Some contributions to high-dimensional statistics and learning with desired properties

Université Gustave Eiffel

Habilitation à Diriger des Recherches

Spécialité : Mathématiques Appliquées

présentée par

Mohamed Hebiri

Soutenue publiquement le 7 janvier 2021 après avis des rapporteurs,

Mme.	**Florentina**	**Bunea**	Cornell University
M.	**Antoine**	**Chambaz**	Université de Paris
M.	**Gábor**	**Lugosi**	ICREA and Pompeu Fabra University

et devant le jury composé de :

M.	**Sylvain**	**Arlot**	Université Paris-Saclay
M.	**Antoine**	**Chambaz**	Université de Paris
M.	**Arnak**	**Dalalyan**	ENSAE-CREST
Mme.	**Claire**	**Lacour**	Université Gustave Eiffel
Mme.	**Florence**	**Merlevède**	Université Gustave Eiffel
M.	**Alexandre**	**Tsybakov**	ENSAE-CREST
M.	**Nicolas**	**Vayatis**	ENS-Cachan

Acknowledgments

My first acknowledgments go to Florentina Bunea, Antoine Chambaz, and Gábor Lugosi. You did me the honor of accepting to review my HDR thesis; thank you for your time and consideration.

J'exprime toute ma reconnaissance à Sylvain Arlot, Antoine Chambaz, Arnak Dalalyan, Claire Lacour, Florence Merlevède, Sacha Tsybakov et Nicolas Vayatis de m'avoir fait l'immense honneur de faire partie de mon jury d'habilitation. Vous êtes des figures bienveillantes, vous m'avez accompagné, de près ou de loin, durant toutes mes années de chercheur. Vous m'avez tant appris !

La carrière d'un enseignant-chercheur est étroitement liée aux rencontres scientifiques et aux relations humaines que celui-ci a la chance de nouer. Je remercie chaleureusement Sara van de Geer de m'avoir accueilli en post-doc à l'ETH, pour sa gentillesse et pour tous les échanges scientifiques et amicaux que nous avons eus. J'en profite pour remercier tous mes amis de l'ETH, Marco, Sarah, Manuel et surtout Fabio et Johannes qui ont fait de cette période à Zürich un des meilleurs épisodes de ma vie.

J'ai pris plaisir à travailler chaque jour entouré de mes collègues et amis : il y a bien entendu les éternels amis Katia, Joseph, Johannes, Arnak, Zaïd, Pierre, Christophe Ch. et Karim. Merci pour votre soutien, et tout particulièrement à toi Katia "ma soeur" ! Vous avez été les protagonistes de mon épanouissement ! A ce noyau dur, se sont greffés peu à peu de nouveaux compagnons d'armes. Christophe d'abord, et ce, depuis ton arrivée à Marne en 2013. Cet événement marque un tournant dans ma vie de chercheur. J'apprécie notre collaboration. D'ailleurs, les recherches que nous avons initiées constituent une partie conséquente de ce manuscrit. Par la suite, Evgenii est arrivé ! Cher Evgenii, tu m'as demandé d'être ton directeur de thèse et pour être honnête, j'ai hésité avant d'accepter. Coup du sort : tu es aujourd'hui une des personnes les plus importantes dans ma vie. Tu es un scientifique exceptionnel, mais surtout un gars en or !
Vous deux, Christophe et Evgenii, êtes devenus des amis proches. Je garderai en mémoire nos pauses café sur la terrasse du 4-ième étage de Copernic et au bâtiment de la bibliothèque (pendant les travaux) à discuter de "*the G function*".

A tous mes amis, j'espère concrétiser encore d'autres projets en votre compagnie. Je profite de ce paragraphe pour témoigner ma gratitude à Massi et Luca, deux fabuleux collaborateurs à l'italienne !

Je remercie également tous mes collègues du LAMA pour les moments partagés ensemble, en particulier avec Luc "mon sosie", Florence, Audrey, Vlad, Romu, Luigi, Pierre-André, Dan, Claire et Chi. Merci à toi Ahmed, pour ta gaieté quotidienne. Bienvenue dans l'équipe !

Je termine par un grand remerciement à mes parents, dont les prières m'accompagnent et me guident tous les jours. Merci à ma famille et à mes amis pour leur soutien et leur bonne humeur. Je remercie, mes filles Iness, Meyssane et Essya. Vous êtes casse-pieds, mais sans vous je ne suis rien ! Enfin, je te remercie infiniment, toi Saniye, ma femme et amie pour ton soutien immuable et tout ce que tu m'as donné...

Vous tous, avez permis à cette habilitation de voir le jour. Merci !

Contents

Introduction

General presentation

The present manuscript is an overview of some of my scientific contributions as *Maître de conférences* at Université Gustave Eiffel. In the early stages of my research career I was mainly active in the field of high-dimensional statistics studying the Lasso and related estimators. Later, I started to explore new research areas connected with learning problems under distribution dependent constraints such as classification and regression with reject option, set-valued classification, and algorithmic fairness. This manuscript is focused on the following three interdependent topics which constitute a major part of my recent research activity.

- **High-dimensional statistics**. I investigated the problem of prediction ability of the Lasso estimator – one of the most used methods in the high dimensional setting that promotes sparsity. I studied how the performance of this method is influenced by the structure of the design matrix and by the calibration of the tuning parameter that controls its sparsity. I was particularly interested in describing the limits of the Lasso and established the particular regimes where fast rates, meaning $\frac{s^*}{n}$ up to log factors, are expected and when such rates are impossible. Here s^* is the underlying sparsity in the model and n is the sample size.

- **Fairness**. Classical machine learning algorithms are mainly oblivious to the basic principles of fairness and equality and often amplify discriminatory biases present in data. A great deal of effort has been devoted to bypass these issues in the last decade. Yet, more often than not, newly developed fairness aware algorithm lack in rigour and their statistical understanding is limited. In my recent research I started to address the issue of building new fairness aware algorithms which, while delivering state-of-the-art performance, come with user-friendly finite sample fairness and performance guarantees.

- **Set-valued classification**. Unlike classical single-class predictor, a set-valued classifier is allowed to output a set of possible class candidates. This framework is particularly appealing for multi-class classification problems with high ambiguity

– a typical scenario in modern days datasets. In recent years my main research area revolves around theoretical and practical study of set-valued classifiers whose expected size can be controlled.

Organization and format. Chapter 1 summarizes my earlier work on high dimensional statistics with an emphasis on the Lasso estimator. It also highlights my recent contributions to the field of fairness aware methods in machine learning; Chapter 2 focuses on my recent contributions to the literature of set-valued classification – the topic of my main research activity at the moment of writing. The results presented in the subsequent chapters are stated in a simple form and proofs are omitted so that the presentation remains as clear as possible. The reader is referred to the original papers for a more complete description of my contributions. The particular choice of the contributions that I decided to cover in this manuscript is solely dictated by my personal preferences and tastes and by my personal view of the importance of the obtained results.

Notation

$a \vee b$	the maximum between $a, b \in \mathbb{R}$
$a \wedge b$	the minimum between $a, b \in \mathbb{R}$
$\lfloor a \rfloor$	the largest non-negative integer that is less than or equal to $a > 0$
$\lceil a \rceil$	the smallest non-negative integer that is greater than or equal to $a > 0$
$a_n \lesssim b_n$	there exists a constant $c > 0$ such that $a_n \leq c b_n$ for all n, where $a_n, b_n : \mathbb{N} \to [0, +\infty)$
$a_n \gtrsim b_n$	there exists a constant $c > 0$ such that $a_n \geq c b_n$ for all n, where $a_n, b_n : \mathbb{N} \to [0, +\infty)$
$[K]$	the set $\{1, \dots, K\}$
2^T	the set of all subsets of a finite set T
$\lvert T \rvert$	the cardinality of a finite set T
$T \triangle T'$	the symmetric difference of two finite sets T, T'
T^c	the complementary set of $T \subset [d]$
$\lVert \boldsymbol{u} \rVert_q$	$(\sum_{j \in [d]} \lvert u_j \rvert^q)^{1/q}$, $0 < q < \infty$
$\lVert \boldsymbol{u} \rVert_0$	$\lvert \{ j \in [d] : u_j \neq 0 \} \rvert$
$\lVert \boldsymbol{u} \rVert_\infty$	$\max_{j \in [d]} \lvert u_j \rvert$
$\boldsymbol{u} \odot \boldsymbol{u}'$	$= (u_1 u'_1, \dots, u_d u'_d)^\top$ the coordinate-wise product of two vectors $\boldsymbol{u}$ and $\boldsymbol{u}'$
$\mathbb{A}_T$	the matrix obtained from $\mathbb{A} \in \mathbb{R}^{n \times d}$ by removing all the columns belonging to $T^c \subset [d]$
$\mathbb{A}^\top$	the transpose of a matrix $\mathbb{A}$
$\mathbb{A}^\dagger$	the Moore-Penrose pseudo inverse of a matrix $\mathbb{A}$
$\mathcal{B}(x, r)$	a Euclidean ball centered at $x \in \mathbb{R}^d$ of radius $r > 0$
$\mathbf{E}$	generic expectation sign
$\mathbf{P}$	generic probability sign
$\mathrm{Leb}(\cdot)$	Lebesgue measure on $\mathbb{R}^d$
$\mathrm{supp}(\mu)$	the support of an absolutely continuous (*w.r.t.* Leb) measure μ on $\mathbb{R}^d$

Chapter 1

Lasso and fairness in regression

In this chapter I summarize my contributions related to the problem of regression. Section 1.1 is focused on the linear regression under sparsity assumption and Section 1.2 reviews my recent contributions to the problem of regression under fairness constraint. Let us first introduce a general setup of regression. Consider a tuple $(Z, Y) \in \mathcal{Z} \times \mathbb{R}$ where $\mathcal{Z}$ is the features space and $\mathbb{R}$ is the space of outcomes. The goal is to study the link between the instance Z and the real-valued output Y, which is defined by the following statistical model

$$Y = f^*(Z) + \xi \ , \tag{1.1}$$

where $\xi \in \mathbb{R}$ is a noise such that $\mathbb{E}[\xi \mid Z] = 0$ almost surely and $f^* : \mathcal{Z} \to \mathbb{R}$ is the regression function, that is, $f^*(z) = \mathbb{E}[Y \mid Z = z]$ for all $z \in \mathcal{Z}$.

1.1 Theoretical study on Lasso prediction

Variable selection in *high dimension* is one of my main research areas that I started studying during my PhD thesis and that I still devote time to. The main concept in high-dimensional statistics is the *curse of dimensionality*, which essentially means that the number of parameters to be estimated is much higher than the number of available observations. If one wishes to obtain meaningful consistency results in high dimension, it is customary to impose additional structural assumptions on the problem at hand. Probably the simplest and the most popular statistical model where this phenomenon brings technical challenges is the sparse Gaussian linear regression model with deterministic design. The popularity of this model can be explained by a wide range of applications that can be tackled with the estimators developed for this scenario. That is, more formally, using the general regression setup in Eq. (1.1) we set $\mathcal{Z} = \mathbb{R}^d$, $Z = \boldsymbol{X}$ and assume that there exists $\boldsymbol{\beta}^* = (\beta_1^*, \dots, \beta_d^*)^\top \in \mathbb{R}^d$ such that $f^*(\mathbf{x}) = \mathbf{x}^\top \boldsymbol{\beta}^*$ for all $\mathbf{x} \in \mathbb{R}^d$. It is assumed that the statistician has collected data $(\boldsymbol{X}_i, Y_i)$ generated according to the relation $Y_i = \boldsymbol{X}_i^\top \boldsymbol{\beta}^* + \xi_i$, for $i = 1, \dots, n$, where $\xi_1, \dots, \xi_n$ are *i.i.d.* centered Gaussian noise random variables with

variance σ^{*2}. The feature vectors $\boldsymbol{X}_i$'s are deterministic and d-dimensional and only a few characteristics in those vectors affect the outputs Y_i's. This is reflected by a classical sparsity assumption saying that the support $J^* = \left\{ j \in \{1, \dots, d\} : \beta_j^* \neq 0 \right\}$ of $\boldsymbol{\beta}^*$ is small (*w.r.t.* n). In vector notation, the model reads as

$$\boldsymbol{y} = \mathbb{X}\boldsymbol{\beta}^* + \boldsymbol{\xi}, \qquad \boldsymbol{\xi} \sim \sigma^* \mathcal{N}_n(0, \mathbf{I}_n) \ , \tag{1.2}$$

where $\boldsymbol{y} := (y_1, \dots, y_n)^\top \in \mathbb{R}^n$ is the response vector, $\mathbb{X} := (\boldsymbol{x}^1, \dots, \boldsymbol{x}^d) = (\boldsymbol{X}_1, \dots, \boldsymbol{X}_n)^\top \in \mathbb{R}^{n \times d}$ is the design matrix, $\boldsymbol{\xi} := (\xi_1, \dots, \xi_n)^\top \in \mathbb{R}^n$ is the noise vector, and $\mathbf{I}_n$ denotes the identity matrix in $\mathbb{R}^{n \times n}$. Without loss of generality, we assume that the d covariates are such that $\|\boldsymbol{x}^j\|_2^2 \leq n$ for all $j \in \{1, \dots, d\}$.

Three main statistical questions can be asked for this sparse linear regression model. One can be interested in estimating $\mathbb{X}\boldsymbol{\beta}^*$, $\boldsymbol{\beta}^*$, or J^*. The estimation of the first quantity is called *prediction* problem, the second one is an *estimation* problem and the last one is a *support recovery* (variable or feature selection) problem. Numerous methods have then been developed to address all three statistical problems. In this part I will exclusively focus on the prediction risk of the Lasso estimator, which is defined for a given regularization parameter $\lambda > 0$ as any solution of the convex optimization problem

$$\hat{\boldsymbol{\beta}}_\lambda^{\text{Lasso}} \in \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^d} \left\{ \frac{1}{2n} \|\boldsymbol{y} - \mathbb{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \right\} \ . \tag{1.3}$$

The popularity of the Lasso estimator can be explained by the availability of computationally efficient solvers, which are able to scale to extremely large in terms of d and n datasets (Bach et al., 2012; Efron et al., 2004; Massias, Gramfort, and Salmon, 2018; Salmon, 2017).

The magnitude of the tuning parameter $\lambda > 0$ determines the amount of penalization and, therefore, has a crucial influence on the performance of the Lasso. The aim of my research is to study the statistical limits of the prediction performance of the Lasso estimator depending on the choice of $\lambda > 0$. Note that while in high dimension ($d > n$) Eq. (1.3) might have multiple solutions (Tibshirani, 2013) it holds that $\mathbb{X}\hat{\boldsymbol{\beta}}_\lambda = \mathbb{X}\hat{\boldsymbol{\beta}}'_\lambda$ for any two such solutions $\hat{\boldsymbol{\beta}}_\lambda$, $\hat{\boldsymbol{\beta}}'_\lambda$, that is, the prediction is unique.

In what follows, we investigate the magnitude of the (in-sample) prediction risk $\frac{1}{n}\|\mathbb{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2^2$. Several non-convex or computationally demanding methods (*e.g.*, MCP (Zhang, 2010), SCAD (Fan and Li, 2001), sparsity pattern aggregation (Rigollet and Tsybakov, 2012)) can reach the rate $\frac{s^*}{n}$ (commonly called the *fast* rate) up to log factors for prediction error *irrespectively* of the design matrix $\mathbb{X}$ (Bunea, Tsybakov, and Wegkamp, 2007a; Dalalyan and Tsybakov, 2007, 2012a; Rigollet and Tsybakov, 2011). When I started working on the subject, it was still unclear whether the Lasso can always achieve fast rates of convergence for this problem for any design matrix $\mathbb{X}$. State-of-the art results have already established such bounds for highly correlated designs as well as for weakly

correlated ones. The latter setting has received the most attention. The goal is to understand how far one can deviate from the orthogonal case and still obtain fast rates (see for instance Bickel, Ritov, and Tsybakov (2009), Bühlmann and van de Geer (2011), Candès and Plan (2009), Koltchinskii (2011), Sun and Zhang (2012), and Van de Geer and Lederer (2013)).

The statistical analysis of the Lasso is based on two main steps. The first step tries to eliminate randomness from the consideration and typically consists in providing a tight high probability bound on $\|\xi^\top \mathbb{X}\|_\infty$. Once the randomness is removed from the problem, the second step consists in relating the ℓ_1-norm of a sparse vector to the prediction error by means of linear algebra and, possibly, additional structural assumptions on the design matrix $\mathbb{X}$. My contributions and in particular the papers **[MH-Journal8]**-**[MH-Journal6]** have been devoted to gaining new insight into the prediction performance of the Lasso and to establishing sharper theoretical guarantees for Lasso prediction. In particular, we have provided answers to the following questions: **i)** can Lasso achieve fast rate of convergence irrespectively of the design matrix $\mathbb{X}$? **ii)** how can we fill the gap between the fast rate results of correlated and uncorrelated designs?

