. Again, as in previous sections the constrained classification framework is well suited in this case. Namely, we consider some set of multi-label classifiers $\mathcal{G}_\theta$ and target the following Bayes rule

$$g^* \in \arg\min \left\{ \mathcal{R}^{\mathrm{fn}}(g) \ : \ g \in \mathcal{G}_\theta \right\} \ .$$

The main question now is what kind of classes $\mathcal{G}_\theta$ we can consider and what kind of algorithms we can construct.

The first possible constraint is motivated by recommendation systems, where we can recommend only a *fixed* number of possible objects. For this situation, a reasonable approach is given by the following sparse Bayes rule

$$g^* \in \arg\min \left\{ \mathcal{R}^{\mathrm{fn}}(g) \ : \ \sum_{l=1}^{L} \mathbb{1}_{\{g^l(X)=1\}} = K, \ \text{a.s.} \right\} \ ,$$

for some $K \in [L]$ chosen by a practitioner. Note, that unlike all previous examples, here the constraint is almost surely and not in expectation.

A possible drawback of this approach is the fact that the almost sure control on the sparsity of g does not give any control on the false positive mistakes. That is, in some situation the value $K \in [L]$ might be over optimistic, and lower values can deliver comparable predictions. In such a case we might consider the setting with an almost sure control over false positive discoveries and target the following Bayes rule

$$g^* \in \arg\min \left\{ \mathcal{R}^{\mathrm{fn}}(g) \ : \ \sum_{l=1}^{L} \mathbb{P}(Y^l = 0, g^l(X) = 1 | X) \leq \beta, \ \text{a.s.} \right\} \ ,$$

for some $\beta > 0$ to be specified. In Chapter 4 we consider these constrained settings in details and provide rates of convergence in both cases. Interestingly, it is shown in Chapter 4 that the problem with almost sure control over false negative discoveries is in some sense hopeless if we do not impose additional, rather restrictive, assumptions on the distribution of (X, Y).

1.4 Organization of the manuscript

This manuscript is partitioned into three interconnected parts: binary classification (Chapter 2); multi-class classification (Chapter 3); multi-label classification (Chapter 4). Each chapter is self-contained, that is, notation and the proofs are provided with a broad discussion on the relevance of the contribution. The following works are at the core of the present manuscript.

- Chapter 2
  - E. Chzhen. "Optimal rates for F-score binary classification". *Submitted*, 2019; [Chzhen, 2019a].
  - E. Chzhen, C. Denis, M. Hebiri, L. Oneto, and M. Pontil. "Leveraging Labeled and Unlabeled Data for Consistent Fair Binary Classification". *NeurIPS*, 2019; [Chzhen et al., 2019b].
- Chapter 3
  - E. Chzhen, C. Denis, and M. Hebiri. "Minimax semi-supervised confidence sets for multi-class classification". *Submitted*, 2019; [Chzhen et al., 2019a].
- Chapter 4
  - E. Chzhen. "Classification of sparse binary vectors". *Submitted*, 2019; [Chzhen, 2019b].
- Not included in the manuscript
  - E. Chzhen, M. Hebiri, and J. Salmon. "On Lasso refitting strategies". *Bernoulli*, 2019; [Chzhen et al., 2019c].
  - E. Chzhen, C. Denis, M. Hebiri, and J. Salmon. "On the benefits of output sparsity for multi-label classification". *Technical report*, 2017; [Chzhen et al., 2017].

1.5 Resumé en français

La contribution de ce manuscrit consiste en quatre parties distinctes interconnectées. Plus précisément, nous considérons quatre problèmes de classification différents, que nous abordons d'un point de vue théorique. Le point commun entre les divers cadres d'études est le type d'approche que nous utilisons pour résoudre le problème. En effet, nous proposons des algorithmes de type *plug-in* pour chacun problème étudié.

Rappelons brièvement la cadre générale du problème de classification. Nous considérons un couple $(X, Y) \in \mathbb{R}^d \times \mathcal{Y}$ lorsque $\mathcal{Y}$ est un ensemble fini. Le vecteur $X \in \mathbb{R}^d$ est appelé vecteur des caractéristiques et $Y \in \mathcal{Y}$ est l'étiquette (ou *label*) associée à $X \in \mathbb{R}^d$. De plus, on suppose que le couple (X, Y) est distribué selon une loi inconnue $\mathbb{P}$ sur $\mathbb{R}^d \times \mathcal{Y}$, et on note $\mathbb{P}_X$ la distribution marginale de $X \in \mathbb{R}^d$. En outre, on suppose que pour des entiers $n, N \geq 0$, deux jeux de données sont observés – $\mathcal{D}_n^{\mathrm{L}} = \{(X_i, Y_i)\}_{i=1}^n$ *i.i.d.* de $\mathbb{P}$ et $\mathcal{D}_N^{\mathrm{U}} = \{X_i\}_{i=1}^N$ *i.i.d.* de $\mathbb{P}_X$. L'ensemble $\mathcal{D}_n^{\mathrm{L}}$ est appelé jeu de données étiquetées. On parler alors d'apprentissage supervisé alors que l'ensemble $\mathcal{D}_N^{\mathrm{U}}$ est appelé jeu de données non-étiquetées et on parle alors d'apprentissage non-supervisées.

Dans ce contexte, notre objectif est de construire un algorithme capable, en utilisant les ensembles de données étiquetées et non étiquetées, d'inférer l'étiquette $Y \in \mathcal{Y}$ pour une nouvelle observation $X \in \mathbb{R}^d$. Ainsi, pour chacun des quatre cadres d'étude, nous spécifions en premier lieu une notion de classification adaptée au problème considéré. Deuxièmement, nous introduisons une notion de risque, qui traduit la qualité de la règle envisagée.

Dans ce qui suit, nous décrivons les quatre problèmes considérés dans ce manuscrit et présentons les principaux résultats obtenus pour chaque problème.

1.5.1 F-score

Dans le contexte de la classification avec F-score, l'ensemble $\mathcal{Y} = \{0, 1\}$, l'ensemble des règles de classification est défini comme $\mathcal{G} = \{g : \mathbb{R}^d \to \{0, 1\}\}$, et la fonction de score est définie pour chaque $g \in \mathcal{G}$ comme

$$F_1(g) \coloneqq \frac{2\mathbb{P}(Y = 1, g(X) = 1)}{\mathbb{P}(Y = 1) + \mathbb{P}(g(X) = 1)} \qquad \text{(F-score)} \ .$$

La meilleure règle de classification est celle qui maximise le F-score, c'est-à-dire

$$g^* \in \arg\max \{F_1(g) \ : \ g \in \mathcal{G}\} \ .$$

Concernant ce problème, on considère les questions suivantes.

- Est-ce que g^* admet une écriture explicite ?
- Qu'est-ce qu'un algorithme optimal dans ce contexte ?
- Quelle est la vitesse de convergence minimax ?

La première question a déjà été abordée dans [Zhao et al., 2013] où il est démontré que le classifieur de Bayes g^* peut être défini pour tout $x \in \mathbb{R}^d$ par

$$g^*(x) = \mathbb{1}_{\{\eta(x) > \theta^*\}} \ , \tag{1.12}$$

où $\eta(\cdot) = \mathbb{P}(Y = 1 \,|\, X = \cdot)$ et $\theta^* \in [0, 1]$ est un seuil qui satisfait l'équation

$$\theta^* \mathbb{P}(Y = 1) = \mathbb{E}(\eta(X) - \theta^*)_+ \ .$$

En Section 2.1 du Chapitre 2 nous abordons ce problème statistique plus en profondeur. En particulier, nous proposons un algorithme *semi-supervisé* $\hat{g}$ qui, sous des hypothèses similaires à celles utilisées par Audibert and Tsybakov [2007], satisfait

$$F_1(g^*) - \mathbb{E}_{(\mathcal{D}_n^{\mathrm{L}}, \mathcal{D}_N^{\mathrm{U}})}\left[F_1(\hat{g})\right] \lesssim n^{-\frac{(1+\alpha)\beta}{2\beta+d}} \ .$$

Il est en outre démontré que la vitesse de convergence ci-dessus est optimale au sens minimax. Cette vitesse est obtenue en supposant qu'une version modifiée de l'hypothèse de marge est vérifiée. Avec les notations que nous avons introduites, celle-ci s'écrit de la façon suivante: pour tout $\delta > 0$

$$\mathbb{P}_X(0 < |\eta(X) - \theta^*| \leq \delta) \leq c\delta^\alpha \ ,$$

pour certaines constantes $c, \alpha > 0$.

1.5.2 Classification équitable

Comme dans la partie précédente, nous avons $\mathcal{Y} = \{0, 1\}$, mais le vecteur des caractéristiques est défini ici comme un couple $(X, S) \in \mathbb{R}^d \times \{0, 1\}$. L'ensemble des règles de classification est donné par $\mathcal{G} = \{g : \mathbb{R}^d \times \{0, 1\} \to \{0, 1\}\}$. Dans ce contexte, la variable binaire S est vue comme un attribut sensible et l'objectif est de construire une règle de classification équitable, c'est-à-dire, qui ne discrimine pas par rapport à S au sens de la définition suivante:

Definition 5 (Equal Opportunity [Hardt et al., 2016])**.** *Une règle de classification* $g : \mathbb{R}^d \times \{0, 1\} \to \{0, 1\}$ *est dite équitable en termes d'opportunité ou* equal opportunity*, si*

$$\mathbb{P}(g(X, S) = 1 \,|\, S = 1, Y = 1) = \mathbb{P}(g(X, S) = 1 \,|\, S = 0, Y = 1) \ ,$$

où $\mathbb{P}$ *est la distribution de* (X, S, Y)*.*

En utilisant cette notion de classifieur équitable, notre objectif est d'imiter au mieux la règle de classification équitable optimale définie par

$$g^* \in \arg\min \{\mathcal{R}(g) \ : \ \mathbb{P}(g(X, S) = 1 \,|\, Y = 1, S = 1) = \mathbb{P}(g(X, S) = 1 \,|\, Y = 1, S = 0)\} \ ,$$

où le risque est donné par $\mathcal{R}(g) = \mathbb{P}(Y \neq g(X, S))$. Ceci est un exemple de problème où les contraintes sur les règles de classification dépendent de la distribution. La dépendance en la distribution inconnue ne permet pas en principe de construire un estimateur $\hat{g}$ qui satisferait

$$\mathbb{P}(\hat{g}(X, S) = 1 \,|\, Y = 1, S = 1) = \mathbb{P}(\hat{g}(X, S) = 1 \,|\, Y = 1, S = 0) \ ,$$

presque sûrement. Ainsi, outre l'excès de risque classique, nous voudrions également contrôler l'ampleur de la violation de cette contrainte, c'est-à-dire que notre objectif est de contrôler la différence suivante :

$$\mathbb{E}\left|\mathbb{P}(\hat{g}(X, S) = 1 \,|\, Y = 1, S = 1) - \mathbb{P}(\hat{g}(X, S) = 1 \,|\, Y = 1, S = 0)\right| \ .$$

Cette question est discutée en détail dans la Section 2.2, où une procédure de type *plug-in* semi-supervisée est proposée de même que sa consistance est établie. En outre, les résultats numériques de la Section 2.2 suggèrent qu'avec un estimateur de $\mathbb{P}(Y = 1 \,|\, X, s)$ bien choisi, la procédure proposée permet d'atteindre des performances satisfaisantes.

1.5.3 Ensembles des confiances

Dans cette partie, nous examinons le problème de la classification multi-classes. C'est-à-dire que l'ensemble des étiquettes $Y \in \mathcal{Y} = \{1, \dots, K\}$ pour $K \in \mathbb{N} \setminus \{0, 1\}$. Contrairement aux approches standard, nous considérons une approche par ensembles de confiance. Un ensemble de confiance Γ est défini comme $\Gamma : \mathbb{R}^d \to 2^{\mathcal{Y}}$. Nous introduisons deux caractéristiques de Γ – *l'erreur* et *l'information* définies respectivement par

$$\mathrm{P}(\Gamma) = \mathbb{P}(Y \notin \Gamma(X)), \qquad \mathrm{I}(\Gamma) = \mathbb{E}\,|\Gamma(X)| \quad .$$

Notre point de départ est l'ensemble de confiance oracle défini pour $\beta \in \{1, \dots, K\}$ comme

$$\Gamma^* \in \arg\min \{\mathrm{P}(\Gamma) \,:\, \mathrm{I}(\Gamma) = \beta\} \quad .$$

Comme d'habitude, nous aimerions construire un algorithme $\hat{\Gamma}$ et contrôler son excès de risque. Ce problème a été examiné pour la première fois par Denis and Hebiri [2017] qui ont fourni un algorithme de type minimisation du risque empirique pour imiter Γ^*_β. Dans l'article [Chzhen et al., 2019a], nous apportons une analyse plus fine de ces ensembles de confiance et fournissons des conséquences non triviales sur les performances d'algorithmes semi-supervisés de type *plug-in* dans ce contexte. En particulier, il est démontré qu'aucune méthode *supervisée* ne peut atteindre des vitesses plus rapides que $1/\sqrt{n}$, même sous une hypothèse de marge favorable. Au contraire, certaines approches *semi-supervisées* permettent de contourner cette limite et de récupérer des vitesses de convergence minimax typiques, à condition que la taille de l'ensemble non étiqueté soit suffisamment grande.

Dans la section 3.1, le problème de la classification des ensembles de confiance avec une taille contrôlée *en espérance* est considéré et les algorithmes semi-supervisés sont étudiés du point de vue minimax.

1.5.4 Multi-label

Enfin, dans la dernière partie de ce manuscrit, nous considérons $\mathcal{Y} = \{0, 1\}^L$ pour $L \in \mathbb{N} \setminus \{0, 1\}$ et les règles de classification sont définies par $g : \mathbb{R}^d \to \{0, 1\}^L$. Ce problème s'appelle la classification multi-label.

Il est intéressant de noter que les applications modernes de la classification multi-label sont généralement asymétriques dans le traitement des étiquettes, voir [Prabhu and Varma, 2014] et les références qui y figurent. Pour illustrer ce fait, considérons l'exemple suivant lié à la reconnaissance d'image - pour certains $l \in [L]$, le résultat $Y^l = 1$ signifie que l'objet assigné à la l^{ieme} coordonnée est présent dans une image $X \in \mathbb{R}^d$. Dans ce contexte, il n'est pas raisonnable de s'attendre à ce qu'il y ait des images avec *tous* les objets possibles, c'est-à-dire que le vecteur des étiquettes $Y \in \{0, 1\}^L$ est parcimonieux. En dehors de cela, dans de nombreuses applications pratiques, il est beaucoup plus avantageux de prédire correctement l'occurrence de $Y^l = 1$, que $Y^l = 0$. Cette idée a motivé la communauté à étudier le risque suivant

$$\mathcal{R}^{\mathrm{fn}}(g) = \frac{1}{L} \sum_{l=1}^{L} \mathbb{P}(Y^l = 1, g^l(X) = 0) \quad ,$$

qui ne considère que le taux de faux négatifs d'une règle de classification g donnée. Nous étudions deux types de problèmes dans ce contexte. Dans le premier, nous définissons la

règle de classification optimale comme

$$g^* \in \arg\min \left\{ \mathcal{R}^{\mathrm{fn}}(g) \, : \, \sum_{l=1}^{L} \mathbb{1}_{\left\{g^l(X)=1\right\}} = K, \ \text{p.s.} \right\} \ ,$$

pour un nombre d'étiquettes positives $K \in [L]$ préalablement choisi par un statisticien. Pour le second problème, nous définissons la règle de classification optimale comme

$$g^* \in \arg\min \left\{ \mathcal{R}^{\mathrm{fn}}(g) \, : \, \sum_{l=1}^{L} \mathbb{P}(Y^l = 0, g^l(X) = 1 | X) \leq \beta, \ \text{p.s.} \right\} \ ,$$

qui contrôle le taux de faux positifs par le biais du paramètre réel $\beta > 0$ également préalablement choisi par un statisticien.

Dans le chapitre 4, nous examinons ces problèmes en détail et fournissons des vitesses de convergence dans les deux cas. Un point notable distingue les deux cadres considérés : il est prouvé au chapitre 4 que le problème du contrôle presque sûr du taux de faux négatifs est en quelque sorte désespéré si nous n'imposons pas une hypothèse supplémentaire, et plutôt restrictive, sur la distribution de (X, Y).

Chapter 2

Binary classification

In this chapter we study two problems of binary classification and provide semi-supervised algorithms for each of them. Section 2.1 considers the problem of binary classification with F-score, briefly discussed in Section 1.1.2. Section 2.2 studies the problem of fair binary classification of equal opportunity mentioned in Section 1.1.4.

2.1 F-score

Section overview. We study the minimax settings of binary classification with F-score under the β-smoothness assumptions on the regression function $\eta(x) = \mathbb{P}(Y = 1|X = x)$ for $x \in \mathbb{R}^d$. We propose a classification procedure which under the α-margin assumption achieves the rate $\mathcal{O}(n^{-(1+\alpha)\beta/(2\beta+d)})$ for the excess F-score. It is known that the Bayes optimal classifier for the F-score can be obtained by thresholding the aforementioned regression function η on some level θ^* to be estimated. The proposed procedure is performed in a semi-supervised manner, that is, for the estimation of the regression function we use a labeled dataset of size $n \in \mathbb{N}$ and for the estimation of the optimal threshold θ^* we use an unlabeled dataset of size $N \in \mathbb{N}$. Interestingly, the value of $N \in \mathbb{N}$ does not affect the rate of convergence, which indicates that it is "harder" to estimate the regression function η than the optimal threshold θ^*. This further implies that the binary classification with F-score behaves similarly to the standard settings of binary classification. Finally, we show that the rates achieved by the proposed procedure are optimal in the minimax sense up to a constant factor.

2.1.1 Introduction

The problem of binary classification is among the most basic and well-studied problems in statistics and machine learning [Vapnik, 1998, Yang, 1999, Bartlett and Mendelson, 2002, Audibert, 2004, Massart and Nédélec, 2006, Audibert and Tsybakov, 2007]. Until very recently, theoretical guarantees were almost exclusively formulated in terms of the probability of misclassification (a.k.a accuracy) as the measure of the risk. This way of measuring the risk is well suited in the case of "well-balanced" distributions and datasets, that is, when the occurrence of both classes are similar.

Once this assumption fails to be satisfied, classifiers based on the accuracy might perform poorly in practice and fail to be relevant. A possible approach to treat such an unbalanced situation is to modify the measure to be optimized in an appropriate way. A popular choice for such a measure is the F-score, whose roots can be traced back to the information retrieval

literature van Rijsbergen [1974], Lewis [1995]. From the statistical point of view there are two alternative approaches [Ye et al., 2012, Dembczynski et al., 2017] to the theoretical treatment of the F-score: Population Utility (PU) and Expected Test Utility (ETU). In this chapter we follow the PU approach which, as noted in [Dembczynski et al., 2017], has stronger roots in classical statistics. In contrast, the ETU framework favors classifiers which optimize the expected prediction error over test sets of predefined size and this framework is more related to the statistical machine learning.

Our goal is to provide minimax analysis of the binary classification with F-score under non-parametric assumptions.

2.1.2 The problem formulation

For any two real numbers $a, b \in \mathbb{R}$ we denote by $a \wedge b$ (resp. $a \vee b$) the minimum (resp the maximum) between a and b. The standard Euclidean norm in $\mathbb{R}^d$ is denoted by $\|\cdot\|_2$ and a ball centered at $x \in \mathbb{R}^d$ of radius r is denoted by $\mathcal{B}(x, r)$. For positive real valued sequences $a_n, b_n : \mathbb{N} \mapsto \mathbb{R}_+$ we say that $a_n = \mathcal{O}(b_n)$ if there exists some positive constant $M > 0$ such that for all $n \in \mathbb{N}$ it holds that $a_n / b_n \leq M$. We consider a random couple (X, Y) taking values in $\mathbb{R}^d \times \{0, 1\}$ with joint distribution $\mathbb{P}$. The vector $X \in \mathbb{R}^d$ is the feature vector and the binary variable $Y \in \{0, 1\}$ is the label. As a technical assumption we assume that $\mathbb{P}(Y = 1) \neq 0$ in what follows. We denote by $\mathbb{P}_X$ the marginal distribution of the feature vector $X \in \mathbb{R}^d$ and by $\eta(X) \coloneqq \mathbb{P}(Y = 1|X)$ the regression function. A classifier is any measurable function $g : \mathbb{R}^d \mapsto \{0, 1\}$ and the set of all such functions is denoted by $\mathcal{G}$. We assume that we have access to two datasets: the first dataset $\mathcal{D}_n^{\mathrm{L}} = \{(X_i, Y_i)\}_{i=1}^n$ consists of $n \in \mathbb{N}$ *i.i.d.* copies of $(X, Y) \sim \mathbb{P}$; and the second dataset $\mathcal{D}_N^{\mathrm{U}} = \{X_i\}_{i=n+1}^{n+N}$ consists of $N \in \mathbb{N}$ independent copies of $X \sim \mathbb{P}_X$. Denote by $\mathbb{P}^{\otimes n}$ and $\mathbb{P}_X^{\otimes N}$ the distributions of $\mathcal{D}_n^{\mathrm{L}}$ and $\mathcal{D}_N^{\mathrm{U}}$ respectively. Moreover, we denote by $\mathbb{E}_{(\mathcal{D}_n^{\mathrm{L}}, \mathcal{D}_N^{\mathrm{U}})}$ the expectation with respect to the distribution of $(\mathcal{D}_n^{\mathrm{L}}, \mathcal{D}_N^{\mathrm{U}})$, that is, with respect to $\mathbb{P}^{\otimes n} \otimes \mathbb{P}_X^{\otimes N}$ on the space $\Big(\mathbb{R}^d \times \{0, 1\}\Big)^n \times \Big(\mathbb{R}^d\Big)^N$. We additionally assume that the size of the unlabeled dataset is not smaller that the size of the labeled dataset, that is, $N \geq n$[1]. For a given classifier $g : \mathbb{R}^d \mapsto \{0, 1\}$ we define its F_b-score[2] for any $b > 0$ by

$$F_b(g) \coloneqq \frac{\mathbb{P}(Y = 1, g(X) = 1)}{b^2 \mathbb{P}(Y = 1) + \mathbb{P}(g(X) = 1)} \ .$$

A Bayes-optimal classifier $g^* : \mathbb{R}^d \mapsto \{0, 1\}$ is any classifier that maximizes the F-score over $\mathcal{G}$, that is,

$$g^* \in \arg\max_{g \in \mathcal{G}} F_b(g) \ .$$

As we have already mentioned in Section 1.1.2, it has been established by Zhao et al. [2013] that a maximizer of the F_1-score can be obtained by comparing the regression function $\eta(X)$ with a threshold $\theta^* \in [0, 1]$. Recall that this threshold depends explicitly on the distribution $\mathbb{P}$ and can be obtained as a unique root of

$$\theta \mapsto \theta \mathbb{P}(Y = 1) - \mathbb{E}(\eta(X) - \theta)_+ \ .$$

[1] Note that one can always satisfy this assumption by augmenting $\mathcal{D}_N^{\mathrm{U}}$ using a portion of $\mathcal{D}_n^{\mathrm{L}}$ and erasing labels. Typically, in practice it is easier to gather the unlabeled data then labeled, that is why this assumption is rather a formality.

[2] We decided to divide the classical definition of the F_b-score by the factor $1 + b^2$ to simplify the notation, thus, it is sufficient to multiply the obtained results by $1 + b^2$, to recover the results on the classical definition of the F_b-score.

In this chapter we generalize the result of [Zhao et al., 2013, Section 6] for an arbitrary value of $b > 0$, the proof can be found in Section 2.1.6.

Theorem 4. *A Bayes-optimal classifier g^* can be obtained point-wise for all $x \in \mathbb{R}^d$ as*

$$g^*(x) = \mathbb{1}_{\{\eta(x) > \theta^*\}} \ , \tag{2.1}$$

where $\theta^* \in [0, 1]$ is a threshold which is obtained as a unique solution of

$$b^2 \theta^* \mathbb{P}(Y = 1) = \mathbb{E}(\eta(X) - \theta^*)_+ \ .$$

Moreover, the classifier g^* satisfies $F_b(g^*) = \theta^*$.

Notice that if the optimal threshold $\theta^* \in [0, 1]$ is known a priori, the problem of binary classification with the F-score is no harder than the standard settings of binary classification with the accuracy as a performance measure. Though, in general the threshold $\theta^* \in [0, 1]$ should be estimated using the data as it depends on the distribution $\mathbb{P}$. Theorem 4 also implies an upper bound on the threshold θ^*, indeed, since $\theta^* = F_b(g^*)$ and for any classifier $g \in \mathcal{G}$ the F-score is upper-bounded by $1/(1 + b^2)$ we have $\theta^* \in [0, 1/(1 + b^2)]$.

For any classifier $g : \mathbb{R}^d \mapsto \{0, 1\}$ we define its excess score as

$$\mathcal{E}_b(g) \coloneqq F_b(g^*) - F_b(g), \quad \text{(excess score)} \ .$$

the excess score is the central object of our analysis and one of our goals here is to provide an estimator whose excess score is as small as possible. Using Theorem 4 we can show that the excess score of any classifier $g : \mathbb{R}^d \mapsto \{0, 1\}$ can be written in a simple form.

Lemma 2. *Let $g : \mathbb{R}^d \mapsto \{0, 1\}$ be any classifier and assume that $\mathbb{P}(Y = 1) \neq 0$, then*

$$\mathcal{E}_b(g) = \frac{\mathbb{E}\left[|\eta(X) - \theta^*| \, \mathbb{1}_{\{g^*(X) \neq g(X)\}}\right]}{b^2 \mathbb{P}(Y = 1) + \mathbb{P}(g(X) = 1)} \ .$$

Let us mention that in general the Bayes optimal rule is not unique and Theorem 4 only states that one of the optimal classifiers has the form described by Theorem 4. The function $\theta \mapsto b^2 \theta \mathbb{P}(Y = 1) - \mathbb{E}(\eta(X) - \theta)_+$ has a unique root (see Section 2.1.6 for the proof), however, other thresholds may result in the same Bayes rule. Indeed, consider a simple example $\eta(x) \equiv 1/2$, $b = 1$, then it is easy to see that the solution θ^* of $\theta/2 = (1/2 - \theta)_+$ is exactly $1/3$, and every Bayes optimal classifier predicts one almost surely. Clearly, any threshold $\theta \in [0, 1/2)$ of the regression function η results in the same classifier. Importantly, Lemma 2 and the equality $\arg\max_{g \in \mathcal{G}} F_1(g) = \theta^*$ are valid *only* for the threshold $\theta^* = 1/3$. In this chapter, we shall always refer to θ^* being the solution of $b^2 \theta \mathbb{P}(Y = 1) = \mathbb{E}(\eta(X) - \theta)_+$ and we refer to this threshold as the *optimal* threshold.

Remark 2. *For the rest of this section of Chapter 2, we focus only on the value $b = 1$ to simplify the presentation. It will be clear from our arguments that the generalization of our theoretical results to an arbitrary value $b > 0$ follows straightforwardly from our analysis.*

Interestingly, the results above demonstrate that the problem of binary classification with F-score has a lot in common with the standard settings. Indeed, in both cases the Bayes optimal classifier is obtained via thresholding the regression function and the expression for the excess risk is also similar. Consequently, following Section 1.1.2 in this chapter we address the following questions

Q1.: Is the problem of binary classification with F-score harder than its more known counterpart? In particular, can the minimax analysis of Audibert and Tsybakov [2007] be extended to these settings and what is an optimal algorithm?

Q2.: We wonder whether the introduction of unlabeled dataset can improve classification algorithms in the context of F-score.

Lemma 2 is crucial for our analysis as it allows to use the scheme provided by Audibert and Tsybakov [2007] for the standard setting of the binary classification. However, as the threshold $\theta^* \in [0,1]$ is unknown beforehand, this machinery cannot be applied in a straightforward way and some effort is required. In this chapter, we pose similar assumptions on the distribution $\mathbb{P}$ to the ones used in [Audibert and Tsybakov, 2007].

Assumption 4 (α-margin assumption)**.** *We say that the distribution $\mathbb{P}$ of the pair $(X,Y) \in \mathbb{R}^d \times \{0,1\}$ satisfies the α-margin assumption if there exist constants $C_0 > 0$, $\delta_0 \in (0, 1/12]$ and $\alpha > 0$ such that for every positive $\delta \leq \delta_0$ we have*

$$\mathbb{P}_X(0 < |\eta(X) - \theta^*| \leq \delta) \leq C_0 \delta^\alpha \ .$$

Remark 3. *The case of "$\alpha = \infty$" is understood in the following manner [Massart and Nédélec, 2006]: there exists a constant $\delta_0 \in (0,1]$ such that*

$$\mathbb{P}_X(0 < |\eta(X) - \theta^*| \leq \delta_0) = 0 \ ,$$

typically this is the most advantageous situation for the binary classification, as the regression function η is separated from the optimal threshold θ^*. Assumption 4 specifies the concentration rate of the regression function η around the optimal threshold θ^*. In order to prove the upper bounds for the proposed method we actually need to have the margin assumption for all $\delta > 0$ and not only for $\delta \leq \delta_0$ However, notice that if Assumption 4 is satisfied, it holds that for all $\delta > 0$

$$\mathbb{P}_X(0 < |\eta(X) - \theta^*| \leq \delta) \leq c_0 \delta^\alpha \ ,$$

where $c_0 = C_0 \vee \delta_0^{-\alpha}$.

As already discussed in Section 1.1, this assumption is tightly related to the rate of convergence in the case of the binary classification [Audibert and Tsybakov, 2007, Massart and Nédélec, 2006]. The classification algorithm that is proposed in this chapter is based on a direct estimation of the regression function η and the optimal threshold θ^*. Here, we consider the case of non-parametric estimation, that is, we assume that the regression function $\eta : \mathbb{R}^d \mapsto \{0,1\}$ lies in some class of β-smooth functions and the marginal density $\mathbb{P}_X$ of $X \in \mathbb{R}^d$ admits a density *w.r.t.* to the Lebesgue measure supported on a well-behaved compact set and uniformly lower- and upper-bounded. The precise description of these assumptions is given in Section 2.1.4, where we prove optimality of our rates. As for now, it is sufficient to assume that there exists a good estimator $\hat{\eta}$ based on the labeled set $\mathcal{D}_n^{\mathrm{L}}$ of the regression function η.

Assumption 5 (Existence of estimator)**.** *There exists an estimator $\hat{\eta}$ based on $\mathcal{D}_n^{\mathrm{L}}$ which satisfies for all $t > 0$*

$$\mathbb{P}^{\otimes n}(|\hat{\eta}(x) - \eta(x)| \geq t) \leq C_1 \exp(-C_2 a_n t^2) \ \textit{a.s.} \ \mathbb{P}_X \ ,$$

for some universal constants $C_1, C_2 > 0$ and an increasing sequence $a_n : \mathbb{N} \mapsto \mathbb{R}_+$.

Recall that in the case of β-smooth regression function $\eta : \mathbb{R}^d \mapsto [0,1]$, a typical non-parametric rate is $a_n = n^{2\beta/(2\beta+d)}$ and it can be achieved by the local polynomial estimator, see Theorem 3. Finally, in this chapter we assume that the probability $\mathbb{P}(Y = 1)$ is lower bounded by some constant which can be arbitrary small but fixed.

Assumption 6 (Lower bounded $\mathbb{P}(Y = 1)$)**.** *We assume that there exists a positive constant p such that $p \leq \mathbb{P}(Y = 1)$.*

It is assumed that the constants C_0, C_1, C_2, p are independent of both $n, N \in \mathbb{N}$, however these constants can depend on the dimension of the problem d, on the value of $\alpha > 0$ as well as on each other. The values of the constants p, C_0, C_1, C_2 are not going to impact the rates of convergence, though they might and will enter as numerical constants in front of the rate. In contrast, the value of α in the margin assumption will explicitly appear in the obtained rates.

2.1.3 Related works and contributions

Literature on the binary classification with F-score is rather broad, it spans both applied and theoretical studies of the problem. It should be noted that the contribution of this part of the manuscript falls into the Population Utility (PU) approach [Dembczynski et al., 2017], that is, the expectation is taken in the numerator and the denominator of the F-score simultaneously. This approach should not be confused with the Expected Test Utility (ETU) approach, for which a non-asymptotic behavior can differ vastly. We refer the reader to [Dembczynski et al., 2017, Ye et al., 2012] where the PU and ETU approaches are discussed in depth and their asymptotic equivalence is established. Since it is not the subject of the present manuscript we omit this discussion here. The asymptotic statistical theory of the binary classification with F-score has been studied in the prior literature [Koyejo et al., 2014, Narasimhan et al., 2014, Menon et al., 2013, Ye et al., 2012]. Let us summarize contributions of this work and highlight the improvements with respect to the previous results on the non-asymptotic analysis of the binary classification with F-score.

- We propose a two-step estimator, which first estimates the regression function η and then the optimal threshold θ^*. Such two-step estimators, which involve an explicit thresholds tuning, are well-known in the literature and demonstrate promising empirical performance [Koyejo et al., 2014] [Keerthi et al., 2007]. An important novelty introduced here is the semi-supervised nature of the procedure which can exploit the unlabeled data. It is already a well established fact that the semi-supervised methods might [Singh et al., 2009] or might not [Rigollet, 2007] improve supervised estimation from a statistical point of view. However, from a practical point of view, the most expensive part of the data gathering process is typically the (correct) labeling. Thus, one may assume that the unlabeled dataset $\mathcal{D}_N^{\mathrm{U}}$ is always available in reality and $N \gg n$ holds. Our analysis implies that in the setting of binary classification with F-score the semi-supervised techniques are not superior to the supervised ones. In contrast, in [Chzhen et al., 2019a] the authors showed that in the context of confidence set classification semi-supervised classifiers might outperform their supervised counterparts.

- From a theoretical point of view, the most relevant reference is a recent work by Yan et al. [2018], where the authors have studied a rather broad class of performance measures for the problem of binary classification, namely Karmic measures whose definition

relies on the confusion matrix. This class includes the F-score, considered in the present manuscript. Under similar, though stronger assumptions on the distribution[3] of the pair $(X, Y) \in \mathbb{R}^d \times \{0, 1\}$ they proposed an algorithm whose rate of convergence is at most $\mathcal{O}(a_n^{-(1+1\wedge\alpha)/2})$. This rate is rather counter intuitive, since it suggests that if the constant α in the margin assumption is large it does not affect the rate of convergence. In contrast, here we show that the minimax rate of convergence is of order $\mathcal{O}(a_n^{-(1+\alpha)/2})$. That is, it strictly improves upon the results in [Yan et al., 2018] whenever the constant $\alpha > 1$. However, it should be noted, that the authors of [Yan et al., 2018] have studied a much more general family of score functions and the sub-optimal rate is a consequence of such a generality.

- We show that the constructed estimator is optimal in the minimax sense over the class of Hölder smooth regression functions. Let us mention that the optimality of the bound is expected, as in Section 1.1 it is shown that the minimax risk in the standard binary classification settings is of order $a_n^{-(1+\alpha)/2}$, and it is achieved by a plug-in rule classifier. Clearly, it is hard to expect that the rate in a more difficult situation can be improved. Nevertheless, to the best of our knowledge, the minimax optimality in the context of binary classification with F-score has not been considered before.

Organization of the section

This contribution is organized as follows: in Section 2.1.4 we present the semi-supervised classification algorithm; in Section 2.1.4 we establish an upper bound on the excess F-score under the margin assumption; in Section 2.1.4 we introduce the class of distributions considered in this chapter and establish a minimax lower bound on the excess F-score.

2.1.4 Main results

In this section we describe the proposed procedure $\hat{g}$ to estimate the Bayes optimal classifier g^* in case of the F-score. This procedure is performed in two steps: on the first step we estimate the regression function $\eta : \mathbb{R}^d \mapsto \{0, 1\}$ using the labeled data $\mathcal{D}_n^{\mathrm{L}}$ while on the second step we estimate the optimal threshold θ^* based on the unlabeled data $\mathcal{D}_N^{\mathrm{U}}$ and the estimator $\hat{\eta}$ provided by the first step. To summarize, the classifier $\hat{g}$ is defined as

$$\hat{g}(x) = \mathbb{1}_{\{\hat{\eta}(x) > \hat{\theta}\}} \ ,$$

where $\hat{\eta}$ is any estimator satisfying Assumption 5 and $\hat{\theta}$ is the unique solution of

$$\theta \left(\frac{1}{N} \sum_{X_i \in \mathcal{D}_N^{\mathrm{U}}} \hat{\eta}(X_i) \right) = \frac{1}{N} \sum_{X_i \in \mathcal{D}_N^{\mathrm{U}}} (\hat{\eta}(X_i) - \theta)_+ \ . \tag{2.2}$$

In practice one can use a simple bisection algorithm [Conte and Boor, 1980, Algorithm 3.1] or its more sophisticated modifications (*e.g.,* regula falsi or the secant method) to approximate $\hat{\theta}$ with any given precision. For our theoretical analysis we assume that Equation (2.2) is solved exactly. However a simple modification of our arguments can handle the situation where the threshold $\hat{\theta}$ is known up to an additive factor $\epsilon_n = \mathcal{O}(a_n^{-1/2})$.

[3] The authors additionally require that the random variable $\eta(X)$ on $[0, 1]$ admits bounded density.

Upper bound

The main result of this part of Section 2.1 is an upper bound on the excess-score of our proposed procedure. Here we provide two theorems, the first one provides an upper bound on the expected difference between the optimal threshold θ^* and its estimate $\hat{\theta}$. The second one gives an upper bound on the excess F-score.

Theorem 5. *If there exists an estimator $\hat{\eta}$ of the regression function η which satisfies Assumption 5, then there exists a constant $C > 0$ which depends on C_0, C_1, C_2, p such that, the threshold $\hat{\theta}$ defined in Eq. (2.2) satisfies*

$$\mathbb{E}_{(\mathcal{D}_n^{\mathrm{L}}, \mathcal{D}_N^{\mathrm{U}})} |\theta^* - \hat{\theta}| \leq C \left(a_n^{-1/2} + N^{-1/2} \right) \ .$$

Theorem 6. *If the distribution $\mathbb{P}$ of (X, Y) satisfies the α-margin assumption for some $C_0 > 0$ and $\alpha \geq 0$ and there exists an estimator $\hat{\eta}$ of the regression function η which satisfies Assumption 5, then there exists a constant $C > 0$ which depends on α, C_0, C_1, C_2, p such that*

$$\mathbb{E}_{(\mathcal{D}_n^{\mathrm{L}}, \mathcal{D}_N^{\mathrm{U}})} \mathcal{E}(\hat{g}) \leq C \left(a_n^{-\frac{1+\alpha}{2}} + N^{-\frac{1+\alpha}{2}} \right) \ ,$$

where $\hat{g}(x) = \mathbb{1}_{\{\hat{\eta}(x) > \hat{\theta}\}}$ with the threshold $\hat{\theta}$ defined in Equation (2.2).

Before proceeding to the proofs let us discuss the implications of these results. First of all, there are two regimes in the bound of Theorems 6, the first one is $N \geq a_n$, in this regime, the dominant term is $a_n^{-(1+\alpha)/2}$ which is the classical rate of convergence in the standard settings of binary classification with the α-margin assumption. The second regime is when $N < a_n$, then the dominating term of the bound is $N^{-(1+\alpha)/2}$. However, let us recall that one can always augment the second unlabeled dataset $\mathcal{D}_N^{\mathrm{U}}$ by dividing $\mathcal{D}_n^{\mathrm{L}}$ into two independent parts. It implies that the second regime never occurs in our theoretical analysis of the excess score and the upper bound is actually independent of N. Similar reasoning holds for the case of the optimal threshold estimation in Theorem 5. Once it is clear that the obtained upper-bounds are actually independent of the size of the unlabeled dataset $\mathcal{D}_N^{\mathrm{U}}$ it is interesting to notice that the dependence on n is the same as in standard binary classification [Audibert and Tsybakov, 2007]. That is, similarly to the standard settings, binary classification with F-score can achieve fast (faster than $1/\sqrt{n}$) and even super-fast (faster than $1/n$) rate depending on the value α and the rate a_n.

Proofs of both theorems relies on the following lemma, whose proof can be found in Section 2.1.6, which relates the difference of the threshold to the difference of the empirical cumulative distribution function (CDF) of $\hat{\eta}$ and the CDF of η.

Lemma 3. *Let $\hat{\theta} \in [0, 1]$ be the threshold which satisfies Equation 2.2, then*

$$\left| \hat{\theta} - \theta^* \right| \mathbb{P}(Y = 1) \leq \int_0^1 \left| \mathbb{P}_X(\eta(X) \leq t) - \frac{1}{N} \sum_{X_i \in \mathcal{D}_N^{\mathrm{U}}} \mathbb{1}_{\{\hat{\eta}(X_i) \leq t\}} \right| dt \ .$$

This result is the main reason why our conclusions on the semi-supervised estimation in the context of classification with F-score is different from the ones in [Chzhen et al., 2019a, Singh et al., 2009]. For instance, in [Chzhen et al., 2019a] we also obtain a final decision rule by thresholding on some estimated level in the contex of confidence set classification. However,

in the present contribution of this chapter the difference between θ^* and $\hat{\theta}$ is controlled via ℓ_1-norm of difference of CDF's, whereas in [Chzhen et al., 2019a] (see Chapter 3 for details) we control a similar quantity through Wassertstein infinity distance.

The complete proof of Theorems 5 and 6 can be found in Section 2.1.6, and here, we only sketch the steps which are different from the analysis of Audibert and Tsybakov [2007]. Recall, that due to Lemma 2 we have the following expression for the excess-score $\mathcal{E}$

$$\mathbb{E}_{(\mathcal{D}_n^{\mathrm{L}},\mathcal{D}_N^{\mathrm{U}})}\frac{\mathbb{E}\,|\eta(X)-\theta^*|\,\mathbb{1}_{\{g^*(X)\neq\hat{g}(X)\}}}{\mathbb{P}(Y=1)+\mathbb{P}(\hat{g}(X)=1)} \leq \frac{1}{p}\mathbb{E}_{(\mathcal{D}_n^{\mathrm{L}},\mathcal{D}_N^{\mathrm{U}})}\mathbb{E}\,|\eta(X)-\theta^*|\,\mathbb{1}_{\{g^*(X)\neq\hat{g}(X)\}} \ .$$

First of all, notice that if for some $x\in\mathbb{R}^d$ the event $g^*(x)\neq\hat{g}(x)$ occurs, than we have

$$|\eta(x)-\theta^*| \leq |\eta(x)-\hat{\eta}(x)| + |\theta^*-\hat{\theta}| \ ,$$

which further implies that at least one of the following inequalities hold for this $x\in\mathbb{R}^d$

$$\begin{aligned}|\eta(x)-\theta^*| &\leq 2\,|\eta(x)-\hat{\eta}(x)| \ , \\ &\leq 2|\theta^*-\hat{\theta}| \ .\end{aligned}$$

Thus, we can upper bound the excess risk as

$$\mathcal{E}(\hat{g}) \leq \underbrace{\frac{1}{p}\mathbb{E}\,|\eta(X)-\theta^*|\,\mathbb{1}_{\{|\eta(X)-\theta^*|\leq 2|\eta(X)-\hat{\eta}(X)|\}}}_{T_1} + \underbrace{\frac{1}{p}\mathbb{E}\,|\eta(X)-\theta^*|\,\mathbb{1}_{\left\{|\eta(X)-\theta^*|\leq 2\left|\theta^*-\hat{\theta}\right|\right\}}}_{T_2} \ .$$

The first term on the right hand side (T_1) of the inequality can be handled by the peeling technique used in [Audibert and Tsybakov, 2007, Lemma 3.1.], which implies that, there exists a constant $C'=C'(p,\alpha,C_0,C_1,C_2)>0$ such that

$$\mathbb{E}_{(\mathcal{D}_n^{\mathrm{L}},\mathcal{D}_N^{\mathrm{U}})}T_1 \leq C' a_n^{-\frac{1+\alpha}{2}} \ .$$

Hence, it remains to upper bound the second term on the right hand side (T_2) of the inequality. Using Lemma 3 we can upper bound T_2 as

$$T_2 \leq \frac{1}{p}\mathbb{E}\,|\eta(X)-\theta^*|\,\mathbb{1}_{\{E\}} \ ,$$

with $E=\left\{p\,|\eta(X)-\theta^*| \leq 2\int_0^1 \left|\mathbb{P}_X(\eta(X)\leq t) - \frac{1}{N}\sum_{X_i\in\mathcal{D}_N^{\mathrm{U}}}\mathbb{1}_{\{\hat{\eta}(X_i)\leq t\}}\right| dt\right\}$. Finally, we upper bound the indicator $\mathbb{1}_{\{E\}}$ by the indicators of two events E^1 and E^2 which are defined as

$$\begin{aligned}E^1 &= \left\{p\,|\eta(X)-\theta^*| \leq 4\sup_{t\in[0,1]}\left|\mathbb{P}_X(\hat{\eta}(X)\leq t) - \frac{1}{N}\sum_{X_i\in\mathcal{D}_N^{\mathrm{U}}}\mathbb{1}_{\{\hat{\eta}(X_i)\leq t\}}\right| dt\right\} \ , \\ E^2 &= \left\{p\,|\eta(X)-\theta^*| \leq 4\int_0^1 |\mathbb{P}_X(\hat{\eta}(X)\leq t) - \mathbb{P}_X(\eta(X)\leq t)|\,dt\right\} \ .\end{aligned}$$

Thus, we have the following upper bound on T_2

$$T_2 \leq \underbrace{\frac{1}{p}\mathbb{E}\,|\eta(X)-\theta^*|\,\mathbb{1}_{\{E^1\}}}_{T_2^1} + \underbrace{\frac{1}{p}\mathbb{E}\,|\eta(X)-\theta^*|\,\mathbb{1}_{\{E^2\}}}_{T_2^2} \ ,$$

Notice that thanks to the Dvoretzky-Kiefer-Wolfowitz inequality [Dvoretzky et al., 1956, Massart, 1990] the term

$$\sup_{t\in[0,1]}\left|\mathbb{P}_X(\hat{\eta}(X)\leq t)-\frac{1}{N}\sum_{X_i\in\mathcal{D}_N^{\mathrm{U}}}\mathbb{1}_{\{\hat{\eta}(X_i)\leq t\}}\right| \quad ,$$

conditionally on $\mathcal{D}_n^{\mathrm{L}}$ admits an exponential concentration with the rate $N^{-1/2}$. Hence, using the margin assumption, one can effortlessly show there exists a constant $C''=C''(p,\alpha,C_0)>0$ such that

$$\mathbb{E}_{(\mathcal{D}_n^{\mathrm{L}},\mathcal{D}_N^{\mathrm{U}})}T_2^1\leq C''N^{-\frac{1+\alpha}{2}} \quad .$$

For the second term T_2^2 we proceed as follows

$$T_2^2\leq\frac{4}{p^2}\mathbb{E}\int_0^1\left|\mathbb{P}_X(\hat{\eta}(X)\leq t)-\mathbb{P}_X(\eta(X)\leq t)\right|dt\mathbb{1}_{\{E^2\}} \quad ,$$

thus, using the α-margin assumption we get

$$T_2^2\leq\frac{C_0 4^{1+\alpha}}{p^{2+\alpha}}\left(\int_0^1\left|\mathbb{P}_X(\hat{\eta}(X)\leq t)-\mathbb{P}_X(\eta(X)\leq t)\right|dt\right)^{1+\alpha} \quad ,$$

the integral on the right hand side of the bound corresponds to the 1-Wasserstein distance on the real line, see for instance [Bobkov and Ledoux, 2016, Theorem 2.9] or [Vallender, 1974] for the proof, and can be further upper-bounded by the L_1 norm between $\hat{\eta}$ and η, that is

$$T_2^2\leq\frac{C_0 4^{1+\alpha}}{p^{2+\alpha}}\left(\mathbb{E}_{\mathbb{P}_X}\left|\eta(X)-\hat{\eta}(X)\right|\right)^{1+\alpha} \quad .$$

Since the estimator $\hat{\eta}$ satisfies Assumption 5, one can show that there exists a constant $C'''=C'''(p,\alpha,C_0,C_1,C_2)>0$ such that

$$\mathbb{E}_{(\mathcal{D}_n^{\mathrm{L}},\mathcal{D}_N^{\mathrm{U}})}T_2^2\leq C'''a_n^{-\frac{1+\alpha}{2}} \quad .$$

Combination of all the inequalities yields the result of Theorem 6. Notice that the same reasoning starting from Lemma 3 implies the upper bound on the threshold estimation, that is, Theorem 5.

Lower bound

In the beginning of this part of Section 2.1 by stating the class of joint distribution $\mathcal{P}_\Sigma$ of the random pair $(X,Y)\in\mathbb{R}^d\times\{0,1\}$ that is considered. The first assumption is made on smoothness of the regression function $\eta:\mathbb{R}^d\mapsto[0,1]$.

Definition 6 (Hölder smoothness)**.** *Let $L>0$ and $\beta>0$. The class of function $\Sigma(\beta,L,\mathbb{R}^d)$ consists of all functions $h:\mathbb{R}^d\mapsto[0,1]$ such that for all $x,x'\in\mathbb{R}^d$, we have*

$$|h(x)-h_x(x')|\leq L\left\|x-x'\right\|_2^\beta \quad ,$$

where $h_x(\cdot)$ is the Taylor expansion of h at point x of degree $\lfloor\beta\rfloor$.

Assumption 7 ((β, L)-Hölder regression function)**.** *The distribution $\mathbb{P}$ of the pair $(X, Y) \in \mathbb{R}^d \times \{0, 1\}$ is such that $\eta \in \Sigma(\beta, L, \mathbb{R}^d)$ for some positive β, L.*

Assumption 7 is usually not sufficient to guarantee the existence of an estimator $\hat{\eta}$ satisfying Assumption 5: extra assumptions are required on the marginal distribution $\mathbb{P}_X$ of the vector $X \in \mathbb{R}^d$.

Definition 7. *A Lebesgue measurable set $A \subset \mathbb{R}^d$ is said to be (c_0, r_0)-regular for some constants $c_0 > 0, r_0 > 0$ if for every $x \in A$ and every $r \in (0, r_0]$ we have*

$$\text{Leb}\left(A \cap \mathcal{B}(x, r)\right) \geq c_0 \, \text{Leb}\left(\mathcal{B}(x, r)\right) \quad ,$$

where Leb *is the Lebesgue measure and $\mathcal{B}(x, r)$ is the Euclidean ball of radius r centered at x.*

Assumption 8 (Strong density assumption)**.** *We say that the marginal distribution $\mathbb{P}_X$ of the vector $X \in \mathbb{R}^d$ satisfies the strong density assumption if*

- *$\mathbb{P}_X$ is supported on a compact (c_0, r_0)-regular set $A \subset \mathbb{R}^d$,*
- *$\mathbb{P}_X$ admits a density μ* w.r.t. *to the Lebesgue measure uniformly lower- and upper-bounded by $\mu_{\min} > 0$ and $\mu_{\max} > 0$ respectively.*

If the regression function $\eta : \mathbb{R}^d \mapsto [0, 1]$ is (β, L)-Hölder and the marginal distribution satisfies the strong density assumption, one can specify Theorem 3 for a bit more general case.

Theorem 7 (Audibert and Tsybakov [2007])**.** *Let $\mathcal{P}$ be a class of distributions on $\mathbb{R}^d \times \{0, 1\}$ such that the regression function $\eta \in \Sigma(\beta, L, \mathbb{R}^d)$ and the marginal distribution $\mathbb{P}_X$ satisfies the strong density assumption. Then, there exists an estimator $\hat{\eta}$ of the regression function satisfying*

$$\sup_{\mathbb{P} \in \mathcal{P}} \mathbb{P}^{\otimes n}(|\hat{\eta}(x) - \eta(x)| \geq t) \leq C_1 \exp\left(-C_2 n^{\frac{2\beta}{2\beta + d}} t^2\right) \quad a.s. \ \mathbb{P}_X \quad ,$$

for come constants C_1, C_2 depending on β, d, L, c_0, r_0.

Consider a class of distribution $\mathcal{P}_\Sigma$ for which Assumptions 4, 7, 8, 6 are satisfied, then Theorem 7 and Theorems 5, 6 imply the following corollary.

Corollary 1. *There exist constants $C, B > 0$ which depend only on $\alpha, p, d, C_0, C_1, C_2$ such that for any $n > 1, N > 1$ we have*

$$\inf_{\hat{g}} \sup_{\mathbb{P} \in \mathcal{P}_\Sigma} \mathbb{E}_{(\mathcal{D}_n^{\text{L}}, \mathcal{D}_N^{\text{U}})} \mathcal{E}_1(\hat{g}) \leq C n^{-\frac{(1+\alpha)\beta}{2\beta + d}} \quad , \tag{2.3}$$

$$\inf_{\hat{\theta}} \sup_{\mathbb{P} \in \mathcal{P}_\Sigma} \mathbb{E}_{(\mathcal{D}_n^{\text{L}}, \mathcal{D}_N^{\text{U}})} \left|\theta^* - \hat{\theta}\right| \leq B n^{-\frac{\beta}{2\beta + d}} \quad . \tag{2.4}$$

where the infima are taken over all estimators $\hat{g}$ and $\hat{\theta}$ respectively.

The next theorem states the upper bounds of the previous corollary are optimal up to a constant multiplicative factor.

Theorem 8. *If $\alpha\beta \leq d$, there exists constants $c > 0$ such that for any $n > 1, N > 1$ we have the following lower-bound on the minimax risk*

$$\inf_{\hat{g}} \sup_{\mathbb{P}\in\mathcal{P}_\Sigma} \mathbb{E}_{(\mathcal{D}_n^{\mathrm{L}},\mathcal{D}_N^{\mathrm{U}})}\mathcal{E}_1(\hat{g}) \geq cn^{-\frac{(1+\alpha)\beta}{2\beta+d}} \ , \tag{2.5}$$

where the infimum is taken over all estimators $\hat{g}$.

The proof of the lower bound can be found in Section 2.1.6, it follows standard information-theoretic arguments using reduction of the minimax risk to a Bayes risk. The construction of the distributions is inspired by both [Rigollet and Vert, 2009] and [Audibert and Tsybakov, 2007], and the actual proof relies precisely on [Audibert, 2004, Lemma 5.1.], which is based on the Assouad's lemma, see for instance [Tsybakov, 2009, Lemma 2.12].

2.1.5 Conclusion

In this chapter we proposed a semi-supervised plug-in type algorithm for the problem of binary classification with F-score. The proposed algorithm can leverage an unlabeled dataset for the estimation of the optimal threshold. Under the margin assumption it is shown that the proposed algorithm is optimal in the minimax sense and can achieve fast rates of convergence. Further development of the binary classification with F-score will be devoted to empirical risk minimization rules.

2.1.6 Proofs

Bayes classifier and Lemma 2

For the rest of this part the parameter $b > 0$ is assumed to be fixed and known. Let us first recall the definition of the F_b-score

$$F_b(g) = \frac{\mathbb{P}(Y = 1, g(X) = 1)}{b^2\mathbb{P}(Y = 1) + \mathbb{P}(g(X) = 1)} \ ,$$

and an optimal classifier is defined as

$$g^* \in \arg\max_{g\in\mathcal{G}} F_b(g) \ .$$

In this part we would like to show that a classifier defined for all $x \in \mathbb{R}^d$ as

$$g_*(x) = \mathbb{1}_{\{\eta(x)\geq\theta^*\}} \ ,$$

with θ^* being a root of

$$\theta \mapsto b^2\mathbb{P}(Y = 1)\theta - \mathbb{E}(\eta(X) - \theta)_+ \ ,$$

is an optimal classifier.

Let us first show that θ^* is well-defined, that is, it exists and is unique for every distribution with $\mathbb{P}(Y = 1) \neq 0$. Hence, we would like to study solutions of the following equation

$$b^2\mathbb{P}(Y = 1)\theta = \mathbb{E}(\eta(X) - \theta)_+ \ .$$

Clearly, the mapping $\theta \mapsto b^2\mathbb{P}(Y=1)\theta$ is continuous and strictly increasing on $[0,1]$ with the range $[0, b^2\mathbb{P}(Y=1)]$ and the mapping $\theta \mapsto \mathbb{E}(\eta(X)-\theta)_+$ is non-increasing on $[0,1]$ with the range $[0, \mathbb{P}(Y=1)]$. Thus, it is sufficient to demonstrate that the mapping $\theta \mapsto \mathbb{E}(\eta(X)-\theta)_+$ is continuous, indeed, let $\theta, \theta' \in [0,1]$, then, due to the Lipschitz continuity of $(\cdot)_+$ we can write

$$|\mathbb{E}(\eta(X)-\theta)_+ - \mathbb{E}(\eta(X)-\theta')_+| \leq \mathbb{E}\,|(\eta(X)-\theta)_+ - (\eta(X)-\theta')_+| \leq |\theta-\theta'| \quad .$$

This implies that the mapping $\theta \mapsto \mathbb{E}(\eta(X)-\theta)_+$ is a contraction and thus is continuous. Hence, the threshold θ^* is well-defined, that is, it exists and is unique. Consequently, the classifier $x \mapsto \mathbb{1}_{\{\eta(x)\geq\theta^*\}}$ is well-defined.

Now, we are interested in the value $F_b(g_*)$, we can write

$$\begin{aligned}
F_b(g_*) &= \frac{\mathbb{P}(Y=1, g_*(X)=1)}{b^2\mathbb{P}(Y=1)+\mathbb{P}(g_*(X)=1)} \\
&= \frac{\mathbb{E}[\eta(X)\mathbb{1}_{\{\eta(X)\geq\theta^*\}}]}{b^2\mathbb{P}(Y=1)+\mathbb{P}(\eta(X)\geq\theta^*)} \\
&= \frac{\mathbb{E}[(\eta(X)-\theta^*)\mathbb{1}_{\{\eta(X)\geq\theta^*\}}] + \theta^*\mathbb{E}\mathbb{1}_{\{\eta(X)\geq\theta^*\}}}{b^2\mathbb{P}(Y=1)+\mathbb{P}(\eta(X)\geq\theta^*)} \\
&= \frac{\mathbb{E}(\eta(X)-\theta^*)_+ + \theta^*\mathbb{P}(\eta(X)\geq\theta^*)}{b^2\mathbb{P}(Y=1)+\mathbb{P}(\eta(X)\geq\theta^*)} \quad ,
\end{aligned}$$

using the definition of θ^* we continue as

$$\begin{aligned}
F_b(g_*) &= \frac{\mathbb{E}(\eta(X)-\theta^*)_+ + \theta^*\mathbb{P}(\eta(X)\geq\theta^*)}{b^2\mathbb{P}(Y=1)+\mathbb{P}(\eta(X)\geq\theta^*)} \\
&= \frac{\theta^* b^2\mathbb{P}(Y=1) + \theta^*\mathbb{P}(\eta(X)\geq\theta^*)}{b^2\mathbb{P}(Y=1)+\mathbb{P}(\eta(X)\geq\theta^*)} = \theta^* \quad .
\end{aligned}$$

To conclude the optimality of g_* we prove Lemma 2.

Proof. Fix an arbitrary measurable function $g : \mathbb{R}^d \mapsto \{0,1\}$, then by the definition of the excess score we have

$$\begin{aligned}
\mathcal{E}_b(g) &\coloneqq \frac{\mathbb{P}(Y=1, g^*(X)=1)}{b^2\mathbb{P}(Y=1)+\mathbb{P}(g^*(X)=1)} - \frac{\mathbb{P}(Y=1, g(X)=1)}{b^2\mathbb{P}(Y=1)+\mathbb{P}(g(X)=1)} \\
&= \frac{\mathbb{E}\eta(X)\mathbb{1}_{\{\eta(X)>\theta^*\}}}{b^2\mathbb{P}(Y=1)+\mathbb{P}(g^*(X)=1)} - \frac{\mathbb{E}\eta(X)\mathbb{1}_{\{g(X)=1\}}}{b^2\mathbb{P}(Y=1)+\mathbb{P}(g(X)=1)} \quad .
\end{aligned}$$

Now adding and subtracting $\frac{\mathbb{E}\eta(X)\mathbb{1}_{\{g(X)=1\}}}{b^2\mathbb{P}(Y=1)+\mathbb{P}(g^*(X)=1)}$ on the right hand side of the equality above

we get

$$\begin{aligned}
\mathcal{E}_b(g) &= \frac{\mathbb{E}\eta(X)\mathbb{1}_{\{\eta(X)>\theta^*\}} - \mathbb{E}\eta(X)\mathbb{1}_{\{g(X)=1\}}}{b^2\mathbb{P}(Y=1)+\mathbb{P}(g^*(X)=1)} \\
&\quad + \underbrace{\frac{\mathbb{E}\eta(X)\mathbb{1}_{\{g(X)=1\}}}{b^2\mathbb{P}(Y=1)+\mathbb{P}(g(X)=1)}}_{=F_b(g)} \left(\frac{\mathbb{P}(g(X)=1)-\mathbb{P}(g^*(X)=1)}{b^2\mathbb{P}(Y=1)+\mathbb{P}(g^*(X)=1)}\right) \\
&= \frac{\mathbb{E}(\eta(X)-\theta^*)(\mathbb{1}_{\{\eta(X)>\theta^*\}} - \mathbb{1}_{\{g(X)=1\}}) + \theta^*(\mathbb{P}(g^*(X)=1)-\mathbb{P}(g(X)=1))}{b^2\mathbb{P}(Y=1)+\mathbb{P}(g^*(X)=1)} \\
&\quad + F_b(g)\left(\frac{\mathbb{P}(g(X)=1)-\mathbb{P}(g^*(X)=1)}{b^2\mathbb{P}(Y=1)+\mathbb{P}(g^*(X)=1)}\right) \\
&= \frac{\mathbb{E}\left|\eta(X)-\theta^*\right|\mathbb{1}_{\{g^*(X)\neq g(X)\}}}{b^2\mathbb{P}(Y=1)+\mathbb{P}(g^*(X)=1)} + (\theta^*-F_b(g))\frac{\mathbb{P}(g^*(X)=1)-\mathbb{P}(g(X)=1)}{b^2\mathbb{P}(Y=1)+\mathbb{P}(g^*(X)=1)} \ .
\end{aligned}$$

Using Theorem 4 we know that $\theta^* = F_b(g^*)$, therefore $\theta^* - F_b(g) = F_b(g^*) - F_b(g) = \mathcal{E}_b(g)$ and we get

$$\mathcal{E}_b(g) = \frac{\mathbb{E}\left|\eta(X)-\theta^*\right|\mathbb{1}_{\{g^*(X)\neq g(X)\}}}{b^2\mathbb{P}(Y=1)+\mathbb{P}(g^*(X)=1)} + \mathcal{E}_b(g)\frac{\mathbb{P}(g^*(X)=1)-\mathbb{P}(g(X)=1)}{b^2\mathbb{P}(Y=1)+\mathbb{P}(g^*(X)=1)} \ .$$

We conclude by solving the previous equality for $\mathcal{E}_b(g)$. Thus, g_* is a Bayes optimal classifier and hence can be denoted by g^*. □

Proof of Lemma 3

Proof. To prove this lemma, it is convenient to rewrite Equation 2.2 in terms of CDF. Let μ be an arbitrary probability measure on $\mathbb{R}^d$ and $p : \mathbb{R}^d \mapsto [0,1]$ be any measurable function, then using Fubini's theorem we can write

$$\begin{aligned}
\int p(x)d\mu(x) &= \int\int_0^1 \mathbb{1}_{\{p(x)>t\}}dtd\mu(x) \\
&= \int_0^1 \mu(p(X)>t)dt \ ,
\end{aligned}$$

and for any $\theta \in [0,1]$, since $(p(x)-\theta)_+ \in [0,1]$ we have

$$\begin{aligned}
\int (p(x)-\theta)_+ d\mu(x) &= \int\int_0^1 \mathbb{1}_{\{p(x)-\theta>t\}}dtd\mu(x) \\
&= \int\int_\theta^{1+\theta} \mathbb{1}_{\{p(x)>t\}}dtd\mu(x) \\
&= \int\int_\theta^1 \mathbb{1}_{\{p(x)>t\}}dtd\mu(x) \\
&= \int_\theta^1 \mu(p(X)>t)dt \ .
\end{aligned}$$

Let us denote by $\mathbb{P}_{X,N} = \frac{1}{N}\sum_{X_i\in\mathcal{D}_N^{\mathrm{U}}}\delta_{X_i}$ the empirical measure of the unlabeled dataset $\mathcal{D}_N^{\mathrm{U}}$. Using these equalities, the thresholds $\theta^*, \hat{\theta} \in [0,1]$ satisfy

$$\hat{\theta} = \frac{\int_{\hat{\theta}}^1 \mathbb{P}_{X,N}(\hat{\eta}(X)>t)dt}{\int_0^1 \mathbb{P}_{X,N}(\hat{\eta}(X)>t)dt}, \quad \theta^* = \frac{\int_{\theta^*}^1 \mathbb{P}_X(\eta(X)>t)dt}{\int_0^1 \mathbb{P}_X(\eta(X)>t)dt} \ .$$

Now, we are in position to bound the difference $|\hat{\theta} - \theta^*|$. First, assume that $\theta^* \geq \hat{\theta}$, then

$$\begin{aligned}\theta^* - \hat{\theta} &= \frac{\int_{\theta^*}^1 \mathbb{P}_X(\eta(X) > t)dt}{\int_0^1 \mathbb{P}_X(\eta(X) > t)dt} - \frac{\int_{\hat{\theta}}^1 \mathbb{P}_{X,N}(\hat{\eta}(X) > t)dt}{\int_0^1 \mathbb{P}_{X,N}(\hat{\eta}(X) > t)dt} \\ &\leq \frac{\int_{\hat{\theta}}^1 \mathbb{P}_X(\eta(X) > t)dt}{\int_0^1 \mathbb{P}_X(\eta(X) > t)dt} - \frac{\int_{\hat{\theta}}^1 \mathbb{P}_{X,N}(\hat{\eta}(X) > t)dt}{\int_0^1 \mathbb{P}_{X,N}(\hat{\eta}(X) > t)dt}\end{aligned}$$

Adding and subtracting $\frac{\int_{\hat{\theta}}^1 \mathbb{P}_{X,N}(\hat{\eta}(X)>t)dt}{\int_0^1 \mathbb{P}_X(\eta(X)>t)dt}$ on the right hand side of the above inequality we get

$$\begin{aligned}\theta^* - \hat{\theta} &\leq \frac{\int_{\hat{\theta}}^1 (\mathbb{P}_X(\eta(X) > t) - \mathbb{P}_{X,N}(\hat{\eta}(X) > t))dt}{\int_0^1 \mathbb{P}_X(\eta(X) > t)dt} \\ &\quad + \underbrace{\frac{\int_{\hat{\theta}}^1 \mathbb{P}_{X,N}(\hat{\eta}(X) > t)dt}{\int_0^1 \mathbb{P}_{X,N}(\hat{\eta}(X) > t)dt}}_{=\hat{\theta}} \frac{\int_0^1 (\mathbb{P}_{X,N}(\hat{\eta}(X) > t) - \mathbb{P}_X(\eta(X) > t))dt}{\int_0^1 \mathbb{P}_X(\eta(X) > t)dt} \\ &= \frac{\int_{\hat{\theta}}^1 (\mathbb{P}_X(\eta(X) > t) - \mathbb{P}_{X,N}(\hat{\eta}(X) > t))dt - \hat{\theta}\int_0^1 (\mathbb{P}_X(\eta(X) > t) - \mathbb{P}_{X,N}(\hat{\eta}(X) > t))dt}{\int_0^1 \mathbb{P}_X(\eta(X) > t)dt} \\ &\leq \frac{1}{\mathbb{P}(Y=1)} \int_0^1 |\mathbb{P}_X(\eta(X) > t) - \mathbb{P}_{X,N}(\hat{\eta}(X) > t)| \, dt \ .\end{aligned}$$

Further, if $\hat{\theta} > \theta^*$ we proceed in similarly and write

$$\begin{aligned}\hat{\theta} - \theta^* &= \frac{\int_{\hat{\theta}}^1 \mathbb{P}_{X,N}(\hat{\eta}(X) > t)dt}{\int_0^1 \mathbb{P}_{X,N}(\hat{\eta}(X) > t)dt} - \frac{\int_{\theta^*}^1 \mathbb{P}_X(\eta(X) > t)dt}{\int_0^1 \mathbb{P}_X(\eta(X) > t)dt} \\ &\leq \frac{\int_{\theta^*}^1 \mathbb{P}_{X,N}(\hat{\eta}(X) > t)dt}{\int_0^1 \mathbb{P}_{X,N}(\hat{\eta}(X) > t)dt} - \frac{\int_{\theta^*}^1 \mathbb{P}_X(\eta(X) > t)dt}{\int_0^1 \mathbb{P}_X(\eta(X) > t)dt} \\ &= \frac{\int_{\theta^*}^1 (\mathbb{P}_{X,N}(\hat{\eta}(X) > t) - \mathbb{P}_X(\eta(X) > t))dt}{\int_0^1 \mathbb{P}_X(\eta(X) > t)dt} \\ &\quad + \frac{\int_{\theta^*}^1 \mathbb{P}_{X,N}(\hat{\eta}(X) > t)dt}{\int_0^1 \mathbb{P}_{X,N}(\hat{\eta}(X) > t)dt} \frac{\int_0^1 (\mathbb{P}_X(\eta(X) > t) - \mathbb{P}_{X,N}(\hat{\eta}(X) > t))dt}{\int_0^1 \mathbb{P}_X(\eta(X) > t)dt} \\ &\leq \frac{1}{\mathbb{P}(Y=1)} \int_0^1 |\mathbb{P}_X(\eta(X) > t) - \mathbb{P}_{X,N}(\hat{\eta}(X) > t)| \, dt \ ,\end{aligned}$$

where the last inequality follows the same lines as for the case $\hat{\theta} \leq \theta^*$. □

Proof of the upper bound

Let $\hat{\eta}$ be an estimator of the regression function based on the labeled dataset $\mathcal{D}_n^{\mathrm{L}}$ which satisfies Assumption 5. Recall, that the estimator $\hat{g}$ is defined for every $x \in \mathbb{R}^d$ as

$$\hat{g}(x) = \mathbb{1}_{\{\hat{\eta}(x) > \hat{\theta}\}} \ ,$$

with $\hat{\theta}$ being the unique solution of Eq. (2.2). Unless stated otherwise, we work conditionally on $(\mathcal{D}_n^{\mathrm{L}}, \mathcal{D}_N^{\mathrm{U}})$. Using Lemma 2 we can express the excess score of $\hat{g}$ as

$$\mathcal{E}_1(\hat{g}) = \frac{\mathbb{E}\,|\eta(X) - \theta^*|\,\mathbb{1}_{\{g^*(X) \neq \hat{g}(X)\}}}{\mathbb{P}(Y=1) + \mathbb{P}(\hat{g}(X)=1)} \leq \frac{1}{p}\mathbb{E}\,|\eta(X) - \theta^*|\,\mathbb{1}_{\{g^*(X) \neq \hat{g}(X)\}} \ .$$

On the event $\{g^*(X) \neq \hat{g}(X)\}$ it holds that $\left\{|\eta(X) - \theta^*| \leq |\hat{\eta}(X) - \eta(X)| + \left|\hat{\theta} - \theta^*\right|\right\}$, thus

$$\begin{aligned}
\mathcal{E}_1(\hat{g}) &\leq \frac{1}{p}\mathbb{E}\,|\eta(X) - \theta^*|\,\mathbb{1}_{\left\{|\eta(X)-\theta^*|\leq|\hat{\eta}(X)-\eta(X)|+\left|\hat{\theta}-\theta^*\right|\right\}} \\
&\leq \frac{1}{p}\mathbb{E}\,|\eta(X) - \theta^*|\,\mathbb{1}_{\left\{|\eta(X)-\theta^*|\leq 2\left|\hat{\theta}-\theta^*\right|\right\}} + \frac{1}{p}\mathbb{E}\,|\eta(X) - \theta^*|\,\mathbb{1}_{\{|\eta(X)-\theta^*|\leq 2|\hat{\eta}(X)-\eta(X)|\}} \quad .
\end{aligned}$$

Using, Lemma 3 the excess risk can be further upper-bounded as

$$\begin{aligned}
\mathcal{E}_1(\hat{g}) \leq& \frac{1}{p}\mathbb{E}\,|\eta(X) - \theta^*|\,\mathbb{1}_{\{|\eta(X)-\theta^*|\leq 2|\hat{\eta}(X)-\eta(X)|\}} \\
&+ \frac{1}{p}\mathbb{E}\,|\eta(X) - \theta^*|\,\mathbb{1}_{\left\{|\eta(X)-\theta^*|\leq\frac{2}{p}\int_0^1\left|\mathbb{P}_X(\eta(X)\leq t)-\frac{1}{N}\sum_{X_i\in\mathcal{D}_N^{\mathrm{U}}}\mathbb{1}_{\{\hat{\eta}(X_i)\leq t\}}\right|dt\right\}} \\
\leq& \frac{1}{p}\mathbb{E}\,|\eta(X) - \theta^*|\,\mathbb{1}_{\{|\eta(X)-\theta^*|\leq 2|\hat{\eta}(X)-\eta(X)|\}} \\
&+ \frac{1}{p}\mathbb{E}\,|\eta(X) - \theta^*|\,\mathbb{1}_{\left\{|\eta(X)-\theta^*|\leq\frac{2}{p}\int_0^1|\mathbb{P}_X(\eta(X)\leq t)-\mathbb{P}_X(\hat{\eta}(X)\leq t)|dt\right\}} \\
&+ \frac{1}{p}\mathbb{E}\,|\eta(X) - \theta^*|\,\mathbb{1}_{\left\{|\eta(X)-\theta^*|\leq\frac{2}{p}\int_0^1\left|\frac{1}{N}\sum_{X_i\in\mathcal{D}_N^{\mathrm{U}}}\mathbb{1}_{\{\hat{\eta}(X_i)\leq t\}}-\mathbb{P}_X(\hat{\eta}(X)\leq t)\right|dt\right\}} \quad .
\end{aligned}$$

Notice that $\int_0^1 |\mathbb{P}_X(\eta(X) \leq t) - \mathbb{P}_X(\hat{\eta}(X) \leq t)|\, dt = \|F_\eta - F_{\hat{\eta}}\|_1$, with $F_\eta, F_{\hat{\eta}}$ being the cumulative distribution functions of $\eta, \hat{\eta}$ respectively, corresponds to the 1-Wasserstein distance, see [Bobkov and Ledoux, 2016] for an in-depth discussion. Therefore, we have

$$\int_0^1 |\mathbb{P}_X(\eta(X) \leq t) - \mathbb{P}_X(\hat{\eta}(X) \leq t)|\, dt \leq \mathbb{E}_{X\sim\mathbb{P}_X}\,|\eta(X) - \hat{\eta}(X)| := \|\eta - \hat{\eta}\|_1 \quad ,$$

and introducing notation $\hat{\mathbb{P}}_X := \frac{1}{N}\sum_{X_i\in\mathcal{D}_N^{\mathrm{U}}}\delta_{X_i}$ for the empirical measure of the feature vector X we can write

$$\begin{aligned}
\mathcal{E}_1(\hat{g}) \leq& \frac{1}{p}\mathbb{E}\,|\eta(X) - \theta^*|\,\mathbb{1}_{\{|\eta(X)-\theta^*|\leq 2|\hat{\eta}(X)-\eta(X)|\}} \\
&+ \frac{1}{p}\mathbb{E}\,|\eta(X) - \theta^*|\,\mathbb{1}_{\left\{|\eta(X)-\theta^*|\leq\frac{2}{p}\|\eta-\hat{\eta}\|_1\right\}} \\
&+ \frac{1}{p}\mathbb{E}\,|\eta(X) - \theta^*|\,\mathbb{1}_{\left\{|\eta(X)-\theta^*|\leq\frac{2}{p}\sup_{t\in[0,1]}\left|\hat{\mathbb{P}}_X(\hat{\eta}(X)\leq t)-\mathbb{P}_X(\hat{\eta}(X)\leq t)\right|\right\}} \quad .
\end{aligned}$$

Finally, using the margin Assumption 4 we can write

$$\begin{aligned}
\mathcal{E}_1(\hat{g}) \leq& \frac{1}{p}\mathbb{E}\,|\eta(X) - \theta^*|\,\mathbb{1}_{\{|\eta(X)-\theta^*|\leq 2|\hat{\eta}(X)-\eta(X)|\}} \\
&+ \frac{2}{p^2}\|\eta - \hat{\eta}\|_1\,\mathbb{P}\left(|\eta(X) - \theta^*| \leq \frac{2}{p}\|\eta - \hat{\eta}\|_1\right) \\
&+ \frac{1}{p}\mathbb{E}\,|\eta(X) - \theta^*|\,\mathbb{1}_{\left\{|\eta(X)-\theta^*|\leq\frac{2}{p}\sup_{t\in[0,1]}\left|\hat{\mathbb{P}}_X(\hat{\eta}(X)\leq t)-\mathbb{P}_X(\hat{\eta}(X)\leq t)\right|\right\}} \\
\leq& \frac{1}{p}\mathbb{E}\,|\eta(X) - \theta^*|\,\mathbb{1}_{\{|\eta(X)-\theta^*|\leq 2|\hat{\eta}(X)-\eta(X)|\}} + \frac{2^{\alpha+1}c_0}{p^{2+\alpha}}\|\eta - \hat{\eta}\|_1^{1+\alpha} \\
&+ \frac{1}{p}\mathbb{E}\,|\eta(X) - \theta^*|\,\mathbb{1}_{\left\{|\eta(X)-\theta^*|\leq\frac{2}{p}\sup_{t\in[0,1]}\left|\hat{\mathbb{P}}_X(\hat{\eta}(X)\leq t)-\mathbb{P}_X(\hat{\eta}(X)\leq t)\right|\right\}} \quad . \qquad (2.6)
\end{aligned}$$

Taking expectation on both sides *w.r.t.* the distribution of $(\mathcal{D}_n^{\mathrm{L}}, \mathcal{D}_N^{\mathrm{U}})$ we follow [Audibert and Tsybakov, 2007, Lemma 3.1] to bound the first term on the right hand side. Though this arguments became classical in statistics, we demonstrate how to performs it for convenience of the reader. Recall, that our goal is to bound

$$(\star) := \mathbb{E}_{\mathcal{D}_n^{\mathrm{L}}} \mathbb{E} \left| \eta(X) - \theta^* \right| \mathbb{1}_{\{|\eta(X) - \theta^*| \leq 2|\hat{\eta}(X) - \eta(X)|\}} \ .$$

For some fixed $\delta > 0$ introduce the following peeling of the space $\mathbb{R}^d$

$$\begin{aligned} \mathcal{A}_0 &= \left\{ x \in \mathbb{R}^d \ : \ 0 < |\eta(x) - \theta^*| \leq \delta \right\} \ , \\ \mathcal{A}_j &= \left\{ x \in \mathbb{R}^d \ : \ 2^{j-1}\delta < |\eta(x) - \theta^*| \leq 2^j \delta \right\} \qquad j \in \mathbb{N} \ . \end{aligned}$$

Using this partition we can write the following expression for $(\star)$

$$\begin{aligned} (\star) &= \sum_{j \geq 0} \mathbb{E}_{\mathcal{D}_n^{\mathrm{L}}} \mathbb{E} \left| \eta(X) - \theta^* \right| \mathbb{1}_{\{|\eta(X) - \theta^*| \leq 2|\hat{\eta}(X) - \eta(X)|\}} \mathbb{1}_{\{X \in \mathcal{A}_j\}} \\ &= \mathbb{E}_{\mathcal{D}_n^{\mathrm{L}}} \mathbb{E} \left| \eta(X) - \theta^* \right| \mathbb{1}_{\{|\eta(X) - \theta^*| \leq 2|\hat{\eta}(X) - \eta(X)|\}} \mathbb{1}_{\{X \in \mathcal{A}_0\}} \\ &\quad + \sum_{j \geq 1} \mathbb{E}_{\mathcal{D}_n^{\mathrm{L}}} \mathbb{E} \left| \eta(X) - \theta^* \right| \mathbb{1}_{\{|\eta(X) - \theta^*| \leq 2|\hat{\eta}(X) - \eta(X)|\}} \mathbb{1}_{\{X \in \mathcal{A}_j\}} \ . \end{aligned}$$

The first term on the right hand side of this equality is bounded in a straightforward way as

$$\mathbb{E}_{\mathcal{D}_n^{\mathrm{L}}} \mathbb{E} \left| \eta(X) - \theta^* \right| \mathbb{1}_{\{|\eta(X) - \theta^*| \leq 2|\hat{\eta}(X) - \eta(X)|\}} \mathbb{1}_{\{X \in \mathcal{A}_0\}} \leq \delta \mathbb{P}_X(0 < |\eta(x) - \theta^*| \leq \delta) \leq c_0 \delta^{1+\alpha} \ ,$$

where in the last inequality the margin assumption is used. The rest of the bound goes in the following way. Fix some $j \in \mathbb{N}$ and consider the j^{th} term in the sum appearing in the expression for $(\star)$. For any such term we can write

$$\begin{aligned} &\mathbb{E}_{\mathcal{D}_n^{\mathrm{L}}} \mathbb{E} \left| \eta(X) - \theta^* \right| \mathbb{1}_{\{|\eta(X) - \theta^*| \leq 2|\hat{\eta}(X) - \eta(X)|\}} \mathbb{1}_{\{X \in \mathcal{A}_j\}} \\ &\qquad = \mathbb{E}_{\mathcal{D}_n^{\mathrm{L}}} \mathbb{E} \left| \eta(X) - \theta^* \right| \mathbb{1}_{\{|\eta(X) - \theta^*| \leq 2|\hat{\eta}(X) - \eta(X)|\}} \mathbb{1}_{\{2^{j-1}\delta < |\eta(X) - \theta^*| \leq 2^j \delta\}} \\ &\qquad \leq \mathbb{E} \left| \eta(X) - \theta^* \right| \mathbb{P}^{\otimes n} \left(|\hat{\eta}(X) - \eta(X)| \geq 2^{j-2}\delta \right) \mathbb{1}_{\{0 < |\eta(X) - \theta^*| \leq 2^j \delta\}} \\ &\qquad \leq 2^j \delta \mathbb{E} \mathbb{P}^{\otimes n} \left(|\hat{\eta}(X) - \eta(X)| \geq 2^{j-2}\delta \right) \mathbb{1}_{\{0 < |\eta(X) - \theta^*| \leq 2^j \delta\}} \ . \end{aligned}$$

Using the assumption on the estimator $\hat{\eta}$, we get for almost all $x \in \mathbb{R}^d$ *w.r.t.* $\mathbb{P}_X$

$$\mathbb{P}^{\otimes n} \left(|\hat{\eta}(x) - \eta(x)| \geq 2^{j-2}\delta \right) \leq C_1 \exp\left(-C_2 a_n 2^{2j-4} \delta^2 \right) \ .$$

This implies that for any $j \in \mathbb{N}$ we have

$$\begin{aligned} &\mathbb{E}_{\mathcal{D}_n^{\mathrm{L}}} \mathbb{E} \left| \eta(X) - \theta^* \right| \mathbb{1}_{\{|\eta(X) - \theta^*| \leq 2|\hat{\eta}(X) - \eta(X)|\}} \mathbb{1}_{\{X \in \mathcal{A}_j\}} \\ &\qquad \leq C_1 \exp\left(-C_2 a_n 2^{2j-4} \delta^2 \right) 2^j \delta \mathbb{P}_X \left(0 < |\eta(X) - \theta^*| \leq 2^j \delta \right) \ . \end{aligned}$$

Using the margin assumption again we get for all $j \in \mathbb{N}$

$$\mathbb{E}_{\mathcal{D}_n^{\mathrm{L}}} \mathbb{E} \left| \eta(X) - \theta^* \right| \mathbb{1}_{\{|\eta(X) - \theta^*| \leq 2|\hat{\eta}(X) - \eta(X)|\}} \mathbb{1}_{\{X \in \mathcal{A}_j\}} \leq C_1 c_0 \exp\left(-C_2 a_n 2^{2j-4} \delta^2 \right) 2^{(1+\alpha)j} \delta^{1+\alpha} \ .$$

Therefore, we arrived at the following bound on $(\star)$

$$(\star) \leq c_0 \delta^{1+\alpha} + C_1 c_0 \delta^{1+\alpha} \sum_{j \geq 1} \exp\left(-C_2 a_n 2^{2j-4} \delta^2 \right) 2^{(1+\alpha)j} \ .$$

Fixing $\delta = a_n^{-1/2}$ we conclude that for some $C > 0$

$$\mathbb{E}_{\mathcal{D}_n^{\mathrm{L}}} \mathbb{E} \left| \eta(X) - \theta^* \right| \mathbb{1}_{\{|\eta(X)-\theta^*| \leq 2|\hat{\eta}(X)-\eta(X)|\}} \leq C a_n^{\frac{1+\alpha}{2}} \quad .$$

Let us come back to Eq. (2.6), using Assumption 5 the term $\|\eta - \hat{\eta}\|_1^{1+\alpha}$ can be bounded with the same rate as the previous term. These arguments would imply that there exists $C \geq 0$ such that for all $n, N \geq 1$ it holds that

$$\begin{aligned}
\mathbb{E}_{(\mathcal{D}_n^{\mathrm{L}}, \mathcal{D}_N^{\mathrm{U}})} \mathcal{E}_1(\hat{g}) &\leq C a_n^{-\frac{1+\alpha}{2}} \\
&\quad + \frac{1}{p} \mathbb{E}_{(\mathcal{D}_n^{\mathrm{L}}, \mathcal{D}_N^{\mathrm{U}})} \mathbb{E} \left| \eta(X) - \theta^* \right| \mathbb{1}_{\left\{|\eta(X)-\theta^*| \leq \frac{2}{p} \sup_{t\in[0,1]} \left| \hat{\mathbb{P}}_X(\hat{\eta}(X) \leq t) - \mathbb{P}_X(\hat{\eta}(X) \leq t) \right| \right\}} \\
&\leq C a_n^{-\frac{1+\alpha}{2}} + \frac{2^{\alpha+1} c_0}{p^{2+\alpha}} \mathbb{E}_{(\mathcal{D}_n^{\mathrm{L}}, \mathcal{D}_N^{\mathrm{U}})} \left(\sup_{t\in[0,1]} \left| \hat{\mathbb{P}}_X(\hat{\eta}(X) \leq t) - \mathbb{P}_X(\hat{\eta}(X) \leq t) \right| \right)^{1+\alpha} \quad .
\end{aligned}$$

It remains to upper bound the second term in the bound above, to this end we recall the classical Dvoretzky-Kiefer-Wolfowitz inequality [Massart, 1990]

Lemma 4 (Dvoretzky-Kiefer-Wolfowitz inequality)**.** *Given $N \geq 0$, let $Z_1, \ldots, Z_N$ be* i.i.d. *real-valued random variables with cumulative distribution function F_Z, denote by $\hat{F}_Z$ the cumulative distribution function with respect to the empirical measure, that is, with respect to $\frac{1}{N}\sum_{i=1}^N \delta_{Z_i}$, then for every $t > 0$ we have*

$$\mathbb{P}\left(\sup_{z\in\mathbb{R}} \left| \hat{F}_Z(z) - F_Z(z) \right| \geq t \right) \leq 2 \exp\left(-2Nt^2\right) \quad .$$

Let us apply this lemma to $Z_i := \hat{\eta}(X_i)$, conditionally on $\mathcal{D}_n^{\mathrm{L}}$ these random variables are *i.i.d.* real-valued, thus for all $t > 0$

$$\mathbb{P}\left(\sup_{z\in[0,1]} \left| \hat{\mathbb{P}}_X(\hat{\eta}(X) \leq z) - \mathbb{P}_X(\hat{\eta}(X) \leq z) \right| \geq t \middle| \mathcal{D}_n^{\mathrm{L}} \right) \leq 2 \exp\left(-2Nt^2\right), \quad \mathcal{D}_n^{\mathrm{L}}\text{-a.s.} \quad .$$

Finally, to conclude the upper bound we apply this exponential concentration to upper bound the expectation. Introduce the following notation for the supremum of the empirical process

$$\Delta_{(\mathcal{D}_N^{\mathrm{U}}, \mathcal{D}_n^{\mathrm{L}})} := \sup_{t\in[0,1]} \left| \hat{\mathbb{P}}_X(\hat{\eta}(X) \leq t) - \mathbb{P}_X(\hat{\eta}(X) \leq t \right| \quad .$$

Using this notation we can upper bound the expected supremum as

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}_n^{\mathrm{L}}} \mathbb{E}_{\mathcal{D}_N^{\mathrm{U}}} \left[\left(\Delta_{(\mathcal{D}_N^{\mathrm{U}}, \mathcal{D}_n^{\mathrm{L}})} \right)^{1+\alpha} \middle| \mathcal{D}_n^{\mathrm{L}} \right] &= \mathbb{E}_{\mathcal{D}_n^{\mathrm{L}}} \int_0^\infty \mathbb{P}\left(\Delta_{(\mathcal{D}_N^{\mathrm{U}}, \mathcal{D}_n^{\mathrm{L}})} \geq t^{\frac{1}{1+\alpha}} \middle| \mathcal{D}_n^{\mathrm{L}} \right) dt \\
&\leq \int_0^\infty 2 \exp\left(-2Nt^{\frac{2}{1+\alpha}}\right) dt \\
&= N^{-\frac{1+\alpha}{2}} 2 \int_0^\infty \exp\left(-2t^{\frac{2}{1+\alpha}}\right) dt \\
&\leq C N^{-\frac{1+\alpha}{2}} \quad ,
\end{aligned}$$

Combining all the bounds we conclude.

Proof of the lower bound

Proof. The proof is similar to the one used in [Audibert and Tsybakov, 2007] and in [Rigollet and Vert, 2009] and is based on Assouad lemma. Similarly, we define the regular grid on $\mathbb{R}^d$ as

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

and denote by $n_q(x) \in G_q$ as the closest point to of the grid G_q to the point $x \in \mathbb{R}^d$. Such a grid defines a partition of the unit cube $[0,1]^d \subset \mathbb{R}^d$ denoted by $\mathcal{X}'_1, \dots, \mathcal{X}'_{q^d}$. Besides, denote by $\mathcal{X}'_{-j} \coloneqq \{x \in \mathbb{R}^d : -x \in \mathcal{X}'_j\}$ for all $j = 1, \dots, q^d$. For a fixed integer $m \le q^d$ and for any $j \in \{1, \dots, m\}$ define $\mathcal{X}_i \coloneqq \mathcal{X}'_i$, $\mathcal{X}_{-i} \coloneqq \mathcal{X}'_{-i}$. For every $\sigma \in \{-1,1\}^m$ we define a regression function η_σ as

$$\eta_\sigma(x) = \begin{cases} \frac{1}{4} + \sigma_j \varphi(x), & \text{if } x \in \mathcal{X}_i \\ \frac{1}{4} - \sigma_j \varphi(x), & \text{if } x \in \mathcal{X}_{-i} \\ \frac{1}{4}, & \text{if } x \in \mathcal{B}(0, \sqrt{d}) \setminus \left(\cup_{i=-m, i\neq 0}^m \mathcal{X}_i \right) \\ \tau, & \text{if } x \in \mathbb{R}^d \setminus \mathcal{B}(0, \sqrt{d}+\rho) \\ \xi(x), & \text{if } x \in \mathcal{B}(0, \sqrt{d}+\rho) \setminus \mathcal{B}(0, \sqrt{d}) \end{cases} ,$$

where ρ, φ, ξ, τ are to be specified and $\mathcal{B}(0, \sqrt{d}+\rho), \mathcal{B}(0, \sqrt{d})$ are Euclidean balls of radius $\sqrt{d}+\rho$ and $\sqrt{d}$ respectively. The definition of the function φ is exactly the same as in Audibert and Tsybakov [2007]. That is, $\varphi \coloneqq C_\varphi q^{-\beta} u(q \left\| x - n_q(x) \right\|_2)$ with some non-increasing infinitely differentiable function such that $u(x) = 1$ for $x \in [0, 1/4]$ and $u(x) = 0$ for $x \ge 1/2$. The function ξ is defined as $\xi(x) = (\tau - 1/4) v([\left\| x \right\|_2 - \sqrt{d}]/\rho) + 1/4$, where v is non-decreasing infinitely differentiable function such that $v(x) = 0$ for $x \le 0$ and $v(x) = 1$ for $x \ge 1$. The constant ρ is chosen big enough to ensure that $|\xi(x) - \xi_x(x')| \le L \left\| x - x' \right\|_2^\beta$ for any $x, x' \in \mathbb{R}^d$.

For any $\sigma \in \{-1,1\}^m$ we construct a marginal distribution P_X which is independent of σ and has a density μ *w.r.t.* to the Lebesgue measure on $\mathbb{R}^d$. Fix some $0 < w \le m^{-1}$ and set A_0 a Euclidean ball in $\mathbb{R}^d$ that has an empty intersection with $\mathcal{B}(0, \sqrt{d}+\rho)$ and whose Lebesgue measure is $\mathrm{Leb}(A_0) = 1 - mq^{-d}$. The density μ is constructed as

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

To complete the construction it remains to specify the value of $\tau \in [0,1]$. The idea here is to force the optimal threshold θ^* to be equal to some predefined constant using the additional degree of freedom provided by the parameter τ. To achieve this we would like to set $\theta^* = 1/4$ and we would like to demonstrate that there exists an appropriate choice of τ which ensures that $\theta^* = 1/4$. First, recall that the optimal threshold θ^* for the classification with the F-score must satisfy the equation

$$\theta^* \mathbb{E}\eta(X) = \mathbb{E}(\eta(X) - \theta^*)_+ .$$

Define $b' = \int_{\mathcal{X}_1} \varphi(x)\mu(x)dx / \int_{\mathcal{X}_1} \mu(x)dx$ and put $\theta^* = 1/4$, notice that the left hand side of the last equality for every $\sigma \in \{-1,1\}^m$ is given by

$$\begin{aligned}\mathbb{E}_\mu \eta_\sigma(X) &= \int_{\mathbb{R}^d} \eta(x)d\mu(x) \\ &= \sum_{j=1}^m \int_{\mathcal{X}_j} (1/4 + \sigma_j \xi(x))d\mu(x) + \sum_{j=1}^m \int_{\mathcal{X}_{-j}} (1/4 - \sigma_j \xi(x))d\mu(x) + \int_{A_0} \tau d\mu(x) \\ &= \frac{mw}{2} + \tau(1 - 2mw) \ .\end{aligned}$$

For the right hand side $\mathbb{E}_\mu(\eta_\sigma(X) - 1/4)_+$, there are two cases $\tau > 1/4$ and $0 < \tau \leq 1/4$, one can easily show that as long as $b' \leq 1/8$ no value of $\tau \in (1, 1/4]$ allows to fix $\theta^* = 1/4$. Therefore, $\tau > 1/4$ and we can write for every $\sigma \in \{-1,1\}^m$

$$\begin{aligned}\mathbb{E}_\mu(\eta_\sigma(X) - 1/4)_+ &= \sum_{j=1}^m \int_{\mathcal{X}_j} (\sigma_j \xi(x))_+ d\mu(x) + \sum_{j=1}^m \int_{\mathcal{X}_{-j}} (-\sigma_j \xi(x))_+ d\mu(x) + \int_{A_0} (\tau - 1/4)d\mu(x) \\ &= mwb' + (\tau - 1/4)(1 - 2mw) \ .\end{aligned}$$

Finally, the parameter τ must satisfy the following equality

$$\frac{1}{4}\left(\frac{mw}{2} + \tau(1 - 2mw)\right) = mwb' + (\tau - 1/4)(1 - 2mw) \ ,$$

solving for τ we get

$$\tau = \frac{1}{3} + \left(\frac{1}{12} - \frac{2b'}{3}\right)\left(\frac{2mw}{1 - 2mw}\right) \ .$$

Notice that this choice of τ implies that for all $\sigma \in \{-1,1\}^m$ the optimal threshold is given by $\theta^* = 1/4$. Moreover, if $mw \leq 1/2$ we can ensure that the value of $\tau \in [0,1]$, that is, it is a valid choice for the regression function. Let us demonstrate that (the margin) Assumption 4 holds for an appropriate choice of m and w. Define $x_0 = (1/2q, \ldots, 1/2q)^\top$, then for every $\sigma \in \{-1,1\}$ we have

$$\begin{aligned}P_X(0 < |\eta_\sigma(X) - 1/4| \leq \delta) &= \frac{2mw}{\mathrm{Leb}(\mathcal{B}(0,(4q)^{-1}))} \int_{\mathcal{B}(x_0,(4q)^{-1})} \mathbb{1}_{\left\{C_\varphi q^{-\beta} u(q\|x - n_q(x)\|_2) \leq \delta\right\}} dx \\ &\quad + \frac{1 - 2mw}{\mathrm{Leb}(A_0)} \int_{A_0} \mathbb{1}_{\left\{\frac{1}{3} + \left(\frac{1}{12} - \frac{2b'}{3}\right)\left(\frac{2mw}{1-2mw}\right) - \frac{1}{4} \leq \delta\right\}} dx \\ &= 2mw \mathbb{1}_{\left\{\delta \geq C_\varphi q^{-\beta}\right\}} + \frac{1 - 2mw}{\mathrm{Leb}(A_0)} \int_{A_0} \mathbb{1}_{\left\{\frac{1}{12} + \left(\frac{1}{12} - \frac{2b'}{3}\right)\left(\frac{2mw}{1-2mw}\right) \leq \delta\right\}} dx \ ,\end{aligned}$$

as long as $b' \leq 3/24$ we can continue as

$$\begin{aligned}P_X(0 < |\eta_\sigma(X) - 1/4| \leq \delta) &\leq 2mw \mathbb{1}_{\left\{\delta \geq C_\varphi q^{-\beta}\right\}} + \mathbb{1}_{\left\{\delta \geq \frac{1}{12}\right\}} \\ &\leq 2mw \mathbb{1}_{\left\{\delta \geq C_\varphi q^{-\beta}\right\}} + 12^\alpha \delta^\alpha \ .\end{aligned}$$

Therefore, if mw is of order $q^{-\alpha\beta}$ the margin assumption is satisfied with $\delta_0 = 1/12$. The strong density assumption can be checked similarly to [Audibert and Tsybakov, 2007]. To finish the proof, for every $\sigma \in \{-1,1\}^m$ we denote by P^σ the distribution of (X,Y) with the marginal P_X and the regression function η^σ. Thus, one can write for any $\hat{g}$

$$\sup_{\mathbb{P} \in \mathcal{P}_\Sigma} \mathbb{E}_{(\mathcal{D}_n^{\mathrm{L}}, \mathcal{D}_N^{\mathrm{U}})} \mathcal{E}(\hat{g}) \geq \sup_{\sigma \in \{-1,1\}^m} \frac{1}{2} \mathbb{E}^\sigma_{(\mathcal{D}_n^{\mathrm{L}}, \mathcal{D}_N^{\mathrm{U}})} \sum_{i=-m, i\neq 0}^m \mathbb{E}_{P_X} |\varphi(X)| \, \mathbb{1}_{\{(1+\mathrm{sign}(i)\sigma_i)/2 \neq \hat{g}(X)\}} \mathbb{1}_{\{X \in \mathcal{X}_i\}} \ ,$$

where $\mathbb{E}^{\sigma}_{(\mathcal{D}_n^{\mathrm{L}},\mathcal{D}_N^{\mathrm{U}})}$ is the expectation taken *w.r.t.* to the *i.i.d.* realizations of $\mathcal{D}_n^{\mathrm{L}}$ and $\mathcal{D}_N^{\mathrm{U}}$ from P^{σ} and P_X respectively, and $\mathrm{sign}(i) = 1$ if $i > 0$ and $\mathrm{sign}(i) = -1$ if $i < 0$.

The rest of the proof can be obtained in various ways. First of all we can follow line by line the proof of [Audibert, 2004, Lemma 5.1.] and in particular the chain of inequalities in [Audibert, 2004, Eq. (6.26)]. As noted by Audibert this machinery gives the best known constants in the lower bound. Though this proof is relatively simple, it also requires to introduce a lot of notation. That is why we follow another path, namely, we shall apply Fano's inequality in the form obtained by Birgé [2005]. Let us recall this inequality and other notions that are required, we also provide these definitions in Appendix.

Definition 8. *Given any two probability measures $\mathbb{P}_1, \mathbb{P}_2$ on some space measurable space $(\mathcal{X}, \mathcal{A})$ the Kullback–Leibler divergence between $\mathbb{P}_1$ and $\mathbb{P}_2$ is defined as*

$$\mathrm{KL}(\mathbb{P}_1, \mathbb{P}_2) := \begin{cases} \int_{\mathcal{X}} \log\left(\frac{d\mathbb{P}_1}{d\mathbb{P}_2}\right) d\mathbb{P}_1, & \mathrm{supp}(\mathbb{P}_1) \subset \mathrm{supp}(\mathbb{P}_2) \\ +\infty, & \textit{otherwise} \end{cases} , \tag{2.7}$$

Fano's inequality in the form proved by [Birgé, 2005] is then stated as.

Lemma 5. *Let $\{\mathbb{P}_i\}_{i=0}^m$ be a finite family of probability measures on $(\mathcal{X}, \mathcal{A})$ and let $\{A_i\}_{i=0}^m$ be a finite family of disjoint events such that $A_i \in \mathcal{A}$ for each $i = 0, \dots, m$. Then,*

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

To apply this inequality we need to simplify our problem. To this end, for any algorithm $\hat{g}$ we write

$$\sup_{\mathbb{P} \in \mathcal{P}_\Sigma} \mathbb{E}_{(\mathcal{D}_n^{\mathrm{L}},\mathcal{D}_N^{\mathrm{U}})} \mathcal{E}(\hat{g}) \geq \sup_{\sigma \in \{-1,1\}^m} \frac{1}{2} \mathbb{E}^{\sigma}_{(\mathcal{D}_n^{\mathrm{L}},\mathcal{D}_N^{\mathrm{U}})} \sum_{i=1}^m \mathbb{E}_{P_X} |\varphi(X)| \mathbb{1}_{\left\{\frac{1+\sigma_i}{2} \neq \hat{g}(X)\right\}} \mathbb{1}_{\{X \in \mathcal{X}_i\}} ,$$

where we have dropped the summation over negative indices $i \in \{-m, \dots, -1\}$. Now, for some set $\mathcal{W} \subset \{-1, 1\}^m$ to be specified define the following test statistics

$$\hat{\sigma} \in \arg\min \left\{ \sum_{i=1}^m \mathbb{E}_{P_X} |\varphi(X)| \mathbb{1}_{\left\{\frac{1+\sigma_i}{2} \neq \hat{g}(X)\right\}} \mathbb{1}_{\{X \in \mathcal{X}_i\}} \, : \, \sigma \in \mathcal{W} \right\} .$$

Now, for any estimator $\hat{g}$ and each $\sigma \in \mathcal{W}$ such that $\sigma \neq \hat{\sigma}$ we can write thanks to the definition of $\hat{\sigma}$ and the triangle inequality

$$\begin{aligned}
2\sum_{i=1}^m \mathbb{E}_{P_X} |\varphi(X)| \mathbb{1}_{\left\{\frac{1+\sigma_i}{2} \neq \hat{g}(X)\right\}} \mathbb{1}_{\{X \in \mathcal{X}_i\}} &= \sum_{i=1}^m \mathbb{E}_{P_X} |\varphi(X)| \mathbb{1}_{\left\{\frac{1+\sigma_i}{2} \neq \hat{g}(X)\right\}} \mathbb{1}_{\{X \in \mathcal{X}_i\}} \\
&\quad + \sum_{i=1}^m \mathbb{E}_{P_X} |\varphi(X)| \mathbb{1}_{\left\{\frac{1+\sigma_i}{2} \neq \hat{g}(X)\right\}} \mathbb{1}_{\{X \in \mathcal{X}_i\}} \\
&\geq \sum_{i=1}^m \mathbb{E}_{P_X} |\varphi(X)| \mathbb{1}_{\left\{\frac{1+\sigma_i}{2} \neq \hat{g}(X)\right\}} \mathbb{1}_{\{X \in \mathcal{X}_i\}} \\
&\quad + \sum_{i=1}^m \mathbb{E}_{P_X} |\varphi(X)| \mathbb{1}_{\left\{\frac{1+\hat{\sigma}_i}{2} \neq \hat{g}(X)\right\}} \mathbb{1}_{\{X \in \mathcal{X}_i\}} \\
&\geq \sum_{i=1}^m \mathbb{E}_{P_X} |\varphi(X)| \mathbb{1}_{\{\sigma_i \neq \hat{\sigma}_i\}} \mathbb{1}_{\{X \in \mathcal{X}_i\}} \\
&= \mathbb{E}_{P_X}\left[|\varphi(X)| \mathbb{1}_{\{X \in \mathcal{X}_1\}}\right] \sum_{i=1}^m \mathbb{1}_{\{\sigma_i \neq \hat{\sigma}_i\}}
\end{aligned}$$

Now let us define the set $\mathcal{W}$, this set is provided by classical result in information theory, typically called Varshamov-Gilbert lemma [Gilbert, 1952, Varshamov, 1957].

Lemma 6. *Let $\delta(\sigma, \sigma')$ denote the Hamming distance between $\sigma, \sigma' \in \{-1, 1\}^m$ given by*

$$\delta(\sigma, \sigma') := \sum_{i=1}^{m} \mathbb{1}_{\left\{\sigma_i \neq \sigma'_i\right\}} \ .$$

There exists $\mathcal{W} \subset \{-1, 1\}^m$ such that for all $\sigma \neq \sigma' \in \mathcal{W}$ we have

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

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

Using the set $\mathcal{W}$ provided by the previous result we can state that for any method $\hat{g}$ and any $\sigma \in \mathcal{W}$ such that $\sigma \neq \hat{\sigma}$ we have

$$\sum_{i=1}^{m} \mathbb{E}_{P_X} \left|\varphi(X)\right| \mathbb{1}_{\left\{\frac{1+\sigma_i}{2} \neq \hat{g}(X)\right\}} \mathbb{1}_{\{X \in \mathcal{X}_i\}} \geq \mathbb{E}_{P_X} \left[\left|\varphi(X)\right| | X \in \mathcal{X}_1\right] \mathbb{P}_X(X \in \mathcal{X}_1) \frac{m}{8} = \frac{C_\varphi q^{-\beta} w m}{8} \ .$$

Thus, we demonstrated that for every $\hat{g}$ we have

$$\begin{aligned}
\sup_{\mathbb{P} \in \mathcal{P}_\Sigma} \mathbb{E}_{(\mathcal{D}_n^{\mathrm{L}}, \mathcal{D}_N^{\mathrm{U}})} \mathcal{E}(\hat{g}) &\geq \frac{C_\varphi q^{-\beta} w m}{16} \max_{\sigma \in \mathcal{W}} \mathbb{E}_{(\mathcal{D}_n^{\mathrm{L}}, \mathcal{D}_N^{\mathrm{U}})} \mathbb{1}_{\{\sigma \neq \hat{\sigma}\}} \\
&= \frac{C_\varphi q^{-\beta} w m}{16} \left(1 - \min_{\sigma \in \mathcal{W}} \mathbb{E}_{(\mathcal{D}_n^{\mathrm{L}}, \mathcal{D}_N^{\mathrm{U}})} \mathbb{1}_{\{\sigma = \hat{\sigma}\}}\right)
\end{aligned}$$

Notice that the event $\mathcal{A}_\sigma = \{\sigma = \hat{\sigma}\}$ are disjoint, thus, we can apply Fano's inequality and obtain

$$\sup_{\mathbb{P} \in \mathcal{P}_\Sigma} \mathbb{E}_{(\mathcal{D}_n^{\mathrm{L}}, \mathcal{D}_N^{\mathrm{U}})} \mathcal{E}(\hat{g}) \geq \frac{C_\varphi q^{-\beta} w m}{16} \left(1 - \left(0.71 \bigvee \frac{\frac{1}{|\mathcal{W}|} \sum_{\sigma \in \mathcal{W}} \mathrm{KL}\left(\mathbb{P}_X^{\otimes N} \otimes (\mathbb{P}^\sigma)^{\otimes n}, \mathbb{P}_X^{\otimes N} \otimes (\mathbb{P}^{\bar{\sigma}})^{\otimes n}\right)}{\log(|\mathcal{W}| + 1)}\right)\right) \ ,$$

where $\bar{\sigma}$ is an arbitrary element of $\mathcal{W}$. Therefore, it remains to upper-bound the KL-divergence between any two fixed $\sigma, \bar{\sigma} \in \mathcal{W}$. We can write for product measures

$$\mathrm{KL}\left(\mathbb{P}_X^{\otimes N} \otimes (\mathbb{P}^\sigma)^{\otimes n}, \mathbb{P}_X^{\otimes N} \otimes (\mathbb{P}^{\bar{\sigma}})^{\otimes n}\right) \leq n \, \mathrm{KL}(\mathbb{P}^\sigma, \mathbb{P}^{\bar{\sigma}}) \ ,$$

and for some $C > 0$ we have

$$\mathrm{KL}(\mathbb{P}^\sigma, \mathbb{P}^{\bar{\sigma}}) \leq 2 \sum_{i=-m, i \neq 0}^{m} \mu\left(\varphi(X) \log\left(\frac{1/4 + \varphi(X)}{1/4 - \varphi(X)}\right), X \in \mathcal{X}_i\right) \leq C q^{-2\beta} w m \ .$$

We arrived at the following bound for any $\hat{g}$

$$\sup_{\mathbb{P} \in \mathcal{P}_\Sigma} \mathbb{E}_{(\mathcal{D}_n^{\mathrm{L}}, \mathcal{D}_N^{\mathrm{U}})} \mathcal{E}(\hat{g}) \geq \frac{C_\varphi q^{-\beta} w m}{16} \left(1 - \left(0.71 \bigvee \frac{C q^{-2\beta} w m n}{\log(|W| + 1)}\right)\right) \ .$$

Recall that thanks to Varshamov-Gilbert lemma we know that

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

Therefore, we get for some $C > 0$ and every estimator $\hat{g}$

$$\sup_{\mathbb{P}\in\mathcal{P}_\Sigma} \mathbb{E}_{(\mathcal{D}_n^{\mathrm{L}},\mathcal{D}_N^{\mathrm{U}})}\mathcal{E}(\hat{g}) \geq \frac{C_\varphi q^{-\beta} wm}{16}\left(1 - \left(0.71 \bigvee Cq^{-2\beta}wn\right)\right) \ .$$

Finally, we conclude by setting the parameters m, w, q as

$$q = \lfloor \bar{C} n^{\frac{1}{2\beta+d}} \rfloor, \quad w = C' q^{-d}, \quad m = \lfloor C'' q^{d-\alpha\beta} \rfloor \ .$$

Note that thanks to the condition $\alpha\beta \leq d$ such a choice is always valid for appropriately chosen constants $\bar{C}, C', C''$.

□

2.2 Fair binary classification

Section overview. We study the problem of fair binary classification using the notion of Equal Opportunity that requires the true positive rate to distribute equally across the sensitive groups. Within this setting we show that the fair optimal classifier is obtained by recalibrating the Bayes classifier by a group-dependent threshold, and we provide a constructive expression for the thresholds. This motivates us to devise a plug-in classification procedure based on both unlabeled and labeled datasets. While the latter is used to learn the output conditional probability, the former is used for calibration. The overall procedure can be computed in polynomial time and it is shown to be statistically consistent both in terms of the classification error and fairness measure. Finally, we present numerical experiments which indicate that our method is often superior or competitive with the state-of-the-art methods on benchmark datasets.

2.2.1 Introduction

As machine learning is spreading more and more in our society, the potential risk of using algorithms that behave unfairly is rising. As a result there is growing interest to design learning methods that meet "fairness" requirements, see, e.g., [Barocas et al., 2018, Donini et al., 2018, Dwork et al., 2018, Hardt et al., 2016, Zafar et al., 2017, Zemel et al., 2013, Kilbertus et al., 2017, Kusner et al., 2017, Calmon et al., 2017, Joseph et al., 2016, Chierichetti et al., 2017, Jabbari et al., 2016, Yao and Huang, 2017, Lum and Johndrow, 2016, Zliobaite, 2015] and references therein. A central goal is to make sure that sensitive information does not "unfairly" influence the outcomes of learning methods. For instance, if we wish to predict whether a university student applicant should be offered a scholarship based on curriculum, we would like our model to not unfairly use additional sensitive information such as gender or race.

Several measures of fairness of a classifier have been studied in the literature [Zafar et al., 2019], ranging from Demographic Parity [Calders et al., 2009], Equal Odds and Equal Opportunity [Hardt et al., 2016], Disparate Treatment, Impact, and Mistreatment [Zafar et al., 2017], which were introduced in Section 1.1.4. In this part of Section 2.2, we study the problem of learning a binary classifier which satisfies the Equal Opportunity fairness constraint. It requires that the true positive rate of the classifier is the same across the sensitive groups. This notion has been used extensively in the literature either as a postprocessing step [Hardt et al., 2016] on a learned classifier or directly during training, see for example [Donini et al., 2018] and references therein.

We address the important problem of devising statistically consistent and computationally efficient learning procedures that meet the fairness constraint. Specifically, we make four contributions. First, we derive in Proposition 2 the expression for the optimal equal opportunity classifier, obtained via thresholding of the Bayes classifier. Second, inspired by the above result we propose a semi-supervised plug-in type method, which first estimates the regression function on labeled data and then estimates the unknown threshold using unlabeled data. Third, we establish in Theorem 9 that the proposed procedure is consistent, that is, it asymptotically satisfies the equal opportunity constraint and its risk converges to the risk of the optimal equal opportunity classifier. Finally, we present numerical experiments which indicate that our method is often superior or competitive with the state-of-the-art on benchmark datasets.

We highlight that the proposed learning algorithm can be applied on top of any off-the shelf method which consistently estimates the regression function (class condition probability), under mild additional assumptions which we discuss later on. Furthermore, our calibration procedure is based on solving a simple univariate problem. Hence the generality, statistical consistency and computational efficiency are strengths of our approach.

Organization of the section

This contribution is organized in the following manner. In Section 2.2.2, we introduce the problem and we derive a form of the optimal equal opportunity classifier. Section 2.2.3 is devoted to the description of our method. In Section 2.2.4 we introduce assumptions used throughout this chapter and establish that the proposed learning algorithm is consistent. Finally, Section 2.2.5 presents numerical experiments with our method.

Related work

In this part we review previous contributions on the subject of fair classification of equall opportunity. Works on algorithmic fairness can be divided in three families. Our algorithm falls within the first family, which modifies a pre-trained classifier in order to increase its fairness properties while maintaining as much as possible the classification performance, see [Pleiss et al., 2017, Beutel et al., 2017, Hardt et al., 2016, Feldman et al., 2015] and references therein. Importantly, for our approach the post-processing step requires only unlabeled data, which is often easier to collect than its labeled counterpart. Methods in the second family enforce fairness directly during the training step, e.g. [Oneto et al., 2019b, Donini et al., 2018, Agarwal et al., 2018, Cotter et al., 2018]. The third family of methods implements fairness by modifying the data representation and then employs standard machine learning methods, see e.g. [Donini et al., 2018, Adebayo and Kagal, 2016, Calmon et al., 2017, Kamiran and Calders, 2009, Zemel et al., 2013, Kamiran and Calders, 2012, 2010] as representative examples.

To the best of our knowledge the formula for the optimal fair classifier presented here is novel. In [Hardt et al., 2016] the authors note that the optimal equalized odds or equal opportunity classifier can be derived from the Bayes optimal classifier, however, no explicit expression for this threshold is provided. The idea of recalibrating the Bayes classifier is also discussed in a number of papers, see for example [Pleiss et al., 2017, Menon and Williamson, 2018] and references therein. More importantly, the problem of deriving *efficient* and *consistent* estimators under fairness constraints has received limited attention in the literature. In [Donini et al., 2018], the authors present consistency results under restrictive assumptions

on the model class. Furthermore, they only consider convex approximations of the risk and fairness constraint and it is not clear how to relate their results to the original problem with the misclassification risk. In [Agarwal et al., 2018], the authors reduce the problem of fair classification to a sequence of cost-sensitive problems by leveraging a saddle point formulation. They show that their algorithm is consistent in both risk and fairness constraints. However, similarly to [Donini et al., 2018], the authors of [Agarwal et al., 2018] assume that the family of possible classifiers admits a bounded Rademacher complexity.

Plug-in methods in classification problems are well established and are well studied from statistical perspective, see [Yang, 1999, Audibert and Tsybakov, 2007, Devroye, 1978] and references therein. In particular, let us recall Section 1.1 where we reviewed the works of [Yang, 1999, Audibert and Tsybakov, 2007] who showed that one can build a plug-in type classifier which is optimal in minimax sense. Until very recently, theoretical studies of plug-in methods were reduced to an efficient estimation of the regression function. Indeed, as it is shown Section 1.1, for standard settings of classification the threshold is always known beforehand, thus, all the information about the optimal classifier is wrapped into the distribution of the label conditionally on the feature.

More recently, classification problems with a distribution dependent threshold have emerged, constrained classification considered in this manuscript being a particular example of such a problem. Other prominent examples include classification with non-decomposable measures [Yan et al., 2018, Koyejo et al., 2015, Zhao et al., 2013] (*e.g.,* the F-score setup in Section 2.1), classification with reject option [Denis and Hebiri, 2015a, Lei, 2014], and confidence set setup (a particular instance of the constrained classification frameowork, see Chapter 3) of multi-class classification [Chzhen et al., 2019a, Sadinle et al., 2018, Denis and Hebiri, 2017], among others. A typical estimation algorithm in these scenarios is based on plug-in strategies, which use extra data to estimate the unknown threshold as noted in Section 1.0.3.

2.2.2 Optimal Equal Opportunity classifier

Let (X, S, Y) be a tuple on $\mathbb{R}^d \times \{0,1\} \times \{0,1\}$ having a joint distribution $\mathbb{P}$. Here the vector $X \in \mathbb{R}^d$ is seen as the vector of features, $S \in \{0,1\}$ a binary sensitive variable and $Y \in \{0,1\}$ a binary output label that we wish to predict from the pair (X, S). We also assume that the distribution is non-degenerate in Y and S that is $\mathbb{P}(S = 1) \in (0,1)$ and $\mathbb{P}(Y = 1) \in (0,1)$. A classifier g is a measurable function from $\mathbb{R}^d \times \{0,1\}$ to $\{0,1\}$, and the set of all such functions is denoted by $\mathcal{G}$. In words, each classifier receives a pair $(x, s) \in \mathbb{R}^d \times \{0,1\}$ and outputs a binary prediction $g(x, s) \in \{0,1\}$. For any classifier g we introduce its associated misclassification risk as

$$\mathcal{R}(g) := \mathbb{P}\left(g(X, S) \neq Y\right) \quad . \tag{2.8}$$

In the context of fair classification the ultimate goal is to recover a *fair* optimal classifier which is formally defined as a Bayes classifier of a constrained problem as

$$g^* \in \arg\min_{g \in \mathcal{G}} \left\{\mathcal{R}(g) \,:\, g \text{ is fair}\right\} \quad .$$

There are various definitions of fairness available in the literature, each having its critics and its supporters. In Section 2.2, we employ the following definition introduced in [Hardt et al., 2016]. We refer the reader to that work as well as [Donini et al., 2018, Agarwal et al.,

2018, Menon and Williamson, 2018] for a discussion, motivation of this definition, and a comparison to other fairness definitions. For the convenience of the reader let us again recall the definition of Equall Opportunity discussed in Section 1.1.4.

Definition 9 (Equal Opportunity [Hardt et al., 2016]). *A classifier* $(x, s) \mapsto g(x, s) \in \{0, 1\}$ *is called fair if*

$$\mathbb{P}\left(g(X, S) = 1 \,|\, S = 1, Y = 1\right) = \mathbb{P}\left(g(X, S) = 1 \,|\, S = 0, Y = 1\right) \quad .$$

The set of all fair classifiers is denoted by $\mathcal{F}(\mathbb{P})$.

Again note, that as in general constrained classification framework, the definition of fairness depends on the underlying distribution $\mathbb{P}$ and hence the whole class $\mathcal{F}(\mathbb{P})$ of the fair classifiers should be estimated. Interestingly, in this case the class $\mathcal{F}(\mathbb{P})$ is non-empty for any distribution $\mathbb{P}$ as it always contains the classifier which always outputs the zero label.

Using this notion of fairness and following the idea of the constrained classification we define an optimal equal opportunity classifier as a solution of the optimization problem

$$\min_{g \in \mathcal{G}} \left\{ \mathcal{R}(g) \,:\, \mathbb{P}\left(g(X, S) = 1 \,|\, Y = 1, S = 1\right) = \mathbb{P}\left(g(X, S) = 1 \,|\, Y = 1, S = 0\right) \right\} \quad . \tag{2.9}$$

We now introduce an assumption on the regression function that plays an important role in establishing the form of the optimal fair classifier. This assumption was already discussed in Section 1.1.3, where we introduced general framework for constrained classification.

Assumption 9. *For each* $s \in \{0, 1\}$ *we require the mapping* $t \mapsto \mathbb{P}\left(\eta(X, S) \leq t \,|\, S = s\right)$ *to be continuous on* $(0, 1)$*, where for all* $(x, s) \in \mathbb{R}^d \times \{0, 1\}$*, we let the regression function*

$$\eta(x, s) := \mathbb{P}\left(Y = 1 \,|\, X = x, S = s\right) = \mathbb{E}\left[Y \,|\, X = x, S = s\right] \quad .$$

Moreover, for every $s \in \{0, 1\}$*, we assume that* $\mathbb{P}\left(\eta(X, s) \geq 1/2 \,|\, S = s\right) > 0$.

Let us mention, that the first part of Assumption 9 is achieved by many distributions and has been already introduced in various contexts, see *e.g.,* [Chzhen et al., 2019a, Yan et al., 2018, Sadinle et al., 2018, Denis and Hebiri, 2015a, Lei, 2014] (most of them are examples of constrained classifications). It says that, for every $s \in \{0, 1\}$ the random variable $\eta(X, s)$ does not have atoms, that is, the event $\{\eta(X, s) = t\}$ has probability zero. The second part of the assumption is specific to the problem of fair classification. It states that the regression function $\eta(X, s)$ must surpass the level $1/2$ on a set of non-zero measure. Informally, returning to the scholarship example mentioned in the introduction, this assumption means that there are individuals from *both* groups who are more likely to be offered a scholarship based on their curriculum.

In the following result we establish that the optimal equal opportunity classifier is obtained by recalibrating the Bayes classifier. Let us mention, that one can obtain this result using the general framework of Section 1.1.3, however, for convenience we include the direct proof in Section 2.2.8.