Fast rates: negative result. Let us specify a setting where Lasso fails to achieve fast rates independently from the choice of the tuning parameter $\lambda > 0$. Let $n \geq 2$ be an integer. We set $m = \lfloor \sqrt{2n} \rfloor$ and define the design matrix $\mathbb{X} \in \mathbb{R}^{n \times 2m}$ by

$$\mathbb{X} = \sqrt{\frac{n}{2}} \begin{pmatrix} \mathbf{1}_m^\top & \mathbf{1}_m^\top \\ \mathbf{I}_m & -\mathbf{I}_m \\ \mathbf{0}_{(n-m-1)\times m} & \mathbf{0}_{(n-m-1)\times m} \end{pmatrix} ,$$

where $\mathbf{1}_m$ stands for the vector of $\mathbb{R}^m$ having all coordinates equal to one and $\mathbf{0}_{(n-m-1)\times m}$ is the matrix of $\mathbb{R}^{(n-m-1)\times m}$ having all coordinates equal to zero. In this particular setting, the correlations between covariates $\mathbf{x}^j$'s are far from ± 1, they are fixed and equal to $1/2$ for most of the couples. We further assume that the noise vector is composed of *i.i.d.* Rademacher random variables, that is $\mathbf{P}(\xi = \mathbf{s}) = 2^{-n}$ for every $\mathbf{s} \in \{\pm 1\}^n$ (thus $\sigma^* = 1$). Let the true regression vector be $\boldsymbol{\beta}^* \in \mathbb{R}^{2m}$ such that $\beta_1^* = \beta_{m+1}^* = 1$ and $\beta_j^* = 0$ for every $j \in [2m] \setminus \{1, m+1\}$, meaning that the sparsity level is 2.

Proposition 1.1. *For any $\lambda > 0$, the prediction loss of the Lasso $\hat{\boldsymbol{\beta}}_\lambda^{\text{Lasso}}$ satisfies the inequality*

$$\mathbf{P}\left(\frac{1}{n} \|\mathbb{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2^2 \geq \frac{1}{2\sqrt{2n}} \right) \geq \frac{1}{2} .$$

This example shows that the prediction error of the Lasso is in some cases at best of the order of $n^{-1/2}$, whatever the tuning parameter is. In the literature, other examples on which the Lasso fails to achieve fast rates have been proposed (see Section 2 in Candès and Plan, 2009). However, to the best of our knowledge, this is the first counterexample

in which such a result is analytically proved for fixed sparsity, fixed correlations, any value of λ and a $\boldsymbol{\beta}^*$ independent of n. This example also clearly demonstrates the limits of the Lasso as a method of prediction. While for several other prediction procedures fast rates are valid without any condition on the correlations between the predictors (Bunea, Tsybakov, and Wegkamp, 2007b; Dalalyan and Tsybakov, 2007, 2012a,b; Raskutti, Wainwright, and Yu, 2011; Rigollet and Tsybakov, 2011), some relatively strong assumptions are necessary for the Lasso to achieve fast rates. It should be noted in defense of the Lasso that it presents major advantages in terms of computational complexity.

"Slow" rates meet "fast" rates. Based on the above discussion it is relevant to consider what kind of prediction performance we can hope for in the worst case scenario. Typically, for this purpose, we investigate so-called "slow" rates where the remaining term in the prediction bound contains the tuning parameters to the power one and is then classically of order $n^{-1/2}$ (see for instance (Rigollet and Tsybakov, 2011; Sun and Zhang, 2012)). For this part, we need additional notation. For the design matrix $\mathbb{X}$ and any subset T of $[d]$, we denote by V_T the linear subspace of $\mathbb{R}^n$ spanned by the columns of $\mathbb{X}_T$. A key quantity in the refinement we propose is

$$\rho_T := n^{-1/2} \max_{j \in [d]} \|(\mathbf{I}_n - \Pi_T)\boldsymbol{x}^j\|_2 \ , \tag{1.4}$$

the maximal Euclidean distance between the normalized columns of $\mathbb{X}$ and the set V_T, where Π_T is the orthogonal projector onto V_T. The quantity ρ_T is a geometric description of the correlation between covariates and the more correlated they are the smaller ρ_T is. In particular, for design perfectly collinear, that is when all covariates belong to the linear space spanned by the covariates in $\mathbb{X}_T$, meaning $\{\boldsymbol{x}^j : j \in [d]\} \subset \mathrm{Span}\{\boldsymbol{x}^j : j \in T\}$, the term ρ_T is null.

Theorem 1.1. *Let $T \subset [d]$ be a set of indices and let $\delta > 0$ be a positive constant. If the tuning parameter $\lambda \geq \rho_T \sigma^* \sqrt{\frac{2\log(d/\delta)}{n}}$, the Lasso (1.3) fulfills*

$$\frac{1}{n}\|\mathbb{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2^2 \leq \inf_{\bar{\boldsymbol{\beta}} \in \mathbb{R}^d} \left\{ \frac{1}{n}\|\mathbb{X}(\bar{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2^2 + 4\lambda\|\bar{\boldsymbol{\beta}}\|_1 \right\} + \frac{2\sigma^{*2}(|T| + 2\log(1/\delta))}{n} \ , \tag{1.5}$$

with probability at least $1 - 2\delta$.

This result is a sharp oracle inequality that holds for any the design matrix $\mathbb{X}$. Its specificity is the dependency of the tuning parameter λ on the factor ρ_T that can make the rate much faster than $n^{-1/2}$. In particular, it turns out that, in contrast to what the nomenclature suggests, the above slow rate bound entails fast rates if the correlations are properly incorporated into the tuning parameter (so that ρ_T is smaller than $n^{-1/2}$). The above bound has the same flavor as existing results in the literature such as in Van de

Geer and Lederer (2013). Indeed both improve the $n^{-1/2}$ classical slow rate by bringing into play a tuning parameter that is much smaller than the universal one which is of order $\sqrt{2\log(d)/n}$. On the other hand, the quantity ρ_T governing the choice of λ and the rate of convergence of the prediction risk is, in general, easier to compute than the entropy that appears in earlier results.

The result stated in Eq. (1.5) is general and is appealing since the quantity ρ_T may be easily computed in some applications. An important case is the Least-Squares estimator with total variation penalty (*TV-estimator*), where the design matrix $\mathbb{X}$ and the set T are completely predetermined. In particular, considering monotone or Hölder continuous signals, we can specify Eq. (1.5) and show that the TV-estimator prediction bound has a remaining term which is almost minimax. In the case of Hölder continuous signals, this result improves on results by Mammen and van de Geer (1997) where a suboptimal rate for the TV-estimator is derived.

Remark 1.1 (Effective number of parameters)**.** *The bound* (1.5) *can be further refined by replacing the number of parameters d with an effective number of parameters as described in* **[MH-Journal6]**, *a work in collaboration with J. Lederer. This effective number of parameters can be considerably smaller than d if the correlations are high thus reducing, thanks to geometrical arguments, the bound by a factor up to $\sqrt{\log(d)}$.*

Fast rates under weighted compatibility condition. According to the value of ρ_T, the bound (1.5) leads to any possible rate between the classical slow and the fast rates of order $n^{-1/2}$ and n^{-1} respectively. In particular, this bound is not well tailored for weakly correlated designs. Here, we provide fast rates type bounds in which the remaining term involved the tuning parameter to the power two. The main result of this part holds under additional condition on the design matrix $\mathbb{X}$ similar, in the flavor, to conditions in previous contributions (Bickel, Ritov, and Tsybakov, 2009; Bunea, Tsybakov, and Wegkamp, 2007b; Cai, Wang, and Xu, 2010; Juditsky and Nemirovski, 2011; Sun and Zhang, 2012; Van de Geer and Bühlmann, 2009; Wainwright, 2009; Zhang, 2009). However, the condition introduced here compares favorably to previously considered assumptions.

Fast rate bounds for Lasso prediction usually rely on assumptions on the correlations between covariates such as low coherence (Candès and Plan, 2009), restricted eigenvalues (Bickel, Ritov, and Tsybakov, 2009; Raskutti, Wainwright, and Yu, 2010), restricted isometry (Candes and Tao, 2007), compatibility (Van de Geer, 2007; Van de Geer and Bühlmann, 2009), cone invertibilty (Ye and Zhang, 2010), *etc.* Influenced by this literature, we introduce a new measure of the correlation between covariates. We need an additional notation for the purpose of introducing it. For any subset T of $[d]$, we define

$$\omega_j(T,\mathbb{X}) = \frac{1}{\sqrt{n}} \|(\mathbf{I}_n - \Pi_T)\boldsymbol{x}^j\|_2, \qquad \forall j \in [d] \ . \tag{1.6}$$

Since $\boldsymbol{x}^j$ are normalized to have an ℓ_2-norm at most equal to $\sqrt{n}$, the weights $\omega_j(T, \mathbb{X})$ are all between zero and one. In particular, $\omega_j(T, \mathbb{X}) = 0$ for every $j \in T$. We further define for any $\gamma > 0$ the set

$$\mathcal{C}_0(T, \gamma, \boldsymbol{\omega}) = \left\{ \boldsymbol{\delta} \in \mathbb{R}^d : \|(\mathbf{1}_d - \gamma^{-1}\boldsymbol{\omega})_{T^c} \odot \boldsymbol{\delta}_{T^c}\|_1 < \|\boldsymbol{\delta}_T\|_1 \right\} .$$

The central quantity in this part is the *weighted compatibility factor* defined for every vector $\boldsymbol{\omega} \in \mathbb{R}^d$ with non-negative entries by

$$\bar{\kappa}_{T,\gamma,\boldsymbol{\omega}} = \inf_{\boldsymbol{\delta} \in \mathcal{C}_0(T,\gamma,\boldsymbol{\omega})} \frac{|T| \cdot \|\mathbb{X}\boldsymbol{\delta}\|_2^2}{n\left\{\|\boldsymbol{\delta}_T\|_1 - \|(\mathbf{1}_d - \gamma^{-1}\boldsymbol{\omega})_{T^c} \odot \boldsymbol{\delta}_{T^c}\|_1\right\}^2} .$$

The weighted compatibility factor allows to relate the ℓ_1-norm of the vectors from the cone $\mathcal{C}_0(T, \gamma, \boldsymbol{\omega})$ to the prediction error. The weighted compatibility assumption requires the constant $\bar{\kappa}_{T,\gamma,\boldsymbol{\omega}}$ to be strictly positive. Then, in order to leverage this condition one first demonstrates that $\boldsymbol{\delta} = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \in C_0(T, \gamma, \boldsymbol{\omega})$ and then applies the following obvious relation on $\bar{\kappa}_{T,\gamma,\boldsymbol{\omega}}$ and the prediction error $\|\mathbb{X}\boldsymbol{\delta}\|_2$

$$\|\boldsymbol{\delta}_T\|_1 - \|(\mathbf{1}_d - \gamma^{-1}\boldsymbol{\omega})_{T^c} \odot \boldsymbol{\delta}_{T^c}\|_1 \leq \sqrt{\frac{|T|}{n\bar{\kappa}_{T,\gamma,\boldsymbol{\omega}}}} \cdot \|\mathbb{X}\boldsymbol{\delta}\|_2 .$$

This is, up to my knowledge, the most relaxed assumption in the literature that leads to fast rates. We can derive the following result.

Theorem 1.2. *Let $\delta \in (0,1)$ be a fixed tolerance level. If for some value $\gamma > 1$, the tuning parameter of the Lasso satisfies $\lambda = \gamma\sigma^*\sqrt{2\log(d/\delta)/n}$, then with probability at least $1 - 2\delta$, the following bound holds*

$$\frac{1}{n}\|\mathbb{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2^2 \leq \inf_{\bar{\boldsymbol{\beta}} \in \mathbb{R}^d, T \subset [p]} \left\{ \frac{1}{n}\|\mathbb{X}(\bar{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2^2 + 4\lambda\|\bar{\boldsymbol{\beta}}_{T^c}\|_1 + \frac{4\sigma^{*2}|T|\log(d/\delta)}{n} \cdot r_{n,d,T} \right\} , \tag{1.7}$$

where the remainder term is given by $r_{n,d,T} = \log^{-1}(d/\delta) + 2|T|^{-1} + \gamma^2\bar{\kappa}^{-1}_{T,\gamma,\boldsymbol{\omega}}$.

The main difference between the Oracle inequality of Theorem 1.1 and the one presented above in Theorem 1.2 is that the former contains $\lambda\|\bar{\boldsymbol{\beta}}\|_1$, while the latter involves the same ℓ_1-norm of $\bar{\boldsymbol{\beta}}$ which is crucially restricted on T^c. As an immediate conclusion of Theorem 1.2 we can plug-in $\bar{\boldsymbol{\beta}} = \boldsymbol{\beta}^*$ and recall that $\boldsymbol{\beta}^*_{T^c} = \mathbf{0}$ to derive the fast rate of convergence of order $|T|\log(d)/n$. However, the price for such a strong guarantee is the additional condition on the design matrix $\mathbb{X}$, namely, assuming that $\mathbb{X}$ is such that $\kappa_{T,\gamma,\boldsymbol{\omega}} > 0$. Unlike previous conditions on the design which were mainly focused on the weakly correlated covariates, an important feature of the weighted compatibility factor is that it can be non-zero for even highly correlated designs. This is the case, for instance, in

the scenario of the TV-estimator, yielding a completely new bound on the prediction performance of this estimator for piece-wise constant functions. With the universal choice of the tuning parameter, we show that the TV-estimator is nearly minimax optimal in the class of piece-wise constant functions.

Conclusion on Lasso. We have demonstrated some limits of the Lasso as a method of prediction. On the other hand, our improved bound on the prediction error have nice implications in particular to show minimax rates for the TV-estimator. A further advantage of the Lasso is its numerical efficiency. This is even more emphasized when it is combined with a refitting step. In the Master thesis of E. Chzhen that I was supervising with J. Salmon, we studied both theoretically and numerically several refitting strategies for the Lasso **[MH-Journal10]**.

1.2 Regression of Demographic Parity

The second problem I have been considering in the regression setting deals with learning a real-valued function which meets fairness criteria. A major goal in modern applications is to mitigate bias that is present in the underlying data distribution (Barocas, Hardt, and Narayanan, 2019) in order to meet ethical criteria of the modern society. However, the classical prediction methods of statistics and machine learning are oblivious to such principles partly due to the lack of the mathematical formalism that would allow to address such an issue. In recent years, various authors and communities have been proposing different formal definition of the notions of fairness, equality, and justice. An attractive formalism relies on the idea that a prediction must not discriminate based on some characteristics of an instance (sensitive feature) such as the gender or the ethnicity. We refer the interested reader to (Barocas, Hardt, and Narayanan, 2019; Mehrabi et al., 2019) for a general introduction to the problem of fair prediction and to (Del Barrio, Gordaliza, and Loubes, 2020; Oneto and Chiappa, 2020) for a review of the most recent theoretical advances. In collaboration with E. Chzhen, C. Denis, L. Oneto, and M. Pontil **[MH-Conf4]**-**[MH-Conf5]**, I investigated the problem of meeting the *group fairness* criteria in regression, where the goal is to build prediction rules that are *statistically independent* from the sensitive feature. The group fairness is not to be confused with the *individual fairness* that aims at producing similar treatments to similar individuals, and results in a different type of constraints on the prediction.

Several notions of *group fairness* have been introduced in the literature, including but not limited to *Equalized-Odds*, *Equality of Opportunity*, *Disparate Impact*, or *Demographic Parity*. We refer for instance to (Barocas, Hardt, and Narayanan, 2019; Hardt, Price, and Srebro, 2016) for an overview of these notions. Most of them are well and clearly defined in the classification setting but their translations to the regression setting is still a subject

of discussion. Probably the clearest notion of fairness in regression is the *Demographic Parity* (Agarwal, Dudik, and Wu, 2019; Chiappa et al., 2020; Jiang et al., 2019). More precisely, consider again the model (1.1). In the context of group fairness the input feature $Z = (\boldsymbol{X}, S) \in \mathbb{R}^d \times \mathcal{S}$, where $\boldsymbol{X}$ corresponds to the vector of covariates and S is the sensitive feature (*e.g.*, gender, ethnicity, qualification, birth place, etc.). In order to simplify the presentation, we restrict ourselves to the case of two groups $\mathcal{S} = \{\pm 1\}$, keeping in mind that the results in **[MH-Conf5]** are much more general and can be applied to the case of continuous sensitive features.

Definition 1.1 (Demographic Parity)**.** *A prediction (possibly randomized)* $g : \mathbb{R}^d \times \mathcal{S} \to \mathbb{R}$ *is* fair *if*

$$\sup_{t \in \mathbb{R}} \Big| \mathbf{P}(g(\boldsymbol{X}, S) \leq t \mid S = +1) - \mathbf{P}(g(\boldsymbol{X}, S) \leq t \mid S = -1) \Big| = 0 \ .$$

Demographic Parity means that $(g(\boldsymbol{X}, S) \mid S{=}{+}1) \stackrel{d}{=} (g(\boldsymbol{X}, S) \mid S{=}{-}1)$, that is, there is statistical independence of the prediction from the sensitive attribute – an intuitive criteria of fairness. Due to the complicated nature of the Demographic Parity constraint, several contributions do not actually study it directly. Instead, they relax the requirement of the equality of distributions by enforcing low correlations or equality of low-order moments (Berk et al., 2017; Calders et al., 2013; Donini et al., 2018; Fitzsimons et al., 2019; Pérez-Suay et al., 2017). Unlike them, the goal in **[MH-Conf4]**-**[MH-Conf5]** is to address the problem of learning a real-valued function which satisfies the exact Demographic Parity constraint in Definition 1.1.

Let us emphasize two trivial, but important facts: *i)* if a prediction g is fair for a distribution $\mathbb{P}$, it might be unfair for other distributions – fairness is a problem-dependent notion; *ii)* for any $c \in \mathbb{R}$ and any distribution $\mathbb{P}$ a prediction $g \equiv c$ is fair – constant predictions do not discriminate. Fact *i)* and *ii)* highlight the intrinsic difficulty of building *accurate* and *fair* predictions.