Proposition 2 (Optimal Rule). *Under Assumption 9 an optimal classifier* g^* *can be obtained for all* $(x, s) \in \mathbb{R}^d \times \{0, 1\}$ *as*

$$g^*(x, 1) = \mathbb{1}_{\left\{1 \leq \eta(x,1)\left(2 - \frac{\theta^*}{\mathbb{P}(Y=1, S=1)}\right)\right\}}, \quad g^*(x, 0) = \mathbb{1}_{\left\{1 \leq \eta(x,0)\left(2 + \frac{\theta^*}{\mathbb{P}(Y=1, S=0)}\right)\right\}} \tag{2.10}$$

where $\theta^* \in \mathbb{R}$ is determined from the equation

$$\frac{\mathbb{E}_{X|S=1}\left[\eta(X,1)\mathbb{1}_{\left\{1\leq\eta(X,1)\left(2-\frac{\theta^*}{\mathbb{P}(Y=1,S=1)}\right)\right\}}\right]}{\mathbb{P}\left(Y=1\,|\,S=1\right)} = \frac{\mathbb{E}_{X|S=0}\left[\eta(X,0)\mathbb{1}_{\left\{1\leq\eta(X,0)\left(2+\frac{\theta^*}{\mathbb{P}(Y=1,S=0)}\right)\right\}}\right]}{\mathbb{P}\left(Y=1\,|\,S=0\right)} \ .$$

Furthermore it holds that $|\theta^*| \leq 2$.

Proof sketch. The proof relies on weak duality. The first step of the proof is to write the minimization problem for g^* using a "min-max" problem formulation. We consider the corresponding dual "max-min" problem and show that it can be analytically solved. Then, the continuity part of Assumption 9 allows to demonstrate that the solution of the "max-min" problem gives a solution of the "min-max" problem. The second part of Assumption 9 is used to prove that $|\theta^*| \leq 2$. □

It is interesting to point out that in case of the F-score classification discussed in Section 2.1 we did not require the continuity assumption introduced here. The reason for such a discrepancy was already partially addressed by Remark 1, where we showed a way to relax the continuity assumption in general constrained framework. In particular, recall the condition on the threshold θ^* in case of the F-score reads as

$$\theta^*\mathbb{P}(Y=1) = \mathbb{E}(\eta(X)-\theta^*)_+.$$

As it is shown in Section 2.1.6, such value θ^* *always* exists, thus, the continuity assumption in case of the F-score is not required.

Before proceeding further, let us define a measure of unfairness, which plays a key role in our statistical analysis, it is seen as the violation of constraints discussed in Chapter 1; the following notion is sometimes referred to as the Difference of Equal Opportunity (DEO) in the literature; see e.g., [Donini et al., 2018].

Definition 10 (Unfairness)**.** *For any classifier g we define its unfairness as*

$$\Delta(g,\mathbb{P}) = \left|\mathbb{P}\left(g(X,S)=1\,|\,S=1,Y=1\right) - \mathbb{P}\left(g(X,S)=1\,|\,S=0,Y=1\right)\right| \ .$$

Recall that in constrained classification, a principal goal is to construct a classification algorithm $\hat{g}$ which satisfies

$$\underbrace{\mathbb{E}[\Delta(\hat{g},\mathbb{P})] \to 0}_{\text{asymptotically fair}}, \quad \text{and} \quad \underbrace{\mathbb{E}[\mathcal{R}(\hat{g})] \to \mathcal{R}(g^*)}_{\text{asymptotically optimal}} \ ,$$

where the expectations are taken with respect to the distribution of data samples. As we shall see our estimator is built from independent sets of labeled and unlabeled samples, that is, it is semi-supervised. Hence the convergence above is meant to hold as both samples grow to infinity.

2.2.3 Proposed procedure

In this part of Section 2.2, we present the proposed plug-in procedure. We assume that we have at our disposal two datasets, labeled $\mathcal{D}_n^{\mathrm{L}}$ and unlabeled $\mathcal{D}_N^{\mathrm{U}}$ defined as

$$\mathcal{D}_n^{\mathrm{L}} = \{(X_i,S_i,Y_i)\}_{i=1}^{n} \overset{i.i.d.}{\sim} \mathbb{P}, \quad \text{and} \quad \mathcal{D}_N^{\mathrm{U}} = \{(X_i,S_i)\}_{i=n+1}^{n+N} \overset{i.i.d.}{\sim} \mathbb{P}_{(X,S)} \ ,$$

where $\mathbb{P}_{(X,S)}$ is the marginal distribution of the vector (X, S). We additionally assume that the estimator $\hat{\eta}$ of the regression function is constructed from $\mathcal{D}_n^{\mathrm{L}}$, independently of $\mathcal{D}_N^{\mathrm{U}}$. Let us denote by $\hat{\mathbb{E}}_{X|S=1}$ and $\hat{\mathbb{E}}_{X|S=0}$ the expectation taken *w.r.t.* the conditional empirical distributions induced by $\mathcal{D}_N^{\mathrm{U}}$, that is,

$$\hat{\mathbb{P}}_{X|S=s} = \frac{1}{|\{(X,S) \in \mathcal{D}_N^{\mathrm{U}} \,:\, S = s\}|} \sum_{\{(X,S)\in\mathcal{D}_N^{\mathrm{U}} \,:\, S=s\}} \delta_X \ ,$$

for all $s \in \{0, 1\}$, and by $\hat{\mathbb{E}}_S$ the expectation taken *w.r.t.* the empirical measure of S, that is, $\hat{\mathbb{P}}_S = \frac{1}{N}\sum_{(X,S)\in\mathcal{D}_N^{\mathrm{U}}} \delta_S$.

Remark 4. *In theory, the empirical distributions might not be well defined, since they are* only *valid if the unlabeled dataset $\mathcal{D}_N^{\mathrm{U}}$ is composed of features from* both *groups. We show how to bypass this problem theoretically in Section 2.2.8. Nevertheless, this remark has little to no impact in practice and in most situations these quantities are well defined.*

Based on the estimator $\hat{\eta}$ and the unlabeled sample $\mathcal{D}_N^{\mathrm{U}}$, let us introduce the following estimators for each $s \in \{0, 1\}$

$$\hat{\mathbb{P}}(Y = 1, S = s) := \hat{\mathbb{E}}_{X|S=s}[\hat{\eta}(X, s)]\hat{\mathbb{P}}_S(S = s) \ .$$

Using the above estimators, as suggested in Section 1.0.3, a straightforward procedure to mimic the optimal classifier g^* provided by Proposition 2 is to employ a *plug-in rule* $\hat{g}$, obtained by replacing all the unknown quantities by either their empirical versions or their estimates. Specifically, we define $\hat{g}$ at $(x, s) \in \mathbb{R}^d \times \{0, 1\}$ by

$$\hat{g}(x,1) = \mathbb{1}_{\left\{1\le\hat{\eta}(x,1)\left(2-\frac{\hat{\theta}}{\hat{\mathbb{P}}(Y=1,S=1)}\right)\right\}}, \quad \hat{g}(x,0) = \mathbb{1}_{\left\{1\le\hat{\eta}(x,0)\left(2+\frac{\hat{\theta}}{\hat{\mathbb{P}}(Y=1,S=0)}\right)\right\}} \ . \tag{2.11}$$

It remains now to define the value of $\hat{\theta}$. Clearly it is desirable to mimic the condition that is satisfied by θ^* in Proposition 2. To this end, we make use of the unlabeled dataset $\mathcal{D}_N^{\mathrm{U}}$ and of the estimator $\hat{\eta}$ previously built from the labeled dataset $\mathcal{D}_n^{\mathrm{L}}$. Consequently, we define a data-driven version of unfairness $\Delta(g, \mathbb{P})$, which allows to construct an approximation $\hat{\theta}$ of the true value θ^*.

Definition 11 (Empirical unfairness)**.** *For any classifier g, an estimator $\hat{\eta}$ based on $\mathcal{D}_n^{\mathrm{L}}$, and unlabeled sample $\mathcal{D}_N^{\mathrm{U}}$ the empirical unfairness is defined as*

$$\hat{\Delta}(g, \mathbb{P}) = \left| \frac{\hat{\mathbb{E}}_{X|S=1}\hat{\eta}(X,1)g(X,1)}{\hat{\mathbb{E}}_{X|S=1}\hat{\eta}(X,1)} - \frac{\hat{\mathbb{E}}_{X|S=0}\hat{\eta}(X,0)g(X,0)}{\hat{\mathbb{E}}_{X|S=0}\hat{\eta}(X,0)} \right| \ .$$

Notice that the empirical unfairness $\hat{\Delta}(g, \mathbb{P})$ is data-driven, that is, it does not involve unknown quantities. One might wonder why it is an empirical version of the quantity $\Delta(g, \mathbb{P})$ in Definition 10 and what is the reason to introduce it. The definition reveals itself when we rewrite the population of unfairness $\Delta(g, \mathbb{P})$ using[4] the identity

$$\mathbb{P}\left(g(X,S) = 1 \,|\, S = s, Y = 1\right) = \frac{\mathbb{P}\left(g(X,S) = 1, Y = 1 \,|\, S = s\right)}{\mathbb{P}\left(Y = 1 \,|\, S = s\right)} = \frac{\mathbb{E}_{X|S=s}[\eta(X,s)g(X,s)]}{\mathbb{E}_{X|S=s}[\eta(X,s)]} \ .$$

[4]Note additionally that for all $s \in \{0, 1\}$ we can write $\mathbb{1}_{\{Y=1,g(X,s)=1\}} \equiv Yg(X, s)$, since both Y and g are binary.

Using the above expression we can rewrite

$$\Delta(g, \mathbb{P}) = \left| \frac{\mathbb{E}_{X|S=1}[\eta(X,1)g(X,1)]}{\mathbb{E}_{X|S=1}[\eta(X,1)]} - \frac{\mathbb{E}_{X|S=0}[\eta(X,0)g(X,0)]}{\mathbb{E}_{X|S=0}[\eta(X,0)]} \right| \ .$$

Hence, the passage from the population unfairness to its empirical version in Definition 11 formally reduces to substituting "hats" to all the unknown quantities.

Using Definition 11, a logical estimator $\hat{\theta}$ of θ^* can be obtained as

$$\hat{\theta} \in \underset{\theta \in [-2,2]}{\arg\min}\, \hat{\Delta}(\hat{g}_\theta, \mathbb{P}) \ ,$$

where, for all $\theta \in [-2, 2]$, $\hat{g}_\theta$ is defined at $(x, s) \in \mathbb{R}^d \times \{0, 1\}$ as

$$\hat{g}_\theta(x,1) = \mathbb{1}_{\left\{1 \le \hat{\eta}(x,1)\left(2 - \frac{\theta}{\hat{\mathbb{P}}(Y=1,S=1)}\right)\right\}}, \quad \hat{g}_\theta(x,0) = \mathbb{1}_{\left\{1 \le \hat{\eta}(x,0)\left(2 + \frac{\theta}{\hat{\mathbb{P}}(Y=1,S=0)}\right)\right\}} \ . \tag{2.12}$$

The proposed estimator $\hat{g}$ is then given by $\hat{g} \equiv \hat{g}_{\hat{\theta}}$. Notice that since the quantity $\hat{\Delta}(\hat{g}_\theta, \mathbb{P})$ is empirical, then there might be no θ which delivers zero for the empirical unfairness. This is exactly the reason we perform a minimization of this quantity.

Remark 5. *Even though we believe that the introduction of the unlabeled sample is one of the strong points of this manuscript, this sample may not be available on some benchmark datasets. In this case, we can simply randomly split the data into two parts disregarding labels in one of them, or alternatively we can use the same sample twice. The second path is not directly justified by our theoretical results, yet, let us suggest the following intuitive explanation for this approach. On the first and the second steps, our procedure approximates two* independent *parts of the distribution $\mathbb{P}$ of the random tuple (X, S, Y). Indeed, following the factorization $\mathbb{P} = \mathbb{P}_{Y|X,S} \otimes \mathbb{P}_{(X,S)}$, the first step of our procedure approximates $\mathbb{P}_{Y|X,S}$, whereas the second step is aimed at $\mathbb{P}_{(X,S)}$ which is independent from $\mathbb{P}_{Y|X,S}$. In our experiments, reported in Section 2.2.5, we exploited the same set of data for both $\mathcal{D}_n$ and $\mathcal{D}_N$, since no unlabelled sample were available and splitting the dataset would have reduced the quality of the trained model because the datasets have a small sample size.*

Besides, notice that this procedure shares the same spirit with the procedure introduced in Section 2.1 in the context of binary classification with the F-score. Indeed, the first steps of both procedure is identical, whereas the second step relies on similar ideas of the threshold estimation.

2.2.4 Consistency

In this part we establish that the proposed procedure is consistent in terms of both risk and constraint. To present our theoretical results we impose two assumptions on the estimator $\hat{\eta}$ and demonstrate how to satisfy them in practice.

Assumption 10. *The estimator $\hat{\eta}$ which is constructed from $\mathcal{D}_n^{\mathrm{L}}$ satisfies for all $s \in \{0, 1\}$*

(i) $\mathbb{E}_{\mathcal{D}_n^{\mathrm{L}}} \mathbb{E}_{X|S=s} \left|\eta(X, S) - \hat{\eta}(X, S)\right| \to 0$ *as* $n \to \infty$*;*

(ii) There exists a sequence $c_{n,N} > 0$ satisfying $\frac{1}{c_{n,N}\sqrt{N}} = o_{n,N}(1)$ and $c_{n,N} = o_{n,N}(1)$ such that $\mathbb{E}_{X|S=s}[\hat{\eta}(X, S)] \ge c_{n,N}$ almost surely.

(ii) $\hat{\eta}(x) \geq 0$ *almost surely for almost all* $x \in \mathbb{R}^d$

Remark 6. *There are two parts in Assumption 10, the first one requires a consistent estimator in* ℓ_1 *norm. This first assumption is rather weak, since there are many different available consistent estimators for the regression function in the literature, including the Maximum likelihood estimator [Yan et al., 2018] for Gaussian Generative Model, local polynomial estimator [Audibert and Tsybakov, 2007] for* β*-Hölder smooth regression function* $\eta(\cdot, s)$*, regularized logistic regression [van de Geer, 2008] for Generalized Linear Model,* k*-Nearest Neighbors estimator [Devroye, 1978] for Lipschitz regression function* $\eta(\cdot, s)$*, and random forest type estimators in various settings [Breiman, 2004, Genuer, 2012, Arlot and Genuer, 2014, Scornet et al., 2015].*
The second part of Assumption 10 means that $\mathbb{E}_{X|S=s}[\hat{\eta}(X, s)]$ *is lower bounded by a positive term vanishing as* N, n *grow to infinity. This condition can be introduced artificially to any predefined estimator. Indeed, assume that we have a consistent estimator* $\tilde{\eta}$ *and let* $\hat{\eta}(x, s) = \max\{\tilde{\eta}(x, s), c_{n,N}\}$*, then the second item of the assumption is satisfied in even a stronger form. Moreover, this estimator* $\hat{\eta}$ *remains consistent, since using the triangle inequality and the fact that* $|\hat{\eta}(x, s) - \tilde{\eta}(x, s)| \leq c_{n,N}$ *for all* $x \in \mathbb{R}^d$*, we have*

$$\mathbb{E}_{\mathcal{D}_n^{\mathrm{L}}} \mathbb{E}_{X|S=s} \left|\eta(X, s) - \hat{\eta}(X, s)\right| \leq \mathbb{E}_{\mathcal{D}_n^{\mathrm{L}}} \mathbb{E}_{X|S=s} \left|\eta(X, s) - \tilde{\eta}(X, s)\right| + c_{n,N} \to 0 \ .$$

Additionally, we impose one more condition on the estimator $\hat{\eta}$ which will also be used in the context of confidence set classification described in Chapter 3.

Assumption 11. *The estimator* $\hat{\eta}$ *is such that for all* $s \in \{0, 1\}$ *the mapping*

$$t \mapsto \mathbb{P}\left(\hat{\eta}(X, s) \leq t \,|\, S = s\right) \ ,$$

is continuous on $(0, 1)$ *almost surely.*

In the fairness setup this assumption allows us to show that the value of $\hat{\Delta}(\hat{g}, \mathbb{P})$ cannot be large, that is, the empirical unfairness of the proposed procedure is small or zero. As we shall see, a control on the empirical unfairness $\hat{\Delta}(\hat{g}, \mathbb{P})$ in Definition 11 is crucial in proving that the proposed procedure $\hat{g}$ achieves both asymptotic fairness and risk consistency.

Remark 7. *Assumption 11 is equivalent to say that there are no atoms in the estimated regression function. It can be fulfilled by a simple modification of any preliminary estimator, by adding a* small *deterministic "noise", the amplitude of which must be decreasing with* n, N *in order to preserve statistical consistency.*

Our remarks suggest that both Assumptions 10 and 11 can be easily satisfied in a variety of practical settings and the most demanding part of these assumptions is the consistency of $\hat{\eta}$.

The next result establishes the statistical consistency of the proposed algorithm.

Theorem 9 (Asymptotic properties)**.** *Under Assumptions 9, 10, and 11 the proposed algorithm satisfies*

$$\lim_{n,N \to \infty} \mathbb{E}_{(\mathcal{D}_n^{\mathrm{L}}, \mathcal{D}_N^{\mathrm{U}})}[\Delta(\hat{g}, \mathbb{P})] = 0 \quad \textit{and} \quad \lim_{n,N \to \infty} \mathbb{E}_{(\mathcal{D}_n^{\mathrm{L}}, \mathcal{D}_N^{\mathrm{U}})}[\mathcal{R}(\hat{g})] \leq \mathcal{R}(g^*) \ .$$

Proof sketch. In order to establish statistical consistency of the proposed procedure, we first introduce an intermediate pseudo-estimator $\tilde{g}$ as follows

$$\tilde{g}(x,1) = \mathbb{1}_{\left\{1 \leq \hat{\eta}(x,1)\left(2 - \frac{\tilde{\theta}}{\mathbb{E}_{X|S=1}[\hat{\eta}(X,1)]\mathbb{P}(S=1)}\right)\right\}}, \quad \tilde{g}(x,0) = \mathbb{1}_{\left\{1 \leq \hat{\eta}(x,0)\left(2 + \frac{\tilde{\theta}}{\mathbb{E}_{X|S=0}[\hat{\eta}(X,0)]\mathbb{P}(S=0)}\right)\right\}}, \tag{2.13}$$

where $\tilde{\theta}$ is chosen such that

$$\frac{\mathbb{E}_{X|S=1}\left[\hat{\eta}(X,1)\tilde{g}(X,1)\right]}{\mathbb{E}_{X|S=1}[\hat{\eta}(X,1)]} = \frac{\mathbb{E}_{X|S=0}\left[\hat{\eta}(X,0)\tilde{g}(X,0)\right]}{\mathbb{E}_{X|S=0}[\hat{\eta}(X,0)]} \ . \tag{2.14}$$

Note that by Assumption 11 such a value $\tilde{\theta}$ always exists. Intuitively, the classifier $\tilde{g}$ "*knows*" the marginal distribution of (X,S), that is, it knows both $\mathbb{P}_{X|s}$ and $\mathbb{P}_S$. It is seen as an idealized version of $\hat{g}$, where the uncertainty is only induced by the lack of knowledge of the regression function η.

We express the excess risk as a sum of two terms, $\mathbb{E}_{\mathcal{D}_n^{\mathrm{L}}}[\mathcal{R}(\tilde{g})] - \mathcal{R}(g^*) + \mathbb{E}_{(\mathcal{D}_n^{\mathrm{L}},\mathcal{D}_N^{\mathrm{U}})}[\mathcal{R}(\hat{g}) - \mathcal{R}(\tilde{g})]$. We show that the first can be bounded by the ℓ_1 distance between $\hat{\eta}$ and η, and thanks to the consistency of $\hat{\eta}$ it converges to zero. The handling of the second term is more involved, but we are able to show that it reduces to a study of suprema of empirical processes conditionally on the labeled sample $\mathcal{D}_n^{\mathrm{L}}$. Same strategy will be used in Chapter 3 in the context of confidence set classification.

To demonstrate that the proposed algorithm is asymptotically fair, we first show that

$$\mathbb{E}_{(\mathcal{D}_n^{\mathrm{L}},\mathcal{D}_N^{\mathrm{U}})}[\Delta(\hat{g},\mathbb{P})] \leq \mathbb{E}_{(\mathcal{D}_n^{\mathrm{L}},\mathcal{D}_N^{\mathrm{U}})}[\hat{\Delta}(\hat{g},\mathbb{P})] + o_{n,N}(1) \ .$$

At last, the continuity Assumption 11 and the theory of empirical processes allow us to show that the term $\mathbb{E}_{(\mathcal{D}_n^{\mathrm{L}},\mathcal{D}_N^{\mathrm{U}})}[\hat{\Delta}(\hat{g},\mathbb{P})]$ converges to zero when N grows to infinity. See Section 2.2.8 for complete proof of this result. □

Remark 8. *Let us mention that it is possible to present our result in a finite sample regime, since our proof of consistency is based on non-asymptotic theory of empirical processes. However, the actual rate of convergence depends on the rate of ℓ_1-norm estimation of the regression function η, which can vary significantly from one setup to another. That is why we decided to present our result in the asymptotic sense.*

2.2.5 Experimental results

Here we present numerical experiments with the proposed method.

We follow the protocol outlined in [Donini et al., 2018]. We consider the following datasets: Arrhythmia, COMPAS, Adult, German, and Drug[5] and compare the following algorithms: Linear Support Vector Machines (Lin.SVM), Support Vector Machines with the Gaussian kernel (SVM), Linear Logistic Regression (Lin.LR), Logistic Regression with the Gaussian kernel (LR), Hardt method [Hardt et al., 2016] to all approaches (Hardt), Zafar method [Zafar et al., 2017] implemented with the code provided by the authors for the linear case[6], the Linear (Lin.Donini) and the Non Linear methods (Donini) proposed in [Donini et al., 2018]

[5]For more information about these datasets please refer to [Donini et al., 2018].

[6]Python code for [Zafar et al., 2017]: `https://github.com/mbilalzafar/fair-classification`

Method	Arrhythmia		COMPAS		Adult		German		Drug	
	ACC	DEO	ACC	DEO	ACC	DEO	ACC	DEO	ACC	DEO
Lin.SVM	0.78±0.07	0.13±0.04	0.75±0.01	0.15±0.02	0.80	0.13	0.69±0.04	0.11±0.10	0.81±0.02	0.41±0.06
Lin.LR	0.79±0.06	0.13±0.05	0.76±0.02	0.16±0.02	0.81	0.12	0.67±0.05	0.12±0.11	0.80±0.01	0.42±0.05
Lin.SVM+Hardt	0.74±0.06	0.07±0.04	0.67±0.03	0.21±0.09	0.80	0.10	0.61±0.15	0.15±0.13	0.77±0.02	0.22±0.09
Lin.LR+Hardt	0.75±0.04	0.08±0.05	0.67±0.02	0.18±0.07	0.81	0.09	0.62±0.05	0.13±0.09	0.76±0.01	0.18±0.04
Zafar	0.71±0.03	0.03±0.02	0.69±0.02	0.10±0.06	0.78	0.05	0.62±0.09	0.13±0.11	0.69±0.03	0.02±0.07
Lin.Donini	0.79±0.07	0.04±0.03	0.76±0.01	0.04±0.03	0.77	0.01	0.69±0.04	0.05±0.03	0.79±0.02	0.05±0.03
Lin.SVM+Ours	0.75±0.08	0.04±0.04	0.73±0.01	0.05±0.02	0.79	0.03	0.68±0.04	0.04±0.03	0.78±0.02	0.01±0.02
Lin.LR+Ours	0.75±0.06	0.04±0.05	0.74±0.02	0.06±0.02	0.80	0.03	0.67±0.05	0.04±0.03	0.77±0.03	0.02±0.02
SVM	0.78±0.06	0.13±0.04	0.73±0.01	0.14±0.02	0.82	0.14	0.74±0.03	0.10±0.06	0.81±0.04	0.38±0.03
LR	0.79±0.05	0.12±0.04	0.74±0.01	0.14±0.02	0.81	0.15	0.75±0.03	0.11±0.06	0.82±0.01	0.37±0.03
RF	0.83±0.03	0.09±0.02	0.77±0.02	0.11±0.02	0.86	0.12	0.78±0.02	0.09±0.04	0.86±0.01	0.29±0.02
SVM+Hardt	0.74±0.06	0.07±0.04	0.71±0.02	0.08±0.02	0.82	0.11	0.71±0.03	0.11±0.18	0.75±0.11	0.14±0.08
LR+Hardt	0.73±0.05	0.10±0.04	0.70±0.02	0.09±0.02	0.80	0.12	0.72±0.04	0.09±0.06	0.77±0.03	0.11±0.04
RF+Hardt	0.79±0.03	0.07±0.01	0.76±0.01	0.07±0.02	0.83	0.05	0.76±0.02	0.06±0.04	0.82±0.01	0.09±0.02
Donini	0.79±0.09	0.03±0.02	0.73±0.01	0.05±0.03	0.81	0.01	0.73±0.04	0.05±0.03	0.80±0.03	0.07±0.05
SVM+Ours	0.77±0.07	0.04±0.02	0.72±0.02	0.06±0.02	0.80	0.02	0.73±0.03	0.04±0.06	0.79±0.02	0.05±0.01
LR+Ours	0.77±0.06	0.04±0.02	0.73±0.01	0.06±0.02	0.80	0.02	0.73±0.02	0.04±0.06	0.80±0.01	0.05±0.02
RF+Ours	0.81±0.04	0.03±0.01	0.76±0.02	0.04±0.02	0.85	0.03	0.77±0.02	0.02±0.02	0.83±0.01	0.04±0.02

Table 2.1: Results (average ± standard deviation, when a fixed test set is not provided) for all the datasets, concerning ACC and DEO.

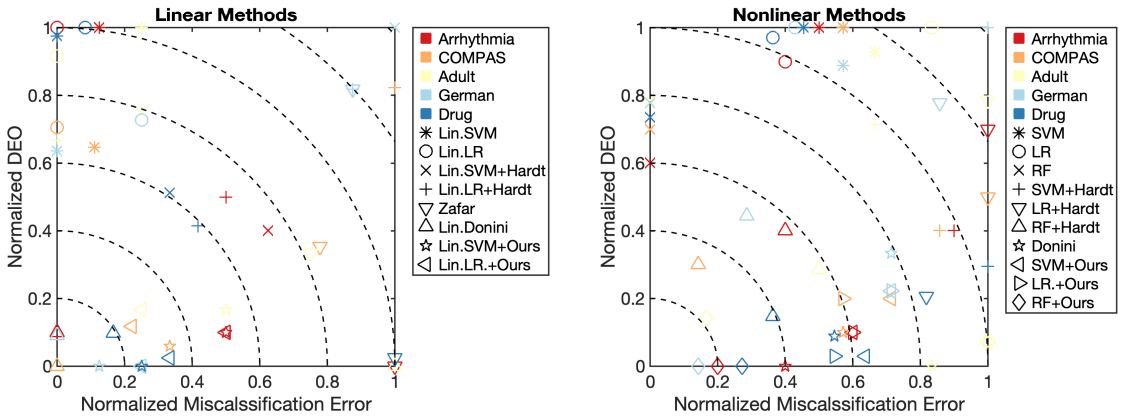

Figure 2.1: Results of Table 2.1 of linear (left) and nonlinear (right) methods when the error and the DEO are normalized in $[0, 1]$ column-wise. Different colors and symbols refer to different datasets and method respectively. The closer a point is to the origin, the better the result is.

and freely available[7], and also Random Forests (RF). Then, since Lin.SVM, SVM, Lin.LR, LR, and RF have also the possibility to output a probability together with the classification, we applied our method in all these cases.

In all experiments, we collect statistics concerning the classification accuracy (ACC), namely probability to correctly classify a sample, and the Difference of Equal Opportunity (DEO) in Definition 9. For Arrhythmia, COMPAS, German and Drug datasets we split the data in two parts (70% train and 30% test), this procedure is repeated 30 times, and we reported the average performance on the test set alongside its standard deviation. For the Adult dataset, we used the provided split of train and test sets. Unless otherwise stated, we employ two steps in the 10-fold CV procedure proposed in [Donini et al., 2018] to select the best hyperparameters with the training set[8]. In the first step, the value of the hyper-

[7]Python code for [Donini et al., 2018]: https://github.com/jmikko/fair_ERM

[8]The regularization parameter (for all method) and the RBF kernel with 30 values, equally spaced in loga-

parameters with the highest accuracy is identified. In the second step, we shortlist all the hyperparameters with accuracy close to the best one (in our case, above 90% of the best accuracy). Finally, from this list, we select the hyperparameters with the lowest DEO.

We also present in Figure 2.1 the results of Table 2.1 for linear (left) and nonlinear (right) methods, when the error (one minus ACC) and the DEO are normalized in $[0, 1]$ column-wise. In the figure, different colors and symbols refer to different datasets and methods, respectively. The closer a point is to the origin, the better the result is.

From Table 2.1 and Figure 2.1 it is possible to observe that the proposed method outperforms all methods except the one of [Donini et al., 2018] for which we obtain comparable performance. Nevertheless, note that our method is more general than the one of [Donini et al., 2018], since it can be applied to any algorithms which return a probability estimator (better if consistent since this will allow us to have a full consistent approach also from the fairness point of view). In fact, on these datasets, RF, which cannot be made trivially fair with the approach proposed in [Donini et al., 2018], outperforms all the available methods.

Note that the results reported in Table 2.1 differ from the one reported in [Donini et al., 2018] since the proposed method requires the knowledge of the sensitive variable at classification time, so Table 2.1 reports just this case. That is, the functional form of the model explicitly depends on the sensitive variable $s \in \{0, 1\}$. Many authors, point out that this may not be permitted in several practical scenarios (see e.g. [Dwork et al., 2018, Roemer and Trannoy, 2015] and reference therein). Yet, removing the sensitive variable from the functional form of the model does not ensure that the sensitive variable is not considered by the model itself. We refer to [Oneto et al., 2019a] for the in-depth discussion on this issue. Further, the method in [Hardt et al., 2016] explicitly requires the knowledge of the sensitive variable for their thresholding procedure. In Section 2.2.6 we show how to modify our method in order to derive a fair optimal classifier without the sensitive variable s in the functional form of the model. Moreover, we propose a modification of our approach which does not use s at decision time and perform additional numerical comparison in this context. We arrive to similar conclusions about the performance of our method as in this part. Yet, the consistency results are not available for this methods and are left for future investigation.

2.2.6 Optimal classifier independent of sensitive feature

In this part we provide guidelines to construct a *plug-in* algorithm which can use the sensitive feature only at training time but cannot use it for future decision making. It is clear that the first step would be to derive fair optimal classifier $g^* : \mathbb{R}^d \to \{0, 1\}$ which is defined as

$$g^* \in \arg\min \left\{ \mathcal{R}(g) \,:\, \mathbb{P}\left(g(X) = 1 \,|\, S = 1, Y = 1\right) = \mathbb{P}\left(g(X) = 1 \,|\, S = 0, Y = 1\right) \right\} \quad ,$$

with $\mathcal{R}(g) := \mathbb{P}(Y \neq g(X))$. Next result establishes this expression.

Proposition 3 (Optimal rule)**.** *Under Assumption 9 an optimal classifier g^* can be obtained for all $x \in \mathbb{R}^d$ as*

$$g^*(x) = \mathbb{1}_{\left\{1 \leq 2\eta(x) + \theta^* \left(\frac{\eta(x,0)}{\mathbb{E}_X[\eta(X,0)]} - \frac{\eta(x,1)}{\mathbb{E}_X[\eta(X,1)]} \right) \right\}} \quad ,$$

rithmic scale between 10^{-4} and 10^4. For RF the number of trees has been set to 1000 and the size of the subset of features optimized at each node has been search in $\{d, \lceil d^{15/16} \rceil, \lceil d^{7/8} \rceil, \lceil d^{3/4} \rceil, \lceil d^{1/2} \rceil, \lceil d^{1/4} \rceil, \lceil d^{1/8} \rceil, \lceil d^{1/16} \rceil, 1\}$ where d is the number of features in the dataset.

Method	Arrhythmia		COMPAS		Adult		German		Drug	
	ACC	DEO	ACC	DEO	ACC	DEO	ACC	DEO	ACC	DEO
Lin.SVM	0.71±0.05	0.10±0.03	0.72±0.01	0.12±0.02	0.78	0.09	0.69±0.04	0.11±0.10	0.79±0.02	0.25±0.04
Lin.LR	0.71±0.04	0.11±0.04	0.73±0.02	0.10±0.03	0.80	0.08	0.68±0.05	0.12±0.09	0.80±0.03	0.23±0.03
Lin.SVM+Hardt	-	-	-	-	-	-	-	-	-	-
Lin.LR+Hardt	-	-	-	-	-	-	-	-	-	-
Zafar	0.67±0.03	0.05±0.02	0.69±0.01	0.10±0.08	0.76	0.05	0.62±0.09	0.13±0.10	0.66±0.03	0.06±0.06
Lin.Donini	0.75±0.05	0.05±0.02	0.73±0.01	0.07±0.02	0.75	0.01	0.69±0.04	0.06±0.03	0.79±0.02	0.10±0.06
Lin.SVM+Ours	0.72±0.05	0.03±0.01	0.72±0.01	0.06±0.02	0.74	0.02	0.68±0.04	0.06±0.04	0.78±0.02	0.12±0.02
Lin.LR+Ours	0.71±0.04	0.04±0.02	0.71±0.02	0.06±0.02	0.76	0.02	0.67±0.05	0.05±0.03	0.79±0.03	0.10±0.01
SVM	0.71±0.05	0.10±0.03	0.73±0.01	0.11±0.02	0.79	0.08	0.74±0.03	0.10±0.06	0.81±0.02	0.22±0.03
LR	0.70±0.06	0.10±0.03	0.74±0.01	0.10±0.02	0.78	0.10	0.75±0.03	0.09±0.05	0.81±0.03	0.21±0.02
RF	0.81±0.02	0.08±0.02	0.76±0.03	0.10±0.02	0.84	0.11	0.77±0.03	0.07±0.04	0.85±0.02	0.19±0.02
SVM+Hardt	-	-	-	-	-	-	-	-	-	-
LR+Hardt	-	-	-	-	-	-	-	-	-	-
RF+Hardt	-	-	-	-	-	-	-	-	-	-
Donini	0.75±0.05	0.05±0.02	0.72±0.01	0.08±0.02	0.77	0.01	0.73±0.04	0.05±0.03	0.79±0.03	0.10±0.05
SVM+Ours	0.71±0.02	0.06±0.02	0.72±0.01	0.05±0.02	0.78	0.02	0.73±0.01	0.06±0.03	0.78±0.02	0.11±0.02
LR+Ours	0.70±0.04	0.06±0.03	0.72±0.01	0.06±0.02	0.77	0.02	0.73±0.02	0.06±0.02	0.77±0.02	0.11±0.02
RF+Ours	0.80±0.03	0.02±0.01	0.76±0.02	0.04±0.02	0.84	0.02	0.76±0.03	0.04±0.02	0.83±0.01	0.06±0.02

Table 2.2: Results (average ± standard deviation, when a fixed test set is not provided) for all the datasets, concerning ACC and DEO. In this case the sensitive feature is not in the functional form of the model.

where θ^* is such that the equality

$$\frac{\mathbb{E}_X\left[\eta(X,1)g^*(X)\right]}{\mathbb{E}_X[\eta(X,1)]} = \frac{\mathbb{E}_X\left[\eta(X,0)g^*(X)\right]}{\mathbb{E}_X[\eta(X,0)]} \ ,$$

is satisfied and $\eta(\cdot) := \mathbb{P}\left(Y=1 \,|\, X=\cdot\right)$.

Observe that to efficiently compute the optimal classifier in this case we need to have access to $\eta(x), \eta(x,s)$ and marginal distribution $\mathbb{P}_X$.

This observation motivates us to propose a plug-in algorithm based on two datasets $\mathcal{D}_n^{\mathrm{L}} = \{(X_i, S_i, Y_i)\}_{i=1}^n$ and $\mathcal{D}_N^{\mathrm{U}} = \{X_i\}_{i=1}^N$. The labeled data $\mathcal{D}_n^{\mathrm{L}}$ allow to estimate $\eta(x), \eta(x,s)$ and the unlabeled data $\mathcal{D}_N^{\mathrm{U}}$ allow to estimate the marginal distribution $\mathbb{P}_X$. Interestingly, we do not need to observe sensitive features in the unlabeled dataset $\mathcal{D}_N^{\mathrm{U}}$.

Formally, our procedure $\hat{g}$ in this case can be defined for all $x \in \mathbb{R}^d$ as

$$\hat{g}(x) = \mathbb{1}_{\left\{1 \leq 2\hat{\eta}(x) + \hat{\theta}\left(\frac{\hat{\eta}(x,0)}{\hat{\mathbb{E}}_X[\hat{\eta}(X,0)]} - \frac{\hat{\eta}(x,1)}{\hat{\mathbb{E}}_X[\eta(X,1)]}\right)\right\}} \ ,$$

where $\hat{\eta}(x), \hat{\eta}(x,s)$ for all $s \in \{0,1\}$ are the estimates of regression functions constructed from $\mathcal{D}_n^{\mathrm{L}}$, and $\hat{\mathbb{E}}_X$ is the empirical expectation based on $\mathcal{D}_N^{\mathrm{U}}$.

Finally, similarly to the previous case the threshold $\hat{\theta}$ is defined as

$$\hat{\theta} \in \arg\min_{\theta} \left| \frac{\hat{\mathbb{E}}_X\left[\hat{\eta}(X,1)\hat{g}_\theta(X)\right]}{\hat{\mathbb{E}}_X[\hat{\eta}(X,1)]} - \frac{\hat{\mathbb{E}}_X\left[\hat{\eta}(X,0)\hat{g}_\theta(X)\right]}{\hat{\mathbb{E}}_X[\hat{\eta}(X,0)]} \right| \ ,$$

with $\hat{g}_\theta$ defined for all $x \in \mathbb{R}^d$ as

$$\hat{g}_\theta(x) = \mathbb{1}_{\left\{1 \leq 2\hat{\eta}(x) + \theta\left(\frac{\hat{\eta}(x,0)}{\hat{\mathbb{E}}_X[\hat{\eta}(X,0)]} - \frac{\hat{\eta}(x,1)}{\hat{\mathbb{E}}_X[\eta(X,1)]}\right)\right\}} \ .$$

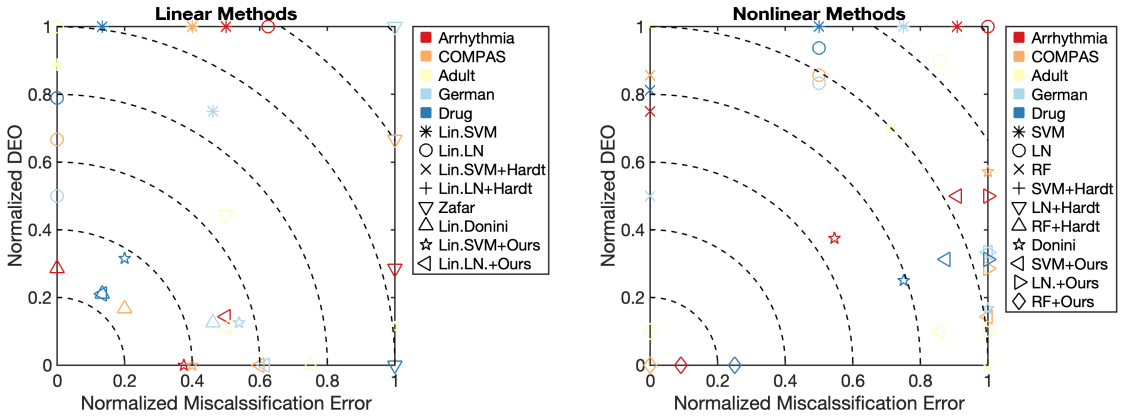

Figure 2.2: Results of Table 2.2 of linear (left) and nonlinear (right) methods when the error and the DEO are normalized in $[0, 1]$ column-wise. Different colors and symbols refer to different datasets and method respectively. The closer a point is to the origin, the better the result is. In this case the sensitive feature is not in the functional form of the model.

Experiments without the sensitive feature

Here we report the equivalent results to those in Table 2.1 and Figure 2.1 into Table 2.2 and Figure 2.2 when the sensitive feature is not in the functional form of the model. Note that the method of Hardt [Hardt et al., 2016] is not able to deal with this setting then there are no results for this case.

From Table 2.2 and Figure 2.2 we can observe analogous results to those in Section 2.2.5. Nevertheless, note that, without the sensitive feature in the functional form of the models, the results are generally less accurate and more fair w.r.t. to the case that the sensitive feature in the functional form of the models. This results is similar to the one reported in [Donini et al., 2018].

2.2.7 Conclusion

Using the notion of equal opportunity we have derived a form of the fair optimal classifier based on group-dependent threshold. Relying on this result we have proposed a semi-supervised plug-in method which enjoys strong theoretical guarantees under mild assumptions. Importantly, our algorithm can be implemented on top of any base classifier which has conditional probabilities as outputs. We have conducted an extensive numerical evaluation comparing our procedure against the state-of-the-art approaches and have demonstrated that our procedure performs well in practice. In future works we would like to extend our analysis to other fairness measures as well as provide consistency results for the algorithm which does not use the sensitive feature at the decision time. Finally, we note that our consistency result is constructive and could be used to derive non-asymptotic rates of convergence for the proposed method, relying upon available rates for the regression function estimator.

2.2.8 Proofs

In this part of Section 2.2, we provide proofs of the results stated in the main body of Section 2.2 and collect some auxiliary results. Specifically, this part contains the proof of Proposition 2, results (with proofs) needed for the proof of Theorem 9.

Optimal classifier

Proof of Proposition 2. Let us study the following minimization problem

$$(*) := \min_{g \in \mathcal{G}} \{\mathcal{R}(g) : \mathbb{P}\left(g(X,S)=1 \,|\, Y=1, S=1\right) = \mathbb{P}\left(g(X,S)=1 \,|\, Y=1, S=0\right)\} \quad .$$

Using the weak duality we can write

$$\begin{aligned}(*) &= \min_{g \in \mathcal{G}} \max_{\lambda \in \mathbb{R}} \{\mathcal{R}(g) + \lambda \left(\mathbb{P}\left(g(X,S)=1 \,|\, Y=1, S=1\right) - \mathbb{P}\left(g(X,S)=1 \,|\, Y=1, S=0\right)\right)\} \\ &\geq \max_{\lambda \in \mathbb{R}} \min_{g \in \mathcal{G}} \{\mathcal{R}(g) + \lambda \left(\mathbb{P}\left(g(X,S)=1 \,|\, Y=1, S=1\right) - \mathbb{P}\left(g(X,S)=1 \,|\, Y=1, S=0\right)\right)\} =: (**) \quad .\end{aligned}$$

We first study the objective function of the max min problem $(**)$, which is equal to

$$\mathbb{P}(g(X,S) \neq Y) + \lambda \left(\mathbb{P}\left(g(X,S)=1 \,|\, Y=1, S=1\right) - \mathbb{P}\left(g(X,S)=1 \,|\, Y=1, S=0\right)\right) \quad .$$

The first step of the proof is to simplify the expression above to linear functional of the classifier g. Notice that we can write for the first term

$$\begin{aligned}\mathbb{P}(g(X,S) \neq Y) &= \mathbb{P}(g(X,S)=0, Y=1) + \mathbb{P}(g(X,S)=1, Y=0) \\ &= \mathbb{P}(g(X,S)=1) + \mathbb{P}(Y=1) - \mathbb{P}(g(X,S)=1, Y=1) - \mathbb{P}(g(X,S)=1, Y=1) \\ &= \mathbb{P}(g(X,S)=1) + \mathbb{P}(Y=1) - 2\mathbb{P}(g(X,S)=1, Y=1) \\ &= \mathbb{P}(Y=1) + \mathbb{E}[g(X,S)] - 2\mathbb{E}\left[\mathbb{1}_{\{g(X,S)=1, Y=1\}} \,|\, S=1\right] \mathbb{P}(S=1) \\ &\quad - 2\mathbb{E}\left[\mathbb{1}_{\{g(X,S)=1, Y=1\}} \,|\, S=0\right] \mathbb{P}(S=0) \\ &= \mathbb{P}(Y=1) + \mathbb{E}[g(X,S)] - 2\mathbb{E}_{X|S=1}[g(X,1)\eta(X,1)]\mathbb{P}(S=1) \\ &\quad - 2\mathbb{E}_{X|S=0}[g(X,0)\eta(X,0)]\mathbb{P}(S=0) \\ &= \mathbb{P}(Y=1) - \mathbb{E}_{X|S=1}[g(X,1)(2\eta(X,1)-1)]\mathbb{P}(S=1) \\ &\quad - \mathbb{E}_{X|S=0}[g(X,0)(2\eta(X,0)-1)]\mathbb{P}(S=0) \quad .\end{aligned}$$

Moreover, we can write for the rest

$$\begin{aligned}\mathbb{P}\left(g(X,S)=1 \,|\, Y=1, S=1\right) &= \frac{\mathbb{P}\left(g(X,S)=1, Y=1 \,|\, S=1\right)}{\mathbb{P}\left(Y=1 \,|\, S=1\right)} = \frac{\mathbb{E}_{X|S=1}[g(X,1)\eta(X,1)]}{\mathbb{P}\left(Y=1 \,|\, S=1\right)} \quad , \\ \mathbb{P}\left(g(X,S)=1 \,|\, Y=1, S=0\right) &= \frac{\mathbb{P}\left(g(X,S)=1, Y=1 \,|\, S=0\right)}{\mathbb{P}\left(Y=1 \,|\, S=0\right)} = \frac{\mathbb{E}_{X|S=0}[g(X,0)\eta(X,0)]}{\mathbb{P}\left(Y=1 \,|\, S=0\right)} \quad .\end{aligned}$$

Using these, the objective of $(**)$ can be simplified as

$$\begin{aligned}\mathbb{P}(Y=1) &+ \mathbb{E}_{X|S=1}\left[g(X,1)\left(\eta(X,1)\left(\frac{\lambda}{\mathbb{P}\left(Y=1 \,|\, S=1\right)} - 2\mathbb{P}(S=1)\right) + \mathbb{P}(S=1)\right)\right] \\ &+ \mathbb{E}_{X|S=0}\left[g(X,0)\left(\eta(X,0)\left(-\frac{\lambda}{\mathbb{P}\left(Y=1 \,|\, S=0\right)} - 2\mathbb{P}(S=0)\right) + \mathbb{P}(S=0)\right)\right] \quad .\end{aligned}$$

Clearly, for every $\lambda \in \mathbb{R}$ a minimizer g^*_λ of the problem $(**)$ can be written for all $x \in \mathbb{R}^d$ as

$$\begin{aligned}g^*_\lambda(x,1) &= \mathbb{1}_{\left\{\eta(X,1)\left(\frac{\lambda}{\mathbb{P}(Y=1 \,|\, S=1)} - 2\mathbb{P}(S=1)\right) + \mathbb{P}(S=1) \leq 0\right\}} = \mathbb{1}_{\left\{1 - \eta(X,1)\left(2 - \frac{\lambda}{\mathbb{P}(Y=1, S=1)}\right) \leq 0\right\}} \\ g^*_\lambda(x,0) &= \mathbb{1}_{\left\{\eta(X,0)\left(-\frac{\lambda}{\mathbb{P}(Y=1 \,|\, S=0)} - 2\mathbb{P}(S=0)\right) + \mathbb{P}(S=0) \leq 0\right\}} = \mathbb{1}_{\left\{1 - \eta(X,0)\left(2 + \frac{\lambda}{\mathbb{P}(Y=1, S=0)}\right) \leq 0\right\}} \quad .\end{aligned}$$

At this moment it is interesting to reflect on this result. Indeed, for $\lambda = 0$ we recover the classical optimal predictor in the context of binary classification. Substituting this classifier into the objective of $(**)$ we arrive at

$$(**) = \mathbb{P}(Y=1) - \min_{\lambda \in \mathbb{R}} \Bigg\{ \mathbb{E}_{X|S=1} \left(\eta(X,1) \left(2\mathbb{P}(S=1) - \frac{\lambda}{\mathbb{P}\left(Y=1 \,|\, S=1\right)} \right) - \mathbb{P}(S=1) \right)_+ \\ + \mathbb{E}_{X|S=0} \left(\eta(X,0) \left(2\mathbb{P}(S=0) + \frac{\lambda}{\mathbb{P}\left(Y=1 \,|\, S=0\right)} \right) - \mathbb{P}(S=0) \right)_+ \Bigg\} \ .$$

It is important to observe that the mappings

$$\lambda \mapsto \mathbb{E}_{X|S=1} \left(\eta(X,1) \left(2\mathbb{P}(S=1) - \frac{\lambda}{\mathbb{P}\left(Y=1 \,|\, S=1\right)} \right) - \mathbb{P}(S=1) \right)_+ \\ \lambda \mapsto \mathbb{E}_{X|S=0} \left(\eta(X,0) \left(2\mathbb{P}(S=0) + \frac{\lambda}{\mathbb{P}\left(Y=1 \,|\, S=0\right)} \right) - \mathbb{P}(S=0) \right)_+ \ ,$$

are convex, therefore we can write the first order optimality conditions as

$$0 \in \partial_\lambda \mathbb{E}_{X|S=1} \left(\eta(X,1) \left(2\mathbb{P}(S=1) - \frac{\lambda}{\mathbb{P}\left(Y=1 \,|\, S=1\right)} \right) - \mathbb{P}(S=1) \right)_+ \\ + \partial_\lambda \mathbb{E}_{X|S=0} \left(\eta(X,0) \left(2\mathbb{P}(S=0) + \frac{\lambda}{\mathbb{P}\left(Y=1 \,|\, S=0\right)} \right) - \mathbb{P}(S=0) \right)_+ \ .$$

Clearly, under Assumption 9 this subgradient is reduced to the gradient almost surely, thus we have the following condition on the optimal value of λ^*

$$\frac{\mathbb{E}_{X|S=1}\left[\eta(X,1) g^*_{\lambda^*}(X,1)\right]}{\mathbb{P}\left(Y=1 \,|\, S=1\right)} = \frac{\mathbb{E}_{X|S=0}\left[\eta(X,0) g^*_{\lambda^*}(X,0)\right]}{\mathbb{P}\left(Y=1 \,|\, S=0\right)} \ ,$$

and the pair $(\lambda^*, g^*_{\lambda^*})$ is a solution of the dual problem $(**)$. Notice that the previous condition can be written as

$$\mathbb{P}\left(g^*_{\lambda^*}(X,S)=1 \,|\, Y=1, S=1\right) = \mathbb{P}\left(g^*_{\lambda^*}(X,S)=1 \,|\, Y=1, S=0\right) \ .$$

This implies that the classifier $g^*_{\lambda^*}$ is fair, that is, it satisfies Definition 9. Finally, it remains to show that $g^*_{\lambda^*}$ is actually an optimal classifier, indeed, since $g^*_{\lambda^*}$ is fair we can write on the one hand

$$\mathcal{R}(g^*_{\lambda^*}) \geq \min_{g \in \mathcal{G}} \left\{ \mathcal{R}(g) \ : \ \mathbb{P}\left(g(X,S)=1 \,|\, Y=1, S=1\right) = \mathbb{P}\left(g(X,S)=1 \,|\, Y=1, S=0\right) \right\} = (*).$$

On the other hand the pair $(\lambda^*, g^*_{\lambda^*})$ is a solution of the dual problem $(**)$, thus we have

$$\begin{aligned}(*) &\geq \mathcal{R}(g^*_{\lambda^*}) + \lambda^* \left(\mathbb{P}\left(g^*_{\lambda^*}(X,S)=1 \,|\, Y=1, S=1\right) - \mathbb{P}\left(g^*_{\lambda^*}(X,S)=1 \,|\, Y=1, S=0\right) \right) \\ &= \mathcal{R}(g^*_{\lambda^*}) \ .\end{aligned}$$

It implies that the classifier $g^*_{\lambda^*}$ is optimal, hence $g^* \equiv g^*_{\lambda^*}$.

Finally, assume that $(2-\theta^*/\mathbb{P}(Y=1,S=1)) \leq 0$, then, clearly $(2+\theta^*/\mathbb{P}(Y=1,S=0)) > 0$, therefore, the condition on θ^* reads as

$$0 = \mathbb{E}_{X|S=0}\left[\eta(X,0)\mathbb{1}_{\left\{1\leq\eta(X,0)\left(2+\frac{\theta^*}{\mathbb{P}(Y=1,S=0)}\right)\right\}}\right] \geq \frac{\mathbb{P}\left(\eta(X,0) \geq \frac{1}{\left(2+\frac{\theta^*}{\mathbb{P}(Y=1,S=0)}\right)} \,|\, S=0\right)}{\left(2+\frac{\theta^*}{\mathbb{P}(Y=1,S=0)}\right)}$$
$$\geq \frac{\mathbb{P}\left(\eta(X,0) \geq 1/2 \,|\, S=0\right)}{\left(2+\frac{\theta^*}{\mathbb{P}(Y=1,S=0)}\right)} > 0 \ ,$$

where the last inequality is due to Assumption 9. We arrive to contradiction, therefore $(2-\theta^*/\mathbb{P}(Y=1,S=1)) > 0$. Similarly, we show that $(2+\theta^*/\mathbb{P}(Y=1,S=0)) > 0$. Combination of both inequalities and the fact that for all $s \in \{0,1\}$ we have $\mathbb{P}(Y=1,S=s) \leq 1$ implies that $|\theta^*| \leq 2$. □

Auxiliary results

Before proceeding to the proof of our main result in Theorem 9, let us first introduce several auxiliary results. We suggest the reader to first understand these results omitting its proofs before proceeding further. We will use $C > 0$ as a generic constant which actually could be different from line to line, yet, this constant is always independent from n, N.

Remark 9. *In this part of the manuscript it is assumed that the unlabeled dataset is sampled* i.i.d. *from $\mathbb{P}_{(X,S)}$, it implies that in theory this dataset could be composed of* only *features belonging to either of the group. Clearly, since $\mathbb{P}(S=1) > 0$ and $\mathbb{P}(S=0) > 0$ then a situation has an extremely small probability of appearing, in terms of N. There are various ways to alleviate this issue. The first one is conditioning on the event that we have at least one sample from each group, however, we have found that this approach unnecessarily over complicates our derivations and does not bring any insights. That is why, we follows another path, which is much simpler, though, might look a little strange at first sight. We actually augment $\mathcal{D}_N^{\mathrm{U}}$ by* four *points $(X_1,1),(X_2,1),(X_3,0),(X_4,0)$ which are sampled as $X_1, X_2 \overset{\text{i.i.d.}}{\sim} \mathbb{P}_{X|S=1}$ and $X_3, X_4 \overset{\text{i.i.d.}}{\sim} \mathbb{P}_{X|S=0}$. Once it is done we can safely assume that $\mathcal{D}_N^{\mathrm{U}}$ consists of* at least *two features from* each *group. The above is simply a technicality which allows to present our result in a correct way.*

The next lemma can be found in [Cribari-Neto et al., 2000].

Lemma 7. *Let Z be a binomial random variable with parameters N, p, then for every $\alpha \in \mathbb{R}$*

$$\mathbb{E}[(1+Z)^\alpha] = \mathcal{O}\left((Np)^\alpha\right) \ .$$

Lemma 8. *For any classifier g we have*

$$\mathcal{R}(g) = \mathbb{E}_{(X,S)}[\eta(X,S)] - \mathbb{E}_{(X,S)}[(2\eta(X,S)-1)g(X,S)] \ .$$

Proof. We can write

$$\begin{aligned}\mathcal{R}(g) &:= \mathbb{P}(Y \neq g(X,S)) = \mathbb{E}[Y(1-g(X,S))] + \mathbb{E}[(1-Y)g(X,S)] \\ &= \mathbb{E}_{(X,S)}\eta(X,S)(1-g(X,S)) + \mathbb{E}_{(X,S)}(1-\eta(X,S))g(X,S) \\ &= \mathbb{E}_{(X,S)}[\eta(X,S)] - \mathbb{E}_{(X,S)}[(2\eta(X,S)-1)g(X,S)] \ .\end{aligned}$$

□

In what follows we shall often use the relations:

$$\begin{aligned}\mathbb{P}(Y=1,S=s) &= \mathbb{P}\left(Y=1\,|\,S=s\right)\mathbb{P}(S=s)\ ,\\ \mathbb{P}\left(Y=1\,|\,S=s\right) &= \mathbb{E}_{X|S=s}[\eta(X,s)]\ .\end{aligned}$$

which holds for all $s\in\{0,1\}$.

Proof of Theorem 9

Below we gather extra tools which are directly related to the proof of our main result, proof are provided later in this section. First lemma gives an upper on the quantity of unfairness $\Delta(g,\mathbb{P})$ in terms of its empirical version in Definition 11.

Lemma 9. *Let g be* any *classifier (data depended or not) and $\hat{\eta}$ be an estimator of the regression function η constructed from $\mathcal{D}_n^{\mathrm{L}}$. Then, almost surely we have*

$$\begin{aligned}\Delta(g,\mathbb{P}) \leq \hat{\Delta}(g,\mathbb{P}) &+ \underbrace{2\frac{\mathbb{E}_{X|S=1}\left|\eta(X,1)-\hat{\eta}(X,1)\right|}{\mathbb{P}\left(Y=1\,|\,S=1\right)}}_{\textit{how good is }\hat{\eta}} + \underbrace{2\frac{\mathbb{E}_{X|S=0}\left|\eta(X,0)-\hat{\eta}(X,0)\right|}{\mathbb{P}\left(Y=1\,|\,S=0\right)}}_{\textit{how good is }\hat{\eta}}\\ &+ \underbrace{\frac{\left|(\mathbb{E}_{X|S=1}-\hat{\mathbb{E}}_{X|S=1})\hat{\eta}(X,1)g(X,1)\right|}{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)}}_{\textit{empirical process}} + \underbrace{\frac{\left|(\mathbb{E}_{X|S=0}-\hat{\mathbb{E}}_{X|S=0})\hat{\eta}(X,0)g(X,0)\right|}{\mathbb{E}_{X|S=0}\hat{\eta}(X,0)}}_{\textit{empirical process}}\\ &+ \frac{\left|\hat{\mathbb{E}}_{X|S=1}\hat{\eta}(X,1)-\mathbb{E}_{X|S=1}\hat{\eta}(X,1)\right|}{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)} + \frac{\left|\hat{\mathbb{E}}_{X|S=0}\hat{\eta}(X,0)-\mathbb{E}_{X|S=0}\hat{\eta}(X,0)\right|}{\mathbb{E}_{X|S=0}\hat{\eta}(X,0)}\ .\end{aligned}$$

The next lemma gives an upper bound on the empirical processes of Lemma 9.

Lemma 10. *There exists a constant $C>0$ that depends only on $\mathbb{P}(S=0)$ and $\mathbb{P}(S=1)$ such that almost surely for all $s\in\{0,1\}$ we have*

$$\mathbb{E}_{\mathcal{D}_N^{\mathrm{U}}}\sup_{t\in[0,1]}\left|(\mathbb{E}_{X|S=s}-\hat{\mathbb{E}}_{X|S=s})\hat{\eta}(X,s)\mathbb{1}_{\{t\leq\hat{\eta}(X,s)\}}\right| \leq C\sqrt{\frac{1}{N}}\ .$$

The next result is obvious, yet, is used several times in our proof.

Lemma 11. *For any function $h_1,h_0:\mathbb{R}^d\to[0,1]$, any $\theta\in\mathbb{R}$, any $a_1,a_0,b_1,b_0\in(0,1)$ we have*

$$\begin{aligned}\mathbb{E}_{X|S=1}&\left[\frac{\theta h_1(X)}{a_1}\mathbb{1}_{\left\{b_1(2h_1(X)-1)-\frac{\theta h_1(X)}{a_1}\geq 0\right\}}\right] = \mathbb{E}_{X|S=1}\left[(2h_1(X)-1)\mathbb{1}_{\left\{b_1(2h_1(X)-1)-\frac{\theta h_1(X)}{a_1}\geq 0\right\}}\right]b_1\\ &- \mathbb{E}_{X|S=1}\left(b_1(2h_1(X)-1)-\frac{\theta h_1(X)}{a_1}\right)_+\ ,\\ \mathbb{E}_{X|S=0}&\left[\frac{\theta h_0(X)}{a_0}\mathbb{1}_{\left\{b_0(2h_0(X)-1)+\frac{\theta h_0(X)}{a_0}\geq 0\right\}}\right] = -\mathbb{E}_{X|S=0}\left[(2h_0(X)-1)\mathbb{1}_{\left\{b_0(2h_0(X)-1)+\frac{\theta h_0(X)}{a_0}\geq 0\right\}}\right]b_0\\ &+ \mathbb{E}_{X|S=0}\left(b_0(2h_0(X)-1)+\frac{\theta h_0(X)}{a_0}\right)_+\ ,\end{aligned}$$

moreover, the expectation $\mathbb{E}_{X|S=s}$ can be replaced by $\hat{\mathbb{E}}_{X|S=s}$ for all $s\in\{0,1\}$.

Proof of asymptotic fairness (Part I of Theorem 9)

Proof. The first step is to show that under Assumption 11 the term $\hat{\Delta}(\hat{g}, \mathbb{P})$ cannot be too big. Indeed, notice that for every $\theta \in [-2, 2]$, thanks to the triangle inequality we can write almost surely

$$\begin{aligned}\hat{\Delta}(\hat{g}_\theta, \mathbb{P}) \leq & \left| \frac{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)\hat{g}_\theta(X,1)}{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)} - \frac{\mathbb{E}_{X|S=0}\hat{\eta}(X,0)\hat{g}_\theta(X,0)}{\mathbb{E}_{X|S=0}\hat{\eta}(X,0)} \right| \\ & + \left| \frac{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)\hat{g}_\theta(X,1)}{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)} - \frac{\hat{\mathbb{E}}_{X|S=1}\hat{\eta}(X,1)\hat{g}_\theta(X,1)}{\hat{\mathbb{E}}_{X|S=1}\hat{\eta}(X,1)} \right| \\ & + \left| \frac{\mathbb{E}_{X|S=0}\hat{\eta}(X,0)\hat{g}_\theta(X,0)}{\mathbb{E}_{X|S=0}\hat{\eta}(X,0)} - \frac{\hat{\mathbb{E}}_{X|S=0}\hat{\eta}(X,0)\hat{g}_\theta(X,0)}{\hat{\mathbb{E}}_{X|S=0}\hat{\eta}(X,0)} \right| . \end{aligned} \tag{2.15}$$

Our goal is to take care of each of the three terms appearing on the right hand side of the inequality. The technique used for the second and the third term is identical, whereas the first term is a bit more involved. Let us start with the second term on the right hand side of Eq. (2.15). For this term we can write almost surely

$$\begin{aligned} & \left| \frac{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)\hat{g}_\theta(X,1)}{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)} - \frac{\hat{\mathbb{E}}_{X|S=1}\hat{\eta}(X,1)\hat{g}_\theta(X,1)}{\hat{\mathbb{E}}_{X|S=1}\hat{\eta}(X,1)} \right| \\ & \leq \left| \frac{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)\hat{g}_\theta(X,1)}{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)} - \frac{\hat{\mathbb{E}}_{X|S=1}\hat{\eta}(X,1)\hat{g}_\theta(X,1)}{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)} \right| \\ & + \left| \frac{\hat{\mathbb{E}}_{X|S=1}\hat{\eta}(X,1)\hat{g}_\theta(X,1)}{\hat{\mathbb{E}}_{X|S=1}\hat{\eta}(X,1)} - \frac{\hat{\mathbb{E}}_{X|S=1}\hat{\eta}(X,1)\hat{g}_\theta(X,1)}{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)} \right| \\ & = \frac{\left|(\mathbb{E}_{X|S=1} - \hat{\mathbb{E}}_{X|S=1})\hat{\eta}(X,1)\hat{g}_\theta(X,1)\right|}{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)} \\ & + \underbrace{\frac{\hat{\mathbb{E}}_{X|S=1}\hat{\eta}(X,1)\hat{g}_\theta(X,1)}{\hat{\mathbb{E}}_{X|S=1}\hat{\eta}(X,1)}}_{\leq 1} \frac{\left|(\mathbb{E}_{X|S=1} - \hat{\mathbb{E}}_{X|S=1})\hat{\eta}(X,1)\mathbb{1}_{\{0 \leq \hat{\eta}(X,1)\}}\right|}{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)} \\ & \leq 2\frac{\sup_{t\in[0,1]}\left|(\mathbb{E}_{X|S=1} - \hat{\mathbb{E}}_{X|S=1})\hat{\eta}(X,1)\mathbb{1}_{\{t \leq \hat{\eta}(X,1)\}}\right|}{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)} , \end{aligned}$$

where the last inequality follows from the fact that $\hat{g}_\theta$ is a thresholding rule. Similarly, we show that the third term in Eq. (2.15) admits the following bound almost surely

$$\begin{aligned} & \left| \frac{\mathbb{E}_{X|S=0}\hat{\eta}(X,0)\hat{g}_\theta(X,0)}{\mathbb{E}_{X|S=0}\hat{\eta}(X,0)} - \frac{\hat{\mathbb{E}}_{X|S=0}\hat{\eta}(X,0)\hat{g}_\theta(X,0)}{\hat{\mathbb{E}}_{X|S=0}\hat{\eta}(X,0)} \right| \\ & \leq 2\frac{\sup_{t\in[0,1]}\left|(\mathbb{E}_{X|S=0} - \hat{\mathbb{E}}_{X|S=0})\hat{\eta}(X,0)\mathbb{1}_{\{t \leq \hat{\eta}(X,0)\}}\right|}{\mathbb{E}_{X|S=0}\hat{\eta}(X,0)} . \end{aligned}$$

Therefore, we arrive at the following bound on $\hat{\Delta}(\hat{g}_\theta, \mathbb{P})$ which holds almost surely

$$\begin{aligned}\hat{\Delta}(\hat{g}_\theta, \mathbb{P}) \leq & \left| \frac{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)\hat{g}_\theta(X,1)}{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)} - \frac{\mathbb{E}_{X|S=0}\hat{\eta}(X,0)\hat{g}_\theta(X,0)}{\mathbb{E}_{X|S=0}\hat{\eta}(X,0)} \right| \\ & + 2\frac{\sup_{t\in[0,1]} \left| (\mathbb{E}_{X|S=1} - \hat{\mathbb{E}}_{X|S=1})\hat{\eta}(X,1)\mathbb{1}_{\{t\leq\hat{\eta}(X,1)\}} \right|}{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)} \\ & + 2\frac{\sup_{t\in[0,1]} \left| (\mathbb{E}_{X|S=0} - \hat{\mathbb{E}}_{X|S=0})\hat{\eta}(X,0)\mathbb{1}_{\{t\leq\hat{\eta}(X,0)\}} \right|}{\mathbb{E}_{X|S=0}\hat{\eta}(X,0)} \ .\end{aligned} \tag{2.16}$$

This is one of the moments when we make use of Assumption 11. Thanks to the continuity we can be sure that for every possible unlabeled sample $\mathcal{D}_N^{\mathrm{U}}$, there exists $\theta'(\mathcal{D}_N^{\mathrm{U}})$ such that

$$\frac{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)\hat{g}_{\theta'(\mathcal{D}_N^{\mathrm{U}})}(X,1)}{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)} = \frac{\mathbb{E}_{X|S=0}\hat{\eta}(X,0)\hat{g}_{\theta'(\mathcal{D}_N^{\mathrm{U}})}(X,0)}{\mathbb{E}_{X|S=0}\hat{\eta}(X,0)} \ .$$

Indeed, for every possible unlabeled sample $\mathcal{D}_N^{\mathrm{U}}$ on the left hand side we have a continuous decreasing of θ function and on the right hand side we have a continuous increasing function of θ. Therefore, such a value $\theta'(\mathcal{D}_N^{\mathrm{U}})$ exists.

Taking infimum over $\theta \in [-2, 2]$ on both sides of Equation (2.16) we obtain

$$\begin{aligned}\hat{\Delta}(\hat{g}, \mathbb{P}) = \hat{\Delta}(\hat{g}_{\hat{\theta}}, \mathbb{P}) \leq & 2\frac{\sup_{t\in[0,1]} \left| (\mathbb{E}_{X|S=1} - \hat{\mathbb{E}}_{X|S=1})\hat{\eta}(X,1)\mathbb{1}_{\{t\leq\hat{\eta}(X,1)\}} \right|}{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)} \\ & + 2\frac{\sup_{t\in[0,1]} \left| (\mathbb{E}_{X|S=0} - \hat{\mathbb{E}}_{X|S=0})\hat{\eta}(X,0)\mathbb{1}_{\{t\leq\hat{\eta}(X,0)\}} \right|}{\mathbb{E}_{X|S=0}\hat{\eta}(X,0)} \ .\end{aligned} \tag{2.17}$$

Using Lemma 9 and applying it to $\hat{g}$ we immediately obtain almost surely

$$\begin{aligned}\Delta(\hat{g}, \mathbb{P}) \leq & 4\frac{\sup_{t\in[0,1]} \left| (\mathbb{E}_{X|S=1} - \hat{\mathbb{E}}_{X|S=1})\hat{\eta}(X,1)\mathbb{1}_{\{t\leq\hat{\eta}(X,1)\}} \right|}{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)} \\ & + 4\frac{\sup_{t\in[0,1]} \left| (\mathbb{E}_{X|S=0} - \hat{\mathbb{E}}_{X|S=0})\hat{\eta}(X,0)\mathbb{1}_{\{t\leq\hat{\eta}(X,0)\}} \right|}{\mathbb{E}_{X|S=0}\hat{\eta}(X,0)} \\ & + 2\frac{\mathbb{E}_{X|S=1} \left| \eta(X,1) - \hat{\eta}(X,1) \right|}{\mathbb{P}\left(Y=1 \,|\, S=1\right)} + 2\frac{\mathbb{E}_{X|S=0} \left| \eta(X,0) - \hat{\eta}(X,0) \right|}{\mathbb{P}\left(Y=1 \,|\, S=0\right)} \ .\end{aligned}$$

Clearly, if $\hat{\eta}$ is a consistent estimator of η then the last two terms on the right hand side are converging to zero in expectation as $n \to \infty$. Therefore, it remain to provide an upper bound for the two empirical processes. Recall, that our goal is to obtain consistency in expectation, thus we take expectation *w.r.t.* $\mathcal{D}_n^{\mathrm{L}}, \mathcal{D}_N^{\mathrm{U}}$ from both sides of the inequality. Thanks to Lemma 10 we have for each $s \in \{0, 1\}$

$$\mathbb{E}_{\mathcal{D}_N^{\mathrm{U}}} \sup_{t\in[0,1]} \left| (\mathbb{E}_{X|S=s} - \hat{\mathbb{E}}_{X|S=s})\hat{\eta}(X,s)\mathbb{1}_{\{t\leq\hat{\eta}(X,s)\}} \right| \leq C\sqrt{\frac{1}{N}} \ .$$

The arguments above imply that there exists an absolute constant $C > 0$ such that

$$\begin{aligned}\mathbb{E}_{(\mathcal{D}_n^{\mathrm{L}}, \mathcal{D}_N^{\mathrm{U}})}[\Delta(\hat{g}, \mathbb{P})] \leq & 2\frac{\mathbb{E}_{\mathcal{D}_n^{\mathrm{L}}}\mathbb{E}_{X|S=1} \left| \eta(X,1) - \hat{\eta}(X,1) \right|}{\mathbb{P}\left(Y=1 \,|\, S=1\right)} + 2\frac{\mathbb{E}_{\mathcal{D}_n^{\mathrm{L}}}\mathbb{E}_{X|S=0} \left| \eta(X,0) - \hat{\eta}(X,0) \right|}{\mathbb{P}\left(Y=1 \,|\, S=0\right)} \\ & + C\sqrt{\frac{1}{N}}\mathbb{E}_{\mathcal{D}_n^{\mathrm{L}}} \frac{1}{\min\{\mathbb{E}_{X|S=1}\hat{\eta}(X,1), \mathbb{E}_{X|S=0}\hat{\eta}(X,0)\}} \ .\end{aligned}$$

Using the second item of Assumption 10, which states that $\min\{\mathbb{E}_{X|S=1}\hat{\eta}(X,1), \mathbb{E}_{X|S=0}\hat{\eta}(X,0)\} \geq c_{n,N}$ almost surely we conclude. □

Proof of asymptotic optimality (Part II of Theorem 9)

In order to show that the risk of the proposed algorithm converges to the risk of the optimal classifier, we follow the similar strategy to the one that we shall use in Chapter 3, that is, we first introduce an intermediate pseudo-estimator $\tilde{g}$ as follows

$$\tilde{g}(x,1) = \mathbb{1}_{\left\{\mathbb{P}(S=1)\leq\hat{\eta}(x,1)\left(2\mathbb{P}(S=1)-\frac{\tilde{\theta}}{\mathbb{E}_{X|S=1}[\hat{\eta}(X,1)]}\right)\right\}} \ , \tag{2.18}$$

$$\tilde{g}(x,0) = \mathbb{1}_{\left\{\mathbb{P}(S=0)\leq\hat{\eta}(x,0)\left(2\mathbb{P}(S=0)+\frac{\tilde{\theta}}{\mathbb{E}_{X|S=0}[\hat{\eta}(X,0)]}\right)\right\}} \ , \tag{2.19}$$

where $\tilde{\theta}$ is a solution of

$$\frac{\mathbb{E}_{X|S=1}\left[\hat{\eta}(X,1)\tilde{g}_\theta(X,1)\right]}{\mathbb{E}_{X|S=1}[\hat{\eta}(X,1)]} = \frac{\mathbb{E}_{X|S=0}\left[\hat{\eta}(X,0)\tilde{g}_\theta(X,0)\right]}{\mathbb{E}_{X|S=0}[\hat{\eta}(X,0)]} \ , \tag{2.20}$$

with $\tilde{g}_\theta$ being defined as for all $x \in \mathbb{R}^d$ as

$$\tilde{g}_\theta(x,1) = \mathbb{1}_{\left\{1\leq\hat{\eta}(X,1)\left(2-\frac{\theta}{\mathbb{E}_{X|S=1}[\hat{\eta}(X,1)]\mathbb{P}(S=1)}\right)\right\}} \ ,$$

$$\tilde{g}_\theta(x,0) = \mathbb{1}_{\left\{1\leq\hat{\eta}(X,0)\left(2+\frac{\theta}{\mathbb{E}_{X|S=0}[\hat{\eta}(X,0)]\mathbb{P}(S=0)}\right)\right\}} \ .$$

Note that thanks to Assumption 11 such a value $\tilde{\theta}$ always exists.

Intuitively, the classifier $\tilde{g}$ knows the marginal distribution of (X,S), that is, it knows both $\mathbb{P}_{X|s}$ and $\mathbb{P}_S$. It is seen as an idealized version of $\hat{g}$, where the uncertainty is only induced by the lack of knowledge of the regression function η. We upper bound the excess risk in two steps. In the first step we upper bound $\mathcal{R}(\tilde{g}) - \mathcal{R}(g^*)$ and on the second we upper bound the difference $\mathcal{R}(\hat{g}) - \mathcal{R}(\tilde{g})$.

Theorem 10 (Bound on the pseudo oracle)**.** *Let $\tilde{g}$ be the pseudo oracle classifier defined in Eq. 2.18 with $\hat{\eta}$ satisfying Assumptions 10 and 11, then*

$$\lim_{n\to\infty} \mathbb{E}_{\mathcal{D}_n^{\mathrm{L}}}[\mathcal{R}(\tilde{g})] - \mathcal{R}(g^*) \leq 0 \ .$$

Proof of Theorem 10. First of all, let us rewrite the equation for θ^* in the following form

$$\mathbb{E}_{X|S=1}\left[\frac{\theta^*\eta(X,1)}{\mathbb{E}_{X|S=1}[\eta(X,1)]}\mathbb{1}_{\left\{\mathbb{P}(S=1)(2\eta(X,1)-1)-\frac{\theta^*\eta(X,1)}{\mathbb{E}_{X|S=1}[\eta(X,1)]}\geq 0\right\}}\right]$$
$$= \mathbb{E}_{X|S=0}\left[\frac{\theta^*\eta(X,0)}{\mathbb{E}_{X|S=0}[\eta(X,0)]}\mathbb{1}_{\left\{\mathbb{P}(S=0)(2\eta(X,0)-1)+\frac{\theta^*\eta(X,0)}{\mathbb{E}_{X|S=0}[\eta(X,0)]}\geq 0\right\}}\right] \ .$$

Using Lemma 11 with $h_s(\cdot) \equiv \eta(\cdot,s), a_s = \mathbb{E}_{X|S=1}[h_s(\cdot)], b_s = \mathbb{P}(S=s)$ for $s \in \{0,1\}$ we get

$$\mathbb{P}(S=1)\mathbb{E}_{X|S=1}[(2\eta(X,1)-1)g^*(X,1)]$$
$$- \mathbb{E}_{X|S=1}\left(\mathbb{P}(S=1)(2\eta(X,1)-1) - \frac{\theta^*\eta(X,1)}{\mathbb{E}_{X|S=1}[\eta(X,1)]}\right)_+$$
$$= -\mathbb{P}(S=0)\mathbb{E}_{X|S=0}[(2\eta(X,0)-1)g^*(X,0)]$$
$$+ \mathbb{E}_{X|S=0}\left(\mathbb{P}(S=0)(2\eta(X,0)-1) + \frac{\theta^*\eta(X,0)}{\mathbb{E}_{X|S=0}[\eta(X,0)]}\right)_+ \ .$$

Rearranging the terms we can arrive at

$$\begin{aligned}
&\mathbb{P}(S=1)\mathbb{E}_{X|S=1}[(2\eta(X,1)-1)g^*(X,1)] + \mathbb{P}(S=0)\mathbb{E}_{X|S=0}[(2\eta(X,0)-1)g^*(X,0)] \\
&= \mathbb{E}_{X|S=1}\left(\mathbb{P}(S=1)(2\eta(X,1)-1) - \frac{\theta^*\eta(X,1)}{\mathbb{E}_{X|S=1}[\eta(X,1)]}\right)_+ \\
&\quad + \mathbb{E}_{X|S=0}\left(\mathbb{P}(S=0)(2\eta(X,0)-1) + \frac{\theta^*\eta(X,0)}{\mathbb{E}_{X|S=0}[\eta(X,0)]}\right)_+ .
\end{aligned}$$

Notice that the left hand side of the above equality can be written as

$$\begin{aligned}
\mathbb{E}_{(X,S)}[(2\eta(X,S)-1)g^*(X,S)] &= \mathbb{E}_{X|S=1}\left(\mathbb{P}(S=1)(2\eta(X,1)-1) - \frac{\theta^*\eta(X,1)}{\mathbb{E}_{X|S=1}[\eta(X,1)]}\right)_+ \\
&\quad + \mathbb{E}_{X|S=0}\left(\mathbb{P}(S=0)(2\eta(X,0)-1) + \frac{\theta^*\eta(X,0)}{\mathbb{E}_{X|S=0}[\eta(X,0)]}\right)_+ .
\end{aligned} \tag{2.21}$$

Thus, combining the previous equality with the expression of the risk from Lemma 8 we get

$$\begin{aligned}
\mathcal{R}(g^*) = \mathbb{E}_{(X,S)}[\eta(X,S)] &- \mathbb{E}_{X|S=1}\left(\mathbb{P}(S=1)(2\eta(X,1)-1) - \frac{\theta^*\eta(X,1)}{\mathbb{E}_{X|S=1}[\eta(X,1)]}\right)_+ \\
&- \mathbb{E}_{X|S=0}\left(\mathbb{P}(S=0)(2\eta(X,0)-1) + \frac{\theta^*\eta(X,0)}{\mathbb{E}_{X|S=0}[\eta(X,0)]}\right)_+ .
\end{aligned} \tag{2.22}$$

Step-wise similar argument yields that for the pseudo-oracle $\tilde{g}$ we can write

$$\begin{aligned}
&\mathbb{E}_{(X,S)}[(2\hat{\eta}(X,S)-1)\tilde{g}(X,S)] \\
&= \mathbb{E}_{X|S=1}\left(\mathbb{P}(S=1)(2\hat{\eta}(X,1)-1) - \frac{\tilde{\theta}\hat{\eta}(X,1)}{\mathbb{E}_{X|S=1}[\hat{\eta}(X,1)]}\right)_+ \\
&\quad + \mathbb{E}_{X|S=0}\left(\mathbb{P}(S=0)(2\hat{\eta}(X,0)-1) + \frac{\tilde{\theta}\hat{\eta}(X,0)}{\mathbb{E}_{X|S=0}[\hat{\eta}(X,0)]}\right)_+ .
\end{aligned} \tag{2.23}$$

Moreover, its risk satisfies

$$\begin{aligned}
\mathcal{R}(\tilde{g}) &= \mathbb{E}_{(X,S)}[\eta(X,S)] - \mathbb{E}_{(X,S)}[(2\eta(X,S)-1)\tilde{g}(X,S)] \\
&\leq \mathbb{E}_{(X,S)}[\eta(X,S)] - \mathbb{E}_{(X,S)}[(2\hat{\eta}(X,S)-1)\tilde{g}(X,S)] + 2\mathbb{E}_{(X,S)}\left|\hat{\eta}(X,S)-\eta(X,S)\right| .
\end{aligned} \tag{2.24}$$

Therefore, combining Eq. (2.22) with Eq. (2.24), we can write for the excess risk

$$\begin{aligned}
\mathcal{R}(\tilde{g}) - \mathcal{R}(g^*) \leq &2\mathbb{E}_{(X,S)}\left|\hat{\eta}(X,S)-\eta(X,S)\right| \\
&+ \mathbb{E}_{X|S=1}\left(\mathbb{P}(S=1)(2\eta(X,1)-1) - \frac{\theta^*\eta(X,1)}{\mathbb{E}_{X|S=1}[\eta(X,1)]}\right)_+ \\
&- \mathbb{E}_{X|S=1}\left(\mathbb{P}(S=1)(2\hat{\eta}(X,1)-1) - \frac{\tilde{\theta}\hat{\eta}(X,1)}{\mathbb{E}_{X|S=1}[\hat{\eta}(X,1)]}\right)_+ \\
&+ \mathbb{E}_{X|S=0}\left(\mathbb{P}(S=0)(2\eta(X,0)-1) + \frac{\theta^*\eta(X,0)}{\mathbb{E}_{X|S=0}[\eta(X,0)]}\right)_+ \\
&- \mathbb{E}_{X|S=0}\left(\mathbb{P}(S=0)(2\hat{\eta}(X,0)-1) + \frac{\tilde{\theta}\hat{\eta}(X,0)}{\mathbb{E}_{X|S=0}[\hat{\eta}(X,0)]}\right)_+ .
\end{aligned}$$

Recall that θ^* is a minimizer of

$$\mathbb{E}_{X|S=1}\left(\mathbb{P}(S=1)(2\eta(X,1)-1)-\frac{\theta\eta(X,1)}{\mathbb{E}_{X|S=1}[\eta(X,1)]}\right)_+ + \mathbb{E}_{X|S=0}\left(\mathbb{P}(S=0)(2\eta(X,0)-1)+\frac{\theta\eta(X,0)}{\mathbb{E}_{X|S=0}[\eta(X,0)]}\right)_+ ,$$

thus we can replace θ^* by $\tilde{\theta}$ and obtain the following upper bound

$$\begin{aligned}\mathcal{R}(\tilde{g})-\mathcal{R}(g^*) \leq & 2\mathbb{E}_{(X,S)}\left|\hat{\eta}(X,S)-\eta(X,S)\right| \\ & + \mathbb{E}_{X|S=1}\left(\mathbb{P}(S=1)(2\eta(X,1)-1)-\frac{\tilde{\theta}\eta(X,1)}{\mathbb{E}_{X|S=1}[\eta(X,1)]}\right)_+ \\ & - \mathbb{E}_{X|S=1}\left(\mathbb{P}(S=1)(2\hat{\eta}(X,1)-1)-\frac{\tilde{\theta}\hat{\eta}(X,1)}{\mathbb{E}_{X|S=1}[\hat{\eta}(X,1)]}\right)_+ \\ & + \mathbb{E}_{X|S=0}\left(\mathbb{P}(S=0)(2\eta(X,0)-1)+\frac{\tilde{\theta}\eta(X,0)}{\mathbb{E}_{X|S=0}[\eta(X,0)]}\right)_+ \\ & - \mathbb{E}_{X|S=0}\left(\mathbb{P}(S=0)(2\hat{\eta}(X,0)-1)+\frac{\tilde{\theta}\hat{\eta}(X,0)}{\mathbb{E}_{X|S=0}[\hat{\eta}(X,0)]}\right)_+ .\end{aligned}$$

Since, for all $x, y \in \mathbb{R}$ we have $(x)_+ - (y)_+ \leq (x-y)_+ \leq |x-y|$ we get

$$\begin{aligned}\mathcal{R}(\tilde{g})-\mathcal{R}(g^*) \leq & 4\mathbb{E}_{(X,S)}\left|\hat{\eta}(X,S)-\eta(X,S)\right| \\ & + \mathbb{E}_{X|S=1}|\tilde{\theta}|\left|\frac{\hat{\eta}(X,1)}{\mathbb{E}_{X|S=1}[\hat{\eta}(X,1)]}-\frac{\eta(X,1)}{\mathbb{E}_{X|S=1}[\eta(X,1)]}\right| \\ & + \mathbb{E}_{X|S=0}|\tilde{\theta}|\left|\frac{\hat{\eta}(X,0)}{\mathbb{E}_{X|S=0}[\hat{\eta}(X,0)]}-\frac{\eta(X,0)}{\mathbb{E}_{X|S=0}[\eta(X,0)]}\right| .\end{aligned}$$

For the same reason why $|\theta^*| \leq 2$ we have $|\tilde{\theta}| \leq 2$, thus for all $s \in \{0,1\}$ we have

$$\begin{aligned}& \mathbb{E}_{X|S=s}|\tilde{\theta}|\left|\frac{\hat{\eta}(X,s)}{\mathbb{E}_{X|S=s}[\hat{\eta}(X,s)]}-\frac{\eta(X,s)}{\mathbb{E}_{X|S=s}[\eta(X,s)]}\right| \\ & \leq 2\mathbb{E}_{X|S=s}\left|\frac{\hat{\eta}(X,s)}{\mathbb{E}_{X|S=s}[\hat{\eta}(X,s)]}-\frac{\eta(X,s)}{\mathbb{E}_{X|S=s}[\eta(X,s)]}\right| \\ & \leq 2\mathbb{E}_{X|S=s}\left|\frac{\hat{\eta}(X,s)}{\mathbb{E}_{X|S=s}[\eta(X,s)]}-\frac{\eta(X,s)}{\mathbb{E}_{X|S=s}[\eta(X,s)]}\right| \\ & + 2\mathbb{E}_{X|S=s}\left|\frac{\hat{\eta}(X,s)}{\mathbb{E}_{X|S=s}[\hat{\eta}(X,s)]}-\frac{\hat{\eta}(X,s)}{\mathbb{E}_{X|S=s}[\eta(X,s)]}\right| \\ & \leq 4\frac{\mathbb{E}_{X|S=s}\left|\eta(X,s)-\hat{\eta}(X,s)\right|}{\mathbb{E}_{X|S=s}[\eta(X,s)]} .\end{aligned}$$

Thanks to Assumption 10, these terms converge to zero in expectation. □

Theorem 11. *Let $\hat{g}$ be the proposed classifier with $\hat{\eta}$ satisfying Assumptions 10 and 11, then*

$$\lim_{n\to\infty} \mathbb{E}_{(\mathcal{D}_n^{\mathrm{L}},\mathcal{D}_N^{\mathrm{U}})}[\mathcal{R}(\hat{g})-\mathcal{R}(\tilde{g})] \leq 0 .$$

Proof. Our goal is to upper bound the quantity $\mathbb{E}_{(\mathcal{D}_n^{\mathrm{L}},\mathcal{D}_N^{\mathrm{U}})}\mathcal{R}(\hat{g}) - \mathcal{R}(\tilde{g})$. We start by providing a bound on $\mathcal{R}(\hat{g}) - \mathcal{R}(\tilde{g})$ which holds almost surely. Recall the equality of Equation (2.23)

$$\begin{aligned}
&\mathbb{E}_{(X,S)}[(2\hat{\eta}(X,S)-1)\tilde{g}(X,S)] \\
&= \mathbb{E}_{X|S=1}\left(\mathbb{P}(S=1)(2\hat{\eta}(X,1)-1) - \frac{\tilde{\theta}\hat{\eta}(X,1)}{\mathbb{E}_{X|S=1}[\hat{\eta}(X,1)]}\right)_+ \\
&\quad + \mathbb{E}_{X|S=0}\left(\mathbb{P}(S=0)(2\hat{\eta}(X,0)-1) + \frac{\tilde{\theta}\hat{\eta}(X,0)}{\mathbb{E}_{X|S=0}[\hat{\eta}(X,0)]}\right)_+ .
\end{aligned}$$

Using this and the expression of the risk given in Lemma 8 we can obtain the following lower bound on the risk of $\tilde{g}$

$$\begin{aligned}
\mathcal{R}(\tilde{g}) &= \mathbb{E}_{(X,S)}[\eta(X,S)] - \mathbb{E}_{(X,S)}[(2\eta(X,S)-1)\tilde{g}(X,S)] \\
&\geq \mathbb{E}_{(X,S)}[\eta(X,S)] - \mathbb{E}_{(X,S)}[(2\hat{\eta}(X,S)-1)\tilde{g}(X,S)] - 2\mathbb{E}_{(X,S)}\left|\hat{\eta}(X,S)-\eta(X,S)\right| \\
&= \mathbb{E}_{(X,S)}[\eta(X,S)] - 2\mathbb{E}_{(X,S)}\left|\hat{\eta}(X,S)-\eta(X,S)\right| \qquad (2.25) \\
&\quad - \mathbb{E}_{X|S=1}\left(\mathbb{P}(S=1)(2\hat{\eta}(X,1)-1) - \frac{\tilde{\theta}\hat{\eta}(X,1)}{\mathbb{E}_{X|S=1}[\hat{\eta}(X,1)]}\right)_+ \\
&\quad - \mathbb{E}_{X|S=0}\left(\mathbb{P}(S=0)(2\hat{\eta}(X,0)-1) + \frac{\tilde{\theta}\hat{\eta}(X,0)}{\mathbb{E}_{X|S=0}[\hat{\eta}(X,0)]}\right)_+ .
\end{aligned}$$

We have thanks to Lemma 11 used with $h_s(\cdot) = \hat{\eta}(\cdot,s), a_s = \hat{\mathbb{E}}_{X|S=s}[h_s(X)], b_s = \hat{\mathbb{P}}(S=s)$ for all $s \in \{0,1\}$

$$\begin{aligned}
\frac{\hat{\mathbb{E}}_{X|S=1}\hat{\theta}\hat{\eta}(X,1)\hat{g}(X,1)}{\hat{\mathbb{E}}_{X|S=1}\hat{\eta}(X,1)} &= \hat{\mathbb{E}}_{X|S=1}[(2\hat{\eta}(X,1)-1)\hat{g}(X,1)]\hat{\mathbb{P}}(S=1) \qquad (2.26) \\
&\quad - \hat{\mathbb{E}}_{X|S=1}\left(\hat{\mathbb{P}}(S=1)(2\hat{\eta}(X,1)-1) - \frac{\hat{\theta}\hat{\eta}(X,1)}{\hat{\mathbb{E}}_{X|S=1}[\hat{\eta}(X,1)]}\right)_+ ,
\end{aligned}$$

and

$$\begin{aligned}
\frac{\hat{\mathbb{E}}_{X|S=0}\hat{\theta}\hat{\eta}(X,0)\hat{g}(X,0)}{\hat{\mathbb{E}}_{X|S=0}\hat{\eta}(X,0)} &= -\hat{\mathbb{E}}_{X|S=0}[(2\hat{\eta}(X,0)-1)\hat{g}(X,0)]\hat{\mathbb{P}}(S=0) \qquad (2.27) \\
&\quad + \hat{\mathbb{E}}_{X|S=0}\left(\hat{\mathbb{P}}(S=0)(2\hat{\eta}(X,0)-1) + \frac{\hat{\theta}\hat{\eta}(X,0)}{\hat{\mathbb{E}}_{X|S=0}[\hat{\eta}(X,0)]}\right)_+ .
\end{aligned}$$

Recall, that thanks to Definition 11 of the empirical unfairness we have

$$|\hat{\theta}|\hat{\Delta}(\hat{g},\mathbb{P}) = \left|\frac{\hat{\mathbb{E}}_{X|S=0}\hat{\theta}\hat{\eta}(X,0)\hat{g}(X,0)}{\hat{\mathbb{E}}_{X|S=0}\hat{\eta}(X,0)} - \frac{\hat{\mathbb{E}}_{X|S=1}\hat{\theta}\hat{\eta}(X,1)\hat{g}(X,1)}{\hat{\mathbb{E}}_{X|S=1}\hat{\eta}(X,1)}\right| .$$

Since, $|\hat{\theta}| \leq 2$, subtracting Eq. (2.27) from Eq. (2.26) and taking absolute value combined

with the triangle inequality we get

$$
\begin{aligned}
&\hat{\mathbb{E}}_{(X,S)}(2\hat{\eta}(X,S)-1)\hat{g}(X,S) \\
&= \hat{\mathbb{E}}_{X|S=0}[(2\hat{\eta}(X,0)-1)\hat{g}(X,0)]\hat{\mathbb{P}}(S=0) + \hat{\mathbb{E}}_{X|S=1}[(2\hat{\eta}(x,1)-1)\hat{g}(X,1)]\hat{\mathbb{P}}(S=1) \quad (2.28)\\
&\geq -2\hat{\Delta}(\hat{g},\mathbb{P}) + \hat{\mathbb{E}}_{X|S=0}\left(\hat{\mathbb{P}}(S=0)(2\hat{\eta}(X,0)-1) + \frac{\hat{\theta}\hat{\eta}(X,0)}{\hat{\mathbb{E}}_{X|S=0}[\hat{\eta}(X,0)]}\right)_+ \\
&\quad + \hat{\mathbb{E}}_{X|S=1}\left(\hat{\mathbb{P}}(S=1)(2\hat{\eta}(X,1)-1) - \frac{\hat{\theta}\hat{\eta}(X,1)}{\hat{\mathbb{E}}_{X|S=1}[\hat{\eta}(X,1)]}\right)_+ .
\end{aligned}
$$

Note that using the bound above we can get the following upper bound on the risk of the proposed classifier

$$
\begin{aligned}
\mathcal{R}(\hat{g}) &= \mathbb{E}_{(X,S)}[\eta(X,S)] - \mathbb{E}_{(X,S)}[(2\eta(X,S)-1)\hat{g}(X,S)] \\
&\leq \mathbb{E}_{(X,S)}[\eta(X,S)] - \mathbb{E}_{(X,S)}[(2\hat{\eta}(X,S)-1)\hat{g}(X,S)] \\
&\quad + 2\mathbb{E}_{(X,S)}\left|\eta(X,S)-\hat{\eta}(X,S)\right| \qquad \text{(replaced } \eta \text{ by } \hat{\eta}\text{)} \\
&\leq \mathbb{E}_{(X,S)}[\eta(X,S)] - \hat{\mathbb{E}}_{(X,S)}[(2\hat{\eta}(X,S)-1)\hat{g}(X,S)] + 2\mathbb{E}_{(X,S)}\left|\eta(X,S)-\hat{\eta}(X,S)\right| \\
&\quad + \left|(\mathbb{E}_{(X,S)} - \hat{\mathbb{E}}_{(X,S)})[(2\hat{\eta}(X,S)-1)\hat{g}(X,S)]\right| \qquad \text{(replaced } \mathbb{E}_{(X,S)} \text{ by } \hat{\mathbb{E}}_{(X,S)}\text{)} \\
&\leq \mathbb{E}_{(X,S)}[\eta(X,S)] - \hat{\mathbb{E}}_{(X,S)}[(2\hat{\eta}(X,S)-1)\hat{g}(X,S)] + 2\mathbb{E}_{(X,S)}\left|\eta(X,S)-\hat{\eta}(X,S)\right| \\
&\quad + \sup_{t\in[0,1]}\left|(\mathbb{E}_{(X,S)} - \hat{\mathbb{E}}_{(X,S)})[(2\hat{\eta}(X,S)-1)\mathbb{1}_{\{t\leq\hat{\eta}(X,S)\}}]\right| \qquad \text{(since } \hat{g} \text{ is thresholding)} \\
&\leq \mathbb{E}_{(X,S)}[\eta(X,S)] + 2\mathbb{E}_{(X,S)}\left|\eta(X,S)-\hat{\eta}(X,S)\right| \\
&\quad + 2\hat{\Delta}(\hat{g},\mathbb{P}) + \sup_{t\in[0,1]}\left|(\mathbb{E}_{(X,S)} - \hat{\mathbb{E}}_{(X,S)})[(2\hat{\eta}(X,S)-1)\mathbb{1}_{\{t\leq\hat{\eta}(X,S)\}}]\right| \\
&\quad - \hat{\mathbb{E}}_{X|S=0}\left(\hat{\mathbb{P}}(S=0)(2\hat{\eta}(X,0)-1) + \frac{\hat{\theta}\hat{\eta}(X,0)}{\hat{\mathbb{E}}_{X|S=0}[\hat{\eta}(X,0)]}\right)_+ \\
&\quad - \hat{\mathbb{E}}_{X|S=1}\left(\hat{\mathbb{P}}(S=1)(2\hat{\eta}(X,1)-1) - \frac{\hat{\theta}\hat{\eta}(X,1)}{\hat{\mathbb{E}}_{X|S=1}[\hat{\eta}(X,1)]}\right)_+ \qquad \text{(after Eq. (2.28))} .
\end{aligned}
$$

Thus, combining this upper bound on $\mathcal{R}(\hat{g})$ with the lower bound on $\mathcal{R}(\tilde{g})$ given in Eq. (2.25) we arrive at

$$
\begin{aligned}
\mathcal{R}(\hat{g}) - \mathcal{R}(\tilde{g}) \leq\ & 4\mathbb{E}_{(X,S)}\left|\eta(X,S)-\hat{\eta}(X,S)\right| + 2\hat{\Delta}(\hat{g},\mathbb{P}) \\
&+ \sup_{t\in[0,1]}\left|(\mathbb{E}_{(X,S)} - \hat{\mathbb{E}}_{(X,S)})[(2\hat{\eta}(X,S)-1)\mathbb{1}_{\{t\leq\hat{\eta}(X,S)\}}]\right| \\
&+ \mathbb{E}_{X|S=1}\left(\mathbb{P}(S=1)(2\hat{\eta}(X,1)-1) - \frac{\tilde{\theta}\hat{\eta}(X,1)}{\mathbb{E}_{X|S=1}[\hat{\eta}(X,1)]}\right)_+ \\
&- \hat{\mathbb{E}}_{X|S=1}\left(\hat{\mathbb{P}}(S=1)(2\hat{\eta}(X,1)-1) - \frac{\hat{\theta}\hat{\eta}(X,1)}{\hat{\mathbb{E}}_{X|S=1}[\hat{\eta}(X,1)]}\right)_+ \\
&+ \mathbb{E}_{X|S=0}\left(\mathbb{P}(S=0)(2\hat{\eta}(X,0)-1) + \frac{\tilde{\theta}\hat{\eta}(X,0)}{\mathbb{E}_{X|S=0}[\hat{\eta}(X,0)]}\right)_+ \\
&- \hat{\mathbb{E}}_{X|S=0}\left(\hat{\mathbb{P}}(S=0)(2\hat{\eta}(X,0)-1) + \frac{\hat{\theta}\hat{\eta}(X,0)}{\hat{\mathbb{E}}_{X|S=0}[\hat{\eta}(X,0)]}\right)_+ .
\end{aligned}
$$

Thanks to Lemma 10 the term $\sup_{t\in[0,1]} \left|(\mathbb{E}_{(X,S)} - \hat{\mathbb{E}}_{(X,S)})[(2\hat{\eta}(X,S)-1)\mathbb{1}_{\{t\leq\hat{\eta}(X,S)\}}]\right|$ converges to zero in expectation[9]. Equation (2.17) with Lemma 10 gives the convergence to zero of $\hat{\Delta}(\hat{g},\mathbb{P})$ in expectation. Assumption 10 tells us that the term $\mathbb{E}_{(X,S)}|\eta(X,S)-\hat{\eta}(X,S)|$ goes to zero in expectation. Thus it only remains to bound the term

$$
\begin{aligned}
(*) =& \mathbb{E}_{X|S=1}\left(\mathbb{P}(S=1)(2\hat{\eta}(X,1)-1) - \frac{\tilde{\theta}\hat{\eta}(X,1)}{\mathbb{E}_{X|S=1}[\hat{\eta}(X,1)]}\right)_+ \\
&- \hat{\mathbb{E}}_{X|S=1}\left(\hat{\mathbb{P}}(S=1)(2\hat{\eta}(X,1)-1) - \frac{\hat{\theta}\hat{\eta}(X,1)}{\hat{\mathbb{E}}_{X|S=1}[\hat{\eta}(X,1)]}\right)_+ \\
&+ \mathbb{E}_{X|S=0}\left(\mathbb{P}(S=0)(2\hat{\eta}(X,0)-1) + \frac{\tilde{\theta}\hat{\eta}(X,0)}{\mathbb{E}_{X|S=0}[\hat{\eta}(X,0)]}\right)_+ \\
&- \hat{\mathbb{E}}_{X|S=0}\left(\hat{\mathbb{P}}(S=0)(2\hat{\eta}(X,0)-1) + \frac{\hat{\theta}\hat{\eta}(X,0)}{\hat{\mathbb{E}}_{X|S=0}[\hat{\eta}(X,0)]}\right)_+ .
\end{aligned}
$$

Notice that (similarly to the case of θ^*) the condition in Eq. (2.20) on $\tilde{\theta}$ is the first order optimality condition for the minimum of the following function

$$
\begin{aligned}
&\mathbb{E}_{X|S=1}\left(\mathbb{P}(S=1)(2\hat{\eta}(X,1)-1) - \frac{\theta\hat{\eta}(X,1)}{\mathbb{E}_{X|S=1}[\hat{\eta}(X,1)]}\right)_+ \\
&+ \mathbb{E}_{X|S=0}\left(\mathbb{P}(S=0)(2\hat{\eta}(X,0)-1) + \frac{\theta\hat{\eta}(X,0)}{\mathbb{E}_{X|S=0}[\hat{\eta}(X,0)]}\right)_+ ,
\end{aligned}
$$

thus, the objective evaluated at minimum, that is, at $\tilde{\theta}$ is less or equal than the one evaluated at $\hat{\theta}$. Which implies that in order to upper bound $(*)$ it is sufficient to provide an upper bound on

$$
\begin{aligned}
(**) =& \mathbb{E}_{X|S=1}\left(\mathbb{P}(S=1)(2\hat{\eta}(X,1)-1) - \frac{\hat{\theta}\hat{\eta}(X,1)}{\mathbb{E}_{X|S=1}[\hat{\eta}(X,1)]}\right)_+ \\
&- \hat{\mathbb{E}}_{X|S=1}\left(\hat{\mathbb{P}}(S=1)(2\hat{\eta}(X,1)-1) - \frac{\hat{\theta}\hat{\eta}(X,1)}{\hat{\mathbb{E}}_{X|S=1}[\hat{\eta}(X,1)]}\right)_+ \\
&+ \mathbb{E}_{X|S=0}\left(\mathbb{P}(S=0)(2\hat{\eta}(X,0)-1) + \frac{\hat{\theta}\hat{\eta}(X,0)}{\mathbb{E}_{X|S=0}[\hat{\eta}(X,0)]}\right)_+ \\
&- \hat{\mathbb{E}}_{X|S=0}\left(\hat{\mathbb{P}}(S=0)(2\hat{\eta}(X,0)-1) + \frac{\hat{\theta}\hat{\eta}(X,0)}{\hat{\mathbb{E}}_{X|S=0}[\hat{\eta}(X,0)]}\right)_+ ,
\end{aligned}
$$

[9] Actually Lemma 10 is stated with $\hat{\eta}(X,S)$, whereas here it is $(2\hat{\eta}(X,S)-1)$. A straightforward modification of the argument used in Lemma 10 yields the desired result.

where we replaced $\tilde{\theta}$ by $\hat{\theta}$ thanks to the optimality of $\tilde{\theta}$. Let us define

$$
\begin{aligned}
(\triangle) =& \mathbb{E}_{X|S=1}\left(\mathbb{P}(S=1)(2\hat{\eta}(X,1)-1) - \frac{\hat{\theta}\hat{\eta}(X,1)}{\mathbb{E}_{X|S=1}[\hat{\eta}(X,1)]}\right)_+ \\
& - \hat{\mathbb{E}}_{X|S=1}\left(\hat{\mathbb{P}}(S=1)(2\hat{\eta}(X,1)-1) - \frac{\hat{\theta}\hat{\eta}(X,1)}{\hat{\mathbb{E}}_{X|S=1}[\hat{\eta}(X,1)]}\right)_+ , \\
(\triangle\triangle) =& \mathbb{E}_{X|S=0}\left(\mathbb{P}(S=0)(2\hat{\eta}(X,0)-1) + \frac{\hat{\theta}\hat{\eta}(X,0)}{\mathbb{E}_{X|S=0}[\hat{\eta}(X,0)]}\right)_+ \\
& - \hat{\mathbb{E}}_{X|S=0}\left(\hat{\mathbb{P}}(S=0)(2\hat{\eta}(X,0)-1) + \frac{\hat{\theta}\hat{\eta}(X,0)}{\hat{\mathbb{E}}_{X|S=0}[\hat{\eta}(X,0)]}\right)_+ .
\end{aligned}
$$

Both bounds are following similar arguments, we demonstrate it for $(\triangle)$, clearly we have

$$
\begin{aligned}
(\triangle) \leq& \hat{\mathbb{E}}_{X|S=1}\left(\mathbb{P}(S=1)(2\hat{\eta}(X,1)-1) - \frac{\hat{\theta}\hat{\eta}(X,1)}{\mathbb{E}_{X|S=1}[\hat{\eta}(X,1)]}\right)_+ \\
& - \hat{\mathbb{E}}_{X|S=1}\left(\hat{\mathbb{P}}(S=1)(2\hat{\eta}(X,1)-1) - \frac{\hat{\theta}\hat{\eta}(X,1)}{\hat{\mathbb{E}}_{X|S=1}[\hat{\eta}(X,1)]}\right)_+ \\
& + \left|(\mathbb{E}_{X|S=1} - \hat{\mathbb{E}}_{X|S=1})\left(\mathbb{P}(S=1)(2\hat{\eta}(X,1)-1) - \frac{\hat{\theta}\hat{\eta}(X,1)}{\mathbb{E}_{X|S=1}[\hat{\eta}(X,1)]}\right)_+\right| .
\end{aligned}
$$

For the first difference on the right hand side of this inequality we can write using the fact that $(x)_+ - (y)_+ \leq |x-y|$ for all $x, y \in \mathbb{R}$ and $|2\hat{\eta}(X,1)-1| \leq 1$ almost surely

$$
\begin{aligned}
& \hat{\mathbb{E}}_{X|S=1}\left(\mathbb{P}(S=1)(2\hat{\eta}(X,1)-1) - \frac{\hat{\theta}\hat{\eta}(X,1)}{\mathbb{E}_{X|S=1}[\hat{\eta}(X,1)]}\right)_+ \\
& - \hat{\mathbb{E}}_{X|S=1}\left(\hat{\mathbb{P}}(S=1)(2\hat{\eta}(X,1)-1) - \frac{\hat{\theta}\hat{\eta}(X,1)}{\hat{\mathbb{E}}_{X|S=1}[\hat{\eta}(X,1)]}\right)_+ \\
& \leq \left|\mathbb{P}(S=1) - \hat{\mathbb{P}}(S=1)\right| + |\hat{\theta}|\left|\frac{\hat{\mathbb{E}}_{X|S=1}[\hat{\eta}(X,1)]}{\mathbb{E}_{X|S=1}[\hat{\eta}(X,1)]} - 1\right|
\end{aligned}
$$

Clearly $\left|\mathbb{P}(S=1) - \hat{\mathbb{P}}(S=1)\right|$ goes to zero in expectation thanks to the law of large numbers or its finite sample variants. Besides, the term $\left|\frac{\hat{\mathbb{E}}_{X|S=0}[\hat{\eta}(X,0)]}{\mathbb{E}_{X|S=0}[\eta(X,0)]} - 1\right|$ can be seen in the following manner: let $Z \in [0,1]$ be a random variable with law $\mathbb{P}_Z$ and $Z_1, \ldots, Z_M$ be its *i.i.d.* realization, then sequentially our question is about

$$
\left|1 - \frac{\bar{Z}}{\mathbb{E}[Z]}\right| ,
$$

with $\bar{Z} = \frac{1}{M}\sum_{i=1}^{M} Z_i$. This term converges to zero in expectation thanks to the multiplicative Chernoff inequality, which is an exponential concentration inequality that allows to obtain

even a rate. Actually, even without the multiplicative Chernoff bound this term goes to zero thanks to the law of large numbers. Therefore, for convergence it remains to study the term

$$(\star) = \left| (\mathbb{E}_{X|S=1} - \hat{\mathbb{E}}_{X|S=1}) \left(\mathbb{P}(S=1)(2\hat{\eta}(X,1)-1) - \frac{\hat{\theta}\hat{\eta}(X,1)}{\mathbb{E}_{X|S=1}[\hat{\eta}(X,1)]} \right)_+ \right| \ .$$

Notice that thanks to the second part of Assumption 10 and the fact that $\hat{\theta} \in [-2,2]$ we have

$$\left| \frac{\hat{\theta}\hat{\eta}(X,1)}{\mathbb{E}_{X|S=1}[\hat{\eta}(X,1)]} \right| \leq \frac{2}{c_{n,N}} \ .$$

Therefore, we can upper bound $(\star)$ as

$$(\star) \leq \sup_{t \in [-2/c_{n,N}, 2/c_{n,N}]} \left| (\mathbb{E}_{X|S=1} - \hat{\mathbb{E}}_{X|S=1}) \left(\mathbb{P}(S=1)(2\hat{\eta}(X,1)-1) + t \right)_+ \right| \ ,$$

where the random quantity has been "supped-out". Introduce,

$$\mathcal{D}_{N_1} = \left\{ X_i \in \mathcal{D}_N^{\mathrm{U}} \ : \ S_i = 1 \right\}$$
$$\mathcal{D}_{N_0} = \left\{ X_i \in \mathcal{D}_N^{\mathrm{U}} \ : \ S_i = 0 \right\} \ ,$$

of size N_1 and N_0 respectively, such that $N_1 + N_0 = N$. Clearly we have $\mathcal{D}_{N_s} \overset{i.i.d.}{\sim} \mathbb{P}_{X|S=s}$ for each $s \in \{0,1\}$. Also recall that Remark 9 implies that neither N_0 nor N_1 are equal to zero, however, both are still random. Besides, denote by $\mathcal{D}_N^S = \left\{ S_i \ : \ (X_i, S_i) \in \mathcal{D}_N^{\mathrm{U}} \right\}$ the which is obtained from $\mathcal{D}_N^{\mathrm{U}}$ by removing features. Thus,

$$\mathbb{E}_{(\mathcal{D}_N^{\mathrm{U}})}(\star) \leq \mathbb{E}_{\mathcal{D}_N^S} \mathbb{E}_{\mathcal{D}_{N_1}} \sup_{t \in [-2/c_{n,N}, 2/c_{n,N}]} \left| (\mathbb{E}_{X|S=1} - \hat{\mathbb{E}}_{X|S=1}) \left((2\hat{\eta}(X,1)-1)\mathbb{P}(S=1) + t \right)_+ \right| \ .$$

Conditionally on $\mathcal{D}_N^S$ we can view N_0 and N_1 as fixed strictly positive integers, moreover, conditionally on $\mathcal{D}_n^{\mathrm{L}}$ the estimator $\hat{\eta}$ is not random as it is built *only* on $\mathcal{D}_n^{\mathrm{L}}$. Thus, we would like to control the following process

$$\mathbb{E}_{\mathcal{D}_{N_1}} \sup_{t \in [-2/c_{n,N}, 2/c_{n,N}]} \left| (\mathbb{E}_{X|S=1} - \hat{\mathbb{E}}_{X|S=1}) \left((2\hat{\eta}(X,1)-1)\mathbb{P}(S=1) + t \right)_+ \right| \ ,$$

conditionally on $\mathcal{D}_N^S, \mathcal{D}_n^{\mathrm{L}}$. First of all we rewrite this process as

$$\frac{1}{c_{n,N}} \mathbb{E}_{\mathcal{D}_{N_1}} \sup_{|t| \leq 1} \left| (\mathbb{E}_{X|S=1} - \hat{\mathbb{E}}_{X|S=1}) \left((2\hat{\eta}(X,1)-1)\mathbb{P}(S=1)c_{n,N} + 2t \right)_+ \right| \ .$$

Thanks to the standard symmetrization argument, which we recall in Theorem 19 of Appendix, we can write

$$\mathbb{E}_{\mathcal{D}_{N_1}} \sup_{|t| \leq 1} \left| (\mathbb{E}_{X|S=1} - \hat{\mathbb{E}}_{X|S=1}) \left((2\hat{\eta}(X,1)-1)\mathbb{P}(S=1)c_{n,N} + 2t \right)_+ \right|$$
$$\leq 2\mathbb{E}_{\mathcal{D}_{N_1}} \mathbb{E}_{\varepsilon} \sup_{|t| \leq 1} \left| \frac{1}{N_1} \sum_{X \in \mathcal{D}_{N_1}} \varepsilon_i f_t(X) \right| \ ,$$

where $f_t(\cdot) = ((2\hat{\eta}(\cdot, 1) - 1)\mathbb{P}(S = 1)c_{n,N} + 2t)_+$. Notice that for each $t, t' \in [-1, 1]$ we have for each $x \in \mathbb{R}^d$

$$|f_t(x) - f_{t'}(x)| \leq 2\,|t - t'| \quad ,$$

that is, the parametrization is 2-Lipschitz. Therefore, standard results in empirical processes (combine [Wellner, 2005, Lemma 6.2] with [Koltchinskii, 2011, Theorem 3.1]) tells us that there exists $C > 0$ such that

$$\mathbb{E}_\varepsilon \sup_{|t| \leq 1} \left| \frac{1}{N_1} \sum_{i=1}^{N_1} \varepsilon_i f_t(X_i) \right| \leq C \sqrt{\frac{1}{N_1}} \quad .$$

Let us briefly sketch the strategy to get the above bound, this result relies on the Dudley's entropy integral. First of all, we need to define a notion of covering number.

Definition 12 (Covering number)**.** *An ϵ-cover of a subset W of a pseudo-metric space (S, d) is a set $\widehat{W} \subset W$ such that for every $w \in W$ there is $\hat{w} \in \widehat{W}$ such that $d(w, \hat{w}) \leq \epsilon$. The ϵ-covering number of W is*

$$\mathcal{N}(\epsilon, W, d) := \min \left\{ \left| \widehat{W} \right| \, : \, \widehat{W} \text{ is an } \epsilon\text{-cover of } W \right\} \quad ,$$

where $\left|\widehat{W}\right|$ is the cardinal of the set $\hat{W}$.

Definition 13 (sub-Gaussian process)**.** *A stochastic process $w \mapsto Z_w$ with indexing set W is sub-Gaussian with respect to a pseudo-metric d on W, if for all $w, w' \in W$ and all $\lambda \in \mathbb{R}$ it holds that*

$$\mathbb{E} \exp\left(\lambda(Z_w - Z_{w'})\right) \leq \exp\left(\frac{\lambda^2 d^2(w, w')}{2}\right) \quad .$$

Now, we are ready to state the Dudley's entropy integral bound.

Theorem 12 (Dudley's entropy integral)**.** *Let $w \mapsto Z_w$ be zero-mean stochastic process that is sub-Gaussian* w.r.t. *a pseudo-metric d on the indexing set W. Then,*

$$\mathbb{E} \sup_{w \in W} Z_w \leq 8\sqrt{2} \int_0^\infty \sqrt{\log \mathcal{N}(\epsilon, W, d)}\, d\epsilon \quad .$$

Let us also point out that the integral can be taken until the diameter of W instead of ∞.

We would like to apply the result[10] above to the process

$$f(t, \mathcal{D}_{N_1}) \mapsto \langle \varepsilon, f(t, \mathcal{D}_{N_1}) \rangle = \sum_{i=1}^{N_1} \varepsilon_i f_t(X_i) \quad ,$$

with

$$f(t, \mathcal{D}_{N_1}) = (f_t(X_1), \ldots, f_t(X_{N_1}))^\top \in \mathbb{R}^{N_1} \quad ,$$
$$W = \left\{ f(t, \mathcal{D}_{N_1}) \in \mathbb{R}^{N_1} \, : \, t \in [-1, 1] \right\} \quad ,$$

[10] Same argument works for $f(t, \mathcal{D}_{N_1}) \mapsto -\langle \varepsilon, f(t, \mathcal{D}_{N_1})$

conditionally on $\mathcal{D}_{N_1}$. We can write

$$\sup_{t\in[-1,1]}\sum_{i=1}^{N_1}\varepsilon_i f_t(X_i) = \sup_{f(t,\mathcal{D}_{N_1})\in W}\langle\varepsilon, f(t,\mathcal{D}_{N_1})\rangle \ .$$

Conditionally on $\mathcal{D}_{N_1}$, the process $f(t,\mathcal{D}_{N_1}) \mapsto \langle\varepsilon, f(t,\mathcal{D}_{N_1})\rangle$ is centered and sub-Gaussian *w.r.t.* the Euclidian distance on $\mathbb{R}^{N_1}$, since ε_i's are the Rademacher variables. Moreover, since, as already established, the parametrization is 2-Lipschitz we have for all $t, t' \in [-1,1]$

$$\sum_{X\in\mathcal{D}_{N_1}} |f_t(X) - f_{t'}(X)|^2 \leq 4N_1 |t-t'|^2 \ .$$

Which implies that for all $\epsilon > 0$ we have

$$\mathcal{N}(\epsilon, W, \|\cdot\|_2) \leq \mathcal{N}\left(\epsilon/\left(2\sqrt{N_1}\right), [-1,1], \|\cdot\|_2\right) \leq \left(1 + 4\sqrt{N_1}/\epsilon\right) \ .$$

The Dudley's entropy integral applied to $f(t,\mathcal{D}_{N_1}) \mapsto \langle\varepsilon, f(t,\mathcal{D}_{N_1})\rangle$ gives us the desired result with $W = \left\{f(t,\mathcal{D}_{N_1}) \in \mathbb{R}^{N_1} : t \in [-1,1]\right\}$ and the above bound on the entropy[11]. This argument reads as

$$\begin{aligned}
\mathbb{E}_\varepsilon \sup_{f(t,\mathcal{D}_{N_1})\in W}\langle\varepsilon, f(t,\mathcal{D}_{N_1})\rangle &\leq 8\sqrt{2}\int_0^\infty \sqrt{\log\mathcal{N}(\epsilon/\left(2\sqrt{N_1}\right), [-1,1], \|\cdot\|_2)}d\epsilon \\
&\leq 16\sqrt{2}\sqrt{N_1}\int_0^2 \sqrt{\log\mathcal{N}(\epsilon, [-1,1], \|\cdot\|_2)}d\epsilon \\
&\leq 16\sqrt{2}\sqrt{N_1}\underbrace{\int_0^2 \sqrt{\log\left(1+\frac{2}{\epsilon}\right)}d\epsilon}_{\text{finite}} \ .
\end{aligned}$$

Finally, division by N_1 on both sides of the inequality yields the desired result.

Now, taking expectation *w.r.t.* $\mathcal{D}_N^s$ from $(\star)$ we get

$$\mathbb{E}_{(\mathcal{D}_N^{\mathrm{U}})}(\star) \leq \frac{C}{c_{n,N}}\mathbb{E}_{\mathcal{D}_N^s}\sqrt{\frac{1}{N_1}} \ ,$$

applying Lemma 7 we get for some $C > 0$ that depends on $\mathbb{P}(S=1)$ that

$$\mathbb{E}_{(\mathcal{D}_N^{\mathrm{U}})}(\star) \leq \frac{C}{c_{n,N}}\sqrt{\frac{1}{N}} \ .$$

Thanks to Assumption 10 we have

$$\frac{1}{c_{n,N}\sqrt{N}} = o(1) \ ,$$

thus, the term $\mathbb{E}_{(\mathcal{D}_N^{\mathrm{U}})}(\star)$ converges to zero. Repeating the same argument for $(\triangle\triangle)$ we conclude. □

[11] We also replaced the integral until ∞ by the integral until 2 as suggested in the formulation of the theorem.