Let us add some more notation. For any univariate probability measure μ, we denote by F_μ its Cumulative Distribution Function (CDF) and by $Q_\mu : [0,1] \to \mathbb{R}$ its quantile function defined for all $t \in (0,1]$ as $Q_\mu(t) = \inf\{y \in \mathbb{R} \,:\, F_\mu(y) \geq t\}$ with $Q_\mu(0) = Q_\mu(0+)$ (Van der Vaart, 2000). For any prediction rule $g : \mathbb{R}^d \times \mathcal{S} \to \mathbb{R}$, we denote by $\nu_{g|s}$ the distribution of $g(\boldsymbol{X}, S)|S = s$, that is, the CDF of $\nu_{g|s}$ is given by

$$F_{\nu_{g|s}}(t) = \mathbb{P}(g(\boldsymbol{X}, S) \leq t \mid S = s) \ . \tag{1.8}$$

To shorten the notation we will write $F_{g|s}$ and $Q_{g|s}$ instead of $F_{\nu_{g|s}}$ and $Q_{\nu_{g|s}}$ respectively. Another interpretation of Demographic Parity is that the Kolmogorov-Smirnov distance between $\nu_{g|+1}$ and $\nu_{g|-1}$ equals zero. Consequently, if g is fair, $\nu_{g|s}$ does not depend on s and to simplify the notation we will write ν_g. One can also note that Definition 1.1 could be stated with any other metric in distribution space. The choice of Kolmogorov-Smirnov

distance is merely a way to quantify finite sample deviations from the equality required by the Demographic Parity.

Demographic Parity can be achieved by infinitely many predictors and, in particular, by the more trivial ones like constant predictions mentioned above. Our goal here is to find a prediction that minimizes the squared risk while being fair in the sense of Demographic Parity:

$$\min_{g \text{ is fair}} \mathbb{E}(Y - g(\boldsymbol{X}, S))^2 \ .$$

In my recent contributions I considered this problem in two papers. In the first one **[MH-Conf4]**, we derive the Lagrangian version of the problem and solve it by exploiting min-max and duality arguments. The resulting optimal rule is based on thresholding of the regression function $f^*(\boldsymbol{X}, S) = \mathbb{E}[Y \mid \boldsymbol{X}, S]$. Exploiting the form of the optimal predictor we build a plug-in estimator for which we derive finite sample results on the risk and on the fairness violation. I do not develop this work here and partly expand on the second work **[MH-Conf5]**. The main objective in **[MH-Conf5]** is to build a method that is **i)** interpretable and explainable, **ii)** computationally efficient, **iii)** comes with strong theoretical guarantees. Previous contributions on regression of Demographic Parity inevitably lose at least one, and more often two, of the aforementioned criteria. Notable exceptions where **ii)** and **iii)** are achieved up to some extent are **[MH-Conf4]** (Agarwal et al., 2018; Oneto, Donini, and Pontil, 2019). To the best of my knowledge **i) + ii) + iii)** are satisfied only by the method in **[MH-Conf5]**.

Our analysis builds a bridge between fair regression and optimal transport theory. Let us first recall the notion of Wasserstein-2 distance on real line, which is a central object of our analysis.

Definition 1.2. *Let μ and ν be two univariate probability measures. The squared Wasserstein-2 distance between μ and ν is defined as*

$$\mathcal{W}_2^2(\mu, \nu) = \inf_{\gamma \in \Gamma_{\mu,\nu}} \int |x - y|^2 \, d\gamma(x, y) \ ,$$

where $\Gamma_{\mu,\nu}$ is the set of distributions (couplings) on $\mathbb{R} \times \mathbb{R}$ such that for all $\gamma \in \Gamma_{\mu,\nu}$ and all measurable sets $A, B \subset \mathbb{R}$ it holds that $\gamma(A \times \mathbb{R}) = \mu(A)$ and $\gamma(\mathbb{R} \times B) = \nu(B)$.

Theorem 1.3 (Characterization of fair optimal prediction)**.** *Assume, for each $s \in \mathcal{S}$, that the univariate measure $\nu_{f^*|s}$ has a density and let $p_s = \mathbb{P}(S = s)$. Then*[1]*,*

$$\min_{g \text{ is fair}} \mathbb{E}(f^*(\boldsymbol{X}, S) - g(\boldsymbol{X}, S))^2 = \min_{\nu} \sum_{s \in \mathcal{S}} p_s \mathcal{W}_2^2(\nu_{f^*|s}, \nu) \ .$$

[1]Since the noise has zero mean, the minimization of $\mathbb{E}(Y - g(\boldsymbol{X}, S))^2$ over g is equivalent to the minimization of $\mathbb{E}(f^*(\boldsymbol{X}, S) - g(\boldsymbol{X}, S))^2$ over g.

Moreover, if g^* and ν^* solve the l.h.s. and the r.h.s. problems respectively, then $\nu^* = \nu_{g^*}$ and

$$g^*(\boldsymbol{x}, s) = p_s f^*(\boldsymbol{x}, s) + (1 - p_s) t^*(\boldsymbol{x}, s) \ , \tag{1.9}$$

where $t^*(\boldsymbol{x}, s) = Q_{f^*|s'} \circ F_{f^*|s}\left(f^*(\boldsymbol{x}, s)\right) = \inf\left\{t \in \mathbb{R} \ : \ F_{f^*|s'}(t) \geq F_{f^*|s}(f^*(\boldsymbol{x}, s))\right\}$ with $s' \neq s$.

This result relies on the classical characterization of optimal coupling in one dimension of the Wasserstein-2 distance (Agueh and Carlier, 2011; Santambrogio, 2015; Villani, 2003). It translates the problem (1.2) into a Wasserstein barycenter problem. Importantly, it provides a closed form expression of the minimizer of the squared risk under the Demographic Parity constraint. We show that a minimizer g^* of the $\mathbb{L}_2$-risk can be used to construct ν^* and vice-versa. This is mainly due to the choice of the $\mathbb{L}_2$-risk which is natural even though arbitrary in the regression setting. However, considering the $\mathbb{L}_2$-risk offers a nice translation from the geometry induced by the $\mathbb{L}_2$ distance in the space of functions into the Wasserstein-2 distance in the space of push-forward measures. We use the above statement to articulate our contributions.

Interpretability. Figure 1.1 summarizes the above result. It has two levels of understanding – population and individual levels. First, the distribution of the optimal fair predictor ν^* is a weighted barycenter of the regression functions $\nu_{f^*|s}$. These weights are determined by the population proportions p_s's. Second, the explicit expression of the optimal fair predictor in Eq. (1.9) gives an interesting interpretation at individuals level. Note that the quantity $t^*(\boldsymbol{x}, s)$ in Eq. (1.9) is determined so that the *ranking* of $f^*(\boldsymbol{x}, s)$ relative to the distribution of $\boldsymbol{X}|S = s$ for group s (*e.g.*, minority) is the same as the ranking of $t^*(\boldsymbol{x}, s)$ relative to the distribution of the group $s' \neq s$ (*e.g.*, majority). In other words, as illustrated in Figure 1.1, the optimal fair predictor $g^*(\boldsymbol{x}, s)$ we assign to an instance $\boldsymbol{x}$ from group s is a convex combination of $f^*(\boldsymbol{x}, s)$, the (unfair) prediction that would have been allocated to that instance in its group, and of $t^*(\boldsymbol{x}, s)$, the (also unfair) prediction that would have received a similar instance, meaning at the same rank, in the other group.

Recently, in the context of causality Plečko and Meinshausen (2019) proposed a practical method, which enforces the Demographic Parity and resembles the strategy of the optimal fair prediction g^* derived in Theorem 1.3. At the time of writing **[MH-Conf5]** we were not aware of their contribution and it should be noted that Plečko and Meinshausen (2019) did not connect their strategy to the optimality in terms of the risk measure. They provide the following interesting interpretation of such a strategy: "*If you are a female better than 60% of the females in the dataset, we assume you would be better than 60% of males had you been male.*" This is exactly the strategy taken by the optimal fair predictor g^*, with the crucial difference that it operates on the level of population instead of the sample.

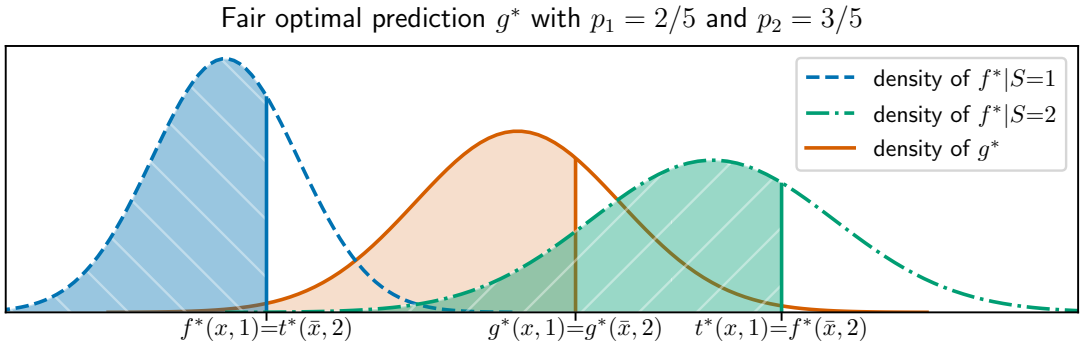

Figure 1.1: For a new point $(\boldsymbol{x},1)$, the value $t^*(\boldsymbol{x},1)$ is chosen such that the shaded `Green Area (//)` $= \mathbb{P}(f^*(\boldsymbol{X},S) \leq t^*(\boldsymbol{x},1)|S=2)$ equals to the shaded `Blue Area (\\)` $= \mathbb{P}(f^*(\boldsymbol{X},S) \leq f^*(\boldsymbol{x},1)|S=1)$. The final prediction $g^*(\boldsymbol{x},1)$ is a convex combination of $f^*(\boldsymbol{x},1)$ and $t^*(\boldsymbol{x},1)$. The same is done for $(\bar{\boldsymbol{x}},2)$.

Data-driven procedure. The above expression of the optimal fair predictor given by Eq. (1.9) suggests a simple post-processing estimation procedure based on the plug-in principle. Having a look at g^*, it turns out that we only need to estimate the regression function f^*, the proportions p_s, as well as the CDF $F_{f^*|s}$ and the quantile function $Q_{f^*|s}$, for $s \in \mathcal{S}$. By doing so, we notice that the estimation procedure is semi-supervised (Vapnik, 1998) since all these quantities, with the exception of f^*, can be estimated using *only unlabeled* data. This is one of the main advantages of our post-processing method since it efficiently makes fair any pre-trained estimator of the regression function at the cost of *only unlabeled* data. This is particularly appealing in applications where the cost of training is large or when one has in hand already a performing estimator of the regression function.

Let us describe an efficient way to collect data for our purpose. For each $s \in \{\pm 1\}$ let $\mathcal{U}^s = \{\boldsymbol{X}_i^s\}_{i=1}^{N_s} \overset{\text{i.i.d.}}{\sim} \mathbb{P}_{\boldsymbol{X}|S=s}$ be a group-wise unlabeled sample. In the following for simplicity we assume that N_{+1} and N_{-1} are *even*.[2]. Let $\mathcal{I}_0^s, \mathcal{I}_1^s \subset [N_s]$ be any fixed partition of $[N_s]$ such that $|\mathcal{I}_0^s| = |\mathcal{I}_1^s| = N_s/2$ and $\mathcal{I}_0^s \cup \mathcal{I}_1^s = [N_s]$. For each $j \in \{0,1\}$ we let $\mathcal{U}_j^s = \left\{\boldsymbol{X}_i^s \in \mathcal{U}^s \,:\, i \in \mathcal{I}_j^s\right\}$ be the restriction of $\mathcal{U}^s$ to $\mathcal{I}_j^s$. We use $\mathcal{U}_0^s$ to estimate $Q_{f^*|s}$ and $\mathcal{U}_1^s$ to estimate $F_{f^*|s}$. For each $f : \mathbb{R}^d \times \mathcal{S} \to \mathbb{R}$ and each $s \in \{\pm 1\}$, we estimate $\nu_{f|s}$ by

$$\hat{\nu}_{f|s}^0 = \frac{1}{|\mathcal{I}_0^s|} \sum_{i \in \mathcal{I}_0^s} \delta\big(f(\boldsymbol{X}_i^s, s) + \zeta_{is} - \cdot\big) \quad \text{and} \quad \hat{\nu}_{f|s}^1 = \frac{1}{|\mathcal{I}_1^s|} \sum_{i \in \mathcal{I}_1^s} \delta\big(f(\boldsymbol{X}_i^s, s) + \zeta_{is} - \cdot\big) \;, \quad (1.10)$$

where δ is the Dirac measure and all ζ_{is} are *i.i.d.* uniform random variables in $[-\sigma, \sigma]$, for some positive σ set[3] by the user. Using the estimators in Eq. (1.10), we define for all

[2]Since we are willing to sacrifice a factor 2 in our bounds, this assumption is without loss of generality.

[3]In practice one should use a very small value for σ (*e.g.*, $\sigma{=}10^{-5}$), which does not alter the statistical

$f : \mathbb{R}^d \times \mathcal{S} \to \mathbb{R}$ estimators of $Q_{f|s}$ and of $F_{f|s}$ as

$$\hat{Q}_{f|s} \equiv Q_{\hat{v}^0_{f|s}} \quad \text{and} \quad \hat{F}_{f|s} \equiv F_{\hat{v}^1_{f|s}} \ . \tag{1.11}$$

That is, $\hat{F}_{f|s}$ and $\hat{Q}_{f|s}$ are the empirical CDF and empirical quantiles of $(f(\boldsymbol{X}, S)+\zeta)|S=s$ based on $\{f(\boldsymbol{X}_i^s, s)+\zeta_{is}\}_{i\in\mathcal{I}_1^s}$ and $\{f(\boldsymbol{X}_i^s, s)+\zeta_{is}\}_{i\in\mathcal{I}_0^s}$ respectively. The noise ζ_{is} serves as a smoothing[4] random variable, since for all $s \in \{\pm 1\}$ and $i \in [N_s]$ the random variables $f(\boldsymbol{X}_i^s, s)+\zeta_{is}$ are *i.i.d.* continuous for any $\mathbb{P}$ and f. In contrast, $f(\boldsymbol{X}_i^s, s)$ might have atoms resulting in a non-zero probability to observe ties in $\{f(\boldsymbol{X}_i^s, s)\}_{i\in\mathcal{I}_j^s}$. This step plays a crucial role in the distribution-free fairness guarantees that we derive in Proposition 1.4.

Finally, let $\mathcal{A}=\{S_i\}_{i=1}^N \overset{\text{i.i.d.}}{\sim} \mathbb{P}_S$ and for each $s \in \mathcal{S}$ let $\hat{p}_s$ be the empirical frequency of $S=s$ evaluated on $\mathcal{A}$. Given a base estimator $\hat{f}$ of f^* constructed from n labeled samples $\mathcal{L}=\{(\boldsymbol{X}_i, S_i, Y_i)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathbb{P}$, we define the final estimator $\hat{g}$ of g^* for all $(\boldsymbol{x}, s) \in \mathbb{R}^d \times \mathcal{S}$ mimicking Eq. (1.9) as

$$\hat{g}(\boldsymbol{x}, s) = \hat{p}_s(\hat{f}(\boldsymbol{x}, s) + \zeta) + (1 - \hat{p}_s)\hat{t}(\boldsymbol{x}, s) \ , \tag{1.12}$$

where $\hat{t}(\boldsymbol{x}, s) = \hat{Q}_{\hat{f}|s'} \circ \hat{F}_{\hat{f}|s}\left(\hat{f}(\boldsymbol{x}, s) + \zeta\right)$ and ζ, uniform on $[-\sigma, \sigma]$, is assumed to be independent from every other random variables.

The numerical performance of the present method has been evaluated on various benchmark datasets with different estimators $\hat{f}$ of the regression function f^* such as (*Linear* and *non Linear*) *Regularized Least-Squares* and *Random Forests*. The experiments indicate that our methods are often superior to or competitive with the state-of-the-art.

Statistical analysis. The first result we state is a *distribution-free* finite sample fairness guarantee for post-processing of *any* base learner with unlabeled data.

Theorem 1.4 (Distribution free fairness guarantees)**.** *For any joint distribution $\mathbb{P}$ of $(\boldsymbol{X}, S, Y)$ and any base estimator $\hat{f}$ constructed on labeled data, the estimator $\hat{g}$ defined in Eq. (1.12) satisfies*

$$\sup_{t\in\mathbb{R}} |\mathbf{P}(\hat{g}(\boldsymbol{X}, S) \le t|S=1) - \mathbf{P}(\hat{g}(\boldsymbol{X}, S) \le t|S=-1)| \le 2\,(N_1 \wedge N_{-1}+2)^{-1}\,\mathbf{1}_{\{N_1 \neq N_{-1}\}} \ , \tag{1.13}$$

$$\mathbf{E}\sup_{t\in\mathbb{R}} |\mathbf{P}(\hat{g}(\boldsymbol{X}, S) \le t|S=1, \mathcal{D}) - \mathbf{P}(\hat{g}(\boldsymbol{X}, S) \le t|S=-1, \mathcal{D})| \le 6\,(N_1 \wedge N_{-1}+1)^{-1/2} \ , \tag{1.14}$$

where $\mathcal{D} = \mathcal{L} \cup \mathcal{A} \cup_{s\in\{\pm 1\}} \mathcal{U}^s$ is the union of all available datasets.

quality of the base estimator $\hat{f}$

[4]The requirement here is that the ζ_{is}'s i) are continuous ii) do not deviate far from zero. Gaussian noise with small variance can also be used though it requires some adaptation to state our finite sample results on the risk.

The first part of Proposition 1.4 shares similarities with contributions on prediction sets (Lei, Robins, and Wasserman, 2013; Lei and Wasserman, 2014) and conformal prediction literature (Vovk, Gammerman, and Shafer, 2005; Zeni, Fontana, and Vantini, 2020) as they also rely on results on rank statistics. Let us point the main difference between the above two results. In Eq. (1.13), we allow to take the expectation over the data inside the supremum. This induces a sharper bound. However, the bound in Eq. (1.14) might be more appealing to the machine learning community as it controls the expected (over data) violation of the fairness constraint with standard parametric rate.