Proofs of auxiliary results

Proof of Lemma 9. We start from the level of unfairness of g, that is, we would like to find an upper bound on

$$\left|\mathbb{P}\left(g(X,S)=1 \,|\, S=1, Y=1\right) - \mathbb{P}\left(g(X,S)=1 \,|\, S=0, Y=1\right)\right| \ ,$$

rewriting the expression above, our goal can be written as

$$\left|\frac{\mathbb{E}_{X|S=1}\eta(X,1)g(X,1)}{\mathbb{E}_{X|S=1}\eta(X,1)} - \frac{\mathbb{E}_{X|S=0}\eta(X,0)g(X,0)}{\mathbb{E}_{X|S=0}\eta(X,0)}\right| \ .$$

Now, we start working with the expression above

$$\begin{aligned}
&\left|\frac{\mathbb{E}_{X|S=1}\eta(X,1)g(X,1)}{\mathbb{E}_{X|S=1}\eta(X,1)} - \frac{\mathbb{E}_{X|S=0}\eta(X,0)g(X,0)}{\mathbb{E}_{X|S=0}\eta(X,0)}\right| \\
&\leq \left|\frac{\mathbb{E}_{X|S=1}\eta(X,1)g(X,1)}{\mathbb{E}_{X|S=1}\eta(X,1)} - \frac{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)g(X,1)}{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)}\right| \\
&\quad + \left|\frac{\mathbb{E}_{X|S=0}\hat{\eta}(X,0)g(X,0)}{\mathbb{E}_{X|S=0}\hat{\eta}(X,0)} - \frac{\mathbb{E}_{X|S=0}\eta(X,0)g(X,0)}{\mathbb{E}_{X|S=0}\eta(X,0)}\right| \\
&\quad + \left|\frac{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)g(X,1)}{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)} - \frac{\mathbb{E}_{X|S=0}\hat{\eta}(X,0)g(X,0)}{\mathbb{E}_{X|S=0}\hat{\eta}(X,0)}\right| \ .
\end{aligned}$$

The first two terms on the right hand side of the inequality can be upper-bounded in a similar way. That is why we only show the bound for the first term, that is, for $S=1$. We have for $(*) = \left|\frac{\mathbb{E}_{X|S=1}\eta(X,1)g(X,1)}{\mathbb{E}_{X|S=1}\eta(X,1)} - \frac{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)g(X,1)}{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)}\right|$

$$\begin{aligned}
(*) &\leq \frac{\mathbb{E}_{X|S=1}\left|\eta(X,1)-\hat{\eta}(X,1)\right|}{\mathbb{P}\left(Y=1\,|\,S=1\right)} + \left|\frac{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)g(X,1)}{\mathbb{E}_{X|S=1}\eta(X,1)} - \frac{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)g(X,1)}{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)}\right| \\
&\leq \frac{\mathbb{E}_{X|S=1}\left|\eta(X,1)-\hat{\eta}(X,1)\right|}{\mathbb{P}\left(Y=1\,|\,S=1\right)} \\
&\quad + \mathbb{E}_{X|S=1}\hat{\eta}(X,1)g(X,1)\left|\frac{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)}{\mathbb{E}_{X|S=1}\eta(X,1)\mathbb{E}_{X|S=1}\hat{\eta}(X,1)} - \frac{\mathbb{E}_{X|S=1}\eta(X,1)}{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)\mathbb{E}_{X|S=1}\eta(X,1)}\right| \\
&\leq \frac{\mathbb{E}_{X|S=1}\left|\eta(X,1)-\hat{\eta}(X,1)\right|}{\mathbb{P}\left(Y=1\,|\,S=1\right)} + \mathbb{E}_{X|S=1}\hat{\eta}(X,1)g(X,1)\frac{\mathbb{E}_{X|S=1}\left|\hat{\eta}(X,1)-\hat{\eta}(X,1)\right|}{\mathbb{E}_{X|S=1}\eta(X,1)\mathbb{E}_{X|S=1}\hat{\eta}(X,1)} \\
&\leq 2\frac{\mathbb{E}_{X|S=1}\left|\eta(X,1)-\hat{\eta}(X,1)\right|}{\mathbb{P}\left(Y=1\,|\,S=1\right)} \ ,
\end{aligned}$$

thus, we have

$$\begin{aligned}
&\left|\mathbb{P}\left(g(X,S)=1 \,|\, S=1, Y=1\right) - \mathbb{P}\left(g(X,S)=1 \,|\, S=0, Y=1\right)\right| \\
&\leq 2\frac{\mathbb{E}_{X|S=1}\left|\eta(X,1)-\hat{\eta}(X,1)\right|}{\mathbb{P}\left(Y=1\,|\,S=1\right)} \\
&\quad + 2\frac{\mathbb{E}_{X|S=0}\left|\eta(X,0)-\hat{\eta}(X,0)\right|}{\mathbb{P}\left(Y=1\,|\,S=0\right)} \\
&\quad + \left|\frac{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)g(X,1)}{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)} - \frac{\mathbb{E}_{X|S=0}\hat{\eta}(X,0)g(X,0)}{\mathbb{E}_{X|S=0}\hat{\eta}(X,0)}\right| \ .
\end{aligned}$$

Finally, it remains to upper bound

$$(**) = \left| \frac{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)g(X,1)}{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)} - \frac{\mathbb{E}_{X|S=0}\hat{\eta}(X,0)g(X,0)}{\mathbb{E}_{X|S=0}\hat{\eta}(X,0)} \right| .$$

Recall that $\hat{\mathbb{E}}_{X|S=1}$ and $\hat{\mathbb{E}}_{X|S=0}$ stands for the expectations taken *w.r.t.* empirical measure induced by $\mathcal{D}_N^{\mathrm{U}}$, and that $\mathcal{D}_N^{\mathrm{U}}$ is independent from $\mathcal{D}_n^{\mathrm{L}}$. Therefore, we can write

$$\begin{aligned}(**) \leq & \left| \frac{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)g(X,1)}{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)} - \frac{\hat{\mathbb{E}}_{X|S=1}\hat{\eta}(X,1)g(X,1)}{\hat{\mathbb{E}}_{X|S=1}\hat{\eta}(X,1)} \right| \\ & + \left| \frac{\mathbb{E}_{X|S=0}\hat{\eta}(X,0)g(X,0)}{\mathbb{E}_{X|S=0}\hat{\eta}(X,0)} - \frac{\hat{\mathbb{E}}_{X|S=0}\hat{\eta}(X,0)g(X,0)}{\hat{\mathbb{E}}_{X|S=0}\hat{\eta}(X,0)} \right| \\ & + \left| \frac{\hat{\mathbb{E}}_{X|S=1}\hat{\eta}(X,1)g(X,1)}{\hat{\mathbb{E}}_{X|S=1}\hat{\eta}(X,1)} - \frac{\hat{\mathbb{E}}_{X|S=0}\hat{\eta}(X,0)g(X,0)}{\hat{\mathbb{E}}_{X|S=0}\hat{\eta}(X,0)} \right| .\end{aligned}$$

Clearly, the last term on the right hand side of the previous inequality corresponds to our empirical criteria since everything can be easily evaluated using data. The first two terms on the right hand side of the inequality can be upper-bounded in a similar fashion, again, we only demonstrate the bound for $S = 1$. We can write

$$\begin{aligned}& \left| \frac{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)g(X,1)}{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)} - \frac{\hat{\mathbb{E}}_{X|S=1}\hat{\eta}(X,1)g(X,1)}{\hat{\mathbb{E}}_{X|S=1}\hat{\eta}(X,1)} \right| \\ & \leq \left| \frac{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)g(X,1)}{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)} - \frac{\hat{\mathbb{E}}_{X|S=1}\hat{\eta}(X,1)g(X,1)}{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)} \right| \\ & \quad + \left| \frac{\hat{\mathbb{E}}_{X|S=1}\hat{\eta}(X,1)g(X,1)}{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)} - \frac{\hat{\mathbb{E}}_{X|S=1}\hat{\eta}(X,1)g(X,1)}{\hat{\mathbb{E}}_{X|S=1}\hat{\eta}(X,1)} \right| .\end{aligned}$$

Notice that for the first term on the right hand side of the inequality we have

$$\left| \frac{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)g(X,1)}{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)} - \frac{\hat{\mathbb{E}}_{X|S=1}\hat{\eta}(X,1)g(X,1)}{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)} \right| \leq \frac{\left| \mathbb{E}_{X|S=1}\hat{\eta}(X,1)g(X,1) - \hat{\mathbb{E}}_{X|S=1}\hat{\eta}(X,1)g(X,1) \right|}{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)} ,$$

whereas for the second term we can write

$$\left| \frac{\hat{\mathbb{E}}_{X|S=1}\hat{\eta}(X,1)g(X,1)}{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)} - \frac{\hat{\mathbb{E}}_{X|S=1}\hat{\eta}(X,1)g(X,1)}{\hat{\mathbb{E}}_{X|S=1}\hat{\eta}(X,1)} \right| \leq \frac{\left| \hat{\mathbb{E}}_{X|S=1}\hat{\eta}(X,1) - \mathbb{E}_{X|S=1}\hat{\eta}(X,1) \right|}{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)} .$$

□

Proof of Lemma 10. Let us first introduce two slices of $\mathcal{D}_N^{\mathrm{U}}$ as

$$\mathcal{D}_{N_1} = \left\{ X_i \in \mathcal{D}_N^{\mathrm{U}} \,:\, S_i = 1 \right\}, \; \mathcal{D}_{N_0} = \left\{ X_i \in \mathcal{D}_N^{\mathrm{U}} \,:\, S_i = 0 \right\}$$

of size N_1 and N_0 respectively, such that $N_1 + N_0 = N$. Clearly we have $\mathcal{D}_{N_s} \overset{i.i.d.}{\sim} \mathbb{P}_{X|S=s}$ for each $s \in \{0, 1\}$. Besides, denote by $\mathcal{D}_N^S = \left\{ S_i \,:\, (X_i, S_i) \in \mathcal{D}_N^{\mathrm{U}} \right\}$ the which is obtained from $\mathcal{D}_N^{\mathrm{U}}$ by removing features. Recalling Remark 9, we have

$$N_1 - 2 \sim \mathrm{Bin}(N, \mathbb{P}(S = 1)), \quad N_0 - 2 \sim \mathrm{Bin}(N, \mathbb{P}(S = 0)) .$$

Clearly, since the proposed algorithm is a thresholding of $\hat{\eta}$ we have

$$\begin{aligned}&\mathbb{E}_{(\mathcal{D}_n^{\mathrm{L}},\mathcal{D}_N^{\mathrm{U}})}\left|(\mathbb{E}_{X|S=0}-\hat{\mathbb{E}}_{X|S=0})\hat{\eta}(X,0)\hat{g}(X,0)\right|\\&\leq \mathbb{E}_{(\mathcal{D}_n^{\mathrm{L}},\mathcal{D}_N^{\mathrm{U}})}\sup_{t\in[0,1]}\left|(\mathbb{E}_{X|S=0}-\hat{\mathbb{E}}_{X|S=0})\hat{\eta}(X,0)\mathbb{1}_{\{t\leq\hat{\eta}(X,0)\}}\right| \ .\end{aligned}$$

Further we work conditionally on $\mathcal{D}_n^{\mathrm{L}}$. Using the classical symmetrization technique [Koltchinskii, 2011, Theorem 2.1.] we get

$$\begin{aligned}&\mathbb{E}_{\mathcal{D}_N^{\mathrm{U}}}\sup_{t\in[0,1]}\left|(\mathbb{E}_{X|S=0}-\hat{\mathbb{E}}_{X|S=0})\hat{\eta}(X,0)\mathbb{1}_{\{t\leq\hat{\eta}(X,0)\}}\right|\\&=\mathbb{E}_{\mathcal{D}_N^S}\mathbb{E}_{\mathcal{D}_{N_0}}\sup_{t\in[0,1]}\left|(\mathbb{E}_{X|S=0}-\hat{\mathbb{E}}_{X|S=0})\hat{\eta}(X,0)\mathbb{1}_{\{t\leq\hat{\eta}(X,0)\}}\right|\\&\leq 2\mathbb{E}_{\mathcal{D}_N^S}\mathbb{E}_{\mathcal{D}_{N_0}}\mathbb{E}_{\varepsilon}\sup_{t\in[0,1]}\left|\frac{1}{N_0}\sum_{X_i\in\mathcal{D}_{N_0}}\varepsilon_i\hat{\eta}(X_i,0)\mathbb{1}_{\{t\leq\hat{\eta}(X_i,0)\}}\right| \ ,\end{aligned}$$

where $\varepsilon_i \overset{i.i.d.}{\sim}$ Rademacher variables. Note that the function class $x\mapsto\mathbb{1}_{\{t\leq\hat{\eta}(x,0)\}}$ has VC-dimension [Vapnik and Chervonenkis, 1971] equal to one. At this moment we will work with

$$\mathbb{E}_{\varepsilon}\sup_{t\in[0,1]}\left|\frac{1}{N_0}\sum_{X_i\in\mathcal{D}_{N_0}}\varepsilon_i\hat{\eta}(X_i,0)\mathbb{1}_{\{t\leq\hat{\eta}(X_i,0)\}}\right| \ ,$$

conditionally on all the data. First of all let us introduce $\mathcal{F}=\left\{f\,:\,\exists t\in[0,1],\,f(x)=\mathbb{1}_{\{t\leq\hat{\eta}(x,0)\}}\right\}$ Thus, our process can be written as

$$\mathbb{E}_{\varepsilon}\sup_{f\in\mathcal{F}}\left|\frac{1}{N_0}\sum_{X_i\in\mathcal{D}_{N_0}}\varepsilon_i\varphi_i(f(X_i))\right| \ ,$$

where $\varphi_i(\cdot)=\eta(X_i,0)\times\cdot$. Clearly, we have $\varphi_i(0)=0$ and for every u,v

$$|\varphi_i(u)-\varphi_i(v)|\leq|u-v| \ .$$

That is, φ_i are contractions, and the contraction theorem [Koltchinskii, 2011, Theorem 2.2.], recalled in Theorem 20 of Appendix, gives

$$\mathbb{E}_{\varepsilon}\sup_{f\in\mathcal{F}}\left|\frac{1}{N_0}\sum_{X_i\in\mathcal{D}_{N_0}}\varepsilon_i\varphi_i(f(X_i))\right|\leq\mathbb{E}_{\varepsilon}\sup_{f\in\mathcal{F}}\left|\frac{1}{N_0}\sum_{X_i\in\mathcal{D}_{N_0}}\varepsilon_i f(X_i)\right| \ .$$

Recall, that the class $\mathcal{F}$ is a VC-class with VC-dimension equal to one. Therefore, it is a known fact [Dvoretzky et al., 1956, Massart, 1990] that there exists $C>0$ such that

$$\mathbb{E}_{\varepsilon}\sup_{f\in\mathcal{F}}\left|\frac{1}{N_0}\sum_{X_i\in\mathcal{D}_{N_0}}\varepsilon_i f(X_i)\right|\leq C\sqrt{\frac{1}{N_0}} \ ,$$

almost surely. The above implies that

$$\mathbb{E}_{\mathcal{D}_N^{\mathrm{U}}}\sup_{t\in[0,1]}\left|(\mathbb{E}_{X|S=0}-\hat{\mathbb{E}}_{X|S=0})\hat{\eta}(X,0)\mathbb{1}_{\{t\leq\hat{\eta}(X,0)\}}\right|\leq C\mathbb{E}_{\mathcal{D}_N^S}\sqrt{\frac{1}{N_0}} \ .$$

The result above is in some sense a non-asymptotic version of the Glivenko-Cantelli theorem for the supremum of the deviation of empirical cumulative distribution function from its true value.

It remains to provide an upper bound on $\mathbb{E}_{\mathcal{D}_N^S}\sqrt{\frac{1}{N_0}}$, to this end we recall that this expectation can be written as

$$\mathbb{E}\sqrt{\frac{1}{2+Z}} \ ,$$

where Z is the binomial random variable with parameters N and $\mathbb{P}(S=0)$. Thus, thanks to Lemma 7 there exists a constant $C>0$ that depends on $\mathbb{P}(S=0)$ such that

$$\mathbb{E}\sqrt{\frac{1}{2+Z}} \leq C\sqrt{\frac{1}{N}} \ .$$

Similarly we get the bound for the case $S=1$. □

Optimal classifier without sensitive feature

Proof of Proposition 3. Let us study the following minimization problem

$$(*) := \min_{g\in\mathcal{G}} \left\{\mathcal{R}(g) \ : \ \mathbb{P}\left(g(X)=1 \,|\, Y=1, S=1\right) = \mathbb{P}\left(g(X)=1 \,|\, Y=1, S=0\right)\right\} \ .$$

Using the weak duality we can write

$$\begin{aligned}(*) &= \min_{g\in\mathcal{G}} \max_{\lambda\in\mathbb{R}} \left\{\mathcal{R}(g) + \lambda\left(\mathbb{P}\left(g(X)=1 \,|\, Y=1, S=1\right) - \mathbb{P}\left(g(X)=1 \,|\, Y=1, S=0\right)\right)\right\} \\ &\geq \max_{\lambda\in\mathbb{R}} \min_{g\in\mathcal{G}} \left\{\mathcal{R}(g) + \lambda\left(\mathbb{P}\left(g(X)=1 \,|\, Y=1, S=1\right) - \mathbb{P}\left(g(X)=1 \,|\, Y=1, S=0\right)\right)\right\} \\ &=: (**) \ .\end{aligned}$$

We first study the objective function of the max min problem $(**)$, which is equal to

$$\mathbb{P}(g(X)\neq Y) + \lambda\left(\mathbb{P}\left(g(X)=1 \,|\, Y=1, S=1\right) - \mathbb{P}\left(g(X)=1 \,|\, Y=1, S=0\right)\right) \ .$$

Using arguments of Lemma 8 we can write

$$\mathbb{P}(g(X)\neq Y) = \mathbb{P}(Y=1) - \mathbb{E}_X[(2\eta(X)-1)g(X)] \ ,$$

where $\eta(\cdot) := \mathbb{P}(Y=1|X=\cdot)$. Moreover, since

$$\mathbb{E}[YS] = \mathbb{E}_S[S\mathbb{E}[Y|S]] = \mathbb{E}_S[S\mathbb{E}_X[\mathbb{E}[Y|X,S]]] = \mathbb{E}_S[S\mathbb{E}_X[\eta(X,S)]] = \mathbb{P}(S=1)\mathbb{E}_X[\eta(X,1)] \ ,$$

we can write for the rest

$$\begin{aligned}\mathbb{P}\left(g(X)=1 \,|\, Y=1, S=1\right) &= \frac{\mathbb{P}\left(g(X)=1, Y=1, S=1\right)}{\mathbb{P}\left(Y=1, S=1\right)} = \frac{\mathbb{E}[g(X)YS]}{\mathbb{E}[YS]} \\ &= \frac{\mathbb{P}(S=1)\mathbb{E}_X[g(X)\eta(X,1)]}{\mathbb{P}(S=1)\mathbb{E}_X[\eta(X,1)]} = \frac{\mathbb{E}_X[g(X)\eta(X,1)]}{\mathbb{E}_X[\eta(X,1)]} \\ \mathbb{P}\left(g(X)=1 \,|\, Y=1, S=0\right) &= \frac{\mathbb{P}\left(g(X)=1, Y=1, S=0\right)}{\mathbb{P}\left(Y=1, S=0\right)} = \frac{\mathbb{E}[g(X)Y(1-S)]}{\mathbb{E}[Y(1-S)]} \\ &= \frac{\mathbb{E}_X[g(X)\eta(X,0)]}{\mathbb{E}_X[\eta(X,0)]} \ .\end{aligned}$$

Using these, the objective of $(**)$ can be simplified as

$$\mathbb{P}(Y=1) - \mathbb{E}_X\left[g(X)\left(2\eta(X) - 1 + \lambda\left(\frac{\eta(X,0)}{\mathbb{E}_X[\eta(X,0)]} - \frac{\eta(X,1)}{\mathbb{E}_X[\eta(X,1)]}\right)\right)\right] \ .$$

Clearly, for every $\lambda \in \mathbb{R}$ a minimizer g^*_λ of the problem $(**)$ can be written for all $x \in \mathbb{R}^d$ as

$$g^*_\lambda(x) = \mathbb{1}_{\left\{2\eta(x)-1+\lambda\left(\frac{\eta(x,0)}{\mathbb{E}_X[\eta(X,0)]} - \frac{\eta(x,1)}{\mathbb{E}_X[\eta(X,1)]}\right)\geq 0\right\}} \ .$$

Similarly to Proposition 2, for $\lambda = 0$ we recover the classical optimal predictor in the context of binary classification. Substituting this classifier into the objective of $(**)$ we arrive at

$$(**) = \mathbb{P}(Y=1) - \min_{\lambda\in\mathbb{R}}\left\{\mathbb{E}_X\left(2\eta(X) - 1 + \lambda\left(\frac{\eta(X,0)}{\mathbb{E}_X[\eta(X,0)]} - \frac{\eta(X,1)}{\mathbb{E}_X[\eta(X,1)]}\right)\right)_+\right\} \ .$$

The mapping

$$\lambda \mapsto \mathbb{E}_X\left(2\eta(X) - 1 + \lambda\left(\frac{\eta(X,0)}{\mathbb{E}_X[\eta(X,0)]} - \frac{\eta(X,1)}{\mathbb{E}_X[\eta(X,1)]}\right)\right)_+ \ ,$$

is convex, therefore we can write the first order optimality conditions as

$$0 \in \partial_\lambda \mathbb{E}_X\left(2\eta(X) - 1 + \lambda\left(\frac{\eta(X,0)}{\mathbb{E}_X[\eta(X,0)]} - \frac{\eta(X,1)}{\mathbb{E}_X[\eta(X,1)]}\right)\right)_+ \ .$$

Clearly, under continuity assumption this subgradient is reduced to the gradient almost surely, thus we have the following condition on the optimal value of λ^*

$$\frac{\mathbb{E}_X\left[\eta(X,1)g^*_{\lambda^*}(X)\right]}{\mathbb{E}_X[\eta(X,1)]} = \frac{\mathbb{E}_X\left[\eta(X,0)g^*_{\lambda^*}(X)\right]}{\mathbb{E}_X[\eta(X,0)]} \ ,$$

and the pair $(\lambda^*, g^*_{\lambda^*})$ is a solution of the dual problem $(**)$. Notice that the previous condition can be written as

$$\mathbb{P}\left(g^*_{\lambda^*}(X) = 1 \,|\, Y=1, S=1\right) = \mathbb{P}\left(g^*_{\lambda^*}(X) = 1 \,|\, Y=1, S=0\right) \ .$$

This implies that the classifier $g^*_{\lambda^*}$ is fair. Finally, it remains to show that $g^*_{\lambda^*}$ is actually an optimal classifier, indeed, since $g^*_{\lambda^*}$ is fair we can write on the one hand

$$\mathcal{R}(g^*_{\lambda^*}) \geq \min_{g\in\mathcal{G}}\left\{\mathcal{R}(g) \,:\, \mathbb{P}\left(g(X)=1 \,|\, Y=1, S=1\right) = \mathbb{P}\left(g(X)=1 \,|\, Y=1, S=0\right)\right\} = (*).$$

On the other hand the pair $(\lambda^*, g^*_{\lambda^*})$ is a solution of the dual problem $(**)$, thus we have

$$\begin{aligned}(*) &\geq \mathcal{R}(g^*_{\lambda^*}) + \lambda^*\left(\mathbb{P}\left(g^*_{\lambda^*}(X)=1 \,|\, Y=1, S=1\right) - \mathbb{P}\left(g^*_{\lambda^*}(X)=1 \,|\, Y=1, S=0\right)\right) \\ &= \mathcal{R}(g^*_{\lambda^*}) \ .\end{aligned}$$

It implies that the classifier $g^*_{\lambda^*}$ is optimal, hence $g^* \equiv g^*_{\lambda^*}$. □

Chapter 3

Multi-class classification

3.1 Confidence set approach

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

3.1.1 Introduction

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

Problem statement

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

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

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

respectively, where $\mathbb{E}_{\mathbb{P}_X}$ stands for the expectation *w.r.t.* the marginal distribution of $X \in \mathbb{R}^d$ and $|\Gamma(x)|$ is the cardinal of Γ at $x \in \mathbb{R}^d$. In this part we write $\mathrm{P}(\cdot)$ instead of $\mathcal{R}(\cdot)$ in order to stress that the quantity $\mathrm{P}(\cdot)$ is not viewed as a risk. Let us recall that, as discussed in Section 1.2.2, the information $\mathrm{I}(\cdot)$ of a confidence set Γ is one of possible ways to measure the size of the confidence set.

In this chapter we study the following instance of the constrained classification framework. For a fixed integer $\beta \in [K]$ a β-Oracle confidence set Γ^*_β is defined as

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

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

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

Assumption 3.1.1 (Continuity of CDF)**.** *For all $k \in [K]$ the cumulative distribution function (CDF) $F_{p_k}(\cdot) := \mathbb{P}_X(p_k(X) \leq \cdot)$ of $p_k(X)$ is continuous on $(0, 1)$.*

Recall that similar continuity assumptions were introduced in Sections 2.2 and 1.1.3 in the contexts of fairness and general constrained binary classification. The next result is a confidence set analogue of Lemma 1, which gave an explicit expression for the Bayes classifier in general constrained binary classification framework.

Proposition 3.1.2 (β-Oracle confidence set)**.** *Fix $\beta \in [K-1]$, 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 3.1.1 a β-Oracle confidence set Γ^*_β can be obtained as

$$\Gamma^*_\beta(x) = \left\{k \in [K] : p_k(x) \geq G^{-1}(\beta)\right\} , \tag{3.1}$$

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

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

Proposition 3.1.3. *Assume that Assumption 3.1.1 is fulfilled, then the β-Oracle defined in Eq. (3.1) is a minimizer of the following risk*

$$\mathcal{R}_\beta(\Gamma) = \mathrm{P}(\Gamma) + G^{-1}(\beta)\, \mathrm{I}(\Gamma) . \tag{3.2}$$

These propositions have been proven in [Denis and Hebiri, 2017, Proposition 4 and Proposition 7], this type of results were already discussed in Chapter 1 (see example from of Eq. (1.1) and the discussion after) where we introduced the problem of constrained classification. Consequently, the accuracy of a confidence set Γ can be for instance quantified according to its excess risk

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

The statistical learning problem is then to estimate Γ^*_β given the data sample $\mathcal{D}_n^{\mathrm{L}}$ and $\mathcal{D}_N^{\mathrm{U}}$. The formulation in Eq. (3.1) of the β-Oracle appears to be closely related to the level set estimation problem [Hartigan, 1987, Polonik, 1995, Tsybakov, 1997, Rigollet and Vert, 2009]. In this setup the estimation of the β-Oracle does not only rely on the regression function but also on the threshold $G^{-1}(\beta)$ which is, as usual, *unknown* beforehand and can be estimated in a semi-supervised way [Denis and Hebiri, 2017]. To better explain these ideas, we give some examples of possible estimation procedures of Γ^*_β.

Confidence set estimators

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

Definition 14 (Supervised and semi-supervised estimators)**.** *A measurable mapping*

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

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

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

Otherwise the estimator is called semi-supervised*. In the sequel, similarly to Chapter 1, for the simplicity of notation we write $\hat{\Gamma}(x)$ instead of $\hat{\Gamma}(x; \mathcal{D}_n^{\mathrm{L}}, \mathcal{D}_N^{\mathrm{U}})$ if no ambiguity is present.*

Intuitively, the supervised estimators do not take into account the information that is provided by the unlabeled sample. Besides, if we denote by $\hat{\Upsilon}$ the set of all estimators, Definition 14 generates a natural partition of $\hat{\Upsilon}$ into two disjoint sets: the *supervised* estimators $\hat{\Upsilon}_{\mathrm{SE}}$ and the *semi-supervised* estimators $\hat{\Upsilon}_{\mathrm{SSE}}$.

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

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

$$\hat{\Gamma}_{\mathrm{top}}(x) = \{\sigma_1(x), \ldots, \sigma_\beta(x)\}, \qquad \forall x \in \mathbb{R}^d \ . \tag{3.3}$$

- *Supervised procedure.* Formally, in this type of methods, we only care about $\mathcal{D}_n^{\mathrm{L}}$ (we forget about $\mathcal{D}_N^{\mathrm{U}}$). We split $\mathcal{D}_n^{\mathrm{L}}$ into two independent samples such that $\mathcal{D}_n^{\mathrm{L}} = \mathcal{D}_{\lfloor n/2 \rfloor}^{\mathrm{L}} \bigcup \mathcal{D}_{\lceil n/2 \rceil}^{\mathrm{L}}$. Consequently, we artificially forget about labels in $\mathcal{D}_{\lceil n/2 \rceil}^{\mathrm{L}}$ and construct $\mathcal{D}_{\lceil n/2 \rceil}^{\mathrm{U}}$ which only consists of feature vectors from $\mathcal{D}_{\lceil n/2 \rceil}^{\mathrm{L}}$. Based on the first sample $\mathcal{D}_{\lfloor n/2 \rfloor}^{\mathrm{L}}$, we consider an estimator $\hat{p}$ of the regression function p. Furthermore, we define

$$\hat{G}(\cdot) = \frac{1}{\lceil n/2 \rceil} \sum_{i \in \mathcal{D}_{\lceil n/2 \rceil}^{\mathrm{U}}} \sum_{k=1}^{K} \mathbb{1}_{\{\{\hat{p}_k(X_i) \geq \cdot\}\}} \ ,$$

and one type of supervised estimator is then defined as follows

$$\hat{\Gamma}_{\mathrm{SE}}(x) = \left\{ k \in [K] : \ \hat{p}_k(x) \geq \hat{G}^{-1}(\beta) \right\}, \qquad \forall x \in \mathbb{R}^d \ . \tag{3.4}$$

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

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

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

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

$$\hat{\Gamma}_{\mathrm{SSE}}(x) = \left\{ k \in [K] : \hat{p}_k(x) \geq \hat{G}^{-1}(\beta) \right\}, \qquad \forall x \in \mathbb{R}^d \ . \tag{3.5}$$

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

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

Q.1: The first question is the statistical performance of these plug-in procedures which is assessed through rates of convergence and their optimality in the minimax sense.

Q.2: The second question focuses on the benefit of the semi-supervised approach. Roughly speaking, are there situations where the semi-supervised approach outperforms the supervised one and how can it be quantified?

Q.3: The third question concentrates on the reason why it is more relevant for this problem to consider more involved estimators than the simple top-β method.

Minimax estimation

As discussed in the introduction of Chapter 1 there is no unique way to address performance of an estimator $\hat{\Gamma}$, in this part we consider three notions of excess risks defined below. For a given family $\mathcal{P}$ of joint distributions on $\mathbb{R}^d \times [K]$, a given estimator $\hat{\Gamma} \in \hat{\Upsilon}$, and fixed integers $K \geq 2$, $\beta \in [K]$, $n, N \in \mathbb{N}$ we are interested in the following maximal risks of $\hat{\Gamma}$

$$\begin{aligned}
\mathcal{E}^{\mathrm{H}}_{n,N}(\hat{\Gamma}; \mathcal{P}) &:= \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{(\mathcal{D}^{\mathrm{L}}_n, \mathcal{D}^{\mathrm{U}}_N)} \mathbb{E}_{\mathbb{P}_X} \left| \hat{\Gamma}(X) \triangle \Gamma^*_\beta(X) \right| && \text{(Hamming risk)} \ , \\
\mathcal{E}^{\mathrm{E}}_{n,N}(\hat{\Gamma}; \mathcal{P}) &:= \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{(\mathcal{D}^{\mathrm{L}}_n, \mathcal{D}^{\mathrm{U}}_N)} \mathcal{R}_\beta(\hat{\Gamma}) - \mathcal{R}_\beta(\Gamma^*_\beta) && \text{(excess risk)} \ , \\
\mathcal{E}^{\mathrm{D}}_{n,N}(\hat{\Gamma}; \mathcal{P}) &:= \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{(\mathcal{D}^{\mathrm{L}}_n, \mathcal{D}^{\mathrm{U}}_N)} \left[\left| \mathrm{P}(\hat{\Gamma}) - \mathrm{P}(\Gamma^*_\beta) \right| + \left| \beta - \mathrm{I}(\hat{\Gamma}) \right| \right] && \text{(discrepancy)} \ ,
\end{aligned}$$

where $\mathbb{E}_{(\mathcal{D}^{\mathrm{L}}_n, \mathcal{D}^{\mathrm{U}}_N)}$ denotes the expectation *w.r.t.* $\mathbb{P}^{\otimes n} \otimes \mathbb{P}_X^{\otimes N}$. These maximal risks are arising in a natural way in the context of confidence set estimation with controlled expected size. The risk $\mathcal{E}^{\mathrm{H}}_{n,N}(\hat{\Gamma}; \mathcal{P})$ corresponds to the estimation of the β-Oracle through the Hamming distance. The second risks is directly connected with Proposition 3.1.2, which gives a description of the β-Oracle as a minimizer of $\mathcal{R}_\beta(\cdot)$. An intuitive goal in this setup is to construct a procedure $\hat{\Gamma}$ that exhibits low error $|\mathrm{P}(\hat{\Gamma}) - \mathrm{P}(\Gamma^*_\beta)|$ and low cardinal $|\beta - \mathrm{I}(\hat{\Gamma})|$ discrepancies. Thus, it is natural to consider $\mathcal{E}^{\mathrm{D}}_{n,N}(\hat{\Gamma}; \mathcal{P})$ which is composed of both.

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

Definition 15 (Minimax rate of convergence)**.** *For a given family $\mathcal{P}$ of joint distributions on $\mathbb{R}^d \times [K]$ and a given family of estimators $\hat{\mathbf{\Gamma}} \subset \hat{\Upsilon}$ the minimax rates are defined as*

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

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

The main families of estimators that we study are the *supervised* $\hat{\Upsilon}_{\mathrm{SE}}$ and the *semi-supervised* $\hat{\Upsilon}_{\mathrm{SSE}}$ estimators. Obviously, since $\hat{\Upsilon} = \hat{\Upsilon}_{\mathrm{SE}} \bigcup \hat{\Upsilon}_{\mathrm{SSE}}$ and $\hat{\Upsilon}_{\mathrm{SE}} \bigcap \hat{\Upsilon}_{\mathrm{SSE}} = \emptyset$, we have the following relation

$$\mathcal{E}^{\square}_{n,N}(\hat{\Upsilon}; \mathcal{P}) = \min \left\{ \mathcal{E}^{\square}_{n,N}(\hat{\Upsilon}_{\mathrm{SE}}; \mathcal{P}) \ , \ \mathcal{E}^{\square}_{n,N}(\hat{\Upsilon}_{\mathrm{SSE}}; \mathcal{P}) \right\} \ .$$

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

Related work

Confidence set approach for classification was pioneered by Vovk [2002b,a], Vovk et al. [2005] by the means of conformal prediction theory. These authors rely on non-conformity measures which are based on some pattern recognition methods, and develop an asymptotic theory. In this chapter, we consider a statistical perspective of confidence set classification and put our focus on non-asymptotic minimax theory.

The problem of confidence set multi-class classification has strong ties with the binary classification with reject option (see Example 1.0.3), also known as binary classification with

abstention in machine learning literature. In the binary classification with rejection, a classifier is allowed to output some special symbol, which indicates the rejection. Such type of classifiers can be seen as confidence sets, which are allowed to output $\emptyset$ or $\{0,1\}$ and are interpreted as reject. This line of research was initiated by Chow [1957, 1970] in the context of information retrieval, where a predefined cost of rejection was considered. An extensive statistical study of this framework was carried in [Herbei and Wegkamp, 2006, Bartlett and Wegkamp, 2008, Wegkamp and Yuan, 2011]. Instead of considering a fixed cost for rejection, which might be too restrictive, one may define two entities: probability of rejection and the probability of misclassification. In the spirit of conformal prediction, Lei [2014] aims at minimizing the probability rejection provided a fixed upper bound on the probability of misclassification. In contrast, Denis and Hebiri [2015a] consider a reversed problem of minimizing the probability of misclassification given a fixed upper bound on the probability of rejection.

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

As for the confidence set approach there are again two possibilities, similar to the binary case. The one that is considered in this chapter was proposed by Denis and Hebiri [2017], where the authors analyse an ERM algorithm and derive oracle inequalities under the margin assumption [Tsybakov, 2004]. More specifically, they consider a convex surrogate of the error $\mathrm{P}(\cdot)$ which relies on a convex real valued loss function ϕ. For a suitable choice of the convex function ϕ they show that, under Assumption 3.1.1, their β-Oracle satisfies

$$\Gamma^*_\beta(\cdot) = \left\{k \in [K] \,:\, f^*_k(\cdot) \geq G^{-1}_{f^*}(\beta)\right\} \quad ,$$

where the vector score function f^* depends on ϕ and the threshold $G^{-1}_{f^*}(\beta)$ is defined similarly to the present manuscript. They propose a two-step estimation procedure of the β-Oracle set based on the ERM algorithm. They first estimate f^* and in the second step they estimate the threshold $G^{-1}_{f^*}(\beta)$ with an unlabeled sample. This procedure is in the same spirit as the semi-supervised procedure (3.5). Furthermore, under mild assumptions, they provide an upper bound on the excess risk and obtain a rate of convergence of order $(n/\log n)^{-\alpha/(\alpha+s)} + N^{-1/2}$, with s being a parameter that depends on the function ϕ and α being the margin parameter. Note that this rate is slower than the rate obtained in the standard classification framework.

Another point of view is rising from the conformal prediction theory [Vovk et al., 2005] which suggests to minimize the information level with a fixed budget on the error level. Statistical properties of this framework were considered in the work of Sadinle et al. [2018]. Their objective is formulated for some $a \in (0,1)$ as

$$\Gamma^*_a \in \arg\min \left\{\mathrm{I}(\Gamma) \,:\, \Gamma \in \Upsilon \;\; s.t. \;\; \mathrm{P}(\Gamma) \leq a\right\} \quad ,$$

and such a confidence set is called a least ambiguous confidence set with bounded error rate. The authors show that under Assumption 3.1.1 this oracle set can be described as a thresholding of the regression function

$$\Gamma^*_a(\cdot) = \left\{k \in [K] \,:\, p_k(\cdot) \geq t_a\right\} \quad ,$$

where the threshold t_a is defined as

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

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

Interestingly, both approaches can be encompassed into the constrained estimation framework [Anbar, 1977, Lepskii, 1990, Brown and Low, 1996], where one would like to construct an estimator with some prescribed properties. These properties are typically reflected by the form of the risk which in our case is the discrepancy measure, that is, the sum of error and information discrepancies. Thus, both frameworks of Sadinle et al. [2018], Denis and Hebiri [2017] can be seen as an extension of the constrained estimation to the classification problems. From the modeling point of view, we believe that the two frameworks can co-exist nicely and a particular choice depends on the considered application. The major difference between the present work and those by Denis and Hebiri [2017] and Sadinle et al. [2018] is the minimax analysis which we provide here and our treatment of semi-supervised techniques.

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

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

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

This discussion would not be complete without classical results on binary classification, as it is directly related to our confidence set setup. As it is mentioned in Section 1.1, non-parametric setting of this problem has been widely studied in literature, with first minimax analysis provided by Yang [1999] and later specified by Audibert and Tsybakov [2007]. Recall, that Audibert and Tsybakov [2007] derive fast rates of convergence for plug-in classifiers based on local polynomial estimators [Stone, 1977, Tsybakov, 1986, Audibert and Tsybakov, 2007] and show their optimality in the minimax sense. One of the aim of present work is to extend these results to the confidence set classification framework.

Another part of our work is to provide a comparison between supervised and semi-supervised procedures. Semi-supervised methods are studied in several papers [Vapnik, 1998,

Rigollet, 2007, Singh et al., 2009, Bellec et al., 2018] and references therein. A simple intuition can be provided on whether one should or not expect a superior performance of the semi-supervised approach. Imagine a situation when the unlabeled sample $\mathcal{D}_N^{\mathrm{U}}$ is so large that one can approximate $\mathbb{P}_X$ up to any desired precision, then, if the optimal decision is independent of $\mathbb{P}_X$, the semi-supervised estimators are not to be considered superior over the supervised estimation. This is the case in a lot of classical problems of statistics, where the inference is solely governed by the behavior of the conditional distribution $\mathbb{P}_{Y|X}$ (for instance regression or binary classification). The situation might be different once the optimal decision relies on the marginal distribution $\mathbb{P}_X$. In this case, as suggested by our findings, the semi-supervised approach might or not outperform the supervised one even in the context of the same problem. Similar conclusions were stated by Singh et al. [2009] in the context of learning under the cluster assumption [Rigollet, 2007].

3.1.2 Main contributions

Bellow we summarize main contributions of this chapter.

- Results of this chapter focus on the case where the regression function p belongs to a Hölder class and satisfy the margin condition. Under these assumptions, we establish lower bounds on the minimax rates, defined in Section 3.1.1 in the confidence set framework.

- As important case study, we first show that top-β type procedures are in general inconsistent. Furthermore, by providing a rigorous definition of the semi-supervised and supervised estimators, we describe the situations when the semi-supervised estimation should be considered superior to its supervised counterpart. Interestingly, analysis introduced in this chapter suggests that these regimes are governed by the interplay of the family of distributions and by the considered measure of performance. Besides, we show that in our settings supervised procedures cannot achieve fast rates, that is, their rates cannot be faster than $n^{-1/2}$. In contrast, recall that, as mentioned in Chapter 1, some other classical settings [Audibert and Tsybakov, 2007, Rigollet and Vert, 2009, Herbei and Wegkamp, 2006] allow to achieve faster rates for supervised methods. Moreover, as we have already seen in Section 2.1 in the setup of binary classification with F-score supervised methods *can* achieve fast rates.

- We provide supervised and semi-supervised estimation procedures, which are optimal or optimal up to an extra logarithmic factor. Importantly, our results show that semi-supervised a plug-in procedure based on local polynomial estimators can achieve fast rates, provided that the size of the unlabeled samples is large enough.

- Finally, we perform a numerical evaluation of the proposed plug-in algorithms against the top-β counterparts. This part supports our theoretical results and empirically demonstrates the reason to consider more involved procedures.

Organization of the chapter

The chapter is organized as follow. In Section 3.1.3, we put some additional notation and introduce the family of distributions $\mathcal{P}$ that we consider. Section 3.1.4 is devoted to the lower bounds on the minimax rates and their implications. In Section 3.1.5 we introduce the

proposed algorithm, establish upper bounds for it, and evaluate its numerical performance. We conclude this chapter by Sections 3.1.6 and 3.1.7 where we discuss and sum-up our results.

3.1.3 Class of confidence sets

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

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

In this chapter the constant $C > 0$ or its lower-cased versions always refer to some constant which might differ from line to line. Importantly, all these constants are independent of n, N but could depend on K, d and other parameters which are assumed to be fixed. Before introducing the families of distributions $\mathcal{P}$ that are considered in this chapter we need the following definitions.

Assumption 3.1.4 (α-margin assumption)**.** *We say that the distribution $\mathbb{P}$ of the pair $(X, Y) \in \mathbb{R}^d \times [K]$ satisfies α-margin assumption if there exists $C_1 > 0$ and $t_0 \in (0, 1)$ such that for every positive $t \leq t_0$*

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

Let us point out an important consequence of Assumption 3.1.1. We have that the condition

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

for all $t \in [0, t_0]$ is equivalent to Assumption 3.1.4. Indeed, the random variables $p_k(X)$'s cannot concentrate at a constant level and in particular at $G^{-1}(\beta)$. Moreover, again due to the continuity Assumption 3.1.1 we have

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

thus the α-margin Assumption 3.1.4 specifies the rate of this convergence. Finally, similarly to the F-score setup discussed in Section 2.1, the restriction of the range of t to $[0, t_0]$ in α-margin Assumption 3.1.4 does not affect its global behavior as for all $t \in [0, 1]$

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

For convenience of the reader, here we recall standard non-parametric assumptions which we have already used in Section 2.1. 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

$$\mathrm{Leb}\left(A \cap \mathrm{Ball}(x, r)\right) \geq c_0 \,\mathrm{Leb}\left(\mathrm{Ball}(x, r)\right), \qquad \forall r \in (0, r_0], \;\; \forall x \in A \;\; .$$

Definition 16 (Strong density). *We say that the probability measure $\mathbb{P}_X$ on $\mathbb{R}^d$ satisfies the $(\mu_{\min}, \mu_{\max}, c_0, r_0)$-strong density assumption if it is supported on a compact (c_0, r_0)-regular set $A \subset \mathbb{R}^d$ and has a density μ w.r.t. the Lebesgue measure such that $\mu(x) = 0$ for all $x \in \mathbb{R}^d \setminus A$ and*

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

Definition 17 (Hölder class, Tsybakov [2009]). *We say that a function $h : \mathbb{R}^d \to \mathbb{R}$ is (γ, L)-Hölder for $\gamma > 0$ and $L > 0$ if h is $\lfloor \gamma \rfloor$ times continuously differentiable and $\forall x, x' \in \mathbb{R}^d$ we have*

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

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

Definition 18. *We denote by $\mathcal{P}(L, \gamma, \alpha)$ a set of joint distributions on $\mathbb{R}^d \times [K]$ which satisfies the following conditions*

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

The family of distributions $\mathcal{P}(L, \gamma, \alpha)$ is also similar to the one considered in [Audibert and Tsybakov, 2007] in the context of binary classification. The major difference is the continuity Assumption 3.1.1, which does not allow to re-use in a straightforward way their construction for lower bounds.

3.1.4 Lower bounds

The main results in the present chapter are the lower bounds which we provide in this section. In particular, we establish in Section 3.1.4 the inconsistency of top-β procedures (see Eq. (3.3) for a definition of the method). Therefore more elaborate methods are required in this framework. As pointed out in the introduction, we distinguish two types of estimators: *supervised* and *semi-supervised* ones for which we provide lower bounds in Section 3.1.4. The obtained rates highlight the benefit of the semi-supervised approach in the context of the confidence set classification.

Before proceeding to our main result, let us first display connections between the different minimax risks. These links are used in the proofs of the lower bounds.

Proposition 3.1.5. *Let Γ be a measurable function from $\mathbb{R}^d$ to $2^{[K]}$, $\beta \in [K]$ and assume that Assumption 3.1.1 is fulfilled, then*

$$\mathrm{P}(\Gamma) - \mathrm{P}(\Gamma^*_\beta) = \mathcal{R}_\beta(\Gamma) - \mathcal{R}_\beta(\Gamma^*_\beta) + G^{-1}(\beta)\left(\beta - \mathrm{I}(\Gamma)\right) \ ,$$

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

Furthermore, if additionally Assumption 3.1.4 is satisfied with $\alpha > 0$, then there exist $C > 0$ which depends only on K, α, C_1 such that for any pair of confidence set classifiers Γ, Γ' it holds that

$$\mathbb{E}_{\mathbb{P}_X} \left| \Gamma(X) \triangle \Gamma'(X) \right| \leq C \left(\mathcal{R}_\beta(\Gamma) - \mathcal{R}_\beta(\Gamma') \right)^{\alpha/(\alpha+1)} \quad . \tag{3.6}$$

Proposition 3.1.6. *For any $K \geq 2$, $\beta \in [K]$ and $n, N \in \mathbb{N}$ the following relation between minimax rates holds:*

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

Proposition 3.1.5, and in particular Eq. (3.6) gives an easy way to establish a lower bound on $\mathcal{E}^{\mathrm{E}}_{n,N}(\hat{\mathbf{\Gamma}}; \mathcal{P})$ via a lower bound on the Hamming distance $\mathcal{E}^{\mathrm{H}}_{n,N}(\hat{\mathbf{\Gamma}}; \mathcal{P})$. This approach allows us to cover the classical non-parametric part of the rate, however, it does not capture the semi-supervised nature of the problem. As we shall see, a different strategy is required to get the correct dependency on the unlabeled sample size. Besides, Proposition 3.1.6 allows to prove a lower bound on the discrepancy $\mathcal{E}^{\mathrm{D}}_{n,N}(\hat{\mathbf{\Gamma}}; \mathcal{P})$ with the correct rate via the lower bound on the excess risk $\mathcal{E}^{\mathrm{E}}_{n,N}(\hat{\mathbf{\Gamma}}; \mathcal{P})$.

Inconsistency of the top-β procedure

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

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

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

Proposition 3.1.7. *Assume that $K \geq 4$, $\beta \in [\lfloor K/2 \rfloor - 1]$ and $\beta \geq 2$, then for all $n, N \in \mathbb{N}$ we have*

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

The proof builds an explicit construction of a distribution $\mathbb{P}$ whose β-Oracle in the sense of Eq. (3.1) satisfies $|\Gamma^*_\beta(x)| > \beta$ for all x in some $A \subset \mathbb{R}^d$ with $\mathbb{P}_X(A) > 0$. Indeed, if such a distribution exists then there is no estimator in $\hat{\Upsilon}_\beta$ that would consistently estimate this β-Oracle. The negative result established in Proposition 3.1.7 is rather instructive by itself as it advocates that a more involved estimation procedure ought to be constructed.

Supervised vs semi-supervised

This section is dedicated to the lower bounds on the supervised and the semi-supervised methods. As already mentioned, estimators which achieve the infimum in the minimax rates are either supervised or semi-supervised. However, a lower bound on $\mathcal{E}^{\square}_{n,N}(\hat{\Upsilon}; \mathcal{P})$ does not discriminate between the supervised and the semi-supervised estimators. For this reason, we consider both of these families of algorithms separately.

Theorem 13 (Supervised estimation)**.** *Let $K \geq 3$, $\beta \in [\lfloor K/2 \rfloor - 1]$. If $2\alpha\lceil \frac{\gamma}{2} \rceil \leq d$, then there exist constants $c, c', c'' > 0$ such that for all $n, N \in \mathbb{N}$*

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

Based on this results we observe that the lower bound for the Hamming risk $\mathcal{E}^{\mathrm{H}}_{n,N}$ is slower than those for the other risks. It is even more significant that the best rate that a supervised estimator can achieve for all of the risks is $n^{-1/2}$ even if the margin assumption holds. This is the major difference with the classical settings where the value of threshold is known (such as classification and level set estimation). Indeed, under the same assumptions on the family of distributions, besides the continuity Assumption 3.1.1, the minimax rate in those frameworks is $n^{-(1+\alpha)\gamma/(2\gamma+d)}$ as proved for instance in [Audibert and Tsybakov, 2007, Rigollet and Vert, 2009]. Next theorem deals with semi-supervised procedures and displays another behavior.

Theorem 14 (Semi-supervised estimation)**.** *Let $K \geq 3$, $\beta \in [\lfloor K/2 \rfloor - 1]$. If $2\alpha\lceil \frac{\gamma}{2} \rceil \leq d$, then there exist constants $c, c', c'' > 0$ such that for all $n, N \in \mathbb{N}$*

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

First, observe that the lower bound for the Hamming distance is, as in the supervised setting, worse than for the other measures of performance. However there is a major difference with the supervised case: as compared to Theorem 13, it is possible for a semi-supervised estimator to achieve rates that are faster than $n^{-1/2}$ if the size of the unlabeled dataset $N \in \mathbb{N}$ is large enough. In particular, when we consider $\mathcal{E}^{\mathrm{E}}_{n,N}$ or $\mathcal{E}^{\mathrm{D}}_{n,N}$ the following relations are necessary to get fast rates

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

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

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

Table 3.1: This table summarizes observations of Corollary 2 for $\mathcal{E}^{\mathrm{E}}_{n,N}$ and $\mathcal{E}^{\mathrm{D}}_{n,N}$. Depending on the relations between α, γ, d and N, n the semi-supervised approach can significantly improve the rates of convergence.

Corollary 2. *Assume that the rates in Theorem 14 (resp. Theorem 13) are minimax, that is, there exist a confidence set $\hat{\Gamma}_{\mathrm{SSE}}$ (resp. $\hat{\Gamma}_{\mathrm{SE}}$) that achieves these rates. Regarding $\mathcal{E}^{\mathrm{E}}_{n,N}$ and $\mathcal{E}^{\mathrm{D}}_{n,N}$ the following conclusions hold*

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

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

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

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

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

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

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

Clearly, similar observation is true for the Hamming risk $\mathcal{E}^{\mathrm{H}}_{n,N}$; however the regime when improvement is possible thanks to semi-supervised approaches is narrowed as $n^{-(1+\alpha)\gamma/(2\gamma+d)} = o\left(n^{-\alpha\gamma/(2\gamma+d)}\right)$. Table 3.1.4 gathers the conclusions of Corollary 2 in a compact form.

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

A final remark we could make before going further concerns the assumption on the parameters α and γ. The condition $2\alpha\lceil\frac{\gamma}{2}\rceil \leq d$ in the lower bounds is slightly more restrictive than the conditions given in [Audibert and Tsybakov, 2007] (they have $\alpha\gamma \leq d$). We believe

that this is an artifact of our proof and could be avoided with a finer choice of hypotheses. Simple modifications of the lower bound of Audibert and Tsybakov [2007] do not work in our settings because their hypotheses are not satisfying Assumption 3.1.1. In contrast, the construction of Rigollet and Vert [2009] satisfies[1] Assumption 3.1.1 but their lower bound is limited by the condition $\alpha\gamma \leq 1$, that is, it does not cover the fast rates as long as the dimension $d > 2$.

Sketch of the proof

In order to prove the lower bounds of Theorems 13 and 14 we actually prove two separate lower bounds on the minimax rates. The two lower bounds that we prove are naturally connected with the proposed two-step estimator in Eq. (3.5). That is, the first lower bound is connected with the problem of non-parametric estimation of p_k for all $k \in [K]$ and the second describes the estimation of the unknown threshold $G^{-1}(\beta)$.

Specifically, the first lower bound is closely related to the one provided in [Audibert and Tsybakov, 2007, Rigollet and Vert, 2009], however, the continuity Assumption 3.1.1 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 of Birgé [2005]. The second lower bound is based on two hypotheses testing and is derived by constructing two different marginal distributions of $X \in \mathbb{R}^d$ which are sufficiently close and a fixed regression function $p(\cdot)$. Crucially, these marginal distributions admit two different values of threshold $G^{-1}(\beta)$ and thus two different β-Oracle. In this part we make use of Pinsker's inequality, see for instance [Tsybakov, 2009].

In order to discriminate the supervised and the semi-supervised procedures we make use of Definition 14. Notice that every supervised procedure thanks to Definition 14 is not "sensitive" to the expectation taken *w.r.t.* the unlabeled dataset $\mathcal{D}_N^{\mathrm{U}}$, that is, randomness is only induced by the labeled dataset $\mathcal{D}_n^{\mathrm{L}}$. This strategy allows to eliminate the dependence of the lower bound on the size of the unlabeled dataset $\mathcal{D}_N^{\mathrm{U}}$ for supervised procedures. Informally, the lower bound on $\mathcal{E}_{n,N}^{\square}(\hat{\Upsilon}_{\mathrm{SE}}; \mathcal{P})$ is obtained from the lower bound on $\mathcal{E}_{n,N}^{\square}(\hat{\Upsilon}_{\mathrm{SSE}}; \mathcal{P})$ by setting $N = 0$.

3.1.5 Upper bounds

In this section, we show that we can build confidence set estimators that achieve, up to a logarithmic factor, the lower bounds stated in Theorems 13-14. In other words, those estimators are *nearly* optimal in the minimax sense. To come straight to the point, we delay the construction of the estimators to Section 3.1.5 and their properties to Section 3.1.5, and focus right now on their upper bounds.

Theorem 15 (Supervised estimation)**.** *Let $K \in \mathbb{N}$, $\beta \in [K-1]$, then there exists a supervised*

[1] Modified properly to fit the classification framework.

estimator $\hat{\Gamma}_{\mathrm{SE}} \in \hat{\Upsilon}_{\mathrm{SE}}$ *and constants* $C, C', C'' > 0$ *such that for all* $n, N \in \mathbb{N}$ *we have*

$$\mathcal{E}^{\mathrm{H}}_{n,N}(\hat{\Gamma}_{\mathrm{SE}}; \mathcal{P}(L,\gamma,\alpha)) \leq C\left(n^{-\frac{\alpha\gamma}{2\gamma+d}} \bigvee n^{-1/2}\right) \ ,$$

$$\mathcal{E}^{\mathrm{E}}_{n,N}(\hat{\Gamma}_{\mathrm{SE}}; \mathcal{P}(L,\gamma,\alpha)) \leq C'\left(\left(\frac{n}{\log n}\right)^{-\frac{(1+\alpha)\gamma}{2\gamma+d}} \bigvee n^{-1/2}\right) \ ,$$

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

Theorem 16 (Semi-supervised estimation)**.** *Let* $K \in \mathbb{N}$, $\beta \in [K-1]$, *then there exists a semi-supervised estimator* $\hat{\Gamma}_{\mathrm{SSE}} \in \hat{\Upsilon}_{\mathrm{SSE}}$ *and constants* $C, C', C'' > 0$ *such that for all* $n, N \in \mathbb{N}$ *we have*

$$\mathcal{E}^{\mathrm{H}}_{n,N}(\hat{\Gamma}_{\mathrm{SSE}}; \mathcal{P}(L,\gamma,\alpha)) \leq C\left(n^{-\frac{\alpha\gamma}{2\gamma+d}} \bigvee (n+N)^{-1/2}\right) \ ,$$

$$\mathcal{E}^{\mathrm{E}}_{n,N}(\hat{\Gamma}_{\mathrm{SSE}}; \mathcal{P}(L,\gamma,\alpha)) \leq C'\left(\left(\frac{n}{\log n}\right)^{-\frac{(1+\alpha)\gamma}{2\gamma+d}} \bigvee (n+N)^{-1/2}\right) \ ,$$

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

First of all, the above upper bounds imply that the lower bounds of Theorems 13-14 are achievable. In particular, in the case of Hamming risk, the upper bounds are optimal; whereas for the excess risk and the discrepancy, the upper bounds fit the lower bounds up to a logarithmic factor. Additionally, these upper bounds support the discussion on the semi-supervised and the supervised estimators provided in Corollary 2. Finally, notice that the upper bounds for the excess risk and for the discrepancy exhibit an extra logarithmic factor. This disagreement with the lower bounds is due to the connection with the ℓ_∞-norm estimation established in Lemma 12. A more detailed discussion on this logarithmic factor is provided in Section 3.1.6.

Construction of the estimators

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

Assumption 3.1.8 (Exponential concentration)**.** *There exist estimators* $\hat{p}_k$ *for all* $k \in [K]$ *based on* $\mathcal{D}_{\lfloor n/2 \rfloor}$ *and positive constants* C'_1, C'_2 *such that for all* $k \in [K]$ *and all* $n \geq 2$ *we have for all* $\delta > 0$

$$\sup_{\mathbb{P} \in \mathcal{P}(L,\gamma,\alpha)} \mathbb{P}^{\otimes \lfloor n/2 \rfloor}\left(|\hat{p}_k(x) - p_k(x)| \geq \delta\right) \leq C'_1 \exp\left(-C'_2 n^{\frac{2\gamma}{2\gamma+d}} \delta^2\right) \ ,$$

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

Assumption 3.1.9 (Continuity of CDF)**.** *For all* $k \in [K]$ *the cumulative distribution function* $F_{\hat{p}_k}(t) := \mathbb{P}_X(\hat{p}_k(X) \leq t)$ *of* $\hat{p}_k(X)$ *is almost surely* $\mathbb{P}^{\otimes \lfloor n/2 \rfloor}$ *continuous on* $(0,1)$.

First let us point out that Assumption 3.1.8 induces that there exists a constant $C > 0$ such that for all $n \geq 2$ and all $\alpha > 0$

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

Assumption 3.1.8 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 et al., 2018]. It is for instance satisfied by the locally polynomial estimator [Stone, 1977, Tsybakov, 1986, Audibert and Tsybakov, 2007]. Assumption 3.1.9 can always be satisfied by slightly processing any estimator $\hat{p}$. Indeed, assume Assumption 3.1.9 fails to be satisfied by some estimator $\hat{p}$. It means that there exists a subset of $\mathbb{R}^d$ of non-zero measure such that at least one $\hat{p}_k$, with $k \in [K]$, is constant on this set. Then, if we add a *deterministic* continuous function with a sufficiently small amplitude[2] to $\hat{p}$ such regions can no longer exist.

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

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

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

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

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

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

for all $x \in \mathbb{R}^d$. Note that $\hat{\Gamma}_{\mathrm{SE}}$ is clearly supervised in the sense of Definition 14, as it is independent of the unlabeled sample $\mathcal{D}^{\mathrm{U}}_N$. In contrast, $\hat{\Gamma}_{\mathrm{SSE}}$ is semi-supervised, since we can find two samples $\mathcal{D}^{\mathrm{U}}_N$ and $\mathcal{D}^{\mathrm{U}'}_N$ which induce different confidence sets.

Properties of the plug-in confidence sets

To show that the estimators introduced in the previous section satisfy the statements of Theorems 15-16 we refine the proof technique used in [Denis and Hebiri, 2017]. That is, we introduce an intermediate quantity

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

[2] It is sufficient to make sure that adding the function preserves its statistical properties, that is, Assumption 3.1.8.

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

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

The confidence set $\tilde{\Gamma}$ assumes knowledge of the marginal distribution $\mathbb{P}_X$ and is seen as an idealized version of both $\hat{\Gamma}_{\mathrm{SE}}$ and $\hat{\Gamma}_{\mathrm{SSE}}$. For the pseudo Oracle the uncertainty is induced only by the lack of knowledge of the regression function. Note however, that the pseudo Oracle $\tilde{\Gamma}$ is not an estimator, since it is not data driven.

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

Lemma 12 (Upper bound on the thresholds)**.** *Let Assumption 3.1.1 be satisfied, then for all* $\beta \in [K]$

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

The proof of Lemma 12 uses elementary properties of the generalized inverse functions which are provided in Section 3.1.8. Besides, let us mention, that the difference $|G^{-1}(\beta) - \tilde{G}^{-1}(\beta)|$ resembles the Wasserstein infinity distance which gives an alternative approach to prove Lemma 12, see [Bobkov and Ledoux, 2016]. Importantly, Lemma 12 explains the extra $\log n$ factor that appears in the upper bound, as the minimax estimation rate in sup norm involves this logarithm, see for instance [Stone, 1982, Tsybakov, 2009]. Another crucial property of the introduced estimators $\hat{\Gamma}_{\mathrm{SE}}$ and $\hat{\Gamma}_{\mathrm{SSE}}$ is obtained via Assumption 3.1.9. It describes the deviation of the information of $\hat{\Gamma}_{\mathrm{SE}}$ and $\hat{\Gamma}_{\mathrm{SSE}}$ from the desired level β.

Proposition 3.1.10 (Denis and Hebiri [2017])**.** *Let $\hat{p}_k$ for all $k \in [K]$ be arbitrary estimators of the regression functions constructed using $\mathcal{D}^{\mathrm{L}}_{\lfloor n/2 \rfloor}$ that satisfies Assumption 3.1.9, then there exist constants $C, C' > 0$ such that for all $n, N \in \mathbb{N}$ it holds that*

$$\mathbb{E}_{(\mathcal{D}^{\mathrm{L}}_n, \mathcal{D}^{\mathrm{U}}_N)} \left| \beta - \mathrm{I}\left(\hat{\Gamma}_{\mathrm{SE}}\right) \right| \leq C n^{-1/2} \quad ,$$
$$\mathbb{E}_{(\mathcal{D}^{\mathrm{L}}_n, \mathcal{D}^{\mathrm{U}}_N)} \left| \beta - \mathrm{I}\left(\hat{\Gamma}_{\mathrm{SSE}}\right) \right| \leq C' (N+n)^{-1/2} \quad .$$

Note that if $\hat{p}_k$ satisfies Assumption 3.1.9 for all $k \in [K]$, then the information of pseudo Oracle $\tilde{\Gamma}$ satisfies $\mathrm{I}(\tilde{\Gamma}) = \beta$. This simple fact is an important step in the proof of Proposition 3.1.10.

Finally, combination of Lemma 12, Proposition 3.1.10, Assumption 3.1.8 with the peeling argument used in [Audibert and Tsybakov, 2007, Lemma 3.1] yields the results of Theorems 15-16.

Simulation study

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

1) Is it more advantageous to go outside of the *classical* multi-class classification settings and consider the confidence set framework? To respond to this question we compute the Bayes optimal multi-class classifier and view it as a confidence set with one label. We compare this Bayes rule with the β-Oracle in terms of the error $\mathrm{P}(\cdot)$ using various values of $\beta \in [K]$ and $K \in \mathbb{N}$.

$K = 10$		
β	β-Oracle	top-β Oracle
2	0.05 (0.01)	0.09 (0.01)
5	0.00 (0.00)	0.01 (0.00)
$K = 100$		
β	β-Oracle	top-β Oracle
2	0.39 (0.01)	0.42 (0.01)
5	0.20 (0.01)	0.22 (0.01)
10	0.09 (0.01)	0.11 (0.01)
20	0.03 (0.01)	0.04 (0.01)

Table 3.2: For each of the $B = 100$ repetitions and each model, we derive the estimated errors P_M of the β-Oracle and of the top-β Oracle *w.r.t.* β. We compute the means and standard deviations (between parentheses) over the $B = 100$ repetitions. Top: the data are generated according to $K = 10$ – Bottom: the data are generated according to $K = 100$.

2) How does the β-Oracle confidence set compares to another "Oracle" (top-β Oracle) which simply includes classes corresponding to the largest values of $p_k(\cdot)$'s?

3) Does the proposed plug-in approach indeed gives a good approximation of the β-Oracle through the error $\mathrm{P}(\cdot)$ and the information $\mathrm{I}(\cdot)$?

4) Despite demonstrating the minimax inconsistency of the top-β approach, we wonder whether in some scenarios it can achieve a comparable performance against our semi-supervised plug-in procedure.

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

i) the label Y is distributed uniformly on $[K]$;

ii) conditional on $Y = k$, the feature X is generated according to a multivariate Gaussian distribution with mean $\mu_k \in \mathbb{R}^{10}$ and identity covariance matrix.

For each $k \in [K]$, the means μ_k are *i.i.d.* realizations of uniform distribution on $[0, 4]^{10}$. For this data generating setup we have the following expression for the regression functions

$$p_k(X) = \frac{\varphi_{\mu_k}(X)}{\sum_{j=1}^{K} \varphi_{\mu_j}(X)},$$

where for each $k \in [K]$, $\varphi_{\mu_k}(X)$ is the density function of a multivariate Gaussian distribution with mean μ_k and identity covariance matrix.

For each K, the *misclassification* error of the *classical* multi-class classification Bayes rule is evaluated based on a sufficiently large dataset. It is valued at 0.22 and at 0.60 for $K = 10$ and for $K = 100$ respectively. These values are relatively high, which suggests that it is reasonable to apply the confidence set approach to this problem.

In the sequel, we aim at reporting estimated errors and information levels of the β-Oracle defined in Eq. (3.1). To this end, for $\beta \in \{2, 5, 10, 20\}$ and each K, we repeat $B = 100$ times the following steps.

β	$K = 10$	$K = 100$
2	2.00 (0.03)	2.00 (0.03)
5	5.00 (0.08)	5.00 (0.06)
10	·	10.00 (0.13)
20	·	20.02 (0.31)

Table 3.3: For each of the $B = 100$ repetitions and each model, we derive the estimated information levels I_M of the β-Oracle set w.r.t. β. We compute the means and standard deviations (in parentheses) over the $B = 100$ repetitions. Left: the data are generated according to $K = 10$ – Right: the data are generated according to $K = 100$.

$K = 10$

	$\hat{\Gamma}_{\mathrm{SSE}}$			top-β		
β	`rforest`	`softmax reg`	`deep learn`	`rforest`	`softmax reg`	`deep learn`
2	0.09 (0.01)	0.06 (0.01)	0.09 (0.01)	0.13 (0.01)	0.10 (0.01)	0.13 (0.02)
5	0.01 (0.00)	0.00 (0.00)	0.01 (0.00)	0.02 (0.00)	0.01 (0.00)	0.02 (0.00)

$K = 100$

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

Table 3.4: For each of the $B = 100$ repetitions and for each model, we derive the estimated errors P of three different $\hat{\Gamma}_{\mathrm{SSE}}$'s *w.r.t.* β. We compute the means and standard deviations (in parentheses) over the $B = 100$ repetitions. For each β and for each N, the $\hat{\Gamma}_{\mathrm{SSE}}$'s, as well as the top procedures are based on, from left to right, `rforest`, `softmax reg` and `deep learn`, which are respectively the random forest, the softmax regression and the deep learning methods. Top: the data are generated according to $K = 10$ – Bottom: the data are generated according to $K = 100$.

i) simulate two datasets $\mathcal{D}_N^{\mathrm{U}}$ and $\mathcal{D}_M^{\mathrm{L}}$ with $N = 10000$ and $M = 1000$;

ii) based on $\mathcal{D}_N^{\mathrm{U}}$, we compute the empirical counterpart of G and provide an approximation of the β-Oracle Γ_β^* given in Eq. (3.1) (we recall that this step requires a dataset which contains only unlabeled features);

iii) finally, over $\mathcal{D}_M$, we compute the empirical counterparts P_M (of $\mathrm{P}(\Gamma_\beta^*)$) and I_M (of $\mathrm{I}(\Gamma_\beta^*)$).

From this estimates, we compute the mean and the standard deviation of P_M and I_M. Tables 3.2 and 3.3 present values of the error and of the information which are achieved by the β-*Oracle* and by the top-β Oracle. Turning to Table 3.2 we confirm the intuition that the error of the β-Oracle decreases as the value of the parameter β increases. Notably, we obtain a satisfactory improvement over the *standard* multi-class classification Bayes rule even for

$K = 10$

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

$K = 100$

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

Table 3.5: For each of the $B = 100$ repetitions and for each model, we derive the estimated information levels I of three different $\hat{\Gamma}_{\mathrm{SSE}}$'s *w.r.t.* β and the sample size N. We compute the means and standard deviations (in parentheses) over the $B = 100$ repetitions. For each β and each N, the $\hat{\Gamma}_{\mathrm{SSE}}$'s are based on, from left to right, `rforest`, `softmax reg` and `deep learn`, which are respectively the random forest, the softmax regression and the deep learning procedures. Top: the data are generated according to $K = 10$ – Bottom: the data are generated according to $K = 100$.

moderate values of β compared to K. For instance, when $K = 10$ and $\beta = 2$ the error of the 2-Oracle confidence set is 0.05, whereas the Bayes classifier has 0.22; likewise, when $K = 100$ and $\beta = 5$ the classification error decreases from 0.60 to 0.20. These observations argue in favor of the confidence set framework. Let us finally point out that Table 3.2 shows that the top-β Oracle is outperformed by the β-Oracle in terms of the error. Nevertheless, top-β Oracle still performs reasonably well.

We now move towards the construction of our semi-supervised plug-in estimators $\hat{\Gamma}_{\mathrm{SSE}}$ defined in Eq. (3.5). For each K and each β, we evaluate the performance of $\hat{\Gamma}_{\mathrm{SSE}}$ according to three different estimations of the regression function: the $\hat{p}_k$'s that are based on random forests, softmax regression and deep learning[3] procedures. Let us point out, that for random forests and softmax regression algorithms, the random variables $\hat{p}_k(X)$ appear to be not continuous, that is, Assumption 3.1.9 is violated. To alleviate this issue, we add to $\hat{p}_k(X)$ an independent small perturbation $|\mathcal{N}(0, e^{-10})|$ for simplicity. The evaluation of the performance of $\hat{\Gamma}_{\mathrm{SSE}}$ relies on the following steps

i) simulate three datasets $\mathcal{D}_n^{\mathrm{L}}$, $\mathcal{D}_N^{\mathrm{U}}$ and $\mathcal{D}_M^{\mathrm{L}}$;

ii) based on $\mathcal{D}_n^{\mathrm{L}}$, we compute the estimators $\hat{p}_k$ of p_k according to the considered procedure;

iii) based on $\mathcal{D}_N^{\mathrm{U}}$ and $\hat{p}_k$ we compute the function $\hat{G}$ and the estimator $\hat{\Gamma}_{\mathrm{SSE}}$ as in Eq. (3.5) (we recall that this step requires a dataset which contains only unlabeled features);

iv) finally, we compute over $\mathcal{D}_M^{\mathrm{L}}$ the empirical counterpart of P and of I for the considered $\hat{\Gamma}_{\mathrm{SSE}}$.

[3]We used H2O package for the implementation of deep learning algorithm available at https://cran.r-project.org/web/packages/h2o/index.html.

Again, during these experiments, we compute means and standard deviations. The parameters K, n, N are fixed as follows: for $K = 10$, we fix $n = 1000$ and $N \in \{100, 10000\}$; for $K = 100$ we fix $n = 10000$ and $N \in \{100, 10000\}$. Finally, the size of $\mathcal{D}_M^{\mathrm{L}}$ is fixed to $M = 1000$. The results are illustrated in Tables 3.4 and 3.5.

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

From Tables 3.3 and 3.5, we observe that the approximation of the information is reasonably good and it gets better with N the number of unlabeled data. Besides, Tables 3.2 and 3.4 demonstrate that our algorithm is sensitive to the choice of the underlying estimator $\hat{p}_k$. Indeed, when $\hat{p}_k$ is estimated via the softmax regression, our algorithm fails to give a good approximation to the error of the β-Oracle.

Table 3.4 provides similar conclusions regarding $\hat{\Gamma}_{\mathrm{SSE}}$, though, unlike the theoretical quantities, there are more scenarios where our method is better than its top-β counterpart. Let us point out, that for $K = 100$ methods that are based on the softmax regression perform poorly in this setup.

3.1.6 Discussion

Around continuity Assumption 3.1.1

The bedrock of this contribution is Assumption 3.1.1. Based on it, we ensure that the β-Oracle confidence set given by Eq. (3.1) is indeed of information β. On top of that, the explicit formulation of excess risk in Proposition 3.1.5 relies on the continuity of function $G(\cdot)$. Should Assumption 3.1.1 fail to be satisfied, then there might be no β-Oracle given by thresholding on some level $\theta \in (0, 1)$. Indeed, assume Assumption 3.1.1 is not satisfied but one can build a β-Oracle having the form $\Gamma_\beta^*(\cdot) = \{k \in [K] \,:\, p_k(\cdot) > \theta\}$ with some θ, then

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

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

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

where the inequality is used in place of the equality. Indeed, under the continuity assumption and thanks to Propositions 3.1.3 and 3.1.5, we have for all confidence sets Γ such that $\mathrm{I}(\Gamma) \leq \beta$

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

This implies that the β-Oracle Γ_β^* defined in Eq. (3.1) is a minimizer of the above constrained problem.

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

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

One might wonder whether extra assumptions on the problem allow a supervised method to get faster rates than $n^{-1/2}$. We give to this question a partial answer following the recent work of Bobkov and Ledoux [2016] and more precisely their Theorem 5.11. Applying this result to our framework, we can state that there exists a positive constant c such that

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

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

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

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

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

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

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

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

Around extra logarithm

Theorems 14 and 16 demonstrate that for the excess risk and the discrepancy, the upper and the lower bounds differ by a logarithmic factor. As we have already pointed out, this factor appears in the upper bounds due to Lemma 12 which relates the difference between two thresholds to the infinity norm. One might hope that if we manage to replace the infinity norm by any other ℓ_q-norm on the right hand side of the inequality in Lemma 12 this logarithm can be eliminated. Unfortunately, it appears that this bound is actually tight, in a sense that one can construct a distribution $\mathbb{P}$ and an estimator $\hat{p}_k$ for all $k \in [K]$ such that an equality is achieved in Lemma 12. These arguments suggest that the obtained upper bound should be optimal. They also imply that the lower bounds could be further refined to get an extra logarithmic factor. Let us also mention that the continuity Assumption 3.1.1 in combination with the margin Assumption 3.1.4 are main obstacles that did not allow us to further refine the lower bounds. In any case, our conclusions shall remain unchanged with or without this log factor in either the lower or the upper bounds.

3.1.7 Conclusion

In this chapter we have studied the minimax settings of confidence set multi-class classification. First of all, we have shown that a top-β type procedures are inconsistent in our settings

and more involved estimators should be proposed. Besides, we have demonstrated that no supervised estimator can achieve rates that are faster than $n^{-1/2}$, which stays in contrast with other classical settings. Additionally, we have shown that fast rates are achievable by semi-supervised methods provided that the size of the unlabeled sample is large enough. Consecutively, we have established that our lower bounds are either optimal or nearly optimal by providing a supervised and a semi-supervised estimators which are tractable in practice. Our future works shall be focused on the Lipschitz condition of $G^{-1}(\cdot)$ discussed in Section 3.1.6, in particular, we want to understand how this extra assumption affects our lower bounds.

3.1.8 Proofs

This section is composed of the following parts: first of all we introduce some technical results used for our proofs; then we provide the proofs of the upper bounds; later we establish main lower bounds; finally, in the end of this section we prove the inconsistency of top-β approaches.

Technical results

In this section we gather several technical results which are used to derive the contributions of this chapter.

Given any two probability measures $\mathbb{P}_1, \mathbb{P}_2$ on some space measurable space $(\mathcal{X}, \mathcal{A})$ total variation distance is defined as

$$\mathrm{TV}(\mathbb{P}_1, \mathbb{P}_2) := \sup_{A \in \mathcal{A}} |\mathbb{P}_1(A) - \mathbb{P}_2(A)| \quad . \tag{3.7}$$

Lemma 13 (Pinsker's inequality)**.** *Given any two probability measures $\mathbb{P}_1, \mathbb{P}_2$ on some measurable space $(\mathcal{X}, \mathcal{A})$ we have*

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

Lemma 14 (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*

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

Proposition 3.1.11 (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 $\beta \in (0, K)$*

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

Then,

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

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

Lemma 15. *Let ε distributed from a uniform distribution on $[K]$ and $Z_1, \ldots, Z_K$, K real valued random variables independent from ε, 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 $Z = \sum_{k=1}^{K} Z_k \mathbb{1}_{\{\{\varepsilon = k\}\}}$ and let U be distributed according to the uniform distribution on $[0,1]$. Then

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

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

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

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

To conclude the proof, we observe that

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

□

Upper bounds

In this section we prove Theorems 15 and 16. It will be clear from our analysis that the proof of Theorem 15 follows directly from Theorem 16 by setting $N = 0$ in the statement of Theorem 16. Thus, in this section for simplicity we omit the subscript SSE from $\hat{\Gamma}_{\text{SSE}}$. Recall that our dataset consists of three parts $\mathcal{D}^{\text{L}}_{\lfloor n/2 \rfloor}, \mathcal{D}^{\text{U}}_{\lceil n/2 \rceil}, \mathcal{D}^{\text{U}}_{N}$. The set $\mathcal{D}^{\text{L}}_{\lfloor n/2 \rfloor}$ is used to construct an estimator $\hat{p}$ of the regression function p, that is, $\hat{p}$ is independent from both $\mathcal{D}^{\text{U}}_{\lceil n/2 \rceil}, \mathcal{D}^{\text{U}}_{N}$. The other two sets $\mathcal{D}^{\text{U}}_{\lceil n/2 \rceil}, \mathcal{D}^{\text{U}}_{N}$ are used in a semi-supervised manner to estimate the threshold, that is, we erase the labels from $\mathcal{D}^{\text{U}}_{\lceil n/2 \rceil}$. Let $\beta \in [K-1]$, and also recall the definition of the proposed semi-supervised estimator for a given $x \in \mathbb{R}^d$

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

with $\hat{p}_k(x)$ satisfying Assumptions 3.1.9, 3.1.8 for all $k \in [K]$. Moreover, $\hat{G}^{-1}(\beta)$ defined as the generalized inverse of

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

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

$$\Gamma_\beta^*(x) = \left\{ k \in [K] \ : \ p_k(x) \geq G^{-1}(\beta) \right\} \ , \tag{3.8}$$

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

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

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

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

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

Lemma 16 (Upper-bound on the thresholds)**.** *Let $X \in \mathbb{R}^d$ and $\mathbb{P}_X$ be a Borel measure on $\mathbb{R}^d$. For two vector functions $p, \hat{p} : \mathbb{R}^d \to [0, 1]^K$, we define*

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

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

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

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

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

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

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

that is, applying Proposition 3.1.11 to the second inequality we get for all $t \in [0, 1]$

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

thus, for $t = G^{-1}(\beta)$ with $\beta \in (0, K)$ thanks to Proposition 3.1.11 we get

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

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

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

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

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

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

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

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

We are in position to prove Theorem 16, let us point out that the most difficult part in Theorem 16 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

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

Recall that thanks to Proposition 3.1.5 we have

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

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

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

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

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

Therefore, for R_1 using Lemma 16 and the observations above we can write

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

finally, using the margin Assumption 3.1.4 we get almost surely data

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

Integrating over the data from both sides and using Assumption 3.1.8 we get for some $C > 0$

$$\mathbb{E}_{(\mathcal{D}_n^{\mathrm{L}}, \mathcal{D}_N^{\mathrm{U}})} R_1 \leq C \left(\frac{n}{\log n}\right)^{-\frac{(1+\alpha)\gamma}{2\gamma+d}} \quad .$$

For R_2 the following trivial upper-bound holds

$$\begin{aligned}
R_2 &= \left(\mathrm{P}(\hat{\Gamma}) - \mathrm{P}(\tilde{\Gamma})\right) + G^{-1}(\beta)\left(\mathrm{I}(\hat{\Gamma}) - \mathrm{I}(\tilde{\Gamma})\right) \\
&= \sum_{k=1}^{K} \mathbb{E}\left(p_k(X) - G^{-1}(\beta)\right)\left(\mathbb{1}_{\left\{k\in\tilde{\Gamma}(X)\right\}} - \mathbb{1}_{\left\{k\in\hat{\Gamma}(X)\right\}}\right) \\
&\leq \underbrace{\sum_{k=1}^{K} \mathbb{E}\left|\mathbb{1}_{\left\{k\in\tilde{\Gamma}(X)\right\}} - \mathbb{1}_{\left\{k\in\hat{\Gamma}(X)\right\}}\right|}_{\mathbb{E}\left|\tilde{\Gamma}(X)\triangle\hat{\Gamma}(X)\right|} \qquad (3.9) \\
&= \sum_{k=1}^{K} \mathbb{E}\left|\mathbb{1}_{\left\{\hat{p}_k(X)\geq\hat{G}^{-1}(\beta)\right\}} - \mathbb{1}_{\left\{\hat{p}_k(X)\geq\tilde{G}^{-1}(\beta)\right\}}\right| \quad ,
\end{aligned}$$

now, thanks to the first property of Proposition 3.1.11 we can write

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

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

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

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

$$\sum_{k=1}^{K} \mathbb{P}_X\left(\left|\hat{G}(\hat{p}_k(X)) - \tilde{G}(\hat{p}_k(X))\right| \geq \left|\tilde{G}(\hat{p}_k(X)) - \beta\right|\right) = \tag{3.10}$$
$$\sum_{k=1}^{K}\sum_{j\geq 0} \mathbb{P}_X\left(\left|\hat{G}(\hat{p}_k(X)) - \tilde{G}(\hat{p}_k(X))\right| \geq \left|\tilde{G}(\hat{p}_k(X)) - \beta\right|, A_j^k\right) \ .$$

Now, we consider ε uniformly distributed on $[K]$ independent of the data and X. Conditional on the data and under Assumption 3.1.9, we apply Lemma 15 with $Z_k = \hat{p}_k(X)$, $Z = \sum_{k=1}^{K} Z_k \mathbb{1}_{\{\{\varepsilon=k\}\}}$ and then obtain that $\tilde{G}(Z)$ is uniformly distributed on $[0, K]$. Therefore, for all $j \geq 0$ and $\delta > 0$, we deduce

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

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

$$\sum_{k=1}^{K} \mathbb{P}_X(A_j^k) \leq \sum_{k=1}^{K} \mathbb{P}_X\left(|\tilde{G}(\hat{p}_k(X)) - \beta| \leq 2^j\delta\right) \leq 2^{j+1}\delta \ . \tag{3.11}$$

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

$$\sum_{k=1}^{K} \mathbb{P}_X\left(|\hat{G}(\hat{p}_k(X)) - \tilde{G}(\hat{p}_k(X))| \geq |\tilde{G}(\hat{p}_k(X)) - \beta|, A_j^k\right) \leq \tag{3.12}$$
$$\sum_{k=1}^{K} \mathbb{P}_X\left(|\hat{G}(\hat{p}_k(X)) - \tilde{G}(\hat{p}_k(X))| \geq 2^{j-1}\delta, A_j^k\right) \ .$$

Thus, we obtain that

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

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

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

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

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

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

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

Therefore, treating A^k_0 separately, we get from inequalities of Eqs. (3.10), (3.11), and (3.12)

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

Finally, choosing $\delta = \dfrac{K}{\sqrt{N+\lceil n/2\rceil}}$ 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 β set $\tilde{\Gamma}$

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

Notice that for the term $\mathbb{E}_{(\mathcal{D}^{\mathrm{L}}_n, \mathcal{D}^{\mathrm{U}}_N)}\mathbb{E}_{X\sim\mathbb{P}_X}\left|\hat{\Gamma}(X)\triangle\tilde{\Gamma}(X)\right|$ we can re-use the proof technique used for the term R_2 in Eq. (3.9). Thus, it remain to upper-bound the term $\mathbb{E}_{(\mathcal{D}^{\mathrm{L}}_n, \mathcal{D}^{\mathrm{U}}_N)}\mathbb{E}_{X\sim\mathbb{P}_X}\left|\tilde{\Gamma}(X)\triangle\Gamma^*_\beta(X)\right|$. The proof on this part closely follows the machinery used in Denis and Hebiri [2017], however, let us mention that they used this method to obtain a bound on the discrepancy which leads to a sub-optimal rate. Nevertheless, their approach gives a correct rate if instead of the discrepancy we bound the Hamming distance. For the sake of completeness we write the principal parts of the proof here.

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

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

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

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

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

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

Thanks to the continuity Assumption 3.1.9 on the estimator and the continuity Assumption 3.1.1 on the distribution we clearly have $\tilde{G}\left(\tilde{G}^{-1}(\beta)\right) = \beta = G\left(G^{-1}(\beta)\right)$. Moreover, we can write

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

Thus, our bound reads as

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

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 3.1.8. 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 3.1.1, 3.1.9 on the distribution and the estimator respectively.

Proof of the lower bounds

This section is devoted to the proof of the lower bounds provided by Theorems 13-14. Before proceeding to the proofs let us briefly sketch the high-level strategy used to prove the lower bounds. In order to prove the lower bounds of Theorems 13-14 we actually prove to separate lower bounds on the minimax risk. Clearly, if some non-negative quantity is lower-bounded by two different values, it is lower-bounded by the maximum between the two. The two lower bounds that we prove are naturally connected with the proposed two-steps estimator, that is, the first lower bound is connected with the problem of non-parametric estimation of p_k for all $k \in [K]$ and the second describes the estimation of the unknown threshold $G^{-1}(\beta)$.

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 3.1.1 makes the proof more involved. In particular, the second lower bound is based on two hypotheses testing and is derived by constructing two different distributions $\mathbb{P}_0, \mathbb{P}_1$ sharing the same regression vector $p(\cdot)$ and having different marginal distributions of $X \in \mathbb{R}^d$. In this part we make use of Pinsker's inequality recalled in Lemma 13.

In order to discriminate the supervised and the semi-supervised procedures we invoke Definition 14. Based on this definition, notice that every supervised procedure is not 'sensitive' to the expectation taken *w.r.t.* the unlabeled dataset $\mathcal{D}_N^{\mathrm{U}}$, that is, randomness is only induced by the labeled dataset $\mathcal{D}_n^{\mathrm{L}}$. This strategy allows to eliminate the dependence of the lower bound on the size of the unlabeled dataset $\mathcal{D}_N^{\mathrm{U}}$ for supervised procedures. Indeed, let $\hat{\Gamma}$ be any supervised estimator in the sense of Definition 14, then for any real valued function of confidence sets h we have

$$\mathbb{E}_{(\mathcal{D}_n^{\mathrm{L}},\mathcal{D}_N^{\mathrm{U}})}[\mathbb{E}_{\mathbb{P}_X} h(\hat{\Gamma}(X;\mathcal{D}_n^{\mathrm{L}},\mathcal{D}_N^{\mathrm{U}}))] = \mathbb{E}_{\mathcal{D}_n^{\mathrm{L}}}[\mathbb{E}_{\mathbb{P}_X} h(\hat{\Gamma}(X;\mathcal{D}_n^{\mathrm{L}},\mathcal{D}_N^{\mathrm{U}'}))] \quad ,$$

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

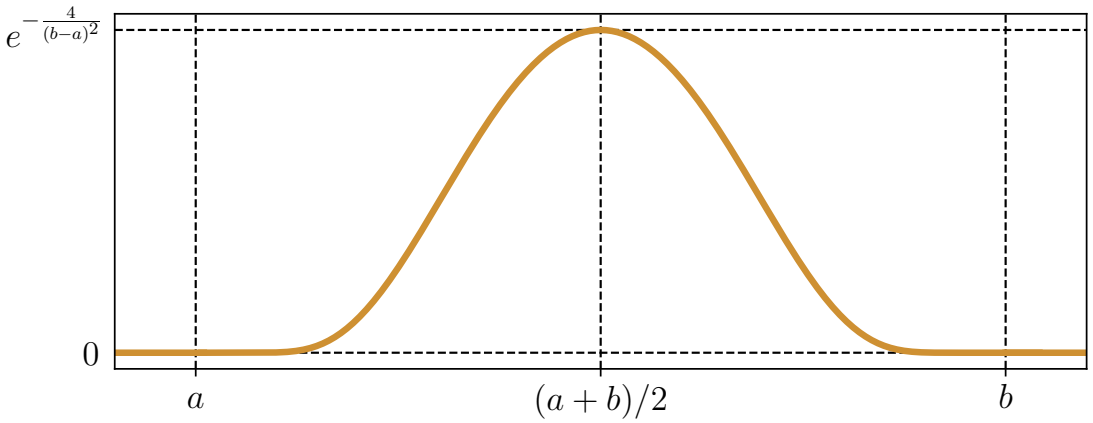

Figure 3.1: Bump function: $x \mapsto \psi_{a,b}(x)$. Importantly, this function is supported on (a, b) and is infinitely smooth.

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

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

$$\kappa_{N,n} = C(N+n)^{-\frac{1}{2}} < 0.1 \ .$$

To prove the lower bound we construct two distribution $\mathbb{P}_0$ and $\mathbb{P}_1$ on $\mathbb{R}^d$ sharing the same regression function $p(\cdot) = (p_1(\cdot), \dots, p_K(\cdot))$ and with different marginals admitting densities μ_0, μ_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

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

Let us denote by $o_i = (r_i + \rho, 0, \dots, 0)^\top$ for $i = 1, 2, 3$ the centers of $\mathcal{X}_1$, $\mathcal{X}_2$ and $\mathcal{X}_3$. Since $\beta < K/2$, one can consider a grid of equidistant points between 0 and $1/K$

$$0 < \frac{1}{4K} < \frac{2K - 4\beta}{4K(K-2\beta)} < \frac{3K - 6\beta}{4K(K-2\beta)} < \frac{1}{K} \ ,$$

and a grid of equidistant points between $1/K$ and $1/2\beta$

$$\frac{1}{K} < \frac{K+6\beta}{8K\beta} < \frac{2K+4\beta}{8K\beta} < \frac{3K+2\beta}{8K\beta} < \frac{1}{2\beta} \ .$$

Next, we aim at building a regression vector and two marginal distributions of X so that we obtain two different thresholds $\frac{K+6\beta}{8K\beta}$ and $\frac{3K+2\beta}{8K\beta}$ sufficiently separated. The regression vector is given by

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

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

In order to define the functions φ_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 3.1 illustrates the behavior of $\psi_{a,b}$ function in one dimension. Note that for every $a, b \in \mathbb{R}$ the function above is infinitely smooth. Using the definition of $\psi_{a,b}$ we define the functions φ_i for $i = 0, \ldots, 4$ as

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

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

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

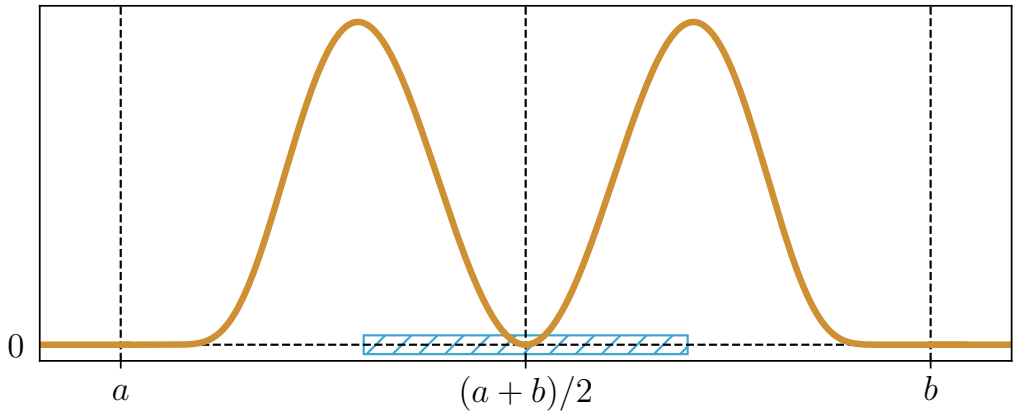

Figure 3.2: Dumped bump function: $x \mapsto \left(x - \frac{a+b}{2}\right)^{2\lceil \frac{\gamma}{2} \rceil} \psi_{a,b}(x)$. Importantly, this function behaves as polynomial of even degree $2\lceil \frac{\gamma}{2} \rceil$ in the affinity of $\frac{a+b}{2}$, while being infinitely smooth and supported on (a, b). It means that if we select a measure which is supported in the affinity of $\frac{a+b}{2}$ (light-blue hatched region) the function on the plot is essentially polynomial *w.r.t.* such a measure.

is infinitely smooth, thus it is (γ, L)-Hölder for a properly chosen C'. Figure 3.2 demonstrates the behavior of the considered construction in one dimension. Note that $\varphi_i(x)$ for $i = 1, 3$ are obtained from the previous mapping by re-scaling which preserves the Hölder constant L. Same reasoning applies to φ_i for $i = 0, 2, 4$.

Now, we define two marginal distributions μ_0, μ_1 by their densities as

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

and both μ_0, μ_1 are equal to zero in unspecified regions. Clearly, the strong density assumption is satisfied on $\mathcal{X}_0$ and $\mathcal{X}_4$ since the density is lower and upper-bounded by a constant independent of both N, n. The parameter ρ is chosen such that the strong density assumption on $\mathcal{X}_i$ for $i = 1, 2, 3$ is satisfied. Notice that

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

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

sets Γ_0^*, Γ_1^* are given as

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

The α-margin assumption: we are in position to check the margin Assumption 3.1.4. Let $t_0 = \frac{1}{2}\left(\frac{K-2\beta}{8K\beta} \bigwedge \frac{1}{4K}\right)$, thus for every $k \in \{2\beta+1,\dots,K\}$ and every $t \leq t_0$ we have

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

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

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

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

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

Therefore we can write using the strong density assumption

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

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

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

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

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

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

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

Notice that the constructed distributions are satisfying Assumption 3.1.1 since the measures are only defined on $\mathcal{X}_0, \mathcal{X}_1, \mathcal{X}_2, \mathcal{X}_3, \mathcal{X}_4$ and the regression functions on these sets are not concentrated around any constant.

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

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

Bound on the KL-divergence: we start by computing the KL-divergence between μ_0 and μ_0

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

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

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

Recall that we are interested in the following quantity

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

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

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

where $(*)$ is defined as

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

thus, for the Hamming risk we can write

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

Now we focus our attention on the sum of two Hamming differences which appears on the right hand side of the above inequality. Since $\mu_0(x) = \mu_1(x)$ for all $x \in \mathcal{X}_1$

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

Substituting this lower bound into the initial inequality we arrive at

$$\begin{aligned}
(*) &\geq 2\beta\kappa_{n,N} \mu_0^{\otimes(n+N)} \otimes \mathbb{P}_{Y|X}^{\otimes n} \left(\frac{d\mu_1^{\otimes(n+N)} \otimes \mathbb{P}_{Y|X}^{\otimes n}}{d\mu_0^{\otimes(n+N)} \otimes \mathbb{P}_{Y|X}^{\otimes n}} \bigwedge 1 \right) \\
&= 2\beta\kappa_{n,N} \left(1 - \mathrm{TV}\left(\mu_0^{\otimes(n+N)} \otimes \mathbb{P}_{Y|X}^{\otimes n}, \mu_1^{\otimes(n+N)} \otimes \mathbb{P}_{Y|X}^{\otimes n} \right) \right) \\
&= 2\beta\kappa_{n,N} \left(1 - \mathrm{TV}\left(\mu_0^{\otimes(n+N)}, \mu_1^{\otimes(n+N)} \right) \right) \\
&\geq 2\beta\kappa_{n,N} \left(1 - \sqrt{\frac{1}{2}\mathrm{KL}\left(\mu_0^{\otimes(n+N)}, \mu_1^{\otimes(n+N)} \right)} \right) \quad \text{(Pinsker's inequality)} \\
&\geq 2\beta\kappa_{n,N} \left(1 - 6\kappa_{n,N}\sqrt{n+N} \right) \ ,
\end{aligned}$$

which implies the desired lower bound on the Hamming risk.

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

$$\mathrm{D}(\hat{\Gamma}, \Gamma_i^*) := \mathcal{R}_\beta(\hat{\Gamma}) - \mathcal{R}_\beta(\Gamma_i^*) = \mu_i \sum_{k=1}^{K} \left| p_k(X) - G^{-1}(\beta) \right| \mathbb{1}_{\left\{k \in \hat{\Gamma}(X) \triangle \Gamma_i^*(X)\right\}} \ .$$

Again, recall that we are interested in

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

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

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

where $(**)$ is defined as

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

we can write

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

and we continue in the same manner as for $\mathrm{H}(\hat{\Gamma}, \Gamma_0^*) + \mathrm{H}(\hat{\Gamma}, \Gamma_1^*)$

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

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

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

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

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

Thus,

$$
\begin{aligned}
(**) &\geq \frac{K-2\beta}{8K} \kappa_{n,N} \left(1 - \mathrm{TV}\left(\mu_0^{\otimes(n+N)} \otimes \mathbb{P}_{Y|X}^{\otimes n}, \mu_1^{\otimes(n+N)} \otimes \mathbb{P}_{Y|X}^{\otimes n} \right) \right) \\
&\geq \frac{K-2\beta}{8K} \kappa_{n,N} \left(1 - \sqrt{\frac{1}{2} \mathrm{KL}\left(\mu_0^{\otimes(n+N)} \otimes \mathbb{P}_{Y|X}^{\otimes n}, \mu_1^{\otimes(n+N)} \otimes \mathbb{P}_{Y|X}^{\otimes n} \right)} \right) \\
&\geq \frac{K-2\beta}{8K} \kappa_{n,N} \left(1 - 6\kappa_{n,N} \sqrt{n+N} \right) \ ,
\end{aligned}
$$

which concludes the first part of the lower bounds.

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

In this section we prove that in case of the Hamming risk $\mathcal{E}^{\mathrm{H}}$ the rate $n^{-\alpha\gamma/(2\gamma+d)}$ is minimax optimal. Notice, that thanks to Proposition 3.1.5 a lower bound of order $n^{-\alpha\gamma/(2\gamma+d)}$ on the Hamming risk $\mathcal{E}^{\mathrm{H}}$ immediately implies a lower bound of order $n^{-(\alpha+1)\gamma/(2\gamma+d)}$ on both $\mathcal{E}^{\mathrm{E}}$ and $\mathcal{E}^{\mathrm{D}}$.

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 5.

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

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

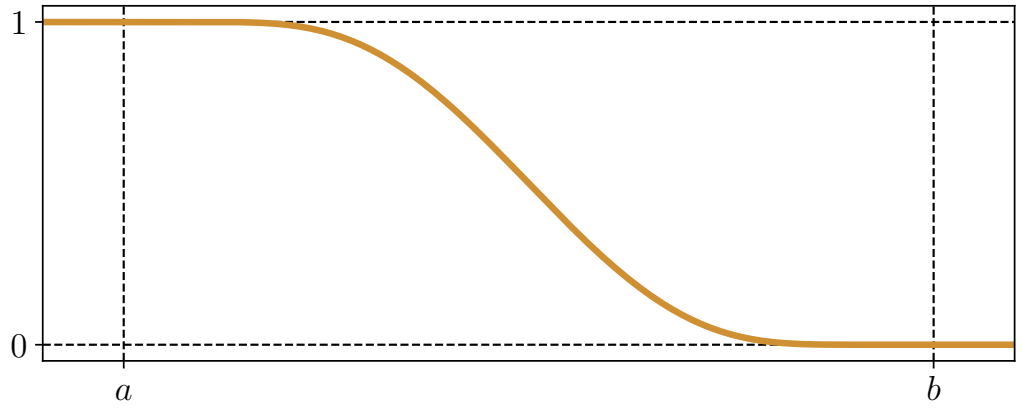

Figure 3.3: Integrated bump: $x \mapsto \frac{\int_x^\infty \psi_{a,b}(t)dt}{\int_a^b \psi_{a,b}(t)dt}$. Importantly, this function is infinitely smooth and is equal to one or zero only outside of the interval (a, b).

and denote by $n_q(x) \in G_q$ as the closest point of the grid G_q to the point $x \in \mathbb{R}^d$. We consider the partition of the unit cube $[0,1]^d \subset \mathbb{R}^d$ defined such that x and y belongs to the same subset if and only if $n_q(x) = n_q(y)$[4]. This partition is denoted by $\mathcal{X}'_0, \mathcal{X}'_1, \ldots, \mathcal{X}'_{q^d-1}$ such that $\mathcal{X}'_0 = \{x \in [0,1]^d \,:\, n_q(x) = (1/2q, \ldots, 1/2q)^\top\}$. Besides, denote by $\mathcal{X}'_{-j} \coloneqq \{x \in \mathbb{R}^d \,:\, -x \in \mathcal{X}'_j\}$ for all $j = 1, \ldots, q^d - 1$. For a fixed integer $m \leq q^d$ and for any $i \in \{1, \ldots, m\}$ define[5] $\mathcal{X}_i \coloneqq \mathcal{X}'_i$, $\mathcal{X}_{-i} \coloneqq \mathcal{X}'_{-i}$. Moreover, we introduce the set $\mathcal{X}_0 = \text{Ball}(0, (4q)^{-1})$. As we shall see, the set $\mathcal{X}_0$ is crucial since it allows to fix the threshold. The set $\bigcup_{i=-m}^m \mathcal{X}_i$ will be used to define the support of the marginal distribution of X.

For every $w \in W \coloneqq \{-1, 1\}^m$ we build the distribution $\mathbb{P}_w \in \mathcal{P}_W$, such that, the marginal distribution $\mathbb{P}_{w,X}$ does not dependent on $w \in \{-1, 1\}^m$ and the regression vector $(p_1^w(x), \ldots, p_K^w(x))$ is constructed as

[4]If $n_q(x)$ is not a singleton, then assign one of them arbitrary.

[5]Note that we dropped $\mathcal{X}'_0$ from the partition.

$$p_1^w(x) = \ldots = p_{\beta-1}^w(x) = v + \frac{c'}{\beta-1} + \frac{g(x)}{\beta-1} \ ,$$

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

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

$$p_{\beta+2}^w(x) = \ldots = p_K^w(x) = v - \frac{c'}{K-\beta-1} - \frac{g(x)}{K-\beta-1} \ ,$$

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

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

The function g is any (γ, L)-Hölder function with sufficiently bounded variation which is not concentrated around any constant, for example

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

for some constant C_g chosen small enough to ensure that it is (γ, L)-Hölder and has a bounded variation by $c'/2 \wedge v/4$. Moreover, the function ξ is constructed as

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

the function $\bar{u}$ is infinitely many times differentiable, is equal to zero on $(-\infty, 0]$ and to one on $[1, +\infty)$. Figure 3.3 shows the behavior of $1 - \bar{u}$. Taking the constant $\rho > 0$ big enough independently of N, n we can ensure that the function ξ is (γ, L)-Hölder.

The function ϕ is constructed similarly to the previous part of the rate, that is, we choose

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

with C_ϕ being sufficiently small such that $\phi(\cdot)$ is (γ, L)-Hölder and also upper-bounded by $c'/2 \wedge v/4$. Finally for the function φ we consider the following construction

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

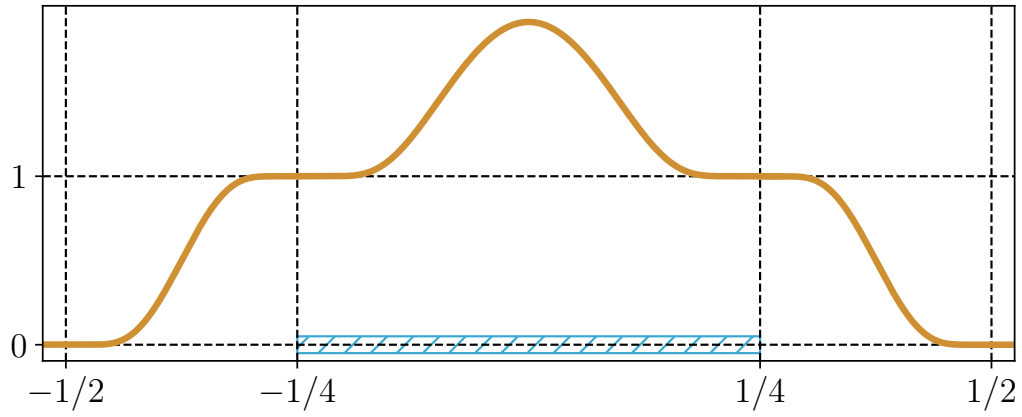

Figure 3.4: The function $x \mapsto u_2(|x|) + \psi_{-\frac{1}{4},\frac{1}{4}}(x)$. Importantly, this function is infinitely smooth and nowhere concentrates at any constant on $(-1/4, 1/4)$. If the marginal measure is supported on the light-blue hatched region the function is lower- and upper-bounded almost surely.

where $u_2(\cdot)$ is defined as

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

Figure 3.4 explains the behavior of this function and helps for better understanding of our results. The constant C_φ is chosen in such a way that the constructed function $\varphi(\cdot)$ is (γ, L)-Hölder and and upper-bounded by $c'/2 \wedge v/4$. Notice that the function $\varphi(x)$ for all $x \in \text{Ball}(0, (4q)^{-1})$ satisfies

$$C_\varphi q^{-\gamma} \leq \varphi(x) \leq C_\varphi q^{-\gamma} \left(1 + \psi_{-\frac{1}{4},\frac{1}{4}}(0)\right) \leq 2C_\varphi q^{-\gamma} \ . \tag{3.13}$$

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 $\text{Ball}(0, \sqrt{d} + \rho)$ and whose Lebesgue measure is $\text{Leb}(A_0) = 1 - mq^{-d}$. The density μ of the marginal distribution of $X \in \mathbb{R}^d$ is constructed as

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

for some τ 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 assumptions

- The functions $p_1^w, \ldots, p_K^w$ are defining some regression function for every $w \in W$. That is, for each $x \in \mathbb{R}^d$ we have $\sum_{k=1}^K p_k^w(x) = 1$ and $0 \leq p_k^w(x) \leq 1$,
- the functions $p_1^w, \ldots, p_K^w$ are (γ, L)-Hölder,

- the function $G_w(t) := \sum_{k=1}^K \int_{\mathbb{R}^d} \mathbb{1}_{\left\{\{p_k^w(x) \geq t\}\right\}} \mu(x)dx$ is continuous,
- the threshold $G^{-1}(\beta)$ is equal to v for every $w \in W$,
- the marginal distribution satisfies the strong density assumption,
- the regression function satisfies α-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

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

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

$$0 < v/4 \leq p_{\beta+1}^w(x) \leq 3v/2 \leq 1, \quad 0 < v/2 \leq p_\beta^w(x) \leq 7v/4 \leq 1 \ ,$$

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

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

due to the choice of c', v we have

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

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

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

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

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

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

The regression function is (γ, L)-Hölder: this implication follows immediately from the construction of φ, ξ, g.

Continuity of $G(t)$: first let us show that $\int_{\mathbb{R}^d} \mathbb{1}_{\left\{p_k^w(x) \geq t\right\}} \mu(x)dx$ is continuous for every $k \in [K]$. For $k = 1, \ldots, \beta-1, \beta+2, \ldots, K$ the continuity follows from the fact that g is never constant. For $k = \beta, \beta+1$ we first write

$$\begin{aligned}\int_{\mathbb{R}^d} \mathbb{1}_{\left\{p_k^w(x) \geq t\right\}} \mu(x)dx = &\sum_{c \in G_q \cup -G_q}^{m} \frac{\tau}{\mathrm{Leb}\left(\mathrm{Ball}(c, (4q)^{-1})\right)} \int_{\mathrm{Ball}(c,(4q)^{-1})} \mathbb{1}_{\left\{p_k^w(x) \geq t\right\}} dx \\ &+ \frac{\tau}{\mathrm{Leb}\left(\mathrm{Ball}(0, (4q)^{-1})\right)} \int_{\mathrm{Ball}(0,(4q)^{-1})} \mathbb{1}_{\left\{p_k^w(x) \geq t\right\}} dx \\ &+ \frac{1 - 2(m+1)\tau}{\mathrm{Leb}(A_0)} \int_{A_0} \mathbb{1}_{\left\{p_k^w(x) \geq t\right\}} dx \ ,\end{aligned}$$

thus for this choice of k the continuity follows from the fact that φ and g are never constant.

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

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

whereas for any $v' < v$, we have $\sum_{k=1}^{K} \mathbb{1}_{\left\{p_k^w(x) \geq v'\right\}} > \beta$ on $\mathcal{X}_0$. Besides, the corresponding β-Oracle sets Γ_w^* are given for every $w \in W$ as

$$\Gamma_w^*(x) = \begin{cases} \{1, \ldots, \beta-1, \beta\}, & x \in \mathcal{X}_i,\ w_i = 1, \\ \{1, \ldots, \beta-1, \beta+1\}, & x \in \mathcal{X}_i,\ w_i = -1, \\ \{1, \ldots, \beta-1, \beta\}, & x \in \mathcal{X}_{-i},\ w_i = -1, \\ \{1, \ldots, \beta-1, \beta+1\}, & x \in \mathcal{X}_{-i},\ w_i = 1, \\ \{1, \ldots, \beta-1, \beta\}, & x \in \mathbb{R}^d \setminus \left(\bigcup_{i=-m, i\neq 0}^{m} \mathcal{X}_i\right) \ . \end{cases}$$

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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$$\sum_{i=-m, i\neq 0}^{m} \mu\left(\varphi(X - n_q(X)) \leq t, X \in \mathcal{X}_i\right) = \frac{2m\tau}{\mathrm{Leb}(\mathrm{Ball}(0,(4q)^{-1}))} \int_{\mathrm{Ball}(0,(4q)^{-1})} \mathbb{1}_{\{\varphi(X)\leq t\}} dx \ .$$

Recall Eq. (3.13) that gives a bound for the function $\varphi(x)$ for all $x \in \mathrm{Ball}(0,(4q)^{-1})$

$$C_\varphi q^{-\gamma} \leq \varphi(x) \leq 2C_\varphi q^{-\gamma} \ .$$

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

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

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

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

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

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

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

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

Thus the margin assumption is satisfied as long as

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

Similarly one can check that the margin assumption is satisfied for $k = \beta + 1$. **Bound on the KL-divergence:** we are in position to upper-bound the KL divergence between any two hypotheses. Fix some $w, w' \in W$, then using the upper bound on $\varphi(\cdot)$ and the convex inequality $x\log(1+x) \leq 2x^2$ for sufficiently small x, we can write for some $C > 0$

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

How many hypotheses to take: let us recall the following result that we already used in the context of the F-score which is a version of Varshamov-Gilbert bound.

Lemma 6. *Let $\delta(\sigma,\sigma')$ denote the Hamming distance between $\sigma,\sigma' \in \{-1,1\}^m$ given by*

$$\delta(\sigma,\sigma') := \sum_{i=1}^{m} \mathbb{1}_{\{\sigma_i \neq \sigma'_i\}} \ .$$

There exists $\mathcal{W} \subset \{-1,1\}^m$ such that for all $\sigma \neq \sigma' \in \mathcal{W}$ we have

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

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

Denote $\mathcal{W} \subset W$ the set provided by Lemma 6 and by $\mathcal{P}_{\mathcal{W}}$ the set of distributions $\mathbb{P}^w$ with $w \in \mathcal{W}$. Taking into account all the above we conclude that $\mathcal{P}_{\mathcal{W}}$ satisfies the assumptions of our result.

Lower bound on the Hamming risk (applying Birgé's Lemma 5): finally, we are in position to lower bound the Hamming risk. Recall that we are interested in the following quantity

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

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

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

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

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

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

$$\begin{aligned} 2\mu \left(\left| \hat{\Gamma}(X) \triangle \Gamma^*_w(X) \right| \right) &\geq \mu \left(\left| \hat{\Gamma}(X) \triangle \Gamma^*_w(X) \right| \right) + \mu \left(\left| \hat{\Gamma}(X) \triangle \Gamma^*_{\hat{w}}(X) \right| \right) \\ &\geq \mu \left(|\Gamma^*_{\hat{w}}(X) \triangle \Gamma^*_w(X)| \right) \geq 2\delta(w,\hat{w}) \mu(\mathcal{X}_1) \\ &= 2\delta(w,\hat{w})\tau \geq \frac{m\tau}{2} \ . \end{aligned}$$

These arguments and Birge's lemma 5 imply that

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

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

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

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

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

One can easily verify that this choice of parameters τ, m, q is possible as long as $2\lceil \frac{\gamma}{2} \rceil \alpha \leq d$ and clearly with our choice we have $\tau m = \mathcal{O}(q^{-\alpha\gamma})$. As already mentioned the lower bound for the excess risk and the discrepancy follows from Propositions 3.1.5 and 3.1.6.

Inconsistency of top-β approach

In this section we prove Proposition 3.1.7. The proof builds an explicit construction of a distribution $\mathbb{P}$ whose β-Oracle satisfies $|\Gamma^*_\beta(x)| > \beta$ for all x in some $A \subset \mathbb{R}^d$ with $\mathbb{P}_X(A) > 0$. Clearly, if such a distribution exists then there is no estimator in $\hat{\Upsilon}_\beta$ that would consistently estimate this β-Oracle. Let $\beta \in [0, \dots, \lfloor K/2 \rfloor - 1]$ be a fixed integer and $K \geq 3$. For the proof of the theorem we shall construct one distribution $\mathbb{P}$ for which none of the estimators with a fixed information can perform well. We start by specifying the marginal distribution of $X \in \mathbb{R}^d$. We start the construction by specifying the density μ of the marginal distribution $\mathbb{P}_X$. Define a disk in $\mathbb{R}^d$ for some positive $r \leq r'$ as $\mathrm{Disk}(r, r') = \left\{x \in \mathbb{R}^d : r \leq \|x\| \leq r'\right\}$. First of all fix some parameters $r_1 < r_2 < 2r_2 < r_3 < 2r_3$ which are independent from n, N. The density μ is supported on $\mathrm{Ball}(0, r_1) \cup \mathrm{Disk}(r_2, 2r_2) \cup \mathrm{Disk}(r_3, 2r_3)$.

Moreover,

- $\mu(x) = \frac{\frac{\beta}{\beta+1} - \mathrm{Leb}(\mathrm{Ball}(0,r_1))}{\mathrm{Leb}(\mathrm{Disk}(r_2,2r_2))}$ for all $x \in \mathrm{Disk}(r_2, 2r_2)$,
- $\mu(x) = \frac{1}{(\beta+1)\,\mathrm{Leb}(\mathrm{Disk}(r_3,2r_3))}$ for all $x \in \mathrm{Disk}(r_3, 2r_3)$,
- $\mu(x) = 1$, for all $x \in \mathrm{Ball}(0, r_1)$,
- $\mu(x) = 0$ otherwise,

where $r_1 > 0$ is chosen small enough to ensure that $\frac{\beta}{\beta+1} - \mathrm{Leb}(\mathrm{Ball}(0, r_1)) > 0$. The regression function $p(\cdot) = (p_1(\cdot), \dots, p_K(\cdot))^\top$ are defined as

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

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

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

$\frac{3}{4}v\left(\frac{\|x\|-2r_2}{r_3-2r_2}\right)$. The above construction defines a distribution $\mathbb{P}$ for which we have

$$G^{-1}(\beta) = \frac{1}{2(\beta+1)}$$
$$\Gamma^*_\beta(x) = \begin{cases} \{1,\dots,\beta+1\}, & x \in \mathrm{Ball}(0,r_1) \bigcup \mathrm{Disk}(r_2,2r_2) \\ \emptyset, & \text{otherwise} \end{cases} .$$

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

$$\begin{aligned}\sum_{k=1}^{K} \int \mathbb{1}_{\{p_k(x) \geq G^{-1}(\beta)\}} \mu(x)dx &= (\beta+1)\left(\int_{\mathrm{Ball}(0,r_1)} \mu(x)dx + \int_{\mathrm{Disk}(r_2,2r_2)} \mu(x)dx\right) \\ &= (\beta+1)\left(\mathrm{Leb}\left(\mathrm{Ball}(0,r_1)\right) + \left(\tfrac{\beta}{\beta+1} - \mathrm{Leb}(\mathrm{Ball}(0,r_1))\right)\right) \\ &= \beta \ .\end{aligned}$$

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

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

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

$$\mathrm{P}(\hat{\Gamma}) - \mathrm{P}(\Gamma^*_\beta) \geq \frac{\frac{\beta}{\beta+1} - \mathrm{Leb}(\mathrm{Ball}(0,r_1))}{4(\beta+1)} \ ,$$

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

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

By construction, the regression vector is (γ, L)-Hölder and the density is lower- and upper-bounded by some positive constants. Hence, it remains to check that the constructed distribution satisfies the α-margin assumption. This can be achieved by an appropriate choice of r_1. Indeed, on the sets $\mathrm{Disk}(r_2,2r_2) \cup \mathrm{Disk}(r_3,2r_3)$ there is a "corridor" of constant size between the regression functions and the threshold $G^{-1}(\beta)$. The threshold $G^{-1}(\beta)$ is only approached by the regression function on the set $\mathrm{Ball}(0,r_1)$. As all the parameters in our construction are independent from $n, N \in \mathbb{N}$ we can find a value r_1 being small enough so that the α-margin assumption is verified for a fixed $\alpha > 0$.

Chapter 4

Multi-label classification

4.1 Constrained approach

Chapter overview. In this chapter we consider a problem of multi-label classification, where each instance is associated with some binary vector. Our focus is to find a classifier which minimizes false negative discoveries under constraints. Depending on the considered set of constraints we propose plug-in methods and provide non-asymptotic analysis under margin type assumptions. Specifically, we analyze two particular examples of constraints that promote sparse predictions: in the first one, we focus on classifiers with ℓ_0-type constraints and in the second one, we address classifiers with bounded false positive discoveries. Both formulations lead to different Bayes rules and, thus, different plug-in approaches. The first considered scenario is the popular multi-label top-K procedure: a label is predicted to be relevant if its score is among the K largest ones. For this case, we provide an excess risk bound that achieves so called "fast" rates of convergence under a generalization of the margin assumption to this settings. The second scenario differs significantly from the top-K settings, as the constraints are distribution dependent. We demonstrate that in this scenario the almost sure control of false positive discoveries is impossible without extra assumptions. To alleviate this issue we propose a sufficient condition for the consistent estimation and provide non-asymptotic upper bound.

4.1.1 Introduction

The goal of the multi-label classification is to annotate an observed object with a set of relevant labels. Several sophisticated algorithms have been recently developed, including tree based algorithms [Jain et al., 2016] and embedding based algorithms [Yu et al., 2014, Bhatia et al., 2015] which are considered to be state-of-the-art. Other contributions have rather focused on efficient implementations of existing multi-label strategies: for instance in [Babbar and Schölkopf, 2017] the authors developed a large-scale distributed framework relying on one-versus-rest strategy applied to linear classifiers, plug-in type classifiers were considered in [Dembczynski et al., 2013].

A consensus on the choice of the performance measure is still missing. Yet, most recent works have pointed out that it is more rewarding to correctly predict a relevant[1] label than to give a correct prediction on irrelevant labels, see [Jain et al., 2016] for a thorough discussion on this topic. Such asymmetry is usually explained by the label space sparsity, that is, there

[1] A label is called relevant for an instance if this instance is tagged with this label.

is only a small set of relevant labels compared to the set of irrelevant ones. It also suggests that the classical Hamming loss is not well tailored for sparse multi-label problems as it treats both false positives and false negatives equally; thus, some modifications ought to be proposed.

To introduce this asymmetric information in a learning algorithm, one can modify the objective loss function to be minimized. For instance, in [Jain et al., 2016] the authors have weighted each label, according to their observed frequency over a dataset. These weights are motivated by the propensity model, which introduces a possibility of non-observing a relevant label. To be more precise, Jain et al. [2016] propose to down-weight the reward for correctly predicting a frequent label, which is motivated by the observation that the frequent labels can be easily predicted by a human. In [Chzhen et al., 2017], the authors proposed to weight false positive (irrelevant labels predicted to be relevant) and false negative (relevant labels predicted to be irrelevant) discoveries separately. The empirical risk minimization procedure was then analyzed thanks to Rademacher's complexity techniques.

Another possible direction is to consider a more complex family of loss functions, which are called non-decomposable, such as F_1-score (see Section 2.1) or AUC among others. A general class of loss functions which can be represented as a ratio of false discoveries is studied in [Koyejo et al., 2015]. Koyejo et al. [2015] showed that the oracle (Bayes optimal) classifier can be obtained by thresholding the regression functions associated with each label, that is, the probability of a label to be relevant. Additionally, the authors proved that algorithms based on plug-in are consistent and have a good empirical performance. In a similar direction, Dembczynski et al. [2013] empirically showed that plug-in algorithms outperform the ones based on the structured loss minimization, in the context of multi-label classification with F_1-score performance measure. Dembczynski et al. [2013] additionally established a statistical consistency of the considered algorithms. Finally, convex empirical risk minimization was studied in [Gao and Zhou, 2011], where authors proved an infinite sample size consistency for convexified Hamming loss and ranking loss. Consistency results are common in the multi-label classification literature. Though, results of non-asymptotic nature, *e.g.,* excess risk bounds, have not received much attention in these settings.

Due to the sparse nature of the problem we propose to focus on classifiers that minimize false negative discoveries and exhibit desirable structural properties. This can be seen as a problem constrained estimation, mainly considered in the settings of regression or parametric estimation [Lepskii, 1990]. In the constrained estimation, similarly to this case, the goal is to find an estimation which inherit some properties desired by a statistician. In this chapter we consider two particular choices of structural constraints. The first type of constraints describes classifiers with a bounded number of predicted labels: for instance, this approach appears naturally in recommendation systems. Bayes optimal classifier in this context is given by the top-K procedure, popular among practitioners: a label is predicted to be relevant if its associated score is among the top-K values. The popularity of this approach is reflected by several recent works [Lapin et al., 2015, Li et al., 2017], where top-K procedures are studied both from applied and asymptotic points of view. In contrast, for this scenario, we establish a non-asymptotic excess risk bound for plug-in based classifiers. The obtained bound can attain "fast" and "super-fast" (faster than $1/n$) rates under a multi-label version margin assumption, similarly to the standard binary classification setup discussed in Section 1.1.

For the second scenario, we consider a set of classifiers with a control over false positive discoveries. This can be relevant when one can tolerate a few false positive discoveries, but needs a parameter which quantifies the level of tolerance. To provide guarantees for this instance, we introduce a different set of assumptions which reflects the label sparsity of a

typical multi-label problem. Under these assumptions, we prove an excess risk bound similar (in terms of rates) to the bound obtained by Denis and Hebiri [2015a], where the authors analyzed a binary classification framework with a control over the probability of rejection.

Organization of the chapter

This chapter is organized in the following way: in Section 4.1.2, we introduce notation used throughout this chapter, formally state the considered framework and lay down important results that we use. Further, Sections 4.1.3 and 4.1.4 are devoted to the theoretical analysis of plug-in rules in the two scenarios mentioned above. We conclude the chapter by a discussion on possible extensions in Section 4.1.5. All the proofs are gathered in Section 4.1.7.