The proof of (1.13) exploits tools from ranks statistics whose use dates back to randomization inference via permutations (Fisher, 1936; Hoeffding, 1952). The main ingredient is the following observation: since we added the randomization, the random variables $\{\hat{f}(\boldsymbol{X}^s, s) + \zeta\} \cup \{\hat{f}(\boldsymbol{X}_i^s, s) + \zeta_{is}\}_{i \in \mathcal{I}_1^s}$ conditionally on labeled data $\mathcal{L}$ are *i.i.d.* and continuous. Then we can show that $\sum_{i \in \mathcal{I}_1^s} \mathbf{1}_{\{\hat{f}(\boldsymbol{X}_i^s, s) + \zeta_{is} \leq \hat{f}(\boldsymbol{X}^s, s) + \zeta_s\}}$, which is involved in an upper bound of Eq. (1.13), is distributed uniformly on $\{0, \dots, |\mathcal{I}_1^s|\}$ (see *e.g.,* (Van der Vaart, 2000, Lemma 13.1)). In order to leverage this result one needs to connect the events $\{\hat{g}(\boldsymbol{X}, S) \leq t | S = 1\}$ and $\{\hat{g}(\boldsymbol{X}, S) \leq t | S = -1\}$ with the events $\{\sum_{i \in \mathcal{I}_1^{+1}} \mathbf{1}_{\{\hat{f}(\boldsymbol{X}_i^{+1}, +1) + \zeta_{i(+1)} \leq \hat{f}(\boldsymbol{X}^{+1}, +1) + \zeta_{+1}\}} \leq t'\}$ and $\{\sum_{i \in \mathcal{I}_1^{-1}} \mathbf{1}_{\{\hat{f}(\boldsymbol{X}_i^{-1}, -1) + \zeta_{i(-1)} \leq \hat{f}(\boldsymbol{X}^{-1}, -1) + \zeta_s\}} \leq t'\}$ for an appropriately chosen t' respectively. The latter is achieved by exploiting the structure of the estimator $\hat{g}$ and in particular the fact that it can be written as a non-linear monotone transformation of $\hat{f}$. The proof of Eq. (1.14) considers classical inverse transform (see *e.g.,* (Van der Vaart, 2000, Sections 13 and 21)) combined with the Dvoretzky-Kiefer-Wolfowitz inequality (Massart, 1990).

As already mentioned, simply achieving fairness can be accomplished by constant predictions, which are generally useless in practice. Let us now derive non-asymptotic risk bounds to complement our fairness guarantees. Unfortunately, finite sample risk guarantees are not achievable without any additional conditions. To this end, we require the following assumption on the distribution $\mathbb{P}$ of $(\boldsymbol{X}, S, Y) \in \mathbb{R}^d \times \mathcal{S} \times \mathbb{R}$.

Assumption 1. *For each $s \in \mathcal{S}$ the univariate measure $\nu_{f^*|s}$ admits a density q_s, which is lower bounded by $\underline{\lambda}_s > 0$ and upper-bounded by $\overline{\lambda}_s \geq \underline{\lambda}_s$.*

Under this assumption we can prove the following finite-sample plug-and-play estimation bound.

Theorem 1.5 (Estimation guarantee)**.** *Let Assumption 1 be satisfied, and set $\sigma \lesssim N_1^{-1/2}$, then the estimator $\hat{g}$ defined in Eq. (1.12) satisfies*

$$\mathbf{E}\|g^* - \hat{g}\|_{\mathbb{L}_1(\mathbb{P}_{\boldsymbol{X},S})} \lesssim \mathbf{E}\|f^* - \hat{f}\|_{\mathbb{L}_1(\mathbb{P}_{\boldsymbol{X},S})} \bigvee \left(\sum_{s \in \{\pm 1\}} p_s N_s^{-1/2} \right) \bigvee \sqrt{\frac{1}{N}} \ ,$$

where the leading constant depends only on $\underline{\lambda}_s, \overline{\lambda}_s$ from Assumptions 1 and $\|f\|_{\mathbb{L}_1(\mathbb{P}_{\boldsymbol{X},S})} = \mathbb{E}|f(\boldsymbol{X}, S)|$.

The proof of this result combines expected deviation of empirical measure from the real measure in terms of Wasserstein distance on real line (Bobkov and Ledoux, 2016) with the already mentioned rank statistics. The first term of the derived bound corresponds to the estimation error of f^* by $\hat{f}$, the second term is the price to pay for not knowing conditional distributions $\boldsymbol{X}|S = s$ while the last term correspond to the price of unknown marginal probabilities of each protected attribute. Notice that if $N_s = p_s N$, which corresponds to the standard *i.i.d.* sampling from $\mathbb{P}_{\boldsymbol{X},S}$ of unlabeled data, the second and the third term are of the same order. If N is sufficiently large, which in most scenarios[5] is without loss of generality, then the rate is dominated by $\mathbf{E}\|f^* - \hat{f}\|_{\mathbb{L}_1(\mathbb{P}_{X,S})}$. Notice that one can find a collection of joint distributions $\mathbb{P}$ such that f^* satisfies demographic parity by simply making the marginal distributions of $\boldsymbol{X}$ and the regression function f^* independent from the sensitive attribute S. Hence, if $\hat{f}$ estimates f^* at the minimax rate *w.r.t.* the $\mathbb{L}_1(\mathbb{P}_{\boldsymbol{X},S})$ norm, then the resulting rate is also minimax optimal for the problem of estimating $g^* \equiv f^*$, provided the unlabeled sample is sufficiently large.

Conclusion on fairness. In summary, we propose a post-processing algorithm which can be applied on top of *any* off-the-shelf estimator of the regression function, in order to transform it into a fair one. The procedure requires *only* unlabeled data and its worst case *training* complexity is $N \log N$ and $\log N$ for the *inference* (with N being the total number of unlabeled data). The resulting algorithm is very effective to impose fairness in terms of the Demographic Parity both empirically and in theory. These results are unique in the fairness literature and we pay a particularly close attention to sharp finite sample controls on both fairness violation and risk on the one hand and interpretability on the other hand. The practical performance of our method demonstrates high efficiency of the developed plug-in approach.

[5]One can achieve it by splitting the labeled dataset $\mathcal{L}$ artificially augmenting the unlabeled one, which ensures that $N > n$. In this case if $\mathbf{E}\|f^* - \hat{f}\|_{\mathbb{L}_1(\mathbb{P}_{X,S})} = O(n^{-1/2})$, then the first term is always dominant in the derived bound.

Chapter 2

Set-valued classification

In the present chapter, we consider the problem of multi-class classification. Let $K \geq 2$ and $(\boldsymbol{X}, Y) \in \mathbb{R}^d \times [K]$ be a random couple following a distribution $\mathbb{P}$ on $\mathbb{R}^d \times [K]$. The vector $\boldsymbol{X} \in \mathbb{R}^d$ is seen as the vector of features with the marginal distribution $\mathbb{P}_{\boldsymbol{X}}$ and $Y \in [K]$ is the corresponding class. The conventional goal in multi-class classification is to predict the class Y for the feature vector $\boldsymbol{X}$ as accurately as possible using the samples. Since Y is a single element of $[K]$, in a classical setting this prediction also consists of a single class. However, in this chapter, we focus on possibly predicting multiple classes, namely a set-valued classifier, $\Gamma : \mathbb{R}^d \to 2^{[K]}$, which belongs to the set Ξ of all measurable functions from $\mathbb{R}^d$ to $2^{[K]}$. The next section provides a motivation for the set-valued framework in general. A particular emphasis is put on the setup where the size of the set-valued classifiers is controlled in expectation by some $s \in \mathbb{N}$. Formally, I will mainly investigate prediction rules that are related to the following optimal set-valued classifier

$$\textbf{s-Oracle:} \qquad \Gamma_s^* \in \arg\min \left\{ \mathbb{P}\left(Y \in \Gamma(\boldsymbol{X})\right) \,:\, \Gamma \text{ s.t. } \mathbb{E}_{\boldsymbol{X}} |\Gamma(\boldsymbol{X})| \leq s \right\} \quad . \tag{2.1}$$

This setting has been introduced in **[MH-Journal9]** and has been the main core of my research these years. I provide in this chapter an overview of my work, based on collaborations with C. Denis **[MH-Journal9]** and with E. Chzhen and C. Denis **[MH-Preprint4]**. In Section 2.3, we sketch some distribution-free results that can be expected in the setting of set-valued classification with controlled expected size and briefly describe my current projects related to the subject. Sections 2.4-2.5 explore the direction of finite sample risk guarantees under classical non-parametric assumptions. In particular, an Empirical Risk Minimization procedure is studied in Section 2.4 and a non-parametric minimax analysis is provided in Section 2.5.

2.1 Introduction to set-valued classification

This section is a summary of a joint work with E. Chzhen, C. Denis, and T. Lorieul where the focus is to provide an overview of various set-valued classification frameworks

(a) MNIST

banana, orange, Granny Smith

impala, gazelle, hartebeest

sports car, convertible

(b) Examples from ImageNet containing either multiple objects (left), or an intrinsically ambiguous image containing a single object associated with a single class (mid), or a single object with multiple matching attributes (right). The official *unique class* associated is shown in bold.

Figure 2.1: Examples from MNIST and ImageNet datasets.

and to highlight their advantages and/or disadvantages in different practical scenarios. Multi-class classification problems emerge in numerous applications ranging from image recognition to medical diagnosis. These applications are highly heterogeneous in terms of their difficulty. Consider Figures 2.1a–2.1b, which provides typical instances from two benchmark datasets of multi-class classification – MNIST and ImageNet. We can notice that the MNIST (Figure 2.1a) is rather simple from the human perspective, since it is effortless to identify the correct class (digit in this case). Actually it appears that standard machine learning algorithms, such as Support Vector Machines or logistic regression can easily achieve up to 95% of accuracy for this task. Meanwhile, using more advanced classification techniques, such as Neural Nets, one can reach up to 99% of accuracy. A simple conclusion from these observations is that the MNIST classification task is *simple* and *unambiguous*. By considering set-valued classifiers we can achieve nearly perfect accuracy at the price of a slight increase in the size. The situation is drastically different for the ImageNet dataset (Figure 2.1b). While this dataset is composed of instances of the form (image, *unique class*) one can observe that the unicity of the class is not intrinsic for this problem – other candidates can equally well explain a given picture. The three samples from the ImageNet illustrate in particular that ambiguity can come from different sources. In the first case, several fruits are present and it is not clear which one should be predicted. In the second, the image of the animal (an impala) does not capture the main feature of it. In the last case, the car fits in two classes (sport and convertible) and both are correct.

Hence, on the one hand, dataset that poses the same level of difficulty as the MNIST dataset can be tackled by classical multi-class predictors. In this sense, probably one of

the most valuable and easy to interpret approaches is to build a single-output classifier $\hat{h} : \mathcal{X} \to \mathcal{Y}$ that mimics the optimal classifier h^* which minimizes the misclassification risk

$$h^* \in \underset{h:\mathcal{X}\to\mathcal{Y}}{\arg\min}\, \mathbb{P}(h(\boldsymbol{X}) \neq Y) \ ,$$

where $\mathbb{P}$ is the joint distribution of $(\boldsymbol{X}, Y)$. One easily shows that h^* is given by

$$h^*(\cdot) = \underset{k\in\mathcal{Y}}{\arg\max}\, p_k(\cdot),$$

where $p_k(\boldsymbol{x}) = \mathbb{P}(Y = k \mid \boldsymbol{X} = \boldsymbol{x})$ for $\boldsymbol{x} \in \mathbb{R}^d$ and $k \in \mathcal{Y}$ is the posterior distribution of classes. On the other hand, when ambiguity occurs, this kind of approaches seems limited since, for instance, several p_k values may be close to the highest value. This is the case in the most modern large scale applications. In such a situation selecting a unique explanatory class among a large list of classes candidates might be less informative and more ambiguous. Therefore predictors that output more than a single class are suitable in this context and we shift to the setting of the *set-valued classification*.

Arguably the most natural set-valued classifier is the one that outputs a fixed amount of candidates for each instance. This type of set-valued classification strategies is called top-k prediction, where k is the amount of candidates predicted. On the population level, the optimal way to output k candidates is to select those that correspond to k highest conditional probabilities $p_1(\boldsymbol{x}), \ldots, p_K(\boldsymbol{x})$. For instance, top-5 score seems to be recommended for the `ImageNet` dataset. However, even in this case, it is unclear why *all* the images should be explained by exactly 5 candidates. Returning to Figure 2.1b we can see that the image with fruits (left) and with an animal (mid) are explained by 3 candidates while the image with a car (right) by 2 – top-5 cannot be a universal strategy since other applications and images might be even more heterogeneous. It seems more meaningful to consider those set-valued classifiers which can adapt to the local difficulty of the instance. Defining set-valued classifiers in a proper way becomes challenging in this case.

Despite the fact that various set-valued classifiers can be considered, we argue that all of these methods have to deal with two parameters that we call the *error rate* and the *set size*. Let us define them rigorously. Recall that a set valued classifier is a mapping $\Gamma : \mathbb{R}^d \to 2^{[K]}$. The *error rate* and the *(expected) size* of a set-valued classifier are defined as

$$\underbrace{\mathrm{L}(\Gamma) = \mathbb{P}\left(Y \notin \Gamma(\boldsymbol{X})\right)}_{\text{error}}, \qquad \underbrace{\mathrm{S}(\Gamma) = \mathbb{E}\left|\Gamma(\boldsymbol{X})\right|}_{\text{size}} \ ,$$

respectively. These two notions appear as fundamental and they have different names depending on the community and the field (for instance, the error rate is often called coverage, recall, or risk). The balance of the set size and the error rate is a common denominator between all set-valued classifiers, and depending on the application and

on the objective to reach, either of the two has to be put forward. While we cannot argue that one framework of set-valued classification is superior to others, we highlight some features in favor of the problem of set-valued classification with expected size set in Eq. (2.1).

Expected size vs. top-k. Probably the most popular and tempting set-valued classifier is referred to as the aforementioned top-k procedure, defined as

$$\textbf{Top-}k\textbf{:} \qquad \Gamma^*_{\text{top(s)}} \in \arg\min \left\{ \mathbb{P}(Y \notin \Gamma(\boldsymbol{X})) \,:\, \forall \boldsymbol{x} \in \mathbb{R}^d \;\; |\Gamma(\boldsymbol{x})| = \mathrm{s} \right\} \;\; ,$$

with some $\mathrm{s} \in [K]$. We refer to Lapin, Hein, and Schiele (2015) and Oh (2017) and references therein for developments in this direction. It is defined with an almost sure type of constraint (hard constraint) in contrast to (2.1) where the constraint is in expectation. Therefore, such an approach does not take into account inhomogeneous structure of the problem. That is, in the regions where the classification is easy, a good set-valued classifier should output less candidates and it should output more candidates for difficult regions. The constraint on the *expected size* allows to bypass this drawback.

Expected size vs. Error-rate constraint. Yet another way to define a set-valued classification framework is proposed by Vovk (2002a,b) and Vovk, Gammerman, and Shafer (2005) and statistically addressed by Sadinle, Lei, and Wasserman (2018), where for a fixed $\alpha \in (0,1)$, they define Γ^*_α as

$$\textbf{Controlled error-rate:} \qquad \Gamma^*_\alpha \in \arg\min \left\{ \mathbb{E}|\Gamma(\boldsymbol{X})| \,:\, \mathbb{P}(Y \notin \Gamma(\boldsymbol{X})) \leq \alpha \right\} \;\; ,$$

that is, Γ^*_α is the "smallest" set-valued classifier with controlled probability of error. Even though this framework is intuitive, in certain situations it suffers from the lack of interpretability and the lack of stability *w.r.t.* the parameter α. Let us construct a simple example to illustrate this phenomenon. Assume that $K \geq 10$ is an even integer and let us define the following distribution $\mathbb{P}$ of the pair $(\boldsymbol{X}, Y)$:

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

$$\underbrace{\begin{cases} p_1(\boldsymbol{x}) := 0.75 \\ \forall k \in \left[2, \frac{K}{2}\right] \cap \mathbb{N}, \;\; p_k(\boldsymbol{x}) := \frac{0.3}{K-2} \\ \forall k > \frac{K}{2}, \;\; p_k(\boldsymbol{x}) := \frac{0.2}{K} \end{cases}}_{\forall \boldsymbol{x} \in [0,1]^d} \;\; , \qquad \underbrace{\begin{cases} p_1(\boldsymbol{x}) := p_2(\boldsymbol{x}) := p_3(\boldsymbol{x}) = 0.25 \\ \forall k \in \left[4, \frac{K}{2}\right] \cap \mathbb{N}, \;\; p_k(\boldsymbol{x}) := \frac{0.3}{K-6} \\ \forall k > \frac{K}{2}, \;\; p_k(\boldsymbol{x}) := \frac{0.2}{K} \end{cases}}_{\forall \boldsymbol{x} \in (1,2]^d} \;\; .$$

The above example is tailored to reflect the inhomogeneous structure of the classification problem. Indeed, on the set $[0,1]^d$ there is a one dominant class and the classification is easy, whereas on the set $(1,2]^d$ the classification is more difficult as three classes can equally well explain the observation. Moreover, if we decide to rely on the previous framework and encounter this type of distributions, then for all values $\alpha \leq 0.1$ the set-valued classifier Γ^*_α satisfies

$$|\Gamma^*_\alpha(\boldsymbol{X})| \geq \frac{K}{2} \text{ almost surely } \mathbb{P}_{\boldsymbol{X}} \ .$$

That is, for the most intuitive range of levels[1] $\alpha \leq 0.1$, the optimal set-valued classifier with controlled error Γ^*_α is too large and its use is limited. This situation occurs when the probability of almost each individual class is very low (*i.e.*, ≈ 0), yet, their sum is relatively high (*i.e.*, ≈ 1). Returning to our example, for large values of K and all $\boldsymbol{x} \in [0,1]^d$ the values of $p_2, \ldots, p_K$ are negligible. However, their total impact, that is $p_2 + \ldots + p_K$, adds up to 0.25. In this situation the optimal set-valued classifier with controlled error Γ^*_α is deemed to output almost all classes if the goal is to control the error. Nevertheless, this framework can be well suited for applications where the control on the error is crucial and even massive outputs are acceptable.