4.1.2 Framework and notation

In this section, we introduce the notation used in this chapter and present the proposed constrained framework. For any positive integer number n we denote by $[n] = \{1, \dots, n\}$ the set of integers between 1 and n. For any vector a in a Euclidean space $\mathbb{R}^n$ and for all $i \in [n]$ we denote by a^i the i^{th} component of the vector a. We denote by $\|\cdot\|_0$ the ℓ_0 norm of a vector, which in case of binary vectors reduces to the number of ones. For every real numbers a, b we denote by $a \wedge b$ the minimum between a and b. Let $(X, Y) \sim \mathbb{P}$, where $X \in \mathcal{X} = \mathbb{R}^d$ and $Y = (Y^1, \dots, Y^L)^\top \in \mathcal{Y} = \{0, 1\}^L$. Denote by $\mathbb{P}_X$ the marginal distribution of X. In this chapter, a classifier $g = (g^1, \dots, g^L)^\top$ is a measurable function from $\mathcal{X}$ to $\mathcal{Y}$, that is $g : \mathcal{X} \mapsto \mathcal{Y}$, and we write $\mathcal{G}(\mathcal{X}, \mathcal{Y})$ for the set of all classifiers (measurable functions). Let $\eta(x) = (\eta^1(x), \dots, \eta^L(x))^\top : \mathcal{X} \mapsto [0, 1]^L$ be the component wise regression function, meaning that for all $l \in [L]$ the l^{th} component of $\eta(x)$ is given by $\eta^l(x) = \mathbb{P}(Y^l = 1 | X = x)$. We denote by $\sigma = (\sigma_1, \dots, \sigma_L)$ a permutation[2] of $[L]$ such that the regression functions is ranked as

$$\eta^{\sigma_1}(x) \geq \ldots \geq \eta^{\sigma_L}(x) \ ,$$

for all $x \in \mathbb{R}^d$. The average false negative risk of a classifier $g \in \mathcal{G}(\mathcal{X}, \mathcal{Y})$ is denoted by

$$\mathcal{R}(g) = \frac{1}{L} \sum_{l=1}^{L} \mathbb{P}\left(g^l(X) = 0, Y^l = 1\right) \ . \tag{4.1}$$

Let us first recall the general setting of constrained classification presented in Chapter 1. For a fixed subset of classifiers $\mathcal{G}_\theta \subset \mathcal{G}(\mathcal{X}, \mathcal{Y})$ parametrized by some abstract θ and specified according to the context, we define a $\mathcal{G}_\theta$-Bayes classifier as

$$g_* \in \arg\min \{\mathcal{R}(g) \ : \ \mathcal{G}_\theta\} \ . \tag{4.2}$$

Importantly, unlike all previous examples in this section we consider those sets $\mathcal{G}_\theta$ which are not written as equalities (inequalities) in expectation. Here we would focus on *almost sure* type constraints, for which plug-in type methods can still be applied.

Again, let us recall that the $\mathcal{G}_\theta$-Bayes rule g_* depends both on the distribution of (X, Y) and on the set of predictors $\mathcal{G}_\theta$. We assume that the minimum is achieved by a classifier $g_* \in \mathcal{G}_\theta$, though we do not assume that this classifier is unique.

Intuitively, this framework aims at minimization of the total number of mistakes on $Y^l = 1$ (relevant labels), over a class of prediction rules $\mathcal{G}_\theta$. For example, the case $\mathcal{G}_\theta = \mathcal{G}(\mathcal{X}, \mathcal{Y})$

[2] we omit the dependence on x and write σ instead of $\sigma(x)$.

leads to an $\mathcal{G}(\mathcal{X},\mathcal{Y})$-oracle $g_* \equiv (1,\dots,1)^\top$, which reflects a complete tolerance over false positive discoveries. This simple example shows, that the choice of $\mathcal{G}_\theta$ is a crucial modeling part of the proposed framework.

Given a data sample $\mathcal{D}_n^{\mathrm{L}} = \{(X_i, Y_i)\}_{i=1}^n$, which consists of *i.i.d.* copies of (X, Y), the goal here is to construct an estimator $\hat{g}$, based on $\mathcal{D}_n^{\mathrm{L}}$, of the $\mathcal{G}_\theta$-Bayes g_*. Estimator $\hat{g}$ is a function that assigns a classifier to every learning sample $\mathcal{D}_n^{\mathrm{L}}$, that is, $\hat{g} : \cup_{n=1}^\infty (\mathcal{X}\times\mathcal{Y})^n \mapsto \mathcal{G}(\mathcal{X},\mathcal{Y})$. We denote by $\mathbb{P}^{\otimes n}$ the product probability measure according to which the data sample $\mathcal{D}_n^{\mathrm{L}}$ is distributed, and by $\mathbb{E}_{\mathbb{P}^{\otimes n}}$ the expectation with respect to $\mathbb{P}^{\otimes n}$. The goal is to provide non-asymptotic bounds on the excess risk $\mathbb{E}_{\mathbb{P}^{\otimes n}}\Big[\mathcal{R}(\hat{g})\Big] - \mathcal{R}(g_*)$. Recall that in the constrained classification framework, we want our estimate $\hat{g}$ to satisfy one of the following conditions:

$$\hat{g} \in \mathcal{G}_\theta,\ \forall n \in \mathbb{N}; \quad \text{or} \quad \hat{g} \xrightarrow[n\to\infty]{} g \in \mathcal{G}_\theta \ , \tag{4.3}$$

where the kind of convergence is to be specified later. Since, the $\mathcal{G}_\theta$-Bayes g_* is typically available in a closed form and depends on an unknown, in practice, regression vector $\eta(x)$, we consider the plug-in type methods. As discussed in Section 1.0.3, a vast amount of literature is focused on the estimation of the regression function $\eta(x)$, that is why this part is not a central object of this particular study. In other words, we are rather interested in describing the performance of a classifier based on an arbitrary estimator $\hat{\eta}(x)$ of the regression function $\eta(x)$ which satisfies for all $l \in [L]$ the following assumption:

Assumption 12 (Exponential bound)**.** *For some positive constants $C_1, C_2 > 0$ and $\gamma > 0$, for all $\delta > 0$ and for all $l \in [L]$ we have:*

$$\mathbb{P}^{\otimes n}\left(|\eta^l(x) - \hat{\eta}^l(x)| \geq \delta\right) \leq C_1 \exp(-C_2 n^\gamma \delta^2) \ \ a.e.\ x \in \mathbb{R}^d \ w.r.t.\ \mathbb{P}_X \ . \tag{4.4}$$

Such a bound holds for various type of estimators and distributions in both parametric [Li et al., 2015] and non-parametric settings [Audibert and Tsybakov, 2007]. In non-parametric settings, typically, the parameter γ depends on the smoothness of η and on the dimension d. Let us notice, that empirical evidences [Dembczynski et al., 2013] suggest to use multinomial logistic regression as an effective estimator for the regression function, though, this estimator might not have the exponential concentration. The rest of the chapter is devoted to theoretical analysis of two specific families $\mathcal{G}_\theta$. In both cases we derive the $\mathcal{G}_\theta$-Bayes classifier g_*, defined in Eq. (4.2). Typically, the $\mathcal{G}_\theta$-Bayes g_* depends on the regression function η, due to the form of the risk considered in Eq. (4.1). Explicit expression for the $\mathcal{G}_\theta$-Bayes, provides with a natural motivation to consider plug-in type rules for the construction of $\hat{g}$. We establish one of the properties in Eq. (4.3) and introduce the set of additional assumptions in order to upper-bound the excess risk.

Let us finish this section with one generic notation used in this chapter. For the estimator $\hat{\eta}(x)$ we denote by $\hat{\sigma} = \hat{\sigma}(x)$ a permutation of $[L]$ such that the following holds for all $x \in \mathbb{R}^d$

$$\hat{\eta}^{\hat{\sigma}_1}(x) \geq \ldots \geq \hat{\eta}^{\hat{\sigma}_L}(x) \ ,$$

we again omit the dependence on x and write $\hat{\sigma}$ instead of $\hat{\sigma}(x)$. We reserve σ and τ for a non-decreasing permutations of $\eta(x)$ and $\hat{\eta}(x)$ respectively.

4.1.3 Control over sparsity

In this section, we consider, the set of K-sparse classifiers, defined for a fixed $K \in [L]$ as:

$$\mathcal{G}_K^{\mathrm{sp}} := \{g \in \mathcal{G}(\mathcal{X},\mathcal{Y}) : \forall x \in \mathbb{R}^d,\ \|g(x)\|_0 \leq K\} \ , \tag{4.5}$$

and put $\mathcal{G}_\theta = \mathcal{G}_K^{\text{sp}}$ (in this case $\theta = K$). Hence, we are interested in a K-sparse classifier, which minimize the total number of mistakes on relevant labels. It is not hard to see that, a $\mathcal{G}_K^{\text{sp}}$-oracle g_* is given by the top-K procedure, this is stated formally in the following lemma:

Lemma 17 ($\mathcal{G}_K^{\text{sp}}$-oracle classifier)**.** *An $\mathcal{G}_K^{sp}$-oracle g_* can be obtained for all $x \in \mathbb{R}^d$ as:*

$$g_*^{\sigma_1}(x) = \ldots = g_*^{\sigma_K}(x) = 1 \ ,$$
$$g_*^{\sigma_{K+1}}(x) = \ldots = g_*^{\sigma_L}(x) = 0 \ .$$

Remark 10. *Observe, that in order to recover the $\mathcal{G}_K^{sp}$-oracle g_* the only information that is needed is $\{\sigma_1(x), \ldots, \sigma_K(x)\}$. In particular, any additional information about the regression vector $\eta(x)$ is not relevant.*

A plug-in strategy $\hat{g}$ in this case can be defined in a straightforward way for all $x \in \mathbb{R}^d$ as:

$$\hat{g}^{\hat{\sigma}_1}(x) = \ldots = \hat{g}^{\hat{\sigma}_K}(x) = 1 \ , \tag{4.6}$$
$$\hat{g}^{\hat{\sigma}_{K+1}}(x) = \ldots = \hat{g}^{\hat{\sigma}_L}(x) = 0 \ . \tag{4.7}$$

Obviously, the plug-in estimator defined above is exactly a K-sparse classifier, that is $\hat{g} \in \mathcal{G}_K^{\text{sp}}$ for every choice of the data sample $\mathcal{D}_n^{\text{L}}$, as required in Eq. (4.3). Since our goal is to predict as relevant the labels with the top-K probabilities, it is natural to restrict our attention to the distributions for which such top-K labels are well separated. In this context, we use a top-K margin assumption in the following form:

Assumption 13 (top-K margin assumption)**.** *We say that the regression vector $\eta(x)$ satisfies top-K margin assumption, if there exist positive constants C, α such that for all $\delta > 0$:*

$$\mathbb{P}_X\{0 < \eta^{\sigma_K}(X) - \eta^{\sigma_{K+1}}(X) \leq \delta\} \leq C\delta^\alpha \ .$$

This assumption is similar to the classical margin assumption used in the context of binary classification. Under Assumption 13 we can obtain the following bound on the excess risk of $\hat{g}$:

Theorem 17. *Under Assumptions 12 and 13, the excess risk of the plug-in classifier in Eq. (4.6) can be bounded as follows:*

$$\mathbb{E}_{\mathbb{P}^{\otimes n}}\left[\mathcal{R}(\hat{g})\right] - \mathcal{R}(g_*) \leq \tilde{C}K\frac{L-K}{L}n^{-\frac{\gamma(\alpha+1)}{2}} \ ,$$

for some universal constant $\tilde{C}$.

The proof of Theorem 17 is based on the following upper bound on the excess risk:

$$\mathcal{R}(\hat{g}) - \mathcal{R}(g_*) \leq \mathbb{E}_{\mathbb{P}_X}\frac{1}{L}\sum_{l=1}^{K}\sum_{j=K+1}^{L}(\eta^{\sigma_l}(X) - \eta^{\sigma_j}(X))\mathbb{1}_{\left\{\hat{g}^{\sigma_l}(X)=0, \hat{g}^{\sigma_j}(X)=1\right\}} \ , \tag{4.8}$$

which in the case $L = 2$ and $K = 1$ reduces to the classical excess risk in binary classification. We notice that there are two interesting consequences of this bound: first, the bound can attain "fast" $(1/n)$ and "super-fast" (faster than $1/n$) rates of convergence in terms of n, depending on γ and α; second, the value of the parameter K (chosen by the practitioner) is often small in applications compared to the total amount of labels L. Hence, the obtained bound illustrates the good performance of the proposed method as it behaves proportionally to K rather than to L. This is crucial when one tries to address scenarios where the total amount of observations n is of the same order as L. Moreover, we expect that the dependence on K and L can be improved or even avoided, since the upper bound on the excess risk in Eq. (4.8) is rather rough.

4.1.4 Control over false positives

In this section, we consider the set of classifiers with controlled false positive discoveries, defined for a fixed $\beta \in [L]$ as:

$$\mathcal{G}_\beta^{\text{fp}} := \left\{ g \in \mathcal{G}(\mathcal{X}, \mathcal{Y}) : \sum_{l=1}^{L} \mathbb{P}\left(g^l(X) = 1, Y^l = 0 | X\right) \leq \beta, \, \mathbb{P}_X\text{-}a.s. \right\} \ , \tag{4.9}$$

and put $\mathcal{G}_\theta = \mathcal{G}_\beta^{\text{fp}}$ (in this case $\theta = \beta$).

Remark 11. *One should note that the following inclusion holds for all $\beta \in [L]$:*

$$\mathcal{G}_\beta^{sp} \subset \mathcal{G}_\beta^{fp} \ ,$$

which indicates that the top-β strategy controls the false positive discoveries. This is intuitive, as the top-β procedure is not making more than β false positive discoveries. However, this control is not optimal in a situation when a larger (compared to β) set of labels could be relevant. In such a scenario, the $\mathcal{G}_\beta^{fp}$-oracle classifier is more advantageous as it is able to output a larger set of potentially relevant labels and still has a controlled false positive discoveries.

As in the previous section, the $\mathcal{G}_\beta^{\text{fp}}$-oracle classifier is given by thresholding the top components of the regression function. However, unlike the previous sparse strategy, in this case the amount of positive components can be different for every $x \in \mathbb{R}^d$.

Lemma 18 (Oracle classifier)**.** *An $\mathcal{G}_\beta^{fp}$-oracle g_* can be obtained for every $x \in \mathbb{R}^d$ as*

$$\begin{aligned} g_*^{\sigma_1}(x) = \ldots = g_*^{\sigma_K}(x) = 1 \ , \\ g_*^{\sigma_{K+1}}(x) = \ldots = g_*^{\sigma_L}(x) = 0 \ , \end{aligned}$$

where $K = K(x)$ is defined as

$$K(x) = \max\left\{ m \in [L] : \sum_{l=1}^{m} (1 - \eta^{\sigma_l}(x)) \leq \beta \right\} \ . \tag{4.10}$$

In this case the optimal strategy can be characterized as top-$K(X)$, where $K(X)$ is a random variable defined in Eq. (4.10). Intuitively, for each feature vector $x \in \mathbb{R}^d$ the threshold $K(x)$ selects labels with high probability to be relevant, the larger the value β (which indicates the higher level of tolerance), the more labels are predicted to be relevant.

Remark 12. *Notice, that unlike the previous scenario, it is not sufficient to recover the ordering of the regression function $\eta(x)$ to obtain the $\mathcal{G}_\beta^{fp}$-oracle. Indeed, due to the definition of $K(X)$, even the knowledge of the whole non-decreasing permutation σ is not sufficient without additional information about the components of the regression vector $\eta(x)$.*

Similarly to the previous section, a natural plug-in strategy $\hat{g}$ reads for all $x \in \mathbb{R}^d$:

$$\hat{g}^{\hat{\sigma}_1}(x) = \ldots = \hat{g}^{\hat{\sigma}_{\hat{K}}}(x) = 1 \ , \tag{4.11}$$

$$\hat{g}^{\hat{\sigma}_{\hat{K}+1}}(x) = \ldots = \hat{g}^{\hat{\sigma}_L}(x) = 0 \ , \tag{4.12}$$

where $\hat{K} = \hat{K}(x)$ is defined as

$$\hat{K}(x) = \max\left\{m \in [L] : \sum_{l=1}^{m}(1 - \hat{\eta}^{\hat{\sigma}_l}(x)) \leq \beta\right\} \ .$$

Ultimately, to recover the $\mathcal{G}^{\text{fp}}_\beta$-oracle g_* we need to estimate both: the non-decreasing permutation σ and the regression vector η. Since, we do not have an access to either of those quantities, we use an estimator $\hat{\eta}$ and its own non-decreasing permutation τ. We define $\hat{\mathcal{G}}^{\text{fp}}_\beta$, replacing $(\eta^l(x))_{l=1}^L$ by $(\hat{\eta}^l(x))_{l=1}^L$ in the definition of $\mathcal{G}^{\text{fp}}_\beta$, in order to prove one of the properties in Eq. (4.3)

Definition 19 (Plug-in $\mathcal{G}^{\text{fp}}_\beta$-set)**.** *For every $\beta \in [L]$ we denote the plug-in β-set as*

$$\hat{\mathcal{G}}^{fp}_\beta := \left\{g \in \mathcal{G}(\mathcal{X}, \mathcal{Y}) : \sum_{l=1}^{L} \mathbb{1}_{\{g^l(X)=1\}}(1 - \hat{\eta}^l(X)) \leq \beta, \, \mathbb{P}_X\text{-}a.s.\right\} \ .$$

Due to the approximation error of $\hat{\eta}$ the plug-in rule $\hat{g}$ does not necessary belong to the set $\mathcal{G}^{\text{fp}}_\beta$ (hence is not comparable to the $\mathcal{G}^{\text{fp}}_\beta$-oracle g_*). However, if the estimator $\hat{\eta}$ of η is consistent, then every classifier $g \in \hat{\mathcal{G}}^{\text{fp}}_\beta$ has asymptotically bounded false positive discoveries on the level β.

Lemma 19 (Embedding of the plug-in set)**.** *There exists $\bar{\beta}$ which satisfies $\bar{\beta} \leq \beta + \sum_{l=1}^L \left\|\eta^l - \hat{\eta}^l\right\|_\infty$ such that for every $g \in \hat{\mathcal{G}}^{fp}_\beta$, we have*

$$g \in \mathcal{G}^{fp}_{\bar{\beta}} \ .$$

Moreover, under Assumption 12, with probability at least $1 - \epsilon$ over the dataset $\mathcal{D}^{\text{L}}_n$ it holds that

$$\hat{\mathcal{G}}^{fp}_\beta \subset \mathcal{G}^{fp}_{\bar{\beta}} \ ,$$

where $\bar{\beta} = \beta + O(Ln^{-\gamma/2}\sqrt{\ln(C_1/\epsilon)})$.

Since, clearly, $\hat{g} \in \hat{\mathcal{G}}^{\text{fp}}_\beta$ by construction, we establish the second requirement in Eq. (4.3), which is a desired property as we want to restrict our attention to the collection of classifiers $\mathcal{G}^{\text{fp}}_\beta$.

Assumption 14 (Local margin assumption)**.** *We say that the regression vector $\eta(x)$ satisfies local margin assumption, if there exist constants $C_0 > 0, \alpha_1 > 0$ such that for all $\delta > 0$, we have*

$$\mathbb{P}_X\left(\eta^{\sigma_k}(X) - \eta^{\sigma_{k+1}}(X) \leq \delta, K(X) = k\right) \leq C_0\delta^{\alpha_1} \ .$$

This assumption states that in the optimal thresholding $K(X) = k$ there is a gap between the k^{th} and $(k+1)^{\text{th}}$ regression function, which is similar to Assumption 13. This is needed in order to recover the permutation σ, at least partially until its K^{th} element. However, since the amount of labels is not fixed a priori and is itself a random variable, the form of this assumption slightly differs from Assumption 13. Additionally one should observe that unlike Assumption 13, the later restricts the possibility of $\eta^{\sigma_k}(X)$ and $\eta^{\sigma_{k+1}}(X)$ to coincide on a set of large measure, which is similar to [Tsybakov, 2004].

Assumption 15 (Sparsity). *We say that the regression vector $\eta(x)$ satisfies sparsity assumption, if for a positive integer S smaller than L, we have*

$$\sum_{l=1}^{L} \mathbb{P}\left(Y^l = 1|X\right) \leq S, \quad \mathbb{P}_X\text{-}a.s.$$

This assumption is similar to the one used in [Chzhen et al., 2017], and aims at leveraging sparsity of most real datasets. It is natural to expect that the value of the sparsity S is smaller than the total amount of labels L. Even though, our analysis does not explicitly assume this relation between S and L, bounds that we obtain are more advantageous for such a scenario. We finally introduce the assumption that is more structural and states that the sum of top regression functions is not too concentrated around β.

Assumption 16 (Global margin assumption). *We say that the regression vector $\eta(x)$ satisfies global margin assumption if, there exists $\alpha_2 > 0$, such that for all $k \geq \beta$, for all $l \in [k]$, and for all $\delta > 0$, we have*

$$\mathbb{P}_X\left(\frac{1}{l}|\sum_{j=k-l+1}^{k}(1-\eta^{\sigma_j}(X)) - \beta| \leq \delta, K(X) = k\right) \leq \beta^{\alpha_2}\delta^{\alpha_2}h(|k-l|) \ ,$$

where $h : \mathbb{R}_+ \mapsto \mathbb{R}_+$ such that $h(0) = 1$ and

$$\sum_{k=\beta}^{L}\sum_{l=1}^{L} h(|k-l|) \leq \tilde{C}(L-\beta) \ ,$$

for some $\tilde{C} > 0$.

The multiplier β^{α_2} is due to the fact that the following inequality must always be satisfied for $\delta = \frac{1}{k}$ and $l = k$:

$$\begin{aligned}\sum_{k=\beta}^{L}\mathbb{P}_X\Big(\frac{1}{k}|\sum_{j=1}^{k}(1-\eta^{\sigma_j}(X)) - \beta| \leq \frac{1}{k}, K(X) = k\Big) &= \sum_{k=\beta}^{L}\mathbb{P}_X\Big(K(X) = k\Big) \\ &= 1 \leq \sum_{k=\beta}^{L}\beta^{\alpha_2}\frac{1}{k^{\alpha_2}}h(0) \ ,\end{aligned}$$

where the first equality holds since on the event $K(X) = k$ the quantity $|\sum_{j=1}^{k}(1-\eta^{\sigma_j}(X)) - \beta|$ is always upper bounded by one. The definition of the function h states that the matrix $H_{k,l} = h(|k-l|)$ is a diagonally dominant matrix such that for all $k \geq \beta$ we have $AH_{k,k} \geq \sum_{l\neq k} H_{k,l}$ for some positive constant A independent from L.

Let us provide a simple intuition for the necessity of Assumption 16. Consider the following multi-label classification problem: $L \geq 2$, $S = 2$, $\beta = 1$, $X \in [0,1]$. And let us define two probability measures $\mathbb{P}_{-1}, \mathbb{P}_{+1}$ which have the marginal distribution $\mathbb{P}_{\pm 1,X} \equiv \text{Leb}$, where Leb is the Lebesgue measure on $[0,1]$. Under both $\mathbb{P}_{-1}, \mathbb{P}_{+1}$ the labels $Y^1, \ldots, Y^L$ are independent, $Y^l \equiv 0$ for all $l = 3, \ldots, L$, $\mathbb{P}_\rho(Y^1 = 1|X) \equiv 3/4$, and $\mathbb{P}_\rho(Y^2 = 1|X) = \eta^2_\rho(X) = 1/4 - \rho\phi_n^{-1}$ for $\rho \in \{-1, 1\}$ and some strictly increasing sequence ϕ_n of $n \in \mathbb{N}$. Assume also that ϕ_n is chosen in such a way that $\phi_1^{-1} \leq 1/8$. One can see, thanks to Lemma 18, that the Oracle classifiers under $\mathbb{P}_{-1}$ and $\mathbb{P}_{+1}$ are given by

$$\begin{aligned} g_*^{-1}(x) &= (1, 1, 0, \ldots, 0)^\top \ , \\ g_*^{+1}(x) &= (1, 0, 0, \ldots, 0)^\top \ , \end{aligned}$$

respectively and the optimal thresholds are $K^{-1}(x) \equiv 2$, $K^{+1}(x) \equiv 1$. Now, let us consider minimax risk[3] over $\mathcal{P} = \{\mathbb{P}_{+1}, \mathbb{P}_{-1}\}$ defined as

$$\inf_{\hat{g}} \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}^{\otimes n}} \left|\mathcal{R}(\hat{g}) - \mathcal{R}(g_*)\right| = \inf_{\hat{g}} \sup_{\rho\in\{-1,1\}} \mathbb{E}_{\mathbb{P}_\rho^{\otimes n}} \left|\mathcal{R}_{\mathbb{P}_\rho}(\hat{g}) - \mathcal{R}_{\mathbb{P}_\rho}(g_*^\rho)\right| \ .$$

Then, for the excess risk $\mathcal{E}_\rho(\hat{g}) = \left|\mathcal{R}_{\mathbb{P}_\rho}(\hat{g}) - \mathcal{R}_{\mathbb{P}_\rho}(g_*^\rho)\right|$ we can write

$$\mathcal{E}_\rho(\hat{g}) = \frac{1}{L}\left|\frac{3}{4}\int_0^1 \mathbb{1}_{\{\hat{g}^1(x)=0\}}dx + \int_0^1 \eta_\rho^2(x)\left(\mathbb{1}_{\{\hat{g}^2(x)=0\}} - \mathbb{1}_{\left\{(g_*^\rho)^2(x)=0\right\}}\right)dx\right| \ ,$$

moreover, using the triangle inequality we can lower bound $\mathcal{E}_{+1}(\hat{g}) + \mathcal{E}_{-1}(\hat{g})$ by

$$\frac{1}{L}\left|-2\phi_n^{-1}\int_0^1 \mathbb{1}_{\{\hat{g}^2(x)=0\}}dx - \int_0^1 \eta_{+1}^2(x)\mathbb{1}_{\left\{(g_*^{+1})^2(x)=0\right\}}dx + \int_0^1 \eta_{-1}^2(x)\mathbb{1}_{\left\{(g_*^{-1})^2(x)=0\right\}}dx\right| \ .$$

Recall that $\{x \in [0,1] : (g_*^{+1})^2(x) = 0\} = [0,1]$ and $\{x \in [0,1] : (g_*^{-1})^2(x) = 0\} = \emptyset$, thus

$$\begin{aligned}
\mathcal{E}_{+1}(\hat{g}) + \mathcal{E}_{-1}(\hat{g}) &\geq \frac{1}{L}\left|-2\phi_n^{-1}\int_0^1 \mathbb{1}_{\{\hat{g}^2(x)=0\}}dx - \frac{1}{4} + \phi_n^{-1}\right| \\
&\geq \frac{1}{4L} - \frac{1}{L}\left|-2\phi_n^{-1}\int_0^1 \mathbb{1}_{\{\hat{g}^2(x)=0\}}dx + \phi_n^{-1}\right| \\
&\geq \frac{1}{4L} - \frac{\phi_n^{-1}}{L}\int_0^1 \left|\mathbb{1}_{\{\hat{g}^2(x)=0\}} - \mathbb{1}_{\{\hat{g}^2(x)=1\}}\right| dx \\
&= \frac{1}{4L} - \frac{\phi_n^{-1}}{L} \geq \frac{1}{8L} \ .
\end{aligned}$$

where the last inequality is due to our choice of ϕ_n. Now, let us define another probability measure $\mathbb{P}_0$ on $[0,1] \times \{0,1\}^L$, such that $\mathbb{P}_{0,X} \equiv$ Leb on $[0,1]$, $Y^1, \ldots, Y^L$ are independent, $Y^l \equiv 0$ for $l = 3, \ldots, L$ and $\mathbb{P}_0(Y^1 = 1|X) \equiv 3/4$, $\mathbb{P}_0(Y^2 = 1|X) \equiv 1/4$. Importantly, both $\mathbb{P}_{-1}$ and $\mathbb{P}_{+1}$ are absolutely continuous *w.r.t.* to $\mathbb{P}_0$. Let us write $(*) = \inf_{\hat{g}} \sup_{\rho\in\{-1,1\}} \mathbb{E}_{\mathbb{P}_\rho^{\otimes n}} \left|\mathcal{R}_{\mathbb{P}_\rho}(\hat{g}) - \mathcal{R}_{\mathbb{P}_\rho}(g_*^\rho)\right|$, thus

$$\begin{aligned}
(*) &\geq \inf_{\hat{g}} \frac{\sum_{\rho\in\{-1,1\}} \mathbb{E}_{\mathbb{P}_{+1}^{\otimes n}} \left|\mathcal{R}_{\mathbb{P}_\rho}(\hat{g}) - \mathcal{R}_{\mathbb{P}_\rho}(g_*^\rho)\right|}{2} \\
&= \inf_{\hat{g}} \frac{\mathbb{E}_{\mathbb{P}_0^{\otimes n}}\left[\frac{d\mathbb{P}_{+1}^{\otimes n}}{d\mathbb{P}_0^{\otimes n}}\mathcal{E}_{+1}(\hat{g})\right] + \mathbb{E}_{\mathbb{P}_0^{\otimes n}}\left[\frac{d\mathbb{P}_{-1}^{\otimes n}}{d\mathbb{P}_0^{\otimes n}}\mathcal{E}_{-1}(\hat{g})\right]}{2} \\
&\geq \inf_{\hat{g}} \frac{\mathbb{E}_{\mathbb{P}_0^{\otimes n}}\left[\min\left\{\frac{d\mathbb{P}_{+1}^{\otimes n}}{d\mathbb{P}_0^{\otimes n}}, \frac{d\mathbb{P}_{-1}^{\otimes n}}{d\mathbb{P}_0^{\otimes n}}\right\}\left(\mathcal{E}_{+1}(\hat{g}) + \mathcal{E}_{-1}(\hat{g})\right)\right]}{2} \\
&\geq \frac{1}{16L}\left(1 - \mathrm{TV}(\mathbb{P}_{+1}^{\otimes n}, \mathbb{P}_{-1}^{\otimes n})\right)
\end{aligned}$$

where $\mathrm{TV}(\cdot,\cdot)$ is the total variation distance between probability measures. Note, however, that for a sufficiently fast decaying[4] ϕ_n we can guarantee that $\mathrm{TV}(\mathbb{P}_{+1}^{\otimes n}, \mathbb{P}_{-1}^{\otimes n}) \leq 1/2$, which implies that

$$\inf_{\hat{g}} \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}^{\otimes n}} \left|\mathcal{R}(\hat{g}) - \mathcal{R}(g_*)\right| \geq \frac{1}{16L} \ .$$

[3] We are forced to put an absolute value in this discussion, since an "estimator" that always outputs the vector $(1, \ldots, 1)^\top$ achieves zero risk and the minimax risk without the absolute value is always non-positive.

[4] One can upper bound the total variation using Pinsker's inequality and observe that the problem reduces to the Kullback-Leibler divergence between two Bernoulli variables.

First of all, observe that the distributions constructed above satisfy Assumptions 14, 15 and the corresponding regression functions are constant. In particular, these regression functions are infinitely many times differentiable, the marginal distribution admits density *w.r.t.* Lebesgue measure supported on $[0, 1]$, and an estimator achieving Assumption 12 exists. However, since the minimax risk is of constant order, it suggests that an extra assumption is necessary for consistency of any estimator. This phenomena occurs due to the behavior of the regression function around the parameter β. The discussion above highlights the fundamental difference between the two frameworks considered in this chapter and motivates the extra Assumption 16, which might seem to be unnatural at the first sight.

Using the assumptions introduced for this model we can state the following result.

Theorem 18. *Assume that the estimator $\hat{\eta}$ satisfies Assumption 12. Therefore, under Assumptions 14–16, the plug-in rule in Eq. (4.11) satisfies*

$$\mathbb{E}_{\mathbb{P}^{\otimes n}}\left[\mathcal{R}(\hat{g})\right] - \mathcal{R}(g_*) \leq \tilde{C}(\beta^{\alpha_2} + S)(L - \beta)n^{-\gamma(\alpha_2 \wedge \alpha_1)/2} \ ,$$

for some universal constant $\tilde{C}$.

The proof of the previous theorem relies on the following Lemma:

Lemma 20 (Partial order)**.** *On the event $\{2\left\|\eta(X) - \hat{\eta}(X)\right\|_\infty < \eta^{\sigma_k}(X) - \eta^{\sigma_{k+1}}(X)\}$, we have for all $l \in [L]$ and all $m \in [L]$ such that $l \leq k < m$:*

$$l' \leq k < m' \ ,$$

where l' and m' are such that $\hat{\sigma}_{l'} = \sigma_l$ and $\hat{\sigma}_{m'} = \sigma_m$.

From the previous result we can conclude that the condition of the lemma yields $\{\hat{\sigma}_1, \ldots, \hat{\sigma}_k\} = \{\sigma_1, \ldots, \sigma_k\}$. To see this it is sufficient to apply Lemma 20 to each $l = 1, \ldots, k$ and $m = k + 1$ and use the fact that l' defined in Lemma 20 is unique and is different for all l. Similarly we can show that $\{\hat{\sigma}_{k+1}, \ldots, \hat{\sigma}_L\} = \{\sigma_{k+1}, \ldots, \sigma_L\}$. Hence, the previous lemma gives an intuitive result: if the estimation $\hat{\eta}$ is accurate enough, then it partially preserves the ordering of η. We point out that the dependence of the obtained bound on the total amount of labels L deteriorates compared to the previous case.

4.1.5 Discussion

The proposed framework is flexible and could be further analyzed. In particular, it is interesting to find other strategies, *i.e.,* other sets $\mathcal{G}_\theta$ that can be of practical interest. In general, we suggest to incorporate any quantity of interest in the set $\mathcal{G}_\theta$ and consider the plug-in approach if the oracle is available explicitly. Notice, that the $\mathcal{G}_K^{\text{sp}}$-oracle does not allow to have an optimal control over false positive discoveries, whereas the $\mathcal{G}_\beta^{\text{fp}}$-oracle does not allow to control sparsity. For instance, one might be interested in both sparsity and false positive discoveries simultaneously, to this end a natural extension is the following set $\mathcal{G}_{\beta,K}$:

$$\mathcal{G}_{\beta,K} = \mathcal{G}_K^{\text{sp}} \cap \mathcal{G}_\beta^{\text{fp}} \ .$$

Each classifier in this set has a controlled number of false positive errors as well as bounded sparsity. Moreover, Remark 11 suggests to choose the value β as ρK, where $\rho \in (0, 1)$. This choice of parameters bounds the output sparsity from below on the level ρK and from above

on the level K, moreover the false positives discoveries are upper-bounded by ρK. It is not hard to show, that an $\mathcal{G}_{\beta,K}$-oracle over such set can be obtained in a similar fashion, that is:

$$g_*^{\sigma_1}(x) = \ldots = g_*^{\sigma_{K_\star}}(x) = 1 \ ,$$
$$g_*^{\sigma_{K_\star+1}}(x) = \ldots = g_*^{\sigma_L}(x) = 0 \ ,$$

where $K_\star = K_\star(x)$ is defined as

$$K_\star(x) = \max\Big\{m \in [L] : \sum_{l=1}^{m}(1 - \eta^{\sigma_l}(x)) \leq \beta\Big\} \wedge K \ .$$

The explicit expression of the $\mathcal{G}_{\beta,K}$-oracle allows to use plug-in approach as before, and we plan to investigate this strategy in future works.

4.1.6 Conclusion

The bound in Theorem 17 is similar to the bound obtained by Audibert and Tsybakov [2007] in the binary classification settings (see Section 1.1) and is known to be minimax optimal in this case. It is important to notice that this bound is independent from the total amount of labels L and only depends on the parameter K. However, we expect that this dependency can be improved, and rates proportional to K/L can be achieved. This intuition is explained by the first step of the proof of Theorem 17, where a rather loose inequality is used.

Theorem 18 is proven under three different assumptions, which are reflecting the structure of the regression vector η. The obtained bound does not have a classical $\gamma(\alpha+1)/2$ rate but $\gamma\alpha/2$ is obtained instead. A simple explanation for this phenomena can be provided: in case the constant α_2 from Assumption 16 is equal to zero, the upper-bound becomes trivial, it is not surprising in view of the discussion provided after Assumption 16. Indeed, the distributions satisfying Assumption 16 with $\alpha_2 = 0$ are the same as the distributions satisfying *only* Assumptions 14, 15 which is not sufficient for upper-bounding the minimax risk. Bound of a similar type can be found in [Denis and Hebiri, 2015a] in the case of binary classification with reject option, where the authors are proposing to control the probability of rejection. We expect that the control over a random quantity, that is the false positive discovery in our case, or the probability of reject in case of [Denis and Hebiri, 2015a], might lead to such rates. We plan to further investigate the behavior obtained in this case and provide minimax lower bounds to show their optimality.

4.1.7 Proofs

Technical lemmas

The following lemma is used throughout this chapter. It ensures that if the estimate $\hat{\eta}$ satisfies the exponential bound in Assumption 12, hence the same bound holds if we replace $l \in [L]$ by σ_j for every $j \in [L]$:

Lemma 21. *Assume that $\hat{\eta}$ satisfies the conditions in Assumption 12, hence for all $j \in [L]$ we have*

$$\mathbb{P}^{\otimes n}\left(|\eta^{\sigma_j}(x) - \hat{\eta}^{\sigma_j}(x)| \geq \delta\right) \leq C_1 \exp(-C_2 n^\gamma \delta^2) \ \textit{for almost every } x \in \mathbb{R}^d \ \textit{w.r.t. } \mathbb{P}_X \ .$$

Proof. A standard disjunction yields:

$$\begin{aligned}\mathbb{P}^{\otimes n}\left(|\eta^{\sigma_j}(x)-\hat{\eta}^{\sigma_j}(x)|\geq\delta\right) &= \sum_{l=1}^{L}\mathbb{P}^{\otimes n}\left(|\eta^{\sigma_j(x)}(x)-\hat{\eta}^{\sigma_j(x)}(x)|\geq\delta\right)\mathbb{1}_{\{\sigma_j(x)=l\}}\\ &= \sum_{l=1}^{L}\mathbb{P}^{\otimes n}\left(|\eta^{l}(x)-\hat{\eta}^{l}(x)|\geq\delta\right)\mathbb{1}_{\{\sigma_j(x)=l\}}\\ &\leq \sum_{l=1}^{L} C_1\exp(-C_2n^{\gamma}\delta^2)\mathbb{1}_{\{\sigma_j(x)=l\}} = C_1\exp(-C_2n^{\gamma}\delta^2)\ ,\end{aligned}$$

and the inequality in Lemma 21 holds for almost every $x\in\mathbb{R}^d$ with respect to $\mathbb{P}_X$. □

Similarly, we can obtain the following bound on the infinity norm of the regression function:

Lemma 22. *Assume that $\hat{\eta}$ satisfies the conditions in Assumption 12, hence we have*

$$\mathbb{P}^{\otimes n}\left(\max_{l\in[L]}\{|\eta^{l}(x)-\hat{\eta}^{l}(x)|\}\geq\delta\right)\leq C_1L\exp(-C_2n^{\gamma}\delta^2)\ \textit{for almost every } x\in\mathbb{R}^d\ \textit{w.r.t. } \mathbb{P}_X\ .$$

Proof of Theorem 17

Proof. We start with the following decomposition of the excess risk:

$$\begin{aligned}\mathcal{E}(\hat{g}) &= \frac{1}{L}\mathbb{E}_{\mathbb{P}^{\otimes n}}\mathbb{E}_{\mathbb{P}_X}\left[\sum_{l=1}^{L}\eta^{\sigma_l}(X)\left(\mathbb{1}_{\{\hat{g}^{\sigma_l}(X)=0\}}-\mathbb{1}_{\left\{g_*^{\sigma_l}(X)=0\right\}}\right)\right]\\ &= \frac{1}{L}\mathbb{E}_{\mathbb{P}^{\otimes n}}\mathbb{E}_{\mathbb{P}_X}\left[\sum_{l=1}^{L}\eta^{\sigma_l}(X)\mathbb{1}_{\left\{\hat{g}^{\sigma_l}(X)=0,g_*^{\sigma_l}(X)=1\right\}}-\sum_{l=1}^{L}\eta^{\sigma_l}(X)\mathbb{1}_{\left\{\hat{g}^{\sigma_l}(X)=1,g_*^{\sigma_l}(X)=0\right\}}\right]\\ &= \frac{1}{L}\mathbb{E}_{\mathbb{P}^{\otimes n}}\mathbb{E}_{\mathbb{P}_X}\left[\sum_{l=1}^{K}\eta^{\sigma_l}(X)\mathbb{1}_{\{\hat{g}^{\sigma_l}(X)=0\}}-\sum_{l=K+1}^{L}\eta^{\sigma_l}(X)\mathbb{1}_{\{\hat{g}^{\sigma_l}(X)=1\}}\right]\ ,\end{aligned}$$

where in the last equality we have used the explicit expression for the oracle from Lemma 17. Now notice that since $\hat{g}$ is exactly K-sparse, hence if in the first sum there are $m\in\{1,\dots,K\}$ non-zero terms, hence there are exactly m non-zero terms in the second sum. Since all the non-zero terms in the first sum are greater than all the non-zero terms in the second sum, we can bound the excess risk by all possible pair-wise differences:

$$\mathcal{E}(\hat{g})\leq\mathbb{E}_{\mathbb{P}^{\otimes n}}\mathbb{E}_{\mathbb{P}_X}\frac{1}{L}\sum_{l=1}^{K}\sum_{j=K+1}^{L}(\eta^{\sigma_l}(X)-\eta^{\sigma_j}(X))\mathbb{1}_{\left\{\hat{g}^{\sigma_l}(X)=0,\hat{g}^{\sigma_j}(X)=1\right\}}\ .$$

On the one hand, according to the plug-in rule definition, on the event $\{\hat{g}^{\sigma_l}(X)=0,\hat{g}^{\sigma_j}(X)=1\}$ we have $\hat{\eta}^{\sigma_j}(X)\geq\hat{\eta}^{\sigma_l}(X)$. On the other hand, due to the definition of σ we have $\eta^{\sigma_l}(X)\geq\eta^{\sigma_j}(X)$ for all $j>l$. Therefore, on the event $\{\hat{g}^{\sigma_l}(X)=0,\hat{g}^{\sigma_j}(X)=1\}$ we have $\eta^{\sigma_l}(X)-\eta^{\sigma_j}(X)\leq|\hat{\Delta}_l|+|\hat{\Delta}_j|$, where $\hat{\Delta}_k=\hat{\eta}^{\sigma_k}(X)-\eta^{\sigma_k}(X)$ for any $k\in[L]$. We can write:

$$\mathcal{E}(\hat{g})\leq\frac{1}{L}\mathbb{E}_{\mathbb{P}^{\otimes n}}\mathbb{E}_{\mathbb{P}_X}\sum_{l=1}^{K}\sum_{j=K+1}^{L}(\eta^{\sigma_l}(X)-\eta^{\sigma_j}(X))\mathbb{1}_{\left\{\eta^{\sigma_l}(X)-\eta^{\sigma_j}(X)\leq|\hat{\Delta}_k|+|\hat{\Delta}_j|\right\}}\ .$$

Denote by $T_{l,j}(X)$ for all $l \in \{1,\dots,K\}$ and $j \in \{K+1,\dots,L\}$ the (l,j)-term in the above sum, that is:

$$\mathcal{E}(\hat{g}) \leq \frac{1}{L}\sum_{l=1}^{K}\sum_{j=K+1}^{L} \mathbb{E}_{\mathbb{P}^{\otimes n}}\mathbb{E}_{\mathbb{P}_X} T_{l,j}(X) \ .$$

Now we restrict our attention on an arbitrary $T_{l,j}(X)$, we can write

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}^{\otimes n}}\mathbb{E}_{\mathbb{P}_X} T_{l,j}(X) &= \mathbb{E}_{\mathbb{P}^{\otimes n}}\mathbb{E}_{\mathbb{P}_X}(\eta^{\sigma_l}(X) - \eta^{\sigma_j}(X))\mathbb{1}_{\left\{\eta^{\sigma_l}(X)-\eta^{\sigma_j}(X)\leq|\hat{\Delta}_l|+|\hat{\Delta}_j|\right\}} \\
&= \sum_{p\geq 0} \mathbb{E}_{\mathbb{P}^{\otimes n}}\mathbb{E}_{\mathbb{P}_X}(\eta^{\sigma_l}(X) - \eta^{\sigma_j}(X))\mathbb{1}_{\left\{\eta^{\sigma_l}(X)-\eta^{\sigma_j}(X)\leq|\hat{\Delta}_l|+|\hat{\Delta}_j|\right\}}\mathbb{1}_{\{X\in A_p\}} \ ,
\end{aligned}$$

where A_p are defined similar to [Audibert and Tsybakov, 2007], that is

$$\begin{aligned}
A_0 &= \{x \in \mathbb{R}^d : 0 < \eta^{\sigma_l}(x) - \eta^{\sigma_j}(x) \leq \delta\} \ , \\
A_p &= \{x \in \mathbb{R}^d : 2^{p-1}\delta < \eta^{\sigma_l}(x) - \eta^{\sigma_j}(X) \leq 2^p\delta\} \text{ for all } p > 0 \ .
\end{aligned}$$

We continue as:

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}^{\otimes n}}\mathbb{E}_{\mathbb{P}_X} T_{l,j}(X) &= \mathbb{E}_{\mathbb{P}^{\otimes n}}\mathbb{E}_{\mathbb{P}_X}(\eta^{\sigma_l}(X) - \eta^{\sigma_j}(X))\mathbb{1}_{\left\{\eta^{\sigma_l}(X)-\eta^{\sigma_j}(X)\leq|\hat{\Delta}_l|+|\hat{\Delta}_j|\right\}}\mathbb{1}_{\{X\in A_0\}} \\
&\quad + \sum_{p\geq 1} \mathbb{E}_{\mathbb{P}^{\otimes n}}\mathbb{E}_{\mathbb{P}_X}(\eta^{\sigma_l}(X) - \eta^{\sigma_j}(X))\mathbb{1}_{\left\{\eta^{\sigma_l}(X)-\eta^{\sigma_j}(X)\leq|\hat{\Delta}_l|+|\hat{\Delta}_j|\right\}}\mathbb{1}_{\{X\in A_p\}} \\
&\leq \mathbb{E}_{\mathbb{P}^{\otimes n}}\mathbb{E}_{\mathbb{P}_X}(\eta^{\sigma_l}(X) - \eta^{\sigma_j}(X))\mathbb{1}_{\left\{0<\eta^{\sigma_l}(X)-\eta^{\sigma_j}(X)\leq\delta\right\}} \\
&\quad + \sum_{p\geq 1} \mathbb{E}_{\mathbb{P}^{\otimes n}}\mathbb{E}_{\mathbb{P}_X}(\eta^{\sigma_l}(X) - \eta^{\sigma_j}(X))\mathbb{1}_{\left\{2^{p-1}\delta\leq|\hat{\Delta}_l|+|\hat{\Delta}_j|\right\}}\mathbb{1}_{\left\{0<\eta^{\sigma_l}(X)-\eta^{\sigma_j}(X)\leq 2^p\delta\right\}} \\
&\leq \delta\mathbb{E}_{\mathbb{P}_X}\mathbb{1}_{\left\{0<\eta^{\sigma_l}(X)-\eta^{\sigma_j}(X)\leq\delta\right\}} \\
&\quad + \sum_{p\geq 1} 2^p\delta\mathbb{E}_{\mathbb{P}_X}\mathbb{P}^{\otimes n}\left(2^{p-1}\delta \leq |\hat{\Delta}_l| + |\hat{\Delta}_j|\right)\mathbb{1}_{\left\{0<\eta^{\sigma_l}(X)-\eta^{\sigma_j}(X)\leq 2^p\delta\right\}} \\
&\leq \delta\mathbb{E}_{\mathbb{P}_X}\mathbb{1}_{\{0<\eta^{\sigma_K}(X)-\eta^{\sigma_{K+1}}(X)\leq\delta\}} \\
&\quad + \sum_{p\geq 1} 2^p\delta\mathbb{E}_{\mathbb{P}_X}\mathbb{P}^{\otimes n}\left(2^{p-1}\delta \leq |\hat{\Delta}_l| + |\hat{\Delta}_j|\right)\mathbb{1}_{\{0<\eta^{\sigma_K}(X)-\eta^{\sigma_{K+1}}(X)\leq 2^p\delta\}} \\
&\leq C\delta^{1+\alpha} + 2\sum_{p\geq 1} 2^p\delta C_1\exp(-C_2a_n2^{2p-2}\delta^2)\mathbb{E}_{\mathbb{P}_X}\mathbb{1}_{\{0<\eta^{\sigma_K}(X)-\eta^{\sigma_{K+1}}(X)\leq 2^p\delta\}} \\
&\leq C\delta^{1+\alpha} + 2CC_1\delta^{\alpha+1}\sum_{p\geq 1} 2^{p(\alpha+1)}\exp(-C_2a_n2^{2p-2}\delta^2) \ ,
\end{aligned}$$

setting $\delta = (a_n)^{\frac{1}{2}}$ we obtain:

$$\mathbb{E}_{\mathbb{P}^{\otimes n}}\mathbb{E}_{\mathbb{P}_X} T_{l,j}(X) \leq \tilde{C}(a_n)^{\frac{1+\alpha}{2}} .$$

We conclude by substituting the obtained bound into the excess risk bound. □

Proof of Lemma 18

The next lemma states that the risk of any classifier $g \in \mathcal{G}_\beta^{\text{fp}}$ can be improved if the classifier is altered pointwise according to the order of the regression vector $\sigma_1,\dots,\sigma_L$. This result allows to prove Lemma 18.

Lemma 23. *Assume that $g \in \mathcal{G}_\beta$. Let $\mathrm{pos}(g(x)) = \{l \in [L] : g^l(x) = 1\}$ and denote by $m(x) = |\mathrm{pos}(g(x))|$ the cardinality of $\mathrm{pos}(g(x))$. For all $x \in \mathbb{R}^d$ define g_m as*

$$g_m^{\sigma_1}(x) = \ldots = g_m^{\sigma_m}(x) = 1 \ ,$$
$$g_m^{\sigma_{m+1}}(x) = \ldots = g_m^{\sigma_L}(x) = 0 \ ,$$

Hence, $g_m \in \mathcal{G}_\beta^{fp}$ and $\mathcal{R}(g_m) \leq \mathcal{R}(g)$.

Proof. First, we show that $g_m \in \mathcal{G}_\beta^{\mathrm{fp}}$, since $g \in \mathcal{G}_\beta$ it holds that

$$\sum_{l \in \mathrm{pos}(g(X))} (1 - \eta^l(X)) \leq \beta, \ \mathbb{P}_X\text{-a.s.} \ ,$$

due to the definition of $\sigma = \sigma(x)$ we have

$$\sum_{l \in \mathrm{pos}(g_m(x))} (1 - \eta^l(x)) \leq \sum_{l \in \mathrm{pos}(g(x))} (1 - \eta^l(x)), \text{ for all } x \in \mathbb{R}^D \ ,$$

indeed, $\sum_{l \in \mathrm{pos}(g_m(x))}^L (1 - \eta^l(x))$ consists of a sum of m smallest values of $(1 - \eta^l(x))_{l=1}^L$, which concludes the first part of the statement. The second part is proven similarly: for all $x \in \mathbb{R}^D$ it obviously holds thanks to the definition of σ that

$$\mathbb{E}\Big[\sum_{l=1}^L \mathbb{1}_{\left\{g_m^l(x)=0, Y^l=1\right\}} | X = x\Big] = \sum_{l \in [L] \setminus \mathrm{pos}(g_m(x))} \eta^l(x) \leq \sum_{l \in [L] \setminus \mathrm{pos}(g(x))} \eta^l(x)$$
$$= \mathbb{E}\Big[\sum_{l=1}^L \mathbb{1}_{\left\{g^l(x)=0, Y^l=1\right\}} | X = x\Big] \ ,$$

which concludes the proof. □

Lemma 18. Let $g \in \mathcal{G}_\beta^{\mathrm{fp}}$ be an oracle. Due to Lemma 23, we can get $g_m \in \mathcal{G}_\beta^{\mathrm{fp}}$ such that $\mathcal{R}(g_m) \leq \mathcal{R}(g)$, so g_m is also an oracle. On the event $\{x \in \mathbb{R}^d : \sum_{l=1}^L \mathbb{1}_{\left\{g_m^l(x)=1\right\}}(1 - \eta^l(x)) \leq \beta\}$ (whose measure is one), it holds that $\mathrm{pos}(g_m(x)) \subset \mathrm{pos}(g_*(x))$ by the construction of $g_*(x)$ and in particular $K(x)$ therefore on this set

$$\mathbb{E}\Big[\sum_{l=1}^L \mathbb{1}_{\left\{g_m^l(x)=0, Y^l=1\right\}} | X = x\Big] \leq \mathbb{E}\Big[\sum_{l=1}^L \mathbb{1}_{\left\{g_*^l(x)=0, Y^l=1\right\}} | X = x\Big] \ .$$

Since the previous inequality holds almost surely $\mathbb{P}_X$, we conclude.

□

Proof of Theorem 18

Lemma 19. Let us fix $g \in \hat{\mathcal{G}}_\beta^{\mathrm{fp}}$. Hence, by the definition of $\hat{\mathcal{G}}_\beta^{\mathrm{fp}}$ we have

$$\sum_{l=1}^L \mathbb{1}_{\left\{g^l(X)=1\right\}}(1 - \hat{\eta}^l(X)) \leq \beta, \ \mathbb{P}_X\text{-a.s.} \ .$$

We introduce the following notation

$$B(z) = \sum_{l=1}^L \mathbb{1}_{\left\{g^l(X)=1\right\}}(1 - z^l) \ .$$

Therefore $B(\hat{\eta}(X)) \leq \beta$, $\mathbb{P}_X$-*a.s.*. The following sequence of inequalities holds

$$\begin{aligned}|B(\hat{\eta}(X)) - B(\eta(X))| &= \left|\sum_{l=1}^{L} \mathbb{1}_{\{g^l(X)=1\}}(\hat{\eta}^l(X) - \eta^l(X))\right| \\ &\leq \sum_{l=1}^{L} \left|(\hat{\eta}^l(X) - \eta^l(X))\right|, \ \mathbb{P}_X\text{-}a.s. \ ,\end{aligned}$$

and this concludes the proof. □

Lemma 20. Let $l \in [L]$ and $m \in [L]$ be such that $l \leq k < m$, hence by of σ we have

$$\eta^{\sigma_l}(X) \geq \eta^{\sigma_k}(X) \geq \eta^{\sigma_{k+1}}(X) \geq \eta^{\sigma_m}(X) \ ,$$

therefore

$$\eta^{\sigma_l}(X) - \eta^{\sigma_m}(X) \geq \eta^{\sigma_k}(X) - \eta^{\sigma_{k+1}}(X) \ .$$

We can write

$$\eta^{\sigma_l}(X) - \eta^{\sigma_m}(X) - \hat{\eta}^{\hat{\sigma}_{l'}}(X) + \hat{\eta}^{\hat{\sigma}_{l'}}(X) - \hat{\eta}^{\hat{\sigma}_{m'}}(X) + \hat{\eta}^{\hat{\sigma}_{m'}}(x) = \eta^{\sigma_l}(X) - \eta^{\sigma_m}(x) \geq \eta^{\sigma_k}(X) - \eta^{\sigma_{k+1}}(X) \ ,$$

which implies

$$\hat{\eta}^{\hat{\sigma}_{l'}}(X) - \hat{\eta}^{\hat{\sigma}_{m'}}(X) + 2\left\|\hat{\eta}(X) - \eta(X)\right\|_\infty \geq \eta^{\sigma_k}(X) - \eta^{\sigma_{k+1}}(X) \ ,$$

and therefore on the event $\{2\left\|\eta(X) - \hat{\eta}(X)\right\|_\infty < \eta^{\sigma_k}(X) - \eta^{\sigma_{k+1}(X)}\}$ we have

$$\hat{\eta}^{\hat{\sigma}_{l'}}(X) \geq \hat{\eta}^{\hat{\sigma}_{m'}}(X) \ ,$$

meaning that $l' < m'$. To conclude that $m' > k$ it is sufficient to notice that the inequality $l' < m'$ holds for at least k different values of l'. Similarly we conclude that $l' \leq k$. □

Theorem 18. Here we denote by $\mathbb{E}$ the expectation $\mathbb{E}_{\mathbb{P}^{\otimes n}}\mathbb{E}_X$ for the sake of simplicity.

$$\begin{aligned}\mathbb{E}\mathcal{R}(\hat{g}) - \mathcal{R}(g_*) &= \frac{1}{L}\mathbb{E}\Big[\sum_{l=1}^{L} \eta^{\sigma_l}(X)\mathbb{1}_{\{\hat{g}^{\sigma_l}(X)=0, g_*^{\sigma_l}(X)=1\}} - \sum_{l=1}^{L} \eta^{\sigma_l}(X)\mathbb{1}_{\{\hat{g}^{\sigma_l}(X)=1, g_*^{\sigma_l}(X)=0\}}\Big] \\ &\leq \frac{1}{L}\mathbb{E}\Big[\sum_{l=1}^{L} \eta^{\sigma_l}(X)\mathbb{1}_{\{\hat{g}^{\sigma_l}(X)=0, g_*^{\sigma_l}(X)=1\}} \sum_{k=\beta}^{L} \mathbb{1}_{\{K(X)=k\}}\Big] \\ &= \underbrace{\frac{1}{L}\mathbb{E}\Big[\sum_{k=\beta}^{L}\Big[\sum_{l=1}^{k} \eta^{\sigma_l}(X)\mathbb{1}_{\{\hat{g}^{\sigma_l}(X)=0\}}\mathbb{1}_{\{K(X)=k\}}\Big]\mathbb{1}_{\left\{2\|\eta(X)-\hat{\eta}(X)\|_\infty \geq \eta^{\sigma_k}(X)-\eta^{\sigma_{k+1}(X)}\right\}}\Big]}_{U_1} \\ &+ \underbrace{\frac{1}{L}\mathbb{E}\Big[\sum_{k=\beta}^{L}\Big[\sum_{l=1}^{k} \eta^{\sigma_l}(X)\mathbb{1}_{\{\hat{g}^{\sigma_l}(X)=0\}}\mathbb{1}_{\{K(X)=k\}}\Big]\mathbb{1}_{\left\{2\|\eta(X)-\hat{\eta}(X)\|_\infty < \eta^{\sigma_k}(X)-\eta^{\sigma_{k+1}(X)}\right\}}\Big]}_{U_2} \\ &= U_1 + U_2 \ .\end{aligned}$$

For U_1, due to Assumption 15 we can write

$$\begin{aligned}
U_1 \leq & \frac{1}{L}\mathbb{E}\Big[\sum_{k=\beta}^{L}\Big[\sum_{l=1}^{k}\eta^{\sigma_l}(X)\mathbb{1}_{\{K(X)=k\}}\mathbb{1}_{\left\{2\|\eta(X)-\hat{\eta}(X)\|_\infty \geq \eta^{\sigma_k}(X)-\eta^{\sigma_{k+1}(X)}\right\}}\Big]\Big] \\
= & \frac{1}{L}\mathbb{E}\Big[\sum_{k=\beta}^{L}\mathbb{1}_{\{K(X)=k\}}\mathbb{1}_{\left\{2\|\eta(X)-\hat{\eta}(X)\|_\infty \geq \eta^{\sigma_k}(X)-\eta^{\sigma_{k+1}(X)}\right\}}\sum_{l=1}^{k}\eta^{\sigma_l}(X)\Big] \\
\leq & \frac{S}{L}\sum_{k=\beta}^{L}\underbrace{\mathbb{E}\Big[\mathbb{1}_{\{K(X)=k\}}\mathbb{1}_{\left\{2\|\eta(X)-\hat{\eta}(X)\|_\infty \geq \eta^{\sigma_k}(X)-\eta^{\sigma_{k+1}(X)}\right\}}\Big]}_{U_1^k} = \frac{S}{L}\sum_{k=\beta}^{L}U_1^k \ .
\end{aligned}$$

define the following sets, similar to the analysis of Audibert and Tsybakov [2007] for binary classification, for all $L \leq k \leq \beta$

$$\begin{aligned}
A_0^k &= \{X \in \mathbb{R}^d : \eta^{\sigma_k}(X) - \eta^{\sigma_{k+1}(X)} \leq \delta\} \ , \\
A_j^k &= \{X \in \mathbb{R}^d : 2^{j-1}\delta < \eta^{\sigma_k}(X) - \eta^{\sigma_{k+1}}(X) \leq 2^j\delta\} \ .
\end{aligned}$$

Therefore, using Assumption 14 for each k such that $L \leq k \leq \beta$ we have

$$\begin{aligned}
U_1^k = & \mathbb{E}\sum_{j\geq 0}\mathbb{1}_{\{K(X)=k\}}\mathbb{1}_{\left\{2\|\eta(X)-\hat{\eta}(X)\|_\infty \geq \eta^{\sigma_k}(X)-\eta^{\sigma_{k+1}(X)}\right\}}\mathbb{1}_{\left\{X\in A_j^k\right\}} \\
= & \mathbb{E}\mathbb{1}_{\{K(X)=k\}}\mathbb{1}_{\left\{2\|\eta(X)-\hat{\eta}(X)\|_\infty \geq \eta^{\sigma_k}(X)-\eta^{\sigma_{k+1}(X)}\right\}}\mathbb{1}_{\left\{X\in A_0^k\right\}} \\
& + \mathbb{E}\sum_{j\geq 1}\mathbb{1}_{\{K(X)=k\}}\mathbb{1}_{\left\{2\|\eta(X)-\hat{\eta}(X)\|_\infty \geq \eta^{\sigma_k}(X)-\eta^{\sigma_{k+1}(X)}\right\}}\mathbb{1}_{\left\{X\in A_j^k\right\}} \\
\leq & \mathbb{E}\mathbb{1}_{\{K(X)=k\}}\mathbb{1}_{\left\{X\in A_0^k\right\}} \\
& + \mathbb{E}\sum_{j\geq 1}\mathbb{1}_{\{K(X)=k\}}\mathbb{1}_{\left\{2\|\eta(X)-\hat{\eta}(X)\|_\infty \geq 2^{j-1}\delta\right\}}\mathbb{1}_{\left\{X\in A_j^k\right\}} \\
\leq & \mathbb{P}\left(0 < \eta^{\sigma_k}(X) - \eta^{\sigma_{k+1}(X)} \leq \delta, K(X) = k\right) \\
& + \mathbb{E}\sum_{j\geq 1}\mathbb{1}_{\{\eta^{\sigma_k}(X)-\eta^{\sigma_{k+1}}(X)\leq 2^j\delta, K(X)=k\}}\mathbb{P}^{\otimes n}\left(2\left\|\eta(X)-\hat{\eta}(X)\right\|_\infty \geq 2^{j-1}\delta\right) \\
\leq & \mathbb{P}\left(0 < \eta^{\sigma_k}(X) - \eta^{\sigma_{k+1}(X)} \leq \delta, K(X) = k\right) \\
& + \sum_{j\geq 1}\mathbb{P}\left(\eta^{\sigma_k}(X) - \eta^{\sigma_{k+1}}(X) \leq 2^j\delta, K(X) = k\right) C_2 L\exp(-C_3 n^\gamma 2^{2j-2}\delta^2) \\
\leq & C_1\delta^{\alpha_1} + \sum_{j\geq 1} C_1 C_2 \delta^\alpha 2^{\alpha_1 j} L\exp(-C_3 n^\gamma 2^{2j-2}\delta^2) \ ,
\end{aligned}$$

let $\delta = n^{-\gamma/2}$, hence

$$U_1^k \leq C_1 n^{-\gamma\alpha_1/2} + C_1 C_2 n^{-\gamma\alpha_1/2}\sum_{j\geq 1}2^{\alpha j}L\exp(-C_3 2^{2j-2}) \leq \tilde{C}_1 L n^{-\gamma\alpha_1/2} \ .$$

Therefore,

$$U_1 \leq \tilde{C}\frac{S}{L}(L-\beta)La_n^{-\alpha/2} = \tilde{C}S(L-\beta)n^{-\gamma\alpha_1/2} \ .$$

For U_2 we can write

$$\begin{aligned}
U_2 = & \frac{1}{L}\mathbb{E}\Big[\sum_{k=\beta}^{L}\Big[\sum_{l=1}^{k}\eta^{\sigma_l}(X)\mathbb{1}_{\{\hat{g}^{\sigma_l}(X)=0\}}\Big]\mathbb{1}_{\{K(X)=k\}}\mathbb{1}_{\left\{2\|\eta(X)-\hat{\eta}(X)\|_\infty < \eta^{\sigma_k}(X)-\eta^{\sigma_{k+1}(X)}\right\}}\Big] \\
= & \frac{1}{L}\sum_{k=\beta}^{L}U_2^k \ ,
\end{aligned}$$

where U_2^k is given as

$$U_2^k = \mathbb{E}\Big[\sum_{l=1}^{k} \eta^{\sigma_l}(X)\mathbb{1}_{\{\hat{g}^{\sigma_l}(X)=0\}}\mathbb{1}_{\{K(X)=k\}}\mathbb{1}_{\left\{2\|\eta(X)-\hat{\eta}(X)\|_\infty<\eta^{\sigma_k}(X)-\eta^{\sigma_{k+1}(X)}\right\}}\Big] \ .$$

For each U_2^k we can write

$$U_2^k \leq \mathbb{E}\Big[\sum_{l=1}^{k} \mathbb{1}_{\{\hat{g}^{\sigma_l}(X)=0\}}\mathbb{1}_{\{K(X)=k\}}\mathbb{1}_{\left\{2\|\eta(X)-\hat{\eta}(X)\|_\infty<\eta^{\sigma_k}(X)-\eta^{\sigma_{k+1}(X)}\right\}}\Big] \ .$$

If $\hat{g}^{\sigma_l}(X) = 0$ for some $l \leq k$, hence by the definition of the plug-in rule we have

$$\sum_{j=1}^{l'}(1-\hat{\eta}^{\hat{\sigma}_j}(X)) > \beta \ ,$$

where l' is such that $\hat{\sigma}_{l'} = \sigma_l$. Additionally, on the event $\{2\left\|\eta(X)-\hat{\eta}(X)\right\|_\infty < \eta^{\sigma_k}(X) - \eta^{\sigma_{k+1}(X)}\}$ according to Lemma 20 we have $\{\hat{\sigma}_1,\ldots,\hat{\sigma}_{l'}\} \subset \{\sigma_1,\ldots,\sigma_k\}$ and hence, on the event $\{K(X)=k\}$ we can write

$$\underbrace{\sum_{j=1}^{l'}(1-\eta^{\hat{\sigma}_j}(X))}_{\text{Sum of } l' \text{ elements}} \leq \underbrace{\sum_{j=k-l'+1}^{k}(1-\eta^{\sigma_j}(X))}_{\text{Sum of the largest } l' \text{ elements}} \leq \beta \ .$$

Therefore on the intersection of the three events

$$\{\hat{g}^{\sigma_l}(X)=0\} \cap \left\{2\left\|\eta(X)-\hat{\eta}(X)\right\|_\infty < \eta^{\sigma_k}(X) - \eta^{\sigma_{k+1}(X)}\right\} \cap \{K(X)=k\}$$

we have

$$l'\left\|\eta(X)-\hat{\eta}(X)\right\|_\infty \geq \left|\sum_{j=1}^{l'}\hat{\eta}^{\hat{\sigma}_j}(X) - \eta^{\hat{\sigma}_j}(X)\right| \geq |\sum_{j=k-l'+1}^{k}(1-\eta^{\sigma_j}(X)) - \beta| \ .$$

Hence,

$$\begin{aligned}
U_2^k &\leq \mathbb{E}\Big[\sum_{l=1}^{k}\mathbb{1}_{\left\{l\|\eta(X)-\hat{\eta}(X)\|_\infty \geq |\sum_{j=k-l+1}^{k}(1-\eta^{\sigma_j}(X))-\beta|\right\}}\mathbb{1}_{\{K(X)=k\}}\mathbb{1}_{\left\{2\|\eta(X)-\hat{\eta}(X)\|_\infty<\eta^{\sigma_k}(X)-\eta^{\sigma_{k+1}}(X)\right\}}\Big] \\
&\leq \mathbb{E}\Big[\sum_{l=1}^{k}\mathbb{1}_{\left\{l\|\eta(X)-\hat{\eta}(X)\|_\infty \geq |\sum_{j=k-l+1}^{k}(1-\eta^{\sigma_j}(X))-\beta|\right\}}\mathbb{1}_{\{K(X)=k\}} \ .
\end{aligned}$$

where the first inequality is obtained by reordering thanks to Lemma 20. With Assumption 16, we can show the following bound, using the same technique as for U_1^k:

$$U_2^k \leq \tilde{C}\beta^{\alpha_2}n^{-\gamma\alpha_2/2}\sum_{l=1}^{L}h(|k-l|) \ ,$$

therefore

$$U_2 \leq \tilde{C}\beta^{\alpha_2}n^{-\gamma\alpha_2/2}\frac{1}{L}\sum_{k=\beta}^{L}\sum_{l=1}^{L}h(|k-l|) \leq \tilde{C}\beta^{\alpha_2}(L-\beta)n^{-\gamma\alpha_2/2} \ .$$

Therefore, we have

$$\begin{aligned}
\mathbb{E}\mathcal{R}(\hat{g}) &\leq \mathcal{R}(g_*) + \tilde{C}(L-\beta)(\beta^{\alpha_2}n^{-\gamma\alpha_2/2} + Sn^{-\gamma\alpha_1/2}) \\
&\leq \mathcal{R}(g_*) + 2\tilde{C}(\beta^{\alpha_2}+S)(L-\beta)n^{-\gamma(\alpha_2\wedge\alpha_1)/2} \ ,
\end{aligned}$$

and the conclusion holds. □

Chapter 5

Concluding remarks

This manuscript presents several results on the plug-in approach to constrained classification. In future it is interesting to further investigate the general settings of constrained classification and provide statistical analysis for this case. In particular, we would like to address the question of minimax rates for this general case and understand whether fast rates are always achievable. Another research direction is the study of semi-supervised methods in the context of the constrained classification, similar to the one perfomed in Section 2.1 and Chapter 3.