The situation is different for the set-valued framework with *controlled expected size* that we consider here as for $s = 2$, the (s-Oracle) Γ^*_s outputs only one class on $[0,1]^d$ and three classes on $(1,2]^d$, while preserving its essential feature of having the controlled size in expectation. Consequently, the considered framework gains in both interpretability and stability of the outcome and diminishes the impact of unlikely classes.

2.2 Set-valued classification with controlled expected size

This section illustrates some properties on the optimal set-valued predictor with controlled expected size Γ^*_s. We recall its definition here.

$$\textbf{s-Oracle:} \qquad \Gamma^*_s \in \arg\min \{L(\Gamma) \ : \ \Gamma \in \Xi \text{ s.t. } S(\Gamma) \leq s\} \ , \tag{2.2}$$

for some $s \in [K]$. Let us point out that the feasible set $\{\Gamma \in \Xi : S(\Gamma) \leq s\}$ of the above problem is distribution dependent. It implies that *a priori* we cannot decide whether a given set-valued classifier is feasible. However, this set only depends on the marginal distribution $\mathbb{P}_{\boldsymbol{X}}$ of the features, which motivates us to introduce unlabeled sample in the observational model. Hence, we are interested in the semi-supervised setup of this problem. That is, in what follows it is assumed that two independent samples are provided – labeled $\mathcal{D}^L_n = \{(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)\} \overset{\text{i.i.d.}}{\sim} \mathbb{P}$ and unlabeled[2]

[1] The level 0.1 is selected only for the sake of simplicity and other values can be considered.

[2] By agreement $N = 0$ means that $\mathcal{D}^U_N = \emptyset$, and the deduced procedure is supervised.

$\mathcal{D}_N^{\mathrm{U}} = \{\boldsymbol{X}_{n+1}, \ldots, \boldsymbol{X}_{n+N}\} \overset{\text{i.i.d.}}{\sim} \mathbb{P}_{\boldsymbol{X}}$ both being independent from $(\boldsymbol{X}, Y)$. The statistical goal is to construct an empirical rule (a set-valued classifier) $\hat{\Gamma} : (\mathbb{R}^d \times [K])^n \times (\mathbb{R}^d)^N \to \Xi$, which mimics the behavior of Γ_{s}^*.

Properties of s-Oracle. Let us start by stating the following mild continuity assumption.

Assumption 2 (Continuity of CDF)**.** *For all $k \in [K]$ the CDF $F_{p_k}(\cdot) := \mathbb{P}_{\boldsymbol{X}}(p_k(\boldsymbol{X}) \leq \cdot)$ of $p_k(\boldsymbol{X})$ is continuous on $(0, 1)$.*

The continuity Assumption 2 is central and will be present in all but one subsequent sections. It allows to express the s-Oracle set-valued classifier Γ_{s}^* in the form of thresholding and to draw the relation of the constrained minimization with its unconstrained counterpart.

Proposition 2.1. *Fix $\mathrm{s} \in (0, K)$, and let the function $G : [0, 1] \to [0, K]$ be defined for all $t \in [0, 1]$ as*

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

then under Assumption 2 an s-Oracle set-valued classifier Γ_{s}^* can be obtained as

$$\Gamma_{\mathrm{s}}^*(\boldsymbol{x}) = \Big\{k \in [K] \ : \ p_k(\boldsymbol{x}) \geq G^{-1}(\mathrm{s})\Big\} \ , \tag{2.3}$$

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

Note that the threshold $G^{-1}(\mathrm{s})$ depends on the joint distribution $\mathbb{P}$ and thus, is unknown beforehand. In the absence of Assumption 2, one needs to consider randomized set-valued classifiers – those that map $\boldsymbol{x}$ into a distribution over $2^{[K]}$ instead of a deterministic prediction. Essentially, Assumption 2 is a sufficient condition under which the s-Oracle is a deterministic set-valued classifier. Additionally, under the continuity assumption, the considered framework is well posed in the sense that the s-Oracle set-valued classifier Γ_{s}^* is unique up to changes on sets of $\mathbb{P}_{\boldsymbol{X}}$ measure zero.

Theorem 2.1. *For every $\mathrm{s} \in (0, K)$, under Assumption 2 the s-Oracle set-valued classifier Γ_{s}^* defined in Proposition 2.1 is unique up to changes on $\mathbb{P}_{\boldsymbol{X}}$ measure zero. That is, for all $\Gamma : \mathbb{R}^d \to 2^{[K]}$ with $\mathrm{S}(\Gamma) \leq \mathrm{s}$ either of the following conditions hold*

- $\mathrm{L}(\Gamma) > \mathrm{L}(\Gamma_{\mathrm{s}}^*)$,
- $\Gamma(\boldsymbol{x}) = \Gamma_{\mathrm{s}}^*(\boldsymbol{x})$ *for almost all* $\boldsymbol{x} \in \mathbb{R}^d$ w.r.t. $\mathbb{P}_{\boldsymbol{X}}$.

The continuity assumption yields another description of the optimal set-valued classifier. Specifically, the next proposition establishes that s-Oracle can be obtained via an unconstrained minimization, which trades-off the error and the size.

Proposition 2.2. *Assume that Assumption 2 is fulfilled, then the* s*-Oracle defined in Eq. (2.3) is a minimizer over all set- valued predictors* Γ *of the following risk*

$$\mathcal{R}_{\mathrm{s}}(\Gamma) = \mathrm{L}(\Gamma) + G^{-1}(\mathrm{s})\,\mathrm{S}(\Gamma) \ . \tag{2.4}$$

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

$$\mathcal{R}_{\mathrm{s}}(\Gamma) - \mathcal{R}_{\mathrm{s}}(\Gamma^*_{\mathrm{s}}) = \sum_{k=1}^{K} \mathbb{E}_{\mathbb{P}_{\boldsymbol{X}}}\left[|p_k(\boldsymbol{X}) - G^{-1}(\mathrm{s})|\mathbf{1}_{\{k\in\Gamma(\boldsymbol{X})\triangle\Gamma^*_{\mathrm{s}}(\boldsymbol{X})\}}\right] \ , \tag{2.5}$$

and the same result holds for $\mathrm{L}_{\mathrm{s}}(\Gamma) - \mathrm{L}_{\mathrm{s}}(\Gamma^*_{\mathrm{s}})$ if we restrict the minimization to set-valued classifiers Γ with expected size s. On can already observe that the above excess risk of any set-valued classifier Γ relies on the behavior of the conditional probabilities p_k around the threshold $G^{-1}(\mathrm{s})$.

Measures of performance. Let us conclude this section by introducing performance measures that we will study in the context of set-valued classification with controlled expected size. For a given estimator $\hat{\Gamma}$ and a joint distribution $\mathbb{P}$ of $(\boldsymbol{X}, Y)$, fixed integers $K \geq 2$, $\mathrm{s} \in [K]$, and $n, N \in \mathbb{N}$, we are interested in the following risks

$$\mathcal{E}^{\mathrm{H}}_{n,N}(\hat{\Gamma};\mathbb{P}) := \mathbf{E}\left|\hat{\Gamma}(\boldsymbol{X})\triangle\Gamma^*_{\mathrm{s}}(\boldsymbol{X})\right| \ , \qquad \text{(Hamming risk)}$$

$$\mathcal{E}^{\mathrm{R}}_{n,N}(\hat{\Gamma};\mathbb{P}) := \mathbf{E}[\mathcal{R}_{\mathrm{s}}(\hat{\Gamma})] - \mathcal{R}_{\mathrm{s}}(\Gamma^*_{\mathrm{s}}) \ , \qquad \text{(Excess risk)}$$

where in the case of $\mathcal{E}^{\mathrm{H}}_{n,N}(\hat{\Gamma};\mathbb{P})$ the symbol $\mathbf{E}$ stands for the joint distribution of $\mathcal{D}^{\mathrm{L}}_n, \mathcal{D}^{\mathrm{U}}_N$, and $\boldsymbol{X}$ and in the case of $\mathcal{E}^{\mathrm{R}}_{n,N}(\hat{\Gamma};\mathbb{P})$ it stands for the joint distribution of $\mathcal{D}^{\mathrm{L}}_n, \mathcal{D}^{\mathrm{U}}_N$. These risks are arising in a natural way in the context of the set-valued classification with controlled expected size. While the risk $\mathcal{E}^{\mathrm{H}}_{n,N}(\hat{\Gamma};\mathbb{P})$ corresponds to the estimation of the s-Oracle through the Hamming distance, the second risk is directly connected with Proposition 2.2, which gives a description of the s-Oracle as a minimizer of $\mathcal{R}_{\mathrm{s}}(\cdot)$. Other measure of performance might be of course considered but these two translate already the main idea.

2.3 Distribution-free size controls

From a data-driven perspective, one of the first questions one may ask is "what can we do without any assumption?". We discuss here so-called *distribution-free* results and start from the point where we have in hand estimators $\hat{p}_k$ of the conditional probabilities p_k whatever their performance and/or nature. In addition we have access to the unlabeled dataset $\mathcal{D}^{\mathrm{U}}_N = \{\boldsymbol{X}_{n+1}, \ldots, \boldsymbol{X}_{n+N}\} \overset{\text{i.i.d.}}{\sim} \mathbb{P}_{\boldsymbol{X}}$ without asking any condition on $\mathbb{P}_{\boldsymbol{X}}$.

Already at this level we are in contradiction with the previous section where we restricted $\mathbb{P}$ to the class of distributions that satisfy Assumption 2, a condition that we put forward as essential to have nice representation of the optimal s-Oracle classifier. While, this condition is indeed central in the theoretical study of the s-Oracle provided in the subsequent sections, here we are willing to consider randomized set-valued estimators and to put aside this assumption for a moment. Randomized estimators allow us to gain distribution-free results in the spirit of conformal prediction theory (Vovk, Gammerman, and Shafer, 2005). The expression of the s-Oracle provided in Eq. (2.3) is particularly attractive but without Assumption 2 we can only guarantee that $\mathrm{S}(\Gamma_{\mathrm{s}}^*) \leq \mathrm{s}$ for all $\mathrm{s} \in [K]$. One can modify the s-Oracle and ensure that $\mathrm{S}(\Gamma_{\mathrm{s}}^*) = \mathrm{s}$ by considering a randomized version of Γ_{s}^* that may differ from Eq. (2.3) only on the event $\{p_k(\boldsymbol{x}) = G^{-1}(\mathrm{s})\}$ that may have a non-zero $\mathbb{P}$-mass. In particular, the randomized version of the s-Oracle would add all classes k in the above event with a certain probability calibrated so that the expected size of the randomized s-Oracle equals s. According to the risk, it seems clear that we cannot expect any control on the Hamming one $\mathcal{E}_{n,N}^{\mathrm{H}}$. Actually, none of the above risks can be controlled unless we assume some kind of consistency of estimators $\hat{p}_k$. With extra consistency assumption we could indeed provide some risk guarantee but this is not the direction we are willing to consider in this section since the goal here is to keep the framework as general as possible.

Distribution-free guarantees via basic plug-in rules. Here we focus on the validity of an estimator in terms of the constraint satisfaction. From a data-driven perspective, an estimator of the s-Oracle can still be built on the plug-in principle: Let $(\zeta_{k,n+i})_{k=1,\dots,K;i=1,\dots,N}$ be *i.i.d.* uniform random variables in $[0,u]$ and consider

$$\hat{G}_u(\cdot) = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} \mathbf{1}_{\{\hat{p}_k(\boldsymbol{X}_{n+i})+\zeta_{k,n+i}>\cdot\}} \ ,$$

be an estimator of the function $G(\cdot)$. Define the randomized set-valued classifier $\hat{\Gamma}_u$ in a point-wise manner as

$$\hat{\Gamma}_u(\boldsymbol{x}) = \left\{k \in [K] \,:\, \hat{p}_k(\boldsymbol{x}) + \zeta_k \geq \hat{G}_u^{-1}(\mathrm{s})\right\} \ , \tag{2.6}$$

where ζ_k is uniform on $[0,u]$ and is independent from all other random variables and $\hat{G}_u^{-1}$ is the generalized inverse of $\hat{G}_u$. The following important property of the introduced estimator $\hat{\Gamma}_u$ describes the deviation of the size of $\hat{\Gamma}_u$ from the desired level s.

Theorem 2.2. *Set $u \leq N^{-1/2}$. For any estimators $\hat{p}_k$ of the conditional probabilities p_k built on*

$\mathcal{D}_n^{\mathrm{L}}$ *and for all* $\mathrm{s} \in [K]$*, it holds*[3] *that for all* $N \in \mathbb{N}$

$$\sup_{\mathbb{P}} \mathbf{E}\left|\mathrm{S}\left(\hat{\Gamma}_u\right) - \mathrm{s}\right| \lesssim \frac{1}{\sqrt{N}} \ , \qquad \text{(size deviation)}$$

where the supremum is taken over all joint distributions of $(\boldsymbol{X}, Y)$ *and the leading constant is universal.*

The obtained rate is parametric and does not depend on the quality of $\hat{p}_k$ as it holds for any estimator. To be more specific, $\hat{p}_k$'s can be any arbitrary simplex valued function of the features. Note that the result of Theorem 2.2 is completely distribution and assumption free. This is achieved thanks to the randomization by $\zeta_{k,n+i}$'s, which ensures that the random variable $\hat{p}_k(\boldsymbol{X}_i) + \zeta_{k,n+i}$ is continuous. Let us conclude this section with a consistency result that is obtained under mild conditions on the p_k's.

Theorem 2.3. *Let* $u := u_{n,N}$ *be a sequence of positive numbers that converges to* 0 *when* $n, N \to \infty$*. Consider a distribution* $\mathbb{P}$ *such that Assumption 2 holds and such that* $\hat{p}_k(X)$ *converges to* $p_k(X)$ *in probability when* $n \to \infty$ *for all* $k \in [K]$*. Then the randomized set-valued classifier* $\hat{\Gamma}_u$ *defined in Eq. (2.6) satisfies*

$$\mathcal{E}^{\mathrm{H}}_{n,N}(\hat{\Gamma}_u; \mathbb{P}) \to 0 \ , \qquad \text{(Hamming risk)}$$
$$\mathcal{E}^{\mathrm{R}}_{n,N}(\hat{\Gamma}_u; \mathbb{P}) \to 0 \ , \qquad \text{(Excess risk)}$$

when $n, N \to \infty$*.*

Distribution-free guarantees via conformal predictions. This paragraph is the product of an on-going work with E. Chzhen and C. Denis. Set $U, U_{n+1}, \ldots, U_{n+N}$ to be *i.i.d.* random variables distributed uniformly on $[K]$. Define the following estimator

$$\hat{\Gamma}_{\mathrm{cp}}(\boldsymbol{x}) = \left\{ k \in [K] \ : \ 1 + \#\left\{ i \ : \ \hat{p}_{U_{n+i}}(\boldsymbol{X}_{n+i}) + \zeta_{U_{n+i},n+i} > \hat{p}_k(\boldsymbol{x}) + \zeta_k \right\} \le \frac{\mathrm{s}(N+1)}{K} \right\} \ .$$

The estimator $\hat{\Gamma}_{\mathrm{cp}}$ is inspired by the theory of conformal prediction (Lei, Robins, and Wasserman, 2013; Vovk, 2002a; Vovk, Gammerman, and Shafer, 2005) and it inherits the fruitful validity property – an essential feature of the conformal inference. Let me give a quick intuition relating the estimator $\hat{\Gamma}_{\mathrm{cp}}$ to the conformal prediction theory. Notice that for any set-valued classifier Γ (data-dependent or not) it holds that

$$\mathbf{E}[\mathrm{S}(\Gamma)] = \mathbf{E} \sum_{k=1}^{K} \mathbf{1}_{\{k \in \Gamma(\boldsymbol{X})\}} = K \cdot \mathbf{P}(U \in \Gamma(\boldsymbol{X})) \ .$$

[3] In this result, the expectation $\mathbb{E}$ in $\mathrm{S}(\Gamma) = \mathbb{E}\,|\Gamma(\boldsymbol{X})|$ is taken over $(\boldsymbol{X}, Y)$ and the perturbations $(\zeta_{k,n+i})_{k=1,\ldots,K; i=1,\ldots,N}$.

Hence, the set-valued classifiers of interest should satisfy $K \cdot \mathbf{P}(U \in \Gamma(\boldsymbol{X})) \leq s$. The conformal prediction theory directly addresses the question of creating a data-driven method $\hat{\Gamma}$ which satisfies this condition in the distribution-free sense.

Proposition 2.3. *Assume that $(N+1)(K-s)/K$ is an integer*[4]*, then for any $\hat{p}_1, \ldots, \hat{p}_K$ independent from $\boldsymbol{X}_{n+1}, \ldots, \boldsymbol{X}_{n+N}$ it holds that*

$$\mathbf{E}[\mathrm{S}(\hat{\Gamma}_{cp})] = s \ .$$

The proof is very short, simple, and instructive. It highlights the reason to randomize the classes via $U_{n+1}, \ldots, U_{n+N}$ as well as it explains the randomization via $\zeta_{n+1}, \ldots, \zeta_{n+N}$.

Proof. Using the definition of $\hat{\Gamma}_{\mathrm{cp}}$ we derive

$$\begin{aligned}
\mathbf{E}|\hat{\Gamma}_{\mathrm{cp}}(\boldsymbol{X})| &= \sum_{k=1}^{K} \mathbf{P}\left(1 + \sum_{i=1}^{N} \mathbf{1}_{\left\{\hat{p}_{U_{n+i}}(\boldsymbol{X}_{n+i}) + \zeta_{U_{n+i},n+i} > \hat{p}_k(\boldsymbol{X}) + \zeta_k\right\}} \leq \frac{s(N+1)}{K}\right) \\
&= K \cdot \mathbf{P}\left(1 + \sum_{i=1}^{N} \mathbf{1}_{\left\{\hat{p}_{U_{n+i}}(\boldsymbol{X}_{n+i}) + \zeta_{U_{n+i},n+i} > \hat{p}_U(\boldsymbol{X}) + \zeta_U\right\}} \leq \frac{s(N+1)}{K}\right) \\
&= K \cdot \mathbf{P}\left((N+1) \cdot \frac{K-s}{K} \leq \mathrm{Rank}\big(\hat{p}_U(\boldsymbol{X}) + \zeta_U\big) - 1\right) \ ,
\end{aligned}$$

where $\mathrm{Rank}\big(\hat{p}_U(\boldsymbol{X}) + \zeta_U\big)$ is the rank of $\hat{p}_U(\boldsymbol{X}) + \zeta_U$ in $\{\hat{p}_{U_{n+i}}(\boldsymbol{X}_{n+i}) + \zeta_{U_{n+i},n+i}\}_{i=1}^{N} \cup \{\hat{p}_U(\boldsymbol{X}) + \zeta_U\}$. Notice that thanks to the fact that $(\zeta_{k,n+i})_{k=1,\ldots,K;i=1,\ldots,N}$ are *i.i.d.* continuous, then the random variables $\{\hat{p}_{U_{n+i}}(\boldsymbol{X}_{n+i}) + \zeta_{U_{n+i},n+i}\}_{i=1}^{N} \cup \{\hat{p}_U(\boldsymbol{X}) + \zeta_U\}$ are exchangeable and continuous. Hence, it holds that the statistic $\mathrm{Rank}\big(\hat{p}_U(\boldsymbol{X}) + \zeta_U\big)$ is distributed uniformly on $\{1, \ldots, N+1\}$. Then, since $(N+1) \cdot \frac{K-s}{K}$ is an integer,

$$K \cdot \mathbf{P}\left((N+1) \cdot \frac{K-s}{K} \leq \mathrm{Rank}\big(\hat{p}_U(\boldsymbol{X}) + \zeta_U\big) - 1\right) = K \cdot \frac{N+1-(N+1) \cdot \frac{K-s}{K}}{N+1} = s \ .$$

□

The previous result seems to be extremely appealing, as it gives a very strong distribution-free guarantee on the size. Let us remark the drawbacks of this result. First of all, the estimator $\hat{\Gamma}_{\mathrm{cp}}$ is highly random due to the uniformly distributed random variables U_{n+i}. One can argue saying that $\hat{\Gamma}_u$ from the previous part is also random. However, the randomness in $\hat{\Gamma}_u$ is more of a technical issue than an intrinsic requirement, whereas the randomness in $\hat{\Gamma}_{\mathrm{cp}}$ is a necessity. Actually in both cases the random variables $\zeta_{n+1}, \ldots, \zeta_{n+N}$ are playing identical roles – they are used to ensure continuity. Secondly, while the result is appealing, it is actually extremely different from that of Theorem 2.2. Indeed, the guarantee on $\hat{\Gamma}_u$ would be more appealing to members of

[4]This condition is only introduced here for the sake of simplicity of the subsequent argument.

the Machine Learning community, while the guarantee on $\hat{\Gamma}_{\text{cp}}$ is closer to the problem of testing in its spirit.

At last, let me provide heuristic arguments indicating the statistical validity of the estimator $\hat{\Gamma}_{\text{cp}}$. Define $\hat{G}(t) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{\left\{\hat{p}_{U_{n+i}}(\boldsymbol{X}_{n+i})+\zeta_{U_{n+i}}, n+i > t\right\}}$, then $\hat{\Gamma}_{\text{cp}}$ can be equivalently written as

$$\hat{\Gamma}_{\text{cp}}(\boldsymbol{x}) = \left\{k \in [K] \,:\, \hat{G}(\hat{p}_k(\boldsymbol{x}) + \zeta_k) \leq \frac{s}{K}\left(1 + \frac{1}{N}\right) - \frac{1}{N}\right\} \ .$$

Note that $\mathbf{E}[\hat{G}(t) \mid \mathcal{D}_n^{\text{L}}] = \frac{1}{K}\tilde{G}(t) := \frac{1}{K}\sum_{k=1}^{K} \mathbf{P}(\hat{p}_k(\boldsymbol{X}) + \zeta_k > t \mid \mathcal{D}_n^{\text{L}})$. Using standard results of empirical process theory one can demonstrate that with high probability it holds for all $t \in [0, 1+u]$ that

$$\hat{G}(t) \approx \tfrac{1}{K}\tilde{G}(t) \pm \frac{1}{\sqrt{N}} \ .$$

This heuristic indicates the following relation between the estimator $\hat{\Gamma}_{\text{cp}}$ and a pseudo-Oracle estimator $\tilde{\Gamma}$, which knows the marginal distribution $\mathbb{P}_{\boldsymbol{X}}$

$$\begin{aligned}\hat{\Gamma}_{\text{cp}}(\boldsymbol{x}) \approx \tilde{\Gamma}(\boldsymbol{x}) &= \left\{k \in [K] \,:\, \tilde{G}(\hat{p}_k(\boldsymbol{x}) + \zeta_k) \leq s + \frac{s}{N} - \frac{K}{N} \pm \frac{K}{\sqrt{N}}\right\} \\ &= \left\{k \in [K] \,:\, \hat{p}_k(\boldsymbol{x}) + \zeta_k \geq \tilde{G}^{-1}\left(s - \frac{K-s}{N} \pm \frac{K}{\sqrt{N}}\right)\right\} \ .\end{aligned}$$

Finally, note that if $\hat{p}_k \approx p_k$, $u \approx 0$, and N is sufficiently large, then

$$\tilde{\Gamma}(\boldsymbol{x}) \approx \Gamma_s^*(\boldsymbol{x}) = \left\{k \in [K] \,:\, p_k(\boldsymbol{x}) \geq G^{-1}(s)\right\} \ .$$

In what follows, we are willing to go further and study the behavior of the excess risks in a finite sample regime without relying on randomized set-valued classifiers. This is the scope of the next two sections where set-valued classifiers are based on empirical risk minimization and plug-in principle respectively.

2.4 Empirical risk minimization based set-valued classification

In this section we focus on set-valued classifiers built upon Empirical Risk Minimization (ERM) methods. In particular, inspired by risk convexification techniques that yield popular methods in the binary classification setting (Bartlett, Jordan, and McAuliffe, 2006; Freund and Schapire, 1997; Friedman, Hastie, and Tibshirani, 2000; Vapnik, 1998; Yuan and Wegkamp, 2010; Zhang, 2004b), we study convex surrogates of the 0/1-loss that are tailored for the multi-class framework. The essential feature of this approach is that by performing ERM we do not directly estimate the posterior probabilities $\boldsymbol{p}(\cdot) = (p_1(\cdot), \ldots, p_K(\cdot))$. Instead, we use a multi-class score vector function $\mathbf{f}(\cdot) = (f_1(\cdot), \ldots, f_K(\cdot))$ which, although it does not necessarily approximate $\boldsymbol{p}(\cdot)$, preserves its essential ordering properties.

Convex surrogate and calibration. Let $\mathbf{f} = (f_1, \dots, f_K) : \mathcal{X} \to \mathbb{R}^K$ be a multi-class score function and set $G_{\mathbf{f}}(\cdot) = \sum_{k=1}^{K} \mathbb{P}_{\boldsymbol{X}}(f_k(\boldsymbol{X}) > \cdot)$. We define a set-valued classifier $\Gamma_{\mathbf{f},\delta}$ associated with some score function $\mathbf{f}$ and some $\delta \in \mathbb{R}$ as

$$\Gamma_{\mathbf{f},\delta}(X) = \{k \in [K] \ : \ f_k(\boldsymbol{X}) \geq -\delta\} \ . \tag{2.7}$$

Analogously to Assumption 2, we require the continuity of $G_{\mathbf{f}}$. This allows us to deduce that for any size $s \in (0, K)$, there exists $\delta \in \mathbb{R}$, such that $G_{\mathbf{f}}(-\delta) = s$. Consequently, this choice of the parameter δ ensures that $\Gamma_{\mathbf{f},\delta}(X)$ meets the size constraint, that is, $S(\Gamma_{\mathbf{f},\delta}) = s$. We aim at solving the problem $\min_{\mathbf{f}\in\mathcal{F}} L(\Gamma_{\mathbf{f},\delta})$ where $\mathcal{F}$ is a class of functions throughout a convex surrogate. Let $\phi : \mathbb{R} \to \mathbb{R}$ be a convex function. We define the ϕ-error of $\mathbf{f}$ by

$$L_\phi(\mathbf{f}) = \mathbb{E}\left[\sum_{k=1}^{K} \phi(Z_k f_k(\boldsymbol{X}))\right] \ , \tag{2.8}$$

where $Z_k = 2 \cdot \mathbf{1}_{\{Y=k\}} - 1$ for all $k = 1, \dots, K$. Therefore, our target score functions are

$$\underbrace{\bar{\mathbf{f}} \in \arg\min_{\mathbf{f}\in\mathcal{F}} L_\phi(\mathbf{f})}_{\text{optimal over } \mathcal{F}} , \qquad \underbrace{\mathbf{f}^* \in \arg\min_{\mathbf{f}} L_\phi(\mathbf{f})}_{\text{overall optimal}} \ , \tag{2.9}$$

for the purpose of building the optimal set-valued classifiers $\Gamma_{\bar{\mathbf{f}},\delta}$ and $\Gamma_{\mathbf{f}^*,\delta}$ respectively.

Now we shift towards the calibration of the loss function ϕ that is used to build these set-valued classifiers (Bartlett, Jordan, and McAuliffe, 2006; Yuan and Wegkamp, 2010; Zhang, 2004b).

Definition 2.1. *We say that the function ϕ is* set-valued calibrated *if for all $s > 0$, there exists $\delta^* \in \mathbb{R}$ such that*

$$\Gamma_{\mathbf{f}^*,\delta^*} = \Gamma_s^* \ ,$$

with $\mathbf{f}^*$ given by Eq. (2.9).

The property of calibration means that the set-valued classifier based on $\mathbf{f}^*$ behaves identically to the s-Oracle Γ_s^*. The following result gives a complete characterization of the set-valued calibration property in terms of the function G.

Proposition 2.4. *The function ϕ is set-valued classifier calibrated if and only if for all $s \in (0, K)$, there exists $\delta^* \in \mathbb{R}$ such that $\phi'(\delta^*)$ and $\phi'(-\delta^*)$ both exists, $\phi'(\delta^*) < 0$, $\phi'(-\delta^*) < 0$ and*

$$G^{-1}(s) = \frac{\phi'(\delta^*)}{\phi'(\delta^*) + \phi'(-\delta^*)} \ ,$$

where ϕ' denotes the derivative of ϕ.

The proof of the proposition follows the lines of Theorem 1 by Yuan and Wegkamp (2010) with minor modifications due to the multi-class setting we consider. Commonly used loss functions like boosting ($x \mapsto \exp(-x)$), least-squares ($x \mapsto (x-1)^2$) and logistic ($x \mapsto \log(1 + \exp(-x))$) are examples of calibrated losses (see for instance (Bartlett, Jordan, and McAuliffe, 2006; Wegkamp and Yuan, 2011)).

Data-driven procedure. The set-valued calibration property in Definition 2.1 and Proposition 2.4 allow us to state a very general consistency result of set-valued classifiers deduced from i) scores functions $\mathbf{f}_n$ such that $G_{\mathbf{f}_n}$ are continuous[5] so that we can find $\delta_n \in \mathbb{R}$ such that $G_{\mathbf{f_n}}(-\delta_n) = s$; ii) the loss function ϕ is well suited set-valued calibrated loss (such as boosting, least-squares, and logistic losses). Indeed, we can show an equivalent of Zhang Lemma (see Theorem 10 by Zhang (2004a)) that says that consistency of the score function $\mathbf{f}_n$ in terms of ϕ-error implies consistency of the corresponding set-valued classifier $\Gamma_{\mathbf{f}_n,\delta_n}$ in terms of error:

$$\Delta L_\phi(\mathbf{f}_n) \xrightarrow{\mathbb{P}} 0 \qquad \Rightarrow \qquad \Delta L(\Gamma_{\mathbf{f}_n,\delta_n}) \xrightarrow{\mathbb{P}} 0 \ ,$$

for all $\mathbb{P}$, where

$$\Delta L(\Gamma_{\mathbf{f},\delta}) = L(\Gamma_{\mathbf{f},\delta}) - L(\Gamma_s^*) \, , \qquad \Delta L_\phi(\mathbf{f}) = L_\phi(\mathbf{f}) - L_\phi(\mathbf{f}^*) \ ,$$

are the excess error and the excess ϕ-error respectively. Yet, we are interested in a more precise description of this link between the two errors. Therefore, let us specify an estimation procedure. Recall that we have in hand a labeled $\mathcal{D}_n^L$ and an unlabeled $\mathcal{D}_N^U$ datasets. The labeled dataset is used to fit the score function $\hat{\mathbf{f}}$ by means of a minimization of the following empirical ϕ-error $\widehat{L}_\phi$ (which is the empirical counterpart of L_ϕ given in (2.8)):

$$\hat{\mathbf{f}} \in \arg\min_{\mathbf{f}\in\mathcal{F}} \widehat{L}_\phi(\mathbf{f}) \, , \qquad \widehat{L}_\phi(\mathbf{f}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} \phi(Z_k^i f_k(\boldsymbol{X}_i)) \ , \tag{2.10}$$

where $Z_k^i = 2 \cdot \mathbf{1}_{\{Y_i=k\}} - 1$ for all $k = 1, \ldots, K$ and $\mathcal{F}$ is a convex set of score functions. At this stage, we need to specify the $\delta \in \mathbb{R}$ such that $\Gamma_{\hat{\mathbf{f}},\delta}$ is of the correct order of size (in expectation). According to Proposition 2.4 this can be described by solving an equation where the only unknown is $G^{-1}(s)$. The latter depends only on $\mathbb{P}_X$ and then can be estimated using only the unlabeled dataset $\mathcal{D}_N^U$. We end up with the following definition of the empirical set-valued predictor based on $\hat{\mathbf{f}}$:

Definition 2.2. *Let $\hat{\mathbf{f}}$ be the minimizer of the empirical ϕ-error given in (2.10) based on $\mathcal{D}_n^L$, and consider the unlabeled dataset $\mathcal{D}_N^U$. Let $s \in (0, K)$. We defined the ERM based set-valued classifier $\hat{\Gamma}_s$ by*

$$\hat{\Gamma}_s(\boldsymbol{x}) = \left\{ k \in [K] : \ \hat{G}(\hat{f}_k(\boldsymbol{x})) \leq s \right\}, \tag{2.11}$$

where

$$\hat{G}(\cdot) = \frac{1}{N} \sum_{\boldsymbol{X}\in\mathcal{D}_N^U} \sum_{k=1}^{K} \mathbf{1}_{\{\hat{f}_k(\boldsymbol{X})\geq\cdot\}}.$$

[5] If this does not hold, a randomized version of $\mathbf{f}_n$ would imply such a continuity.

Statistical analysis. Now we investigate rates of convergence for the empirical set-valued classifiers $\hat{\Gamma}_s$ from Definition 2.2 *w.r.t.* the risk $\mathcal{E}^R_{n,N}$. We need the following continuity assumption.

Assumption 3 (Continuity of scores CDF)**.** *For all $k \in [K]$, conditionally on the data, the CDF $F_{\hat{f}_k}(t) := \mathbb{P}_X(\hat{f}_k(\boldsymbol{X}) \leq t)$ of $\hat{f}_k(\boldsymbol{X})$ is almost surely $\mathbb{P}^{\otimes n}$ continuous on $(0,1)$.*

As in Section 2.3 this condition on the estimator can be satisfied using randomization of the functions $\hat{f}_k$ without changing the statistical performance of the resulting set-valued predictor. As it is common in the study of ERM-type algorithms, we are going to introduce an additional assumption, which allows to derive better rates of convergence. Various and often related forms of additional assumptions leading to faster rates of convergence were proposed in the literature. We adopt the classical margin condition to our framework (Mammen and Tsybakov, 1999; Polonik, 1995; Tsybakov, 2004).

Assumption 4 (α-margin assumption)**.** *We say that the distribution $\mathbb{P}$ of the pair $(\boldsymbol{X}, Y) \in \mathbb{R}^d \times [K]$ satisfies α-margin assumption if there exists $C_1 > 0$ and $t_0 \in (0,1)$ such that for every positive $t \leq t_0$*

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

The exponent α will directly specify the rates of convergence. The classification problem gets easier as α grows. It is important to note that since we assume that the distribution functions of $p_k(\boldsymbol{X})$ are continuous for each k, we have $\mathbb{P}_X\left(0 < |p_k(\boldsymbol{X}) - G^{-1}(s)| \leq t\right) \to 0$ with $t \to 0$. Therefore, the margin condition only specifies the rate of this decay to 0. The second condition relies on the modulus of convexity of L_ϕ induced by the convex function ϕ *w.r.t.* the $\mathbb{L}_2(\mathbb{P}_X)$-norm.

Assumption 5 (ε^2-modulus of convexity)**.** *Consider the modulus of convexity of L_ϕ induced by the convex function ϕ* w.r.t. *the $\mathbb{L}_2(\mathbb{P}_X)$-norm defined by*

$$\omega(\varepsilon) = \inf\left\{\frac{L_\phi(\mathbf{f}) + L_\phi(\mathbf{g})}{2} - L_\phi\left(\frac{f+g}{2}\right) \ : \ \sum_{k=1}^K \mathbb{E}_X\left[(f_k - g_k)^2(\boldsymbol{X})\right] \geq \varepsilon^2\right\},$$

We assume that L_ϕ has modulus of convexity ε^2, that is, there exists $C_2 > 0$ such that $\omega(\varepsilon) \geq C_2\varepsilon^2$.

The parameter ε^2 characterizes the convexity of the loss function. We refer to (Bartlett, Jordan, and McAuliffe, 2006; Bartlett and Mendelson, 2006) for more details on modulus of convexity for classification risk. Finally, we assume more regularity on the loss function ϕ.

Assumption 6 (Calibration and Lipschitzness)**.** *We assume that ϕ is classification calibrated and L-Lipschitz for $L > 0$, that is: for all $t, t' \in \mathbb{R}$ it holds that*

$$\left|\phi(t) - \phi(t')\right| \leq L\left|t - t'\right| \ .$$

Introduce the marginal conditional excess ϕ-loss on $\mathbf{f} = (f_1, \ldots, f_K)$ defined as

$$\Delta \mathrm{L}_\phi^k(\mathbf{f}(\boldsymbol{X})) = p_k(\boldsymbol{X})(\phi(f_k(\boldsymbol{X})) - \phi(f_k^*(\boldsymbol{X}))) + (1 - p_k(\boldsymbol{X}))(\phi(-f_k(\boldsymbol{X})) - \phi(-f_k^*(\boldsymbol{X}))) \ ,$$

for $k = 1, \ldots, K$. We can now state our error control for the empirical set-valued classifiers defined by (2.11).

Theorem 2.4. *Assume that $\|f\|_\infty \leq B$ for all $f \in \mathcal{F}$. Let $M_n = \mathcal{N}(1/n, L_\infty, \mathcal{F})$ be the covering number of $\mathcal{F}$ w.r.t. L_∞-norm with closed balls with radius $1/n$. Under Assumptions 2-3-4-5-6, and if there exist constants $C > 0$ and $r > 1$ such that* [6]

$$\left| p_k(\boldsymbol{X}) - G^{-1}(\mathrm{s}) \right|^r \leq C \Delta \mathrm{L}_\phi^k(-\delta^*) \ , \tag{2.12}$$

with δ^* is provided in Proposition 2.4, it holds that

$$\mathcal{E}_{n,N}^{\mathrm{R}}(\hat{\Gamma}_{\mathrm{s}}; \mathbb{P}) \lesssim \left\{ \inf_{\mathbf{f} \in \mathcal{F}} \Delta \mathrm{L}_\phi(\mathbf{f}) + \frac{\log(M_n)}{n} \right\}^{\alpha/(\alpha+r)} + \frac{1}{\sqrt{N}} \ ,$$

where the leading constant depends only on L, B, r and α.

We conclude that rate of convergence for the excess error is $\left(\frac{\log(M_n)}{n}\right)^{\alpha/(\alpha+r)} + \frac{1}{\sqrt{N}}$. At the time it was proved, this result was the first bound, up to my knowledge, that provides a control on the excess risk for set-valued classifiers in multi-class setting. Compared to the literature, the exponent $\alpha/(\alpha + r)$ is not classical and it is not clear whether it is improvable. The second part of the rates which is of order $N^{-1/2}$ relies on the estimation of the function $\widetilde{G}(t) = \sum_{k=1}^K (1 - F_{\hat{\mathbf{f}}_k}(t))$, which serves as a pseudo-oracle CDF that knows the marginal distribution $\mathbb{P}_{\boldsymbol{X}}$. This part of the estimation is established under the mild Assumptions 2 and 3.

Numerical analysis. We apply our methodology with different choices of loss functions ϕ and then build set-valued classifiers $\hat{\Gamma}_{\mathrm{s}}$ based on the random forest, the softmax regression, the support vector machines, and the k nearest neighbors (with $k = 11$) procedures. Additionally, in accordance with Theorem 2.4 we develop an aggregated set-valued classifiers by considering the class $\mathcal{F}$ defined as the convex hull of the above 4 procedures and the boosting loss.

We evaluate the performance of the procedure on two real datasets: the *Forest type mapping* dataset and the *one-hundred plant species leaves* dataset coming from the UCI database. We refer to these two datasets as `Forest` and `Plant` respectively. The `Forest` dataset consists of $K = 4$ classes and 523 labeled observations (we gather the train and test sets) with 27 features. Here the classes are unbalanced. In the `Plant` dataset, there

[6] With abuse of notation, we write $\Delta \mathrm{L}_\phi^k(-\delta^*)$ instead of $\Delta \mathrm{L}_\phi^k((-\delta^*, \ldots, -\delta^*))$ since no confusion can occur.

Forest ($K = 4$)						
		s-set				
s		rforest	softmax reg	svm	kknn	CV
2	L	0.02 (0.02)	0.06 (0.02)	0.02 (0.01)	0.05 (0.03)	0.02 (0.01)
	S	2.00 (0.09)	2.00 (0.08)	2.00 (0.09)	2.00 (0.08)	2.00 (0.08)

Plant ($K = 100$)						
		s-set				
s		rforest	softmax reg	svm	kknn	CV
2	L	0.18 (0.03)	0.77 (0.02)	0.32 (0.04)	0.20 (0.03)	0.17 (0.03)
	S	2.00 (0.09)	2.02 (0.18)	1.99 (0.10)	2.00 (0.08)	2.00 (0.08)
10	L	0.02 (0.01)	0.42 (0.04)	0.03 (0.02)	0.08 (0.03)	0.02 (0.01)
	S	9.95 (0.38)	10.06 (0.58)	9.98 (0.22)	9.98 (0.23)	9.96 (0.37)

Table 2.1: For each of the $B = 100$ repetitions and for each dataset, we derive the estimated errors L and sizes S of the different set-valued classifiers *w.r.t.* s. We compute the means and standard deviations (between parentheses) over the $B = 100$ repetitions. For each s, the set-valued classifiers are based on—from left to right—`rforest`, `softmax reg` and `svm`, `kknn` and `CV` which are respectively the random forest, the softmax regression, support vector machines, k nearest neighbors and the aggregation procedure. Top: the dataset is the `Forest` – the dataset is the `Plant`.

are $K = 100$ classes and 1600 labeled observations. This dataset is balanced so that each class consists of 16 observations. The original dataset contains 3 covariates (each covariate consists of 64 features). In order to make the problem more challenging, we drop 2 covariates.

To get an indication of the statistical significance of the aggregated procedure (referred to as `CV`) we compare it to the set-valued classifiers that result from each component of the library in terms of errors and sizes. Without going in deep details, we mention that we split the dataset in three: the labeled $\mathcal{D}_n^{\mathrm{L}}$ and an unlabeled $\mathcal{D}_N^{\mathrm{U}}$ datasets of size n and N respectively to train the set-valued classifiers and a third labeled dataset to compute errors and sizes consists of M samples. We set the sizes of the samples as $n = 200$, $N = 100$ and $M = 223$ for the `Forest` dataset, and $n = 1000$, $N = 200$ and $M = 400$ for the `Plant` one. As a benchmark, we note that the misclassification error $\mathbb{P}(Y \neq h(\boldsymbol{X}))$ of the best classifier h from the library for the `Forest` dataset is evaluated at 0.15 , whereas in the `Plant` dataset, it is evaluated at 0.40. The performance of the classical single-output classifier is rather weak in the latter dataset.

Results are reported in Table 2.1, and confirm our expectations. A general observation is that the size constraint is quite well satisfied for all the methods thanks to the unlabeled sample. Also, even low values of the set size s lead to a drastic improvements in terms of the error when compared to the benchmarks evaluated on single-outputs classifiers. For instance, for the `Plant`, the error rate of the set-valued classifier with $s = 2$ based on

random forests is 0.18 whereas the misclassification error rate of the best component in the library is 0.40 – a considerable improvement for the price of only one extra class in average. Overall, the aggregated set-valued classifier (`CV`) outperforms all components of the library in all of the experiments which motivates the use of aggregation procedure.

2.5 Minimax set-valued classification

In this section, we put the focus on non-asymptotic minimax analysis of set-valued classifiers with controlled expected size. In contrast to the previous section, we build the data-driven procedure based on plug-in principle and then we exploit classical non-parametric theory tools to derive upper and lower bounds on the excess-risk (Audibert and Tsybakov, 2007; Györfi et al., 2002; Mammen and Tsybakov, 1999; Massart and Nédélec, 2006; Yang, 1999). From the technical point of view, our work is close in spirit to the one by Audibert and Tsybakov (2007) who study the statistical performance of plug-in classification rules under assumptions which involve the smoothness of the regression function and the margin condition. They derive fast rates of convergence (faster than $n^{-1/2}$) for plug-in classifiers based on local polynomial estimators (Audibert and Tsybakov, 2007; Stone, 1977; Tsybakov, 1986) and show their optimality in the minimax sense.

The central notion we manipulate in this section is the minimax rate of convergence in the semi-supervised setting.

Definition 2.3 (Minimax rate of convergence)**.** *For a given family $\mathcal{P}$ of joint distributions on $\mathbb{R}^d \times [K]$ the minimax rates are defined as*

$$\mathcal{E}^{\square}_{n,N}(\mathcal{P}) := \inf_{\hat{\Gamma}} \sup_{\mathbb{P}\in\mathcal{P}} \mathcal{E}^{\square}_{n,N}(\hat{\Gamma}; \mathcal{P}) \ ,$$

where $\square$ is H *or* R *and the infimum is taken over all set-valued classifiers based on $\mathcal{D}^{\mathrm{L}}_n$ and $\mathcal{D}^{\mathrm{U}}_N$. The sequence $\mathcal{E}^{\square}_{n,0}(\mathcal{P})$ corresponds to the* supervised *regime, while the sequence $\mathcal{E}^{\square}_{n,N}(\mathcal{P})$ for $N \geq 1$ corresponds to the* semi-supervised *regime.*

Our work has the objective to gain new insight into the minimax analysis of the set-valued classifiers framework with controlled expected size. In particular, we address the following questions: **i)** what is a minimax setup in this problem and what are the minimax rates of convergence?; **ii)** can we statistically justify the introduction of the unlabeled data $\mathcal{D}^{\mathrm{U}}_N$ from the minimax perspective? To be more precise, we would like to understand whether the rates of convergence are affected by N – the size of the unlabeled sample. Neither of these natural questions have been considered and answered in the previous literature.

Assumptions. In this part we state all the assumptions used in this work and state the family of distributions $\mathcal{P}$ which drives the minimax rates. The first assumption is the margin assumption Assumption 4 that we already introduced in the previous Section 2.3.

The second assumption restricts the set of possible marginal distributions of the feature vectors. Following (Audibert and Tsybakov, 2007), we first introduce the notion of regular set. Let c_0 and r_0 be two positive constants. We say that a Borel set $A \subset \mathbb{R}^d$ is a (c_0, r_0)-regular set if

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

Definition 2.4 (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 $\mu(\boldsymbol{x}) = 0$ for all $\boldsymbol{x} \in \mathbb{R}^d \setminus A$ and*

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

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

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

Definition 2.5 (Hölder class, Tsybakov, 2008)**.** *We say that a function $h : \mathbb{R}^d \to \mathbb{R}$ is (β, L)-Hölder for $\beta > 0$ and $L > 0$ if h is $\lfloor \beta \rfloor$ times continuously differentiable and $\forall \boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^d$ we have*

$$\left| h(\boldsymbol{x}') - h_{\boldsymbol{x}}(\boldsymbol{x}') \right| \leq L \left\| \boldsymbol{x} - \boldsymbol{x}' \right\|^{\beta} \ ,$$

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

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

Definition 2.6. *We denote by $\mathcal{P}(L, \beta, \alpha)$ the set of joint distributions on $\mathbb{R}^d \times [K]$ which satisfy the following conditions*

- *the marginal $\mathbb{P}_X$ satisfies the $(\mu_{\min}, \mu_{\max}, c_0, r_0)$-strong density,*

- *for all $k \in [K]$ the k^{th} regression function $p_k(\cdot) = \mathbb{P}(Y = k \mid \boldsymbol{X} = \cdot)$ belongs to the $(\beta, L, \mathbb{R}^d)$-Hölder class, that is, $p_k \in \mathcal{H}(\beta, L, \mathbb{R}^d)$ for all $k \in [K]$,*
- *for all $k \in [K]$ the regression function p_k satisfy the α-Margin assumption,*
- *for all $k \in [K]$, the cumulative distribution function F_{p_k} of $p_k(\boldsymbol{X})$ is continuous.*

The family of distributions $\mathcal{P}(L, \beta, \alpha)$ resembles the one considered in (Audibert and Tsybakov, 2007) in the context of binary classification. The major difference is the continuity Assumption 2, which poses certain restrictions and does not allow to re-use in a straightforward way their construction for lower bounds.

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

Theorem 2.5. *Let $K \geq 3$, $s \in [\lfloor K/2 \rfloor - 1]$. If $2\alpha\lceil \frac{\beta}{2} \rceil \leq d$, then for all $n, N \in \mathbb{N}$ it holds that*

$$\mathcal{E}^{\mathrm{H}}_{n,N}(\mathcal{P}(L, \beta, \alpha)) \gtrsim n^{-\frac{\alpha\beta}{2\beta+d}} \bigvee (n+N)^{-1/2} \qquad \textit{(Hamming risk)} \ ,$$
$$\mathcal{E}^{\mathrm{R}}_{n,N}(\mathcal{P}(L, \beta, \alpha)) \gtrsim n^{-\frac{(1+\alpha)\beta}{2\beta+d}} \bigvee (n+N)^{-1/2} \qquad \textit{(Excess risk)} \ .$$

First of all, based on these results, we observe that the lower bound for the Hamming risk $\mathcal{E}^{\mathrm{H}}_{n,N}$ is slower than those of the other risk, which is explained by the structure of the Hamming risk. Secondly, the above lower bounds imply that the best rate in the supervised regime is $n^{-1/2}$ across all the risk. Therefore, even if the margin assumption is very strong ($\alpha \gg 1$) supervised methods ($N = 0$) *cannot* achieve fast rates. This fact is the major difference with classical setups where the value of threshold is known (such as classification and level set estimation). Indeed, under the same assumptions on the family of distributions, with the continuity Assumption 2, the minimax rate in those frameworks is $n^{-(1+\alpha)\beta/(2\beta+d)}$ as proved for instance in (Audibert and Tsybakov, 2007; Rigollet and Vert, 2009) and unlabeled data *cannot* improve it. In contrast, this limitation can be neglected in the semi-supervised regime. Indeed, for sufficiently large unlabeled sample, the dominant term in the lower bound is of order $n^{-(1+\alpha)\beta/(2\beta+d)}$, which can be faster than $n^{-1/2}$. To be more precise, when we consider $\mathcal{E}^{\mathrm{R}}_{n,N}$ or $\mathcal{E}^{\mathrm{D}}_{n,N}$ the following relations are *necessary* to get fast rates of convergence

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

The condition on the left hand side ensures that we have enough unlabeled data to eliminate the impact of not knowing threshold $G^{-1}(s)$ in Eq. (2.3). Whereas, the condition on the right hand side ensures that the classification problem with "known" threshold admits fast rates. The above discussion suggests that the lack of knowledge of the threshold

$G^{-1}(s)$ is significant, and the considered framework is more difficult from the statistical perspective than its more classical counterparts. The condition $2\alpha\lceil\frac{\beta}{2}\rceil \leq d$ in the lower bounds is slightly more restrictive than the conditions given in (Audibert and Tsybakov, 2007) (they have $\alpha\beta \leq d$). It might be an artifact of the proof and probably could be avoided with a finer choice of hypotheses.

Upper bounds. In this part, we show that we can build a set-valued classifier that achieves, up to a logarithmic factor, the lower bounds stated in Theorem 2.5. In other words, our estimator is *nearly* optimal in the minimax sense. To come straight to the point, we delay the construction of the estimator to the next paragraph and focus right now on the upper bounds.

Theorem 2.6. *Let $K \in \mathbb{N}$, $s \in (0, K)$, then there exists an estimator $\hat{\Gamma}$ such that for all $n, N \in \mathbb{N}$ we have*

$$\sup_{\mathbb{P}\in\mathcal{P}(L,\beta,\alpha)} \mathcal{E}^{\mathrm{H}}_{n,N}(\hat{\Gamma};\mathbb{P}) \lesssim n^{-\frac{\alpha\beta}{2\beta+d}} \bigvee (n+N)^{-1/2} \qquad \textit{(Hamming risk)} \ ,$$

$$\sup_{\mathbb{P}\in\mathcal{P}(L,\beta,\alpha)} \mathcal{E}^{\mathrm{R}}_{n,N}(\hat{\Gamma};\mathbb{P}) \lesssim \left(\frac{n}{\log n}\right)^{-\frac{(1+\alpha)\beta}{2\beta+d}} \bigvee (n+N)^{-1/2} \qquad \textit{(Excess risk)} \ .$$

An immediate conclusion from the above result is that the lower bounds of Theorems 2.5 are achievable. In particular, in the case of Hamming risk, the rates are optimal; whereas for the Excess risk the upper bounds match the lower bounds up to a logarithmic factor. Thus, the necessary conditions for the fast rates of semi-supervised methods are also *sufficient*. Let us mention that the presence of the logarithmic factor in these upper bounds is due to ℓ_∞-norm estimation and will be discussed later (see Lemma 2.1). It is also instructive to consider the case $N = +\infty$, which formally corresponds to the classification with known marginal distribution $\mathbb{P}_X$. In this case, the rate of convergence depends only on the size of the labeled sample and the obtained rates resemble classical results in classification and level-set estimation (Audibert and Tsybakov, 2007; Rigollet and Vert, 2009).

Construction of the estimator. The construction of $\hat{\Gamma}$ that reaches the rates in the former upper bounds involves a preliminary estimators $\hat{p}_k$ of the regression functions p_k, $k \in [K]$. These estimators $\hat{p}_k$ are constructed using an arbitrary half $\mathcal{D}_{\lfloor n/2 \rfloor}$ of the labeled dataset $\mathcal{D}^{\mathrm{L}}_n$ and they must satisfy the following assumptions.

Assumption 7 (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 and $\delta_0 \geq 0$ such that for all $k \in [K]$ and all $n \geq 2$ we have*

for all $\delta > \delta_0 n^{-\beta/(2\beta+d)}$

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

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

Assumption 8 (Continuity of CDF)**.** *For all* $k \in [K]$*, conditionally on the data, the CDF* $F_{\hat{p}_k}(t) := \mathbb{P}_{\boldsymbol{X}}(\hat{p}_k(\boldsymbol{X}) \leq t)$ *of* $\hat{p}_k(\boldsymbol{X})$ *is almost surely* $\mathbb{P}^{\otimes\lfloor n/2\rfloor}$ *continuous on* $(0,1)$.

First let us point out that Assumption 7 induces that for all $n \geq 2$ and all $\alpha > 0$

$$\sup_{\mathbb{P}\in\mathcal{P}(L,\beta,\alpha)} \mathbf{E}\,\|\boldsymbol{p} - \hat{\boldsymbol{p}}\|_{\infty,\mathbb{P}_{\boldsymbol{X}}}^{1+\alpha} \lesssim \left(\frac{n}{\log n}\right)^{-\frac{(1+\alpha)\beta}{2\beta+d}} \ ,$$

where $\boldsymbol{p}(\cdot) = (p_1(\cdot), \ldots, p_K(\cdot))$, $\hat{\boldsymbol{p}}(\cdot) = (\hat{p}_1(\cdot), \ldots, \hat{p}_K(\cdot))$, and $\|\boldsymbol{p}\|_{\infty,\mathbb{P}_{\boldsymbol{X}}} := \inf\{C \geq 0 : \max_{k\in[K]} |p_k(\boldsymbol{x})| \leq C,\ a.e.\ \boldsymbol{x} \in \mathbb{R}^d\ w.r.t.\ \mathbb{P}_{\boldsymbol{X}}\}$. Assumption 7 is commonly used in the statistical community when dealing with rates of convergence in the classification settings (Audibert and Tsybakov, 2007; Lei, 2014; Sadinle, Lei, and Wasserman, 2018). It is for instance satisfied by the locally polynomial estimator (Audibert and Tsybakov, 2007; Stone, 1977; Tsybakov, 1986) with $\delta_0 = 0$. In addition, Assumption 8 can always be satisfied by slightly processing any estimator $\hat{\boldsymbol{p}}$ (see for instance Section 2.3 for a possible way to process by randomization). Given these considerations, neither Assumption 7 nor Assumption 8 are restrictive, as both of them are satisfied by either classical estimators or by their slightly modified versions thereof.

Recall that according to Proposition 2.1 the optimal set-valued classifier Γ_s^* involves the threshold $G^{-1}(s)$ in its expression. Moreover, $G^{-1}(s)$ is distribution dependent, and thus ought to be estimated using data. A straightforward estimation procedure can be constructed using the unlabeled dataset $\mathcal{D}_N^{\mathrm{U}}$. To make our presentation mathematically correct we introduce[7] the following notation $\mathcal{D}_n^{\mathrm{L}} = \mathcal{D}_{\lfloor n/2\rfloor} \bigcup \mathcal{D}_{\lceil n/2\rceil}$, where $\mathcal{D}_{\lfloor n/2\rfloor}$ is the dataset used to build the estimators $\hat{p}_k$ for $k \in [K]$. Now, all the labels are removed from $\mathcal{D}_{\lceil n/2\rceil}$. That is, $\mathcal{D}_{\lceil n/2\rceil}$ consists of $\lceil n/2\rceil$ *i.i.d.* samples from $\mathbb{P}_{\boldsymbol{X}}$. Consequently, the semi-supervised estimator of $G(\cdot)$ is defined as

$$\hat{G}(\cdot) = \frac{1}{\lceil n/2\rceil + N} \sum_{\boldsymbol{X}\in\mathcal{D}_N^{\mathrm{U}}\bigcup\mathcal{D}_{\lceil n/2\rceil}} \sum_{k=1}^{K} \mathbf{1}_{\{\hat{p}_k(\boldsymbol{X})>\cdot\}} \ .$$

Finally, the plug-in based set-valued classifier $\hat{\Gamma}$ is defined point-wise as

$$\hat{\Gamma}(\boldsymbol{x}) = \left\{k \in [K] \ :\ \hat{p}_k(\boldsymbol{x}) \geq \hat{G}^{-1}(s)\right\} \ . \tag{2.13}$$

[7] We split the data to ensure that the unlabeled sample is larger than the labeled one. From practical perspective this splitting is unnecessary as long as $N \geq n$.

Proof sketch of Theorem 2.6 To show that the estimators introduced in this section satisfy the statement of Theorem 2.6 we significantly refine the proof technique used for Theorem 2.4 in the ERM case, where we obtained suboptimal rates. We start from the same point introducing the intermediate pseudo-oracle function

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

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

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

The classifier $\tilde{\Gamma}$ assumes knowledge of the marginal distribution $\mathbb{P}_{\boldsymbol{X}}$. It is an idealized version of $\hat{\Gamma}$, and formally corresponds to the case $N = +\infty$. This is why the pseudo Oracle $\tilde{\Gamma}$ is not an estimator and only serves as an intermediate quantity in the proof. Besides, thanks to Assumption 8, this pseudo Oracle satisfies the size constraints, that is $\mathrm{S}(\tilde{\Gamma}) = \mathrm{s}$ almost surely.

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

Lemma 2.1 (Upper bound on the thresholds)**.** *Let Assumption 2 be satisfied, then for all* $\mathrm{s} \in (0, K)$

$$\left| G^{-1}(\mathrm{s}) - \tilde{G}^{-1}(\mathrm{s}) \right| \leq \| \boldsymbol{p} - \hat{\boldsymbol{p}} \|_{\infty, \mathbb{P}_{\boldsymbol{X}}} , \quad \textit{almost surely } \mathbb{P}^{\otimes n} \otimes \mathbb{P}_{\boldsymbol{X}}^{\otimes N} \ .$$

The difference $|G^{-1}(\mathrm{s}) - \tilde{G}^{-1}(\mathrm{s})|$ resembles the Wasserstein infinity distance which gives an alternative approach to prove Lemma 2.1, see (Bobkov and Ledoux, 2016). Lemma 2.1 explains the extra $\log n$ factor that appears in the upper bound, as the minimax estimation in sup norm contains the $\log n$ factor, see for instance (Stone, 1982; Tsybakov, 2008).

It is intuitively clear that if, on top of Lemma 2.1, we manage to control the difference $|\tilde{G}^{-1}(\mathrm{s}) - \hat{G}^{-1}(\mathrm{s})|$ then the proof of the upper bound would simply follow the arguments of Audibert and Tsybakov (2007). Yet, such a control is not feasible under our assumptions. To see this, notice that conditionally on $\mathcal{D}_n$ the quantity $|\tilde{G}^{-1}(\mathrm{s}) - \hat{G}^{-1}(\mathrm{s})|$ resembles the deviation of quantile from its empirical version. However, classical result[8] on asymptotic normality of sample quantiles (Ma and Robinson, 1998, Theorem 2) tells that in order to have a central limit theorem with $(n+N)^{-1/2}$ rate it is necessary and sufficient to require $\tilde{G}'(\tilde{G}^{-1}(\mathrm{s})) > 0$. From the minimax perspective, this condition cannot be satisfied since we do not require any lower bound on the derivative of $G(\cdot)$.

[8] We can arrive to a similar conclusion from (Bobkov and Ledoux, 2016, Theorem 5.11)

To circumvent this drawback, we construct the estimator $\hat{\boldsymbol{p}}$ which satisfies continuity Assumption 8. This construction allows to avoid the direct control of $|\tilde{G}^{-1}(s) - \hat{G}^{-1}(s)|$ and leverage standard properties of generalized inverse of a continuous and monotone function. In the next paragraph we demonstrate that the upper bound can be improved if we assume that the derivative of $G(\cdot)$ is uniformly lower bounded, that is, $G^{-1}(\cdot)$ has some regularity.

Upper bound under extra assumptions. Our previous setting illustrates the importance of unlabeled data since semi-supervised set-valued classifiers have shown their superiority over supervised ones. We aim at understanding whether the above phenomena occurs under even stronger assumptions on the joint distribution of $(\boldsymbol{X}, Y)$. The above puts forward that the statistical of set-valued classifiers highly depends on the control of the quantile function. We then impose smoothness condition on this quantity:

Assumption 9. *There exists constant $L', \rho > 0$ such that the function $G^{-1}(\cdot)$ belongs to the $(\rho, L', \mathbb{R})$-Hölder class.*

From now on we consider the family of joint distributions $\mathcal{P}(L, \beta, \alpha, L', \rho)$ of $(\boldsymbol{X}, Y)$ which satisfies all assumptions in Definition 2.6 and satisfies Assumption 9 with some constants $L', \rho > 0$. Let us point out that the family $\mathcal{P}(L, \beta, \alpha, L', \rho)$ is much smaller than $\mathcal{P}(L, \beta, \alpha)$. Indeed, $\mathcal{P}(L, \beta, \alpha, L', \rho)$ excludes all those G functions which are locally constants. Moreover, as L' is assumed to be independent of the sample size, there are no G functions whose slope is locally decreasing with the data. It actually means that if we take all functions G which are generated by the distributions from $\mathcal{P}(L, \beta, \alpha, L', \rho)$ then these functions are not dense (in infinity norm) in the set of all functions G generated by the distributions from $\mathcal{P}(L, \beta, \alpha)$. Thus, the family $\mathcal{P}(L, \beta, \alpha, L', \rho)$ is considerably smaller than the one without the Hölder assumption $\mathcal{P}(L, \beta, \alpha)$.

The main result of this section is uniform over $\mathcal{P}(L, \beta, \alpha, L', \rho)$ upper-bound on the excess risk of the estimator $\hat{\Gamma}$ given by Eq. (2.13).

Theorem 2.7. *Let $K \in \mathbb{N}$, $s \in (0, K)$, then the estimator $\hat{\Gamma}$ defined in (2.13) satisfies for all $n, N \in \mathbb{N}$*

$$\sup_{\mathbb{P} \in \mathcal{P}(L,\beta,\alpha,L',\rho)} \mathcal{E}^{\mathrm{R}}_{n,N}(\hat{\Gamma}; \mathbb{P}) \lesssim \left(\frac{n}{\log n}\right)^{-\frac{(1+\alpha)\beta}{2\beta+d}} \bigvee \left\{ (N+n)^{-\frac{(1+\alpha)\rho}{2}} \bigwedge (N+n)^{-\frac{1}{2}} \right\} .$$

As compared to Theorem 2.6, this upper bound includes the additional term $(N+n)^{-\frac{(1+\alpha)\rho}{2}}$ that makes this upper-bound faster than its counterpart in Theorem 2.6. This term also helps to understand the different regimes of convergence *w.r.t.* the smoothness ρ of the function G^{-1}: i) if $(1+\alpha)\rho \leq 1$, the function G^{-1} is not regular enough and the term $(N+n)^{-\frac{(1+\alpha)\rho}{2}}$ is still negligible as compared to $(N+n)^{-1/2}$, then $(N+n)^{-1/2}$ is the

dominant term and unlabeled data can improve the rate; ii) if $\rho \geq 2\beta/(2\beta + d)$ then all terms that involve N become smaller as compared to the first term and unlabeled data cannot help. iii) between these two extreme regimes the rate is governed by the interplay between α, β and ρ. The proof of Theorem 2.7 combines results from (Bobkov and Ledoux, 2016) on the Wasserstein infinity distance between empirical and real distributions and the argument of Audibert and Tsybakov (2007). Let us point out the difference between proofs of Theorems 2.6 and 2.7. The proof of Theorem 2.6 introduces the pseudo-oracle $\tilde{\Gamma}$ and partly relies on the continuity Assumption 8 posed on the estimator $\hat{\boldsymbol{p}}$. In contrast, the proof of Theorem 2.7 does not require this assumption, yet, this bound is valid for a much smaller family of distributions, since the function G^{-1} is assumed to be Hölder. Instead of the pseudo-oracle $\tilde{\Gamma}$, which "knows" the marginal distribution of the features, the proof of Theorem 2.7 is based on $G_{n,N}$, which is analogous to $\tilde{\Gamma}$, but can be seen as a pseudo-oracle that "knows" the conditional distribution of the labels instead (instead of the marginal distribution $\mathbb{P}_X$). The reason of such a discrepancy is dictated by the difficulty of estimating the conditional distribution of Y given $\boldsymbol{X}$. Indeed, on $\mathcal{P}(L, \beta, \alpha)$ the dominant term of the upper bound is connected with the estimation of the *marginal distribution* of the features. Meanwhile, on $\mathcal{P}(L, \beta, \alpha, L', \rho)$ the dominant term is connected with the estimation of the *conditional distribution* of labels. We do not provide the minimax lower bound on the family $\mathcal{P}(L, \beta, \alpha, L', \rho)$, however, it can be recovered from the proof of Theorem 2.5 (only the first part of the proof) by straightforward but cumbersome modification.

Lastly, let us provide a simple intuition describing the role of the Hölder condition on G^{-1}. This type of conditions is well known in the analysis of order statistics and sample quantiles. For example, consider a random variable $X \sim \frac{1}{2}\mathcal{U}([-2, -1]) + \frac{1}{2}\mathcal{U}([1, 2])$ and denote by $F : \mathbb{R} \to [0, 1]$ the CDF of X. Consider $(X_1, \ldots, X_n)$ *i.i.d.* copies[9] of X and let F_n be the empirical CDF of $(X_1, \ldots, X_n)$. Our goal is to understand the statistical properties of $F_n^{-1}(1/2) = X_{\left(\frac{n}{2}\right)}$, specifically, how to control its deviation from $F^{-1}(1/2) = -1$. However, the quantile function of X is not even continuous around $1/2$, thus this control is impossible. Indeed, notice that if there are at least $\frac{n}{2} + 1$ realizations of $(X_1, \ldots, X_n)$ that end up in $[1, 2]$, then we have $|F_n^{-1}(1/2) - F^{-1}(1/2)| \geq 1/2$. Therefore, we have

$$\mathbb{P}\left(|F_n^{-1}(1/2) - F^{-1}(1/2)| \geq 1/2\right) \geq \mathbb{P}\left(V > \frac{n}{2}\right) \longrightarrow 1/2 \ ,$$

where V is the binomial random variable with parameters $(n, 1/2)$.

2.6 Bibliography

Classification with reject option. The problem of set-valued classifier classification has strong ties with the binary classification with reject option, also known as binary classi-

[9]For simplicity n assumed to be even.

fication 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 set-valued classifiers, which are allowed, in addition to singletons, 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 out in (Bartlett and Wegkamp, 2008; Herbei and Wegkamp, 2006; Lei, 2014; Wegkamp and Yuan, 2011) and also in a work in collaboration with C. Denis **[MH-Journal11]**. Among these references, the procedures presented in this chapter are partially inspired by **[MH-Journal11]** where we also consider a semi-supervised approach to build set-valued classifiers invoking cumulative distribution functions. The similarity is however rather limited since most of the technical tools we used in our proofs in the multi-class setting are quite different. In particular, the analysis in **[MH-Journal11]** leads to suboptimal rates for set-valued predictor based on plug-in principle.

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

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

$$\Gamma^*_\alpha \in \arg\min \{\mathrm{S}(\Gamma) \,:\, \Gamma \in \Xi \text{ s.t. } \mathrm{L}(\Gamma) \leq \alpha\} \quad ,$$

and such a set-valued classifier is called a least ambiguous set-valued classifier with bounded error rate. The authors show that under Assumption 2 this oracle set can be described as a thresholding of the regression function

$$\Gamma^*_\alpha(\cdot) = \{k \in [K] \,:\, p_k(\cdot) \geq t_\alpha\} \quad ,$$

where the threshold t_α is defined as

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

As in the framework of this chapter, the optimal set-valued predictor can by derived via thresholding of the conditional probabilities p_k's. Notice that Sadinle, Lei, and Wasserman (2018) also proceed in two steps as we do, that is, they first estimate the conditional

probabilities $\boldsymbol{p}$ and estimate the threshold t_α after. However, in addition to the instability drawback we pointed in Section 2.1, set-valued predictor defined with the error-rate constraint requires a second *labeled* dataset for the estimator of t_a, due to the presence of $\mathbb{P}(Y = k)$, the marginal distribution of the labels in the constraint. Besides, their theoretical analysis is carried out under a different set of assumptions on the joint distribution $\mathbb{P}$. Apart from the standard margin assumption, they require a so-called detectability, that