Chapter 6

Appendix

General constrained framework

Proposition 1. *Let $L_{\mathbb{P}} : \{0,1\} \times \{0,1\} \to \mathbb{R}$ be a measurable loss function and $C_{\mathbb{P}} : \{0,1\} \times \{0,1\} \to \mathbb{R}$ a measurable constraint function, and consider the following problem*

$$\min \left\{ \mathbb{E}_{(Z,Y)\sim\mathbb{P}}[L_{\mathbb{P}}(g(Z),Y)] \ : \ \mathbb{E}_{(Z,Y)\sim\mathbb{P}}[C_{\mathbb{P}}(g(Z),Y)] = 0 \right\} \ ,$$

then this constrained binary classification formulation admits the representation from Equations (1.10), (1.11), with

$$\begin{aligned}
A_{\mathbb{P}}(Z) &= L_{\mathbb{P}}(0,1)\mathbb{P}(Y=1|Z) + L_{\mathbb{P}}(0,0)\mathbb{P}(Y=0|Z) \ , \\
B_{\mathbb{P}}(Z) &= (L_{\mathbb{P}}(1,1) - L_{\mathbb{P}}(0,1))\mathbb{P}(Y=1|Z) + (L_{\mathbb{P}}(1,0) - L_{\mathbb{P}}(0,0))\mathbb{P}(Y=0|Z) \ , \\
\bar{A}_{\mathbb{P}}(Z) &= C_{\mathbb{P}}(0,1)\mathbb{P}(Y=1|Z) + C_{\mathbb{P}}(0,0)\mathbb{P}(Y=0|Z) \ , \\
\bar{B}_{\mathbb{P}}(Z) &= (C_{\mathbb{P}}(0,1) - C_{\mathbb{P}}(1,1))\mathbb{P}(Y=1|Z) + (C_{\mathbb{P}}(0,0) - C_{\mathbb{P}}(1,0))\mathbb{P}(Y=0|Z) \ .
\end{aligned}$$

Proof of Proposition 1. For simplicity we omit the index $\mathbb{P}$ from both $F_{\mathbb{P}}, \bar{F}_{\mathbb{P}}$. Using the properties of conditional expectations and the notation $\eta(Z) := \mathbb{P}(Y=1|Z)$ we can write

$$\begin{aligned}
\mathbb{E}_{(Z,Y)\sim\mathbb{P}}[L_{\mathbb{P}}(g(Z),Y)] &= \mathbb{E}_{(Z,Y)\sim\mathbb{P}}[L_{\mathbb{P}}(g(Z),1)\mathbb{1}_{\{Y=1\}} + L_{\mathbb{P}}(g(Z),0)\mathbb{1}_{\{Y=0\}}] \\
&= \mathbb{E}_{Z\sim\mathbb{P}_Z}[L_{\mathbb{P}}(g(Z),1)\eta(Z) + L_{\mathbb{P}}(g(Z),0)(1-\eta(Z))] \\
&= \mathbb{E}_{Z\sim\mathbb{P}_Z}[L_{\mathbb{P}}(1,1)\mathbb{1}_{\{g(Z)=1\}}\eta(Z) + L_{\mathbb{P}}(1,0)\mathbb{1}_{\{g(Z)=1\}}(1-\eta(Z))] \\
&\quad + \mathbb{E}_{Z\sim\mathbb{P}_Z}[L_{\mathbb{P}}(0,1)\eta(Z)(1-\mathbb{1}_{\{g(Z)=1\}}) + L_{\mathbb{P}}(0,0)(1-\mathbb{1}_{\{g(Z)=1\}})(1-\eta(Z))] \\
&= L_{\mathbb{P}}(0,1)\mathbb{P}(Y=1) + L_{\mathbb{P}}(0,0)\mathbb{P}(Y=0) \\
&\quad + \mathbb{E}_{Z\sim\mathbb{P}_Z}\left[((L_{\mathbb{P}}(1,1) - L_{\mathbb{P}}(0,1))\eta(Z) + (L_{\mathbb{P}}(1,0) - L_{\mathbb{P}}(0,0))(1-\eta(Z)))\,\mathbb{1}_{\{g(Z)=1\}}\right] \ .
\end{aligned}$$

Same derivations hold for $C_{\mathbb{P}}$. □

Lemma 1 (Bayes rule)**.** *Under Assumption 2 a Bayes optimal classifier g^* can be obtained for all $z \in \mathcal{Z}$ as*

$$g_{\lambda^*}(z) = \mathbb{1}_{\left\{\lambda^*\bar{B}_{\mathbb{P}}(z) - B_{\mathbb{P}}(z) > 0\right\}} \ ,$$

where λ^* is determined as a root of

$$\lambda \mapsto \mathbb{E}_{Z\sim\mathbb{P}_Z}\left[\bar{B}_{\mathbb{P}}(Z)\mathbb{1}_{\left\{\lambda\bar{B}_{\mathbb{P}}(Z) - B_{\mathbb{P}}(Z) > 0\right\}}\right] - \mathbb{E}_{Z\sim\mathbb{P}_Z}[\bar{A}_{\mathbb{P}}(Z)] \ .$$

Moreover, for every classifier g we can write

$$\mathcal{R}(g) - \mathcal{R}(g^*) = \mathbb{E}_{Z\sim\mathbb{P}_Z}\left|B_{\mathbb{P}}(Z) - \lambda^*\bar{B}_{\mathbb{P}}(Z)\right|\mathbb{1}_{\{g(Z)\neq g^*(Z)\}} - \lambda^*(\mathbb{E}_{Z\sim\mathbb{P}_Z}[\bar{A}_{\mathbb{P}}(Z) - \bar{B}_{\mathbb{P}}(Z)\mathbb{1}_{\{g(Z)=1\}}]) \ .$$

Proof of Lemma 1. To prove this result we demonstrate that the minmax theorem holds in this case by direct computation. Let us study the minimization problem

$$(*) := \min \left\{ \mathcal{R}(g) \, : \, \mathbb{E}_{Z\sim\mathbb{P}_Z}[\bar{A}_{\mathbb{P}}(Z) - \bar{B}_{\mathbb{P}}(Z)\mathbb{1}_{\{g(Z)=1\}}] = 0 \right\} \ .$$

Using the weak duality argument we can write

$$\begin{aligned}(*) &\geq \max_{\lambda\in\mathbb{R}} \min_{g} \left\{ \mathbb{E}_{Z\sim\mathbb{P}_Z}\left[A_{\mathbb{P}}(Z) + B_{\mathbb{P}}(Z)\mathbb{1}_{\{g(Z)=1\}}\right] + \lambda \mathbb{E}_{Z\sim\mathbb{P}_Z}[\bar{A}_{\mathbb{P}}(Z) - \bar{B}_{\mathbb{P}}(Z)\mathbb{1}_{\{g(Z)=1\}}] \right\} \\ &= \max_{\lambda\in\mathbb{R}} \min_{g} \left\{ \mathbb{E}_{Z\sim\mathbb{P}_Z}\left[(A_{\mathbb{P}}(Z) + \lambda\bar{A}_{\mathbb{P}}(Z)) + (B_{\mathbb{P}}(Z) - \lambda\bar{B}_{\mathbb{P}}(Z))\mathbb{1}_{\{g(Z)=1\}}\right] \right\} \ .\end{aligned}$$

For every $\lambda \in \mathbb{R}$ we denote by g_λ a minimizer of the problem

$$\min_{g} \left\{ \mathbb{E}_{Z\sim\mathbb{P}_Z}\left[(A_{\mathbb{P}}(Z) + \lambda\bar{A}_{\mathbb{P}}(Z)) + (B_{\mathbb{P}}(Z) - \lambda\bar{B}_{\mathbb{P}}(Z))\mathbb{1}_{\{g(Z)=1\}}\right] \right\} \ .$$

Clearly, for every fixed $\lambda \in \mathbb{R}$ a minimizer g_λ can be given for all $z \in \mathcal{Z}$ as

$$g_\lambda(z) = \mathbb{1}_{\left\{B_{\mathbb{P}}(z) - \lambda\bar{B}_{\mathbb{P}}(z) < 0\right\}} \ .$$

Note that this choice of g_λ minimizes the expression under the expectation point-wise and the measurability of both $B_{\mathbb{P}}(z)$, $\bar{B}_{\mathbb{P}}(z)$ ensures that this choice is measurable, that is, g_λ is a classifier. Therefore, we can write

$$\begin{aligned}(*) &\geq \max_{\lambda\in\mathbb{R}} \left\{ \mathbb{E}_{Z\sim\mathbb{P}_Z}[A_{\mathbb{P}}(Z) + \lambda\bar{A}_{\mathbb{P}}(Z)] - \mathbb{E}_{Z\sim\mathbb{P}_Z}(\lambda\bar{B}_{\mathbb{P}}(Z) - B_{\mathbb{P}}(Z))_+ \right\} \\ &= -\min_{\lambda\in\mathbb{R}} \left\{ \mathbb{E}_{Z\sim\mathbb{P}_Z}(\lambda\bar{B}_{\mathbb{P}}(Z) - B_{\mathbb{P}}(Z))_+ - \mathbb{E}_{Z\sim\mathbb{P}_Z}[A_{\mathbb{P}}(Z) + \lambda\bar{A}_{\mathbb{P}}(Z)] \right\} \ .\end{aligned}$$

It is important to observe that the mapping

$$\lambda \mapsto \mathbb{E}_{Z\sim\mathbb{P}_Z}(\lambda\bar{B}_{\mathbb{P}}(Z) - B_{\mathbb{P}}(Z))_+ - \mathbb{E}_{Z\sim\mathbb{P}_Z}[A_{\mathbb{P}}(Z) + \lambda\bar{A}_{\mathbb{P}}(Z)] \ ,$$

is convex. Indeed, the function $\lambda\bar{B}_{\mathbb{P}}(z) - B_{\mathbb{P}}(z)$ is affine in λ for every $z \in \mathcal{Z}$ and hence is convex in λ for every $z \in \mathcal{Z}$. Moreover, $(\lambda\bar{B}_{\mathbb{P}}(z) - B_{\mathbb{P}}(z))_+$ is convex as a maximum between two convex functions for every $z \in \mathcal{Z}$, besides, $\mathbb{E}_{Z\sim\mathbb{P}_Z}(\lambda\bar{B}_{\mathbb{P}}(Z) - B_{\mathbb{P}}(Z))_+$ is convex as the convexity is closed under weighted addition. Finally, $-\mathbb{E}_{Z\sim\mathbb{P}_Z}[A_{\mathbb{P}}(Z) + \lambda\bar{A}_{\mathbb{P}}(Z)]$ is affine in λ, thus we established the convexity of the objective. Therefore, we can write necessary and sufficient conditions for the optimality in λ for non-differentiable functions

$$\begin{aligned}&0 \in \partial_\lambda \left(\mathbb{E}_{Z\sim\mathbb{P}_Z}(\lambda\bar{B}_{\mathbb{P}}(Z) - B_{\mathbb{P}}(Z))_+ - \mathbb{E}_{Z\sim\mathbb{P}_Z}[A_{\mathbb{P}}(Z) + \lambda\bar{A}_{\mathbb{P}}(Z)] \right) \ . \\ &0 \in \partial_\lambda \left(\mathbb{E}_{Z\sim\mathbb{P}_Z}(\lambda\bar{B}_{\mathbb{P}}(Z) - B_{\mathbb{P}}(Z))_+ \right) - \mathbb{E}_{Z\sim\mathbb{P}_Z}[\bar{A}_{\mathbb{P}}(Z)] \ .\end{aligned}$$

Under Assumption 2 and using the dominant convergence theorem we have at the optimum λ^* we have

$$0 = \mathbb{E}_{Z\sim\mathbb{P}_Z}\left[\bar{B}_{\mathbb{P}}(Z)\mathbb{1}_{\left\{\lambda^*\bar{B}_{\mathbb{P}}(Z) - B_{\mathbb{P}}(Z) > 0\right\}} \right] - \mathbb{E}_{Z\sim\mathbb{P}_Z}[\bar{A}_{\mathbb{P}}(Z)] \ .$$

Let us denote by g_{λ^*} a classifier given for all $z \in \mathcal{Z}$ by

$$g_{\lambda^*}(z) = \mathbb{1}_{\left\{\lambda^*\bar{B}_{\mathbb{P}}(z) - B_{\mathbb{P}}(z) > 0\right\}} \ .$$

Note that by the definition of λ^* we have

$$\mathbb{E}_{Z\sim\mathbb{P}_Z}\left[\bar{B}_{\mathbb{P}}(Z)\mathbb{1}_{\{g_{\lambda^*}(Z)=1\}}\right] - \mathbb{E}_{Z\sim\mathbb{P}_Z}[\bar{A}_{\mathbb{P}}(Z)] = 0 \ , \tag{6.1}$$

thus this classifier g_{λ^*} satisfies the desired constraints, which means that

$$\begin{aligned}
\mathcal{R}(g_{\lambda^*}) \geq (*) &\geq \max_{\lambda\in\mathbb{R}} \min_g \left\{\mathbb{E}_{Z\sim\mathbb{P}_Z}\left[(A_{\mathbb{P}}(Z) + \lambda\bar{A}_{\mathbb{P}}(Z)) + (B_{\mathbb{P}}(Z) - \lambda\bar{B}_{\mathbb{P}}(Z))\mathbb{1}_{\{g(Z)=1\}}\right]\right\} \\
&= \mathbb{E}_{Z\sim\mathbb{P}_Z}\left[(A_{\mathbb{P}}(Z) + B_{\mathbb{P}}(Z)\mathbb{1}_{\{g_{\lambda^*}(Z)=1)\}}) + \lambda^*(\bar{A}_{\mathbb{P}}(Z) - \bar{B}_{\mathbb{P}}(Z)\mathbb{1}_{\{g_{\lambda^*}(Z)=1)\}}\right] \\
&= \mathbb{E}_{Z\sim\mathbb{P}_Z}\left[A_{\mathbb{P}}(Z) + B_{\mathbb{P}}(Z)\mathbb{1}_{\{g_{\lambda^*}(Z)=1)\}}\right] = \mathcal{R}(g_{\lambda^*}) \ ,
\end{aligned}$$

where the last equality is due to the definition of the risk and the one before is thanks Eq. (6.1). All this implies that

$$\mathcal{R}(g_{\lambda^*}) = \min\left\{\mathcal{R}(g) \ : \ \mathbb{E}_{Z\sim\mathbb{P}_Z}[\bar{A}_{\mathbb{P}}(Z) - \bar{B}_{\mathbb{P}}(Z)\mathbb{1}_{\{g(Z)=1\}}] = 0\right\} \ ,$$

thus,

$$g_{\lambda^*} \in \arg\min\left\{\mathcal{R}(g) \ : \ \mathbb{E}_{Z\sim\mathbb{P}_Z}[\bar{A}_{\mathbb{P}}(Z) - \bar{B}_{\mathbb{P}}(Z)\mathbb{1}_{\{g(Z)=1\}}] = 0\right\} \ .$$

Now, let us derive the excess risk for any classifier g satisfying the constraints in Equation (1.11). Fix an arbitrary classifier g satisfying the constraints in Equation (1.11), clearly since $\mathbb{E}_{Z\sim\mathbb{P}_Z}[\bar{A}_{\mathbb{P}}(Z) - \bar{B}_{\mathbb{P}}(Z)\mathbb{1}_{\{g(Z)=1\}}] = 0$ we have

$$\begin{aligned}
\mathcal{R}(g) - \mathcal{R}(g^*) &= \mathbb{E}_{Z\sim\mathbb{P}_Z}\left[B_{\mathbb{P}}(Z)\mathbb{1}_{\{g(Z)=1\}} - B_{\mathbb{P}}(Z)\mathbb{1}_{\{g^*(Z)=1\}}\right] \\
&= \mathbb{E}_{Z\sim\mathbb{P}_Z}\left[B_{\mathbb{P}}(Z)\mathbb{1}_{\{g(Z)=1\}} - B_{\mathbb{P}}(Z)\mathbb{1}_{\{g^*(Z)=1\}}\right] \\
&\quad + \lambda^*(\mathbb{E}_{Z\sim\mathbb{P}_Z}[\bar{A}_{\mathbb{P}}(Z) - \bar{B}_{\mathbb{P}}(Z)\mathbb{1}_{\{g(Z)=1\}}]) - \lambda^*(\mathbb{E}_{Z\sim\mathbb{P}_Z}[\bar{A}_{\mathbb{P}}(Z) - \bar{B}_{\mathbb{P}}(Z)\mathbb{1}_{\{g^*(Z)=1\}}]) \\
&= \mathbb{E}_{Z\sim\mathbb{P}_Z}\left[(B_{\mathbb{P}}(Z) - \lambda^*\bar{B}_{\mathbb{P}}(Z))\mathbb{1}_{\{g(Z)=1\}} - (B_{\mathbb{P}}(Z) - \lambda^*\bar{B}_{\mathbb{P}}(Z))\mathbb{1}_{\{g^*(Z)=1\}}\right] \\
&= \mathbb{E}_{Z\sim\mathbb{P}_Z}\left|B_{\mathbb{P}}(Z) - \lambda^*\bar{B}_{\mathbb{P}}(Z)\right|\mathbb{1}_{\{g(Z)\neq g^*(Z)\}} \ ,
\end{aligned}$$

where the last equality holds thanks to the form of the Bayes optimal classifier g^*. □

Some technical results

Let us first introduce the notion of Kullback–Leibler divergence of two probability measures.

Definition 8. *Given any two probability measures $\mathbb{P}_1, \mathbb{P}_2$ on some space measurable space $(\mathcal{X}, \mathcal{A})$ the Kullback–Leibler divergence between $\mathbb{P}_1$ and $\mathbb{P}_2$ is defined as*

$$\mathrm{KL}(\mathbb{P}_1, \mathbb{P}_2) := \begin{cases} \int_{\mathcal{X}} \log\left(\frac{d\mathbb{P}_1}{d\mathbb{P}_2}\right) d\mathbb{P}_1, & \mathrm{supp}(\mathbb{P}_1) \subset \mathrm{supp}(\mathbb{P}_2) \\ +\infty, & \textit{otherwise} \end{cases} \ , \tag{2.7}$$

The next result is used in the context of the F-score and in the context of confidence set classification. It typically allows to fix the size of the hypothesis set when proving lower bounds and describes how reach the Hamming hypercube is.

Lemma 6. *Let $\delta(\sigma, \sigma')$ denote the Hamming distance between $\sigma, \sigma' \in \{-1, 1\}^m$ given by*

$$\delta(\sigma, \sigma') := \sum_{i=1}^m \mathbb{1}_{\{\sigma_i \neq \sigma'_i\}} \ .$$

There exists $\mathcal{W} \subset \{-1,1\}^m$ *such that for all* $\sigma \neq \sigma' \in \mathcal{W}$ *we have*

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

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

The next lemma is a version of Fano's inequality derived by Birgé [2005], it is often used to derive lower bounds for a variety of statistical problems.

Lemma 5. *Let* $\{\mathbb{P}_i\}_{i=0}^m$ *be a finite family of probability measures on* $(\mathcal{X}, \mathcal{A})$ *and let* $\{A_i\}_{i=0}^m$ *be a finite family of disjoint events such that* $A_i \in \mathcal{A}$ *for each* $i = 0, \dots, m$. *Then,*

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

The next two theorems became classical tools in the theory of empirical processes. In the context of the present manuscript, these results are used in Section 2.2 when we discuss the setup of fair binary classification.

Theorem 19 (Symmetrization (see also Theorem 2.1 in [Koltchinskii, 2011]))**.** *Let* $Z_1, \dots, Z_n$ *be* i.i.d. *copies of a real valued random variable* $Z \sim \mathbb{P}$ *and* $\mathcal{F} = \{f : \mathbb{R} \to \mathbb{R}\}$ *be a class of* $\mathbb{P}$*-integrable functions. Denote by* $\varepsilon_1, \dots, \varepsilon_n$ i.i.d. *Rademacher variables independent from* $Z, Z_1, \dots, Z_n$, *that is, for each* $i \in [n]$ *we have* $\mathbb{P}(\varepsilon = -1) = \mathbb{P}(\varepsilon = +1) = 1/2$, *then*

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(Z_i) - \mathbb{E}[f(Z)] \right| \leq 2\mathbb{E} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(Z_i) \right| \ ,$$

where on the left hand side the expectation is take w.r.t. *the distribution of* $Z_1, \dots, Z_n$ *and on the right hand side* w.r.t. *the distribution of* $Z_1, \dots, Z_n, \varepsilon_1, \dots, \varepsilon_n$.

Theorem 20 (Contraction theorem (see also Theorem 2.2 in [Koltchinskii, 2011]))**.** *Let* $\mathcal{T} \subset \mathbb{R}^n$ *and let* $\varphi_i : \mathbb{R} \to \mathbb{R}$ *for* $i \in [n]$ *be functions satisfying* $\varphi_i(0) = 0$ *and*

$$|\varphi_i(u) - \varphi_i(v)| \leq |u - v| \ \text{ for all } u, v \in \mathbb{R} \ ,$$

that is, each φ_i *is a contraction. Denote by* $\varepsilon_1, \dots, \varepsilon_n$ i.i.d. *Rademacher variables, then we have*

$$\mathbb{E} \sup_{t \in \mathcal{T}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \varphi_i(t_i) \right| \leq \mathbb{E} \sup_{t \in \mathcal{T}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i t_i \right| \ ,$$

where for all $i \in [n]$ *we denote by* t_i *the* i^{th} *coordinate of the vector* $t \in \mathcal{T}$ *and the expectation is taken* w.r.t. *the Rademacher variables* $\varepsilon_1, \dots, \varepsilon_n$.

Bibliography

R. Adams. *Sobolev spaces / Robert A. Adams.* Academic Press New York, 1975. ISBN 0120441500. (page 15)

J. Adebayo and L. Kagal. Iterative orthogonal feature projection for diagnosing bias in black-box models. In *Conference on Fairness, Accountability, and Transparency in Machine Learning*, 2016. (page 56)

A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach. A reductions approach to fair classification. *arXiv preprint arXiv:1803.02453*, 2018. (page 56, 57)

R. Agrawal, A. Gupta, Y. Prabhu, and M. Varma. Multi-label learning with millions of labels: Recommending advertiser bid phrases for web pages. In *Proceedings of the International World Wide Web Conference*, May 2013. (page 27)

D. Anbar. A modified robbins-monro procedure approximating the zero of a regression function from below. *Ann. Statist.*, 5(1):229–234, 01 1977. (page 95)

S. Arlot and R. Genuer. Analysis of purely random forests bias. *arXiv preprint arXiv:1407.3939*, 2014. (page 62)

J-Y. Audibert. Aggregated estimators and empirical complexity for least square regression. *Ann. Inst. H. Poincaré Probab. Statist.*, 40(6):685–736, 2004. (page 34, 44, 53)

J-Y. Audibert. Fast learning rates in statistical inference through aggregation. *Ann. Statist.*, 37(4):1591–1646, 2009. (page 10)

J-Y. Audibert and A. Tsybakov. Fast learning rates for plug-in classifiers. *Ann. Statist.*, 35 (2):608–633, 2007. (page 10, 16, 17, 18, 20, 23, 26, 31, 34, 37, 40, 41, 43, 44, 49, 51, 52, 57, 62, 95, 96, 98, 100, 101, 102, 104, 105, 115, 118, 131, 139, 146, 148, 151)

R. Babbar and B. Schölkopf. Dismec: Distributed sparse machines for extreme multi-label classification. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pages 721–729, 2017. (page 136)

S. Barocas and A. Selbst. Big data's disparate impact. *Calif. L. Rev.*, 104:671, 2016. (page 24)

S. Barocas, M. Hardt, and A. Narayanan. *Fairness and Machine Learning.* fairmlbook.org, 2018. http://www.fairmlbook.org. (page 24, 55)

P. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: risk bounds and structural results. *J. Mach. Learn. Res.*, 3(Spec. Issue Comput. Learn. Theory):463–482, 2002. ISSN 1532-4435. (page 34)

P. Bartlett and M. Wegkamp. Classification with a reject option using a hinge loss. *J. Mach. Learn. Res.*, 9:1823–1840, 2008. (page 94)

P. Bartlett, O. Bousquet, and S. Mendelson. Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005. (page 11)

P. Bartlett, M. Jordan, and J. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006. (page 12)

Z. Barutcuoglu, R. E. Schapire., and O. G. Troyanskaya. Hierarchical multi-label prediction of gene function. *Bioinformatics*, 22(7):830–836, 2006. (page 27)

P. Bellec, A. Dalalyan, E. Grappin, and Q. Paris. On the prediction loss of the lasso in the partially labeled setting. *Electron. J. Statist.*, 12(2):3443–3472, 2018. (page 96)

O. Besov, V. Ilin, and S. Nikolskii. *Integralnye predstavleniya funktsii i teoremy vlozheniya*. Fizmatlit "Nauka", Moscow, second edition, 1996. ISBN 5-02-014532-7. In Russian. (page 15)

A. Beutel, J. Chen, Z. Zhao, and E. H. Chi. Data decisions and theoretical implications when adversarially learning fair representations. In *Conference on Fairness, Accountability, and Transparency in Machine Learning*, 2017. (page 56)

A. Beygelzimer, J. Langford, Yu. Lifshits, G. Sorkin, and A. Strehl. Conditional probability tree estimation analysis and algorithms. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 51–58. AUAI Press, 2009. (page 26)

K. Bhatia, H. Jain, P. Kar, M. Varma, and P. Jain. Sparse local embeddings for extreme multi-label classification. In *NIPS*. 2015. (page 136)

L. Birgé. A new lower bound for multiple hypothesis testing. *IEEE Transactions on Information Theory*, 51(4), April 2005. (page 53, 102, 126, 157)

A. Blumer, A. Ehrenfeucht, D. Haussler, and M. Warmuth. Learnability and the vapnik-chervonenkis dimension. *Journal of the ACM (JACM)*, 36(4):929–965, 1989. (page 6, 11)

S. Bobkov and M. Ledoux. One-dimensional empirical measures, order statistics and Kantorovich transport distances. 2016. to appear in the Memoirs of the Amer. Math. Soc. (page 42, 48, 105, 110, 113)

O. Bousquet and A. Elisseeff. Stability and generalization. *Journal of machine learning research*, 2(Mar):499–526, 2002. (page 12)

L. Breiman. Consistency for a simple model of random forests. Technical report, Statistics Department University Of California At Berkeley, 2004. (page 62)

L Brown and M Low. A constrained risk inequality with applications to nonparametric functional estimation. *The annals of Statistics*, 24(6):2524–2535, 1996. (page 95)

T. Calders, F. Kamiran, and M. Pechenizkiy. Building classifiers with independency constraints. In *IEEE international conference on Data mining*, 2009. (page 25, 55)

F. Calmon, D. Wei, B. Vinzamuri, K. N. Ramamurthy, and K. R. Varshney. Optimized pre-processing for discrimination prevention. In *Neural Information Processing Systems*, 2017. (page 55, 56)

F. Chierichetti, R. Kumar, S. Lattanzi, and S. Vassilvitskii. Fair clustering through fairlets. In *Neural Information Processing Systems*, 2017. (page 55)

C.-K. Chow. An optimum character recognition system using decision functions. *IRE Transactions on Electronic Computers*, (4):247–254, 1957. (page 5, 8, 13, 94)

C.-K. Chow. On optimum error and reject trade-off. *IEEE Transactions on Information Theory*, 16:41–46, 1970. (page 5, 13, 94)

E. Chzhen. Optimal rates for F-score binary classification. Submitted to JMLR, May 2019a. (page 29)

E. Chzhen. Classification of sparse binary vectors. Submitted to J. Stat. Plan. Inference, 2019b. (page 29)

E. Chzhen, C. Denis, M Hebiri, and J. Salmon. On the benefits of output sparsity for multi-label classification. Technical report, 2017. (page 29, 137, 143)

E. Chzhen, C. Denis, and M. Hebiri. Minimax semi-supervised confidence sets for multi-class classification. Submitted to Ann. Stat., April 2019a. (page 27, 29, 32, 38, 40, 41, 57, 58)

E. Chzhen, C. Denis, M. Hebiri, L. Oneto, and M. Pontil. Leveraging Labeled and Unlabeled Data for Consistent Fair Binary Classification. Submitted to NeurIPS19, 2019b. (page 29)

E. Chzhen, M. Hebiri, and J. Salmon. On lasso refitting strategies. *Bernouilli (to appear)*, 2019c. (page 29)

S. D. Conte and C. W. D. Boor. *Elementary Numerical Analysis: An Algorithmic Approach*. McGraw-Hill Higher Education, 3rd edition, 1980. (page 39)

A. Cotter, M. Gupta, H. Jiang, N. Srebro, K. Sridharan, S. Wang, B. Woodworth, and S. You. Training well-generalizing classifiers for fairness metrics and other data-dependent constraints. *arXiv preprint arXiv:1807.00028*, 2018. (page 56)

F. Cribari-Neto, N. Garcia, and K. Vasconcellos. A note on inverse moments of binomial variates. *Brazilian Review of Econometrics*, 20(2):269–277, 2000. (page 70)

B. de Finetti. *Probability, induction and statistics. The art of guessing*. John Wiley & Sons, London-New York-Sydney, 1972. Wiley Series in Probability and Mathematical Statistics. (page 123)

B. de Finetti. *Theory of probability: a critical introductory treatment. Vol. 1*. John Wiley & Sons, London-New York-Sydney, 1974. (page 123)

O. Dekel and O. Shamir. Multiclass-multilabel classification with more classes than examples. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 137–144, 2010. (page 26)

K. Dembczyński, W. Waegeman, W. Cheng, and E. Hüllermeier. On label dependence and loss minimization in multi-label classification. *Machine Learning*, 88(1-2):5–45, 2012. (page 28)

K. Dembczynski, A. Jachnik, W. Kotlowski, W. Waegeman, and E. Hüllermeier. Optimizing the f-measure in multi-label classification: Plug-in rule approach versus structured loss minimization. In *ICML*, pages 1130–1138, 2013. (page 136, 137, 139)

K. Dembczynski, W. Kotłowski, O. Koyejo, and N. Natarajan. Consistency analysis for binary classification revisited. In *ICML*, pages 961–969. JMLR. org, 2017. (page 35, 38)

C. Denis and M. Hebiri. Confidence sets for classification. In *Statistical Learning and Data Sciences*, pages 301–312. Springer International Publishing, 2015a. (page 57, 58, 94, 138, 146)

C. Denis and M. Hebiri. Confidence sets for classification. In *International Symposium on Statistical Learning and Data Sciences*, pages 301–312. Springer, 2015b. (page 13)

C. Denis and M. Hebiri. Confidence sets with expected sizes for multiclass classification. *JMLR*, 18(1):3571–3598, 2017. (page 27, 32, 57, 89, 91, 94, 95, 104, 105, 112, 117, 118)

L. Devroye. The uniform convergence of nearest neighbor regression function estimators and their application in optimization. *IEEE Transactions on Information Theory*, 24(2): 142–151, 1978. (page 57, 62)

L. Devroye. Any discrimination rule can have an arbitrarily bad probability of error for finite sample size. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (2): 154–157, 1982. (page 10)

L. Devroye, L. Györfi, and G. Lugosi. *A probabilistic theory of pattern recognition*, volume 31 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, 1996. (page 4, 12, 17)

V Dinh, Lam Si Tung Ho, Nguyen Viet Cuong, Duy Nguyen, and Binh T. Nguyen. Learning from non-iid data: Fast rates for the one-vs-all multiclass plug-in classifiers. In Rahul Jain, Sanjay Jain, and Frank Stephan, editors, *Theory and Applications of Models of Computation*, pages 375–387, Cham, 2015. Springer International Publishing. (page 26)

M. Donini, L. Oneto, S. Ben-David, J. S. Shawe-Taylor, and M. Pontil. Empirical risk minimization under fairness constraints. In *Neural Information Processing Systems*, 2018. (page 55, 56, 57, 59, 63, 64, 65, 67)

R. Dudley. The sizes of compact subsets of hilbert space and continuity of gaussian processes. *Journal of Functional Analysis*, 1(3):290–330, 1967. (page 11)

A. Dvoretzky, J. Kiefer, and J. Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Ann. Math. Statist.*, 27 (3):642–669, 09 1956. (page 42, 86)

C. Dwork, N. Immorlica, A. T. Kalai, and M. D. M. Leiserson. Decoupled classifiers for group-fair and efficient machine learning. In *Conference on Fairness, Accountability and Transparency*, 2018. (page 55, 65)

D. Eisenstat and D. Angluin. The vc dimension of k-fold union. *Information Processing Letters*, 101(5):181–184, 2007. (page 11)

M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. In *International Conference on Knowledge Discovery and Data Mining*, 2015. (page 56)

S. Gao, W. Wu, C. H. Lee, and T. S. Chua. A MFoM learning approach to robust multiclass multi-label text categorization. In *ICML*, pages 329–336, 2004. (page 27)

W. Gao and Z-H. Zhou. On the consistency of multi-label learning. In *COLT*, pages 341–358, 2011. (page 137)

R. Genuer. Variance reduction in purely random forests. *Journal of Nonparametric Statistics*, 24(3):543–562, 2012. (page 62)

D. Gil, J. Girela, J. De Juan, J. Gomez-Torres, and M. Johnsson. Predicting seminal quality with artificial intelligence methods. *Expert Systems with Applications*, 39(16):12564–12573, 2012. (page 16)

E. Gilbert. A comparison of signalling alphabets. *The Bell system technical journal*, 31(3): 504–522, 1952. (page 54)

E. Giné and R. Nickl. *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2015. doi: 10.1017/CBO9781107337862. (page 14, 15)

H A. Güvenir, G. Demiröz, and N. Ilter. Learning differential diagnosis of erythemato-squamous diseases using voting feature intervals. *Artificial Intelligence in Medicine*, 13: 147, 1998. (page 16)

M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. In *Neural Information Processing Systems*, 2016. (page 9, 23, 24, 25, 31, 55, 56, 57, 58, 63, 65, 67)

J. A. Hartigan. Estimation of a convex density contour in two dimensions. *J. Amer. Statist. Assoc.*, 82(397):267–270, 1987. (page 91, 95)

D. Haussler. Sphere packing numbers for subsets of the boolean n-cube with bounded vapnik-chervonenkis dimension. *Journal of Combinatorial Theory, Series A*, 69(2):217–232, 1995. (page 11)

R. Herbei and M. Wegkamp. Classification with reject option. *Canad. J. Statist.*, 34(4): 709–721, 2006. (page 5, 94, 96)

W. Hoeffding. Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, 58(301):13–30, 1963. (page 111)

M. Hristache, A. Juditsky, J. Polzehl, and V. Spokoiny. Structure adaptive approach for dimension reduction. *The Annals of Statistics*, 29(6):1537–1566, 2001a. (page 16)

M. Hristache, A. Juditsky, and V. Spokoiny. Direct estimation of the index coefficient in a single-index model. *Annals of Statistics*, pages 595–623, 2001b. (page 16)

I. Ibragimov and R. Khasminskii. *Statistical Estimation: Asymtotic Theory*. Applications of Mathematics Series. Springer-Verlag, 1981. ISBN 9780387905235. (page 15)

S. Jabbari, M. Joseph, M. Kearns, J. Morgenstern, and A. Roth. Fair learning in markovian environments. In *Conference on Fairness, Accountability, and Transparency in Machine Learning*, 2016. (page 55)

H. Jain, Y. Prabhu, and M. Varma. Extreme multi-label loss functions for recommendation, tagging, ranking & other missing label applications. In *KDD*, pages 935–944, 2016. (page 136, 137)

H. Jiang. Uniform convergence rates for kernel density estimation. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1694–1703. JMLR. org, 2017. (page 16)

M. Joseph, M. Kearns, J. H. Morgenstern, and A. Roth. Fairness in learning: Classic and contextual bandits. In *Neural Information Processing Systems*, 2016. (page 55)

A. Juditsky, O. Lepski, and A. Tsybakov. Nonparametric estimation of composite functions. *The Annals of Statistics*, 37(3):1360–1404, 2009. (page 16)

F. Kamiran and T. Calders. Classifying without discriminating. In *International Conference on Computer, Control and Communication*, 2009. (page 56)

F. Kamiran and T. Calders. Classification with no discrimination by preferential sampling. In *Machine Learning Conference*, 2010. (page 56)

F. Kamiran and T. Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, 2012. (page 56)

S. S. Keerthi, V. Sindhwani, and O. Chapelle. An efficient method for gradient-based adaptation of hyperparameters in svm models. In *NIPS*, pages 673–680. 2007. (page 38)

N. Kilbertus, M. Rojas-Carulla, G. Parascandolo, M. Hardt, D. Janzing, and B. Schölkopf. Avoiding discrimination through causal reasoning. In *Neural Information Processing Systems*, 2017. (page 55)

A.N. Kolmogorov and V.M. Tikhomirov. eps-entropy and eps-capacity of sets in functional spaces. *Uspekhi Matematicheskikh Nauk, [N. S.]*, 17, 01 1961. doi: 10.1007/978-94-017-2973-4_7. (page 15)

V. Koltchinskii. *Oracle inequalities in empirical risk minimization and sparse recovery problems*, volume 2033 of *Lecture Notes in Mathematics*. Springer, Heidelberg, 2011. (page 82, 86, 157)

A. Korostelëv and A. Tsybakov. *Minimax theory of image reconstruction*, volume 82 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 1993. (page 16)

O. Koyejo, N. Natarajan, P. Ravikumar, and I. Dhillon. Consistent binary classification with generalized performance metrics. In *NIPS*, pages 2744–2752. 2014. (page 38)

O. Koyejo, N. Natarajan, P. Ravikumar, and I. Dhillon. Consistent multilabel classification. In *NIPS*, pages 3321–3329. 2015. (page 57, 137)

M. J. Kusner, J. Loftus, C. Russell, and R. Silva. Counterfactual fairness. In *Neural Information Processing Systems*, 2017. (page 55)

M. Lapin, M. Hein, and B. Schiele. Top-k multiclass svm. In *NIPS*, pages 325–333, 2015. (page 137)

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. (page 26)

M. Ledoux and M. Talagrand. *Probability in Banach spaces*, volume 23 of *Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)]*. Springer-Verlag, Berlin, 1991. ISBN 3-540-52013-9. Isoperimetry and processes. (page 11)

J. Lei. Classification with confidence. *Biometrika*, 101(4):755–769, 2014. (page 57, 58, 94, 104)

O. Lepskii. Asymptotic minimax estimation with prescribed properties. *Theory of Probability & Its Applications*, 34(4):604–615, 1990. (page 95, 137)

D. Lewis. Evaluating and optimizing autonomous text classification systems. In *ACM*, pages 246–254. ACM Press, 1995. (page 19, 35)

T. Li, A. Prasad, and P. Ravikumar. Fast classification rates for high-dimensional gaussian generative models. In *NIPS*, pages 1054–1062, 2015. (page 139)

X. Li, F. Zhao, and Y. Guo. Multi-label image classification with a probabilistic label enhancement model. pages 430–439, 2014. (page 27)

Y. Li, Y. Song, and J. Luo. Improving pairwise ranking for multi-label image classification. *CVPR*, pages 1837–1845, 2017. (page 137)

F. Louzada, A. Ara, and G. Fernandes. Classification methods applied to credit scoring: Systematic review and overall comparison. *Surveys in Operations Research and Management Science*, 21(2):117–134, 2016. (page 16)

K. Lum and J. Johndrow. A statistical framework for fair predictive algorithms. *arXiv preprint arXiv:1610.08077*, 2016. (page 55)

E. Mammen and A. Tsybakov. Smooth discrimination analysis. *Ann. Statist.*, 27(6):1808–1829, 1999. (page 11, 17)

P. Massart. The tight constant in the dvoretzky-kiefer-wolfowitz inequality. *Ann. Probab.*, 18(3):1269–1283, 07 1990. (page 42, 50, 86)

P. Massart and É Nédélec. Risk bounds for statistical learning. *Ann. Statist.*, 34(5):2326–2366, 10 2006. (page 11, 18, 34, 37)

J. Matoušek. *Lectures on discrete geometry*, volume 108. Springer, 2002. (page 11)

C. McDiarmid. *On the method of bounded differences*, page 148–188. London Mathematical Society Lecture Note Series. Cambridge University Press, 1989. (page 12)

A. Menon and R. C. Williamson. The cost of fairness in binary classification. In *Conference on Fairness, Accountability and Transparency*, 2018. (page 56, 58)

A. Menon, H. Narasimhan, S. Agarwal, and S. Chawla. On the statistical consistency of algorithms for binary classification under class imbalance. In *ICML*, volume 28, pages 603–611. PMLR, 17–19 Jun 2013. (page 38)

H. Narasimhan, R. Vaish, and S. Agarwal. On the statistical consistency of plug-in classifiers for non-decomposable performance measures. In *NIPS*, pages 1493–1501. 2014. (page 38)

L. Oneto, M. Donini, A. Elders, and M. Pontil. Taking advantage of multitask learning for fair classification. In *AAAI/ACM Conference on AI, Ethics, and Society*, 2019a. (page 65)

L. Oneto, M. Donini, and M. Pontil. General fair empirical risk minimization. *arXiv preprint arXiv:1901.10080*, 2019b. (page 56)

I. Partalas, A. Kosmopoulos, M. Baskiotis, T. Artières, G. Paliouras, É. Gaussier, I. Androutsopoulos, M.-R. Amini, and P. Gallinari. LSHTC: A benchmark for large-scale text classification. *CoRR*, abs/1503.08581, 2015. (page 27)

G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Weinberger. On fairness and calibration. In *Neural Information Processing Systems*, 2017. (page 56)

W. Polonik. Measuring mass concentrations and estimating density contour clusters-an excess mass approach. *Ann. Statist.*, 23(3):855–881, 06 1995. (page 17, 91, 95)

Y. Prabhu and M. Varma. Fastxml: A fast, accurate and stable tree-classifier for extreme multi-label learning. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 263–272. ACM, 2014. (page 28, 32)

Y. Prabhu, A. Kag, S. Gopinath, K. Dahiya, S. Harsola, R. Agrawal, and M. Varma. Extreme multi-label learning with label features for warm-start tagging, ranking and recommendation. In *Proceedings of the ACM International Conference on Web Search and Data Mining*, February 2018. (page 27)

H. Ramaswamy, A. Tewari, and S. Agarwal. Consistent algorithms for multiclass classification with an abstain option. *Electronic Journal of Statistics*, 12(1):530–554, 2018. (page 94)

P. Rigollet. Generalization error bounds in semi-supervised classification under the cluster assumption. *Journal of Machine Learning Research*, 8(Jul):1369–1392, 2007. (page 14, 38, 96)

P. Rigollet and R. Vert. Optimal rates for plug-in estimators of density level sets. *Bernoulli*, 2009. (page 10, 16, 44, 51, 91, 95, 96, 100, 102, 118)

J. E. Roemer and A. Trannoy. Equality of opportunity. In *Handbook of income distribution*, 2015. (page 65)

W. Rudin. *Real and complex analysis*. McGraw-Hill Book Co., New York, third edition, 1987. (page 15)

M. Sadinle, J. Lei, and L. Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, pages 1–12, 2018. (page 57, 58, 94, 95, 104)

M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz. A bayesian approach to filtering junk E-mail. In *Learning for Text Categorization: Papers from the 1998 Workshop*, Madison, Wisconsin, 1998. AAAI Technical Report WS-98-05. (page 16)

J. Schmidt-Hieber. Nonparametric regression using deep neural networks with relu activation function. *arXiv preprint arXiv:1708.06633*, 2017. (page 16)

E. Scornet, G. Biau, and J.-P. Vert. Consistency of random forests. *Ann. Statist.*, 43(4): 1716–1741, 08 2015. (page 62)

S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Learnability, stability and uniform convergence. *Journal of Machine Learning Research*, 11(Oct):2635–2670, 2010. (page 12)

J. Simon. Sobolev, besov and nikolskii fractional spaces: imbeddings and comparisons for vector valued spaces on an interval. *Annali di Matematica Pura ed Applicata*, 157(1): 117–148, 1990. (page 15)

A. Singh, R. Nowak, and J. Zhu. Unlabeled data: Now it helps, now it doesn't. In *NIPS*, pages 1513–1520. 2009. (page 38, 40, 96)

S. Sobolev. *Some Applications of Functional Analysis in Mathematical Physics*. Translations of mathematical monographs. American Mathematical Society, 1991. ISBN 9780821845493. URL https://books.google.fr/books?id=WNinIHsaqXsC. (page 15)

C. Stone. Consistent nonparametric regression. *Ann. Statist.*, pages 595–620, 1977. (page 95, 104)

C. Stone. Optimal global rates of convergence for nonparametric regression. *The annals of statistics*, pages 1040–1053, 1982. (page 16, 17, 105)

V. Sudakov. Geometric problems of the theory of infinite-dimensional probability distributions. *Trudy Mat. Inst. Steklov.*, 141:191, 1976. ISSN 0371-9685. (page 11)

G. Tsoumakas, I. Katakis, and I. Vlahavas. Mining multi-label data. In *Data mining and knowledge discovery handbook*, pages 667–685. Springer, 2009. (page 27)

A. Tsybakov. Robust reconstruction of functions by the local-approximation method. *Problemy Peredachi Informatsii*, 22(2):69–84, 1986. (page 16, 95, 104)

A. Tsybakov. On nonparametric estimation of density level sets. *Ann. Statist.*, 25(3):948–969, 06 1997. (page 17, 91, 95)

A. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Ann. Statist.*, 32(1): 135–166, 02 2004. (page 11, 94, 95, 142)

A. Tsybakov. *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York, 2009. (page 14, 15, 17, 44, 98, 102, 105)

A. Tsybakov and S. van de Geer. Square root penalty: adaptation to the margin in classification and in edge estimation. *The Annals of Statistics*, 33(3):1203–1224, 2005. (page 11)

S. Vallender. Calculation of the wasserstein distance between probability distributions on the line. *Theory of Probability & Its Applications*, 18(4):784–786, 1974. (page 42)

S. van de Geer. High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, 36(2):614–645, 2008. (page 62)

A. van der Vaart. *Asymptotic statistics*, volume 3 of *Camb. Ser. Stat. Probab. Math.* Cambridge University Press, Cambridge, 1998. (page 112)

C. J. van Rijsbergen. Foundation of evaluation. *Journal of documentation*, 30(4):365–373, 1974. (page 19, 35)

V. Vapnik. *Statistical learning theory*. Wiley, 1998. (page 4, 6, 34, 95)

V. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. In *Measures of complexity*, pages 11–30. Springer, 1971. (page 6, 11, 86)

R. Varshamov. Estimate of the number of signals in error correcting codes. *Dokl. Akad. Nauk SSSR*, 117:739—-741, 1957. (page 54)

V. Vovk. On-line confidence machines are well-calibrated. In *Proceedings of the Forty-Third Annual Symposium on Foundations of Computer Science*, pages 187–196. CA. IEEE Computer Society, Los Alamitos, 2002a. (page 93)

V. Vovk. Asymptotic optimality of transductive confidence machine. In *Algorithmic learning theory*, volume 2533 of *Lecture Notes in Comput. Sci.*, pages 336–350. Springer, Berlin, 2002b. (page 93)

V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic learning in a random world*. Springer, New York, 2005. (page 93, 94)

M. Wegkamp and M. Yuan. Support vector machines with a reject option. *Bernoulli*, 17(4): 1368–1385, 2011. (page 94)

J. Wellner. Empirical processes: Theory and applications. Technical report, Delft University of Technology, 2005. (page 82)

B. Yan, S. Koyejo, K. Zhong, and P. Ravikumar. Binary classification with karmic, threshold-quasi-concave metrics. In *ICML*, volume 80. PMLR, 10–15 Jul 2018. (page 38, 39, 57, 58, 62)

Y. Yang. Minimax nonparametric classification: Rates of convergence. *IEEE Transactions on Information Theory*, 45(7):2271–2284, Nov 1999. (page 16, 17, 18, 34, 57, 95)

S. Yao and B. Huang. Beyond parity: Fairness objectives for collaborative filtering. In *Neural Information Processing Systems*, 2017. (page 55)

N. Ye, K. Chai, W. Lee, and H. Chieu. Optimizing f-measures: A tale of two approaches. In *ICML*, 2012. (page 35, 38)

H. F. Yu, P. Jain, P. Kar, and I. S. Dhillon. Large-scale multi-label learning with missing labels. In *ICML*, pages 593–601, 2014. (page 136)

M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *International Conference on World Wide Web*, 2017. (page 25, 55, 63)

M. B. Zafar, I. Valera, M. Gomez-Rodriguez, and K. P. Gummadi. Fairness constraints: A flexible approach for fair classification. *Journal of Machine Learning Research*, 20(75): 1–42, 2019. (page 55)

R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning fair representations. In *International Conference on Machine Learning*, 2013. (page 55, 56)

T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32(1):56–85, 2004. (page 12)

M-J. Zhao, N. Edakunni, A. Pocock, and G. Brown. Beyond fano's inequality: bounds on the optimal f-score, ber, and cost-sensitive risk and their implications. *JMLR*, 14(Apr): 1033–1090, 2013. (page 13, 19, 20, 30, 35, 36, 57)

I. Zliobaite. On the relation between accuracy and fairness in binary classification. *arXiv preprint arXiv:1505.05723*, 2015. (page 55)

Title: Méthodes de type plug-in en classification.

Key words: classification contrainte, classification supervisée, classification semi-supervisée, classification par plug-in, ensembles de confiance, F-score, analyse minimax, classification équitable, classification multi-label.

Abstract. Ce manuscrit étudie plusieurs problèmes de classification sous contraintes. Dans ce cadre de classification, notre objectif est de construire un algorithme qui a des performances aussi bonnes que la meilleure règle de classification ayant une propriété souhaitée. Fait intéressant, les méthodes de classification de type plug-in sont bien appropriées à cet effet. De plus, il est montré que, dans plusieurs configurations, ces règles de classification peuvent exploiter des données non étiquetées, c'est-à-dire qu'elles sont construites de manière semi-supervisée.

Le Chapitre 2 décrit deux cas particuliers de la classification binaire - la classification où la mesure de performance est reliée au F-score, et la classification équitable. A ces deux problèmes, des procédures semi-supervisées sont proposées. En particulier, dans le cas du F-score, il s'avère que cette méthode est optimale au sens minimax sur une classe usuelle de distributions non-paramétriques. Aussi, dans le cas de la classification équitable, la méthode proposée est consistante en terme de risque de classification, tout en satisfaisant asymptotiquement la contrainte dégalité des chances. De plus, la procédure proposée dans ce cadre d'étude surpasse en pratique les algorithmes de pointe.

Le Chapitre 3 décrit le cadre de la classification multi-classes par le biais d'ensembles de confiance. Là encore, une procédure semi-supervisée est proposée et son optimalité presque minimax est établie. Il est en outre établi qu'aucun algorithme supervisé ne peut atteindre une vitesse de convergence dite rapide.

Le Chapitre 4 décrit un cas de classification multi-labels dans lequel on cherche à minimiser le taux de faux-négatifs sous réserve de contraintes de type presque sûres sur les règles de classification. Dans cette partie, deux contraintes spécifiques sont prises en compte : les classifieurs parcimonieux et ceux soumis à un contrôle des erreurs négatives à tort. Pour les premiers, un algorithme supervisé est fourni et il est montré que cet algorithme peut atteindre une vitesse de convergence rapide. Enfin, pour la seconde famille, il est montré que des hypothèses supplémentaires sont nécessaires pour obtenir des garanties théoriques sur le risque de classification.

Title: Plug-in methods in classification.

Key words: constrained classification, supervised classification, semi-supervised classification, plug-in classifiers, confidence sets, F-score, minimax analysis, fairness in classification, multi-label classification.

Abstract. This manuscript studies several problems of constrained classification. In this frameworks of classification our goal is to construct an algorithm which performs as good as the best classifier that obeys some desired property. Plug-in type classifiers are well suited to achieve this goal. Interestingly, it is shown that in several setups these classifiers can leverage unlabeled data, that is, they are constructed in a semi-supervised manner.

Chapter 2 describes two particular settings of binary classification – classification with F-score and classification of equal opportunity. For both problems semi-supervised procedures are proposed and their theoretical properties are established. In the case of the F-score, the proposed procedure is shown to be optimal in minimax sense over a standard non-parametric class of distributions. In the case of the classification of equal opportunity the proposed algorithm is shown to be consistent in terms of the misclassification risk and its asymptotic fairness is established. Moreover, for this problem, the proposed procedure outperforms state-of-the-art algorithms in the field.

Chapter 3 describes the setup of confidence set multi-class classification. Again, a semi-supervised procedure is proposed and its nearly minimax optimality is established. It is additionally shown that no supervised algorithm can achieve a so-called fast rate of convergence. In contrast, the proposed semi-supervised procedure can achieve fast rates provided that the size of the unlabeled data is sufficiently large.

Chapter 4 describes a setup of multi-label classification where one aims at minimizing false negative error subject to almost sure type constraints. In this part two specific constraints are considered – sparse predictions and predictions with the control over false negative errors. For the former, a supervised algorithm is provided and it is shown that this algorithm can achieve fast rates of convergence. For the later, it is shown that extra assumptions are necessary in order to obtain theoretical guarantees in this case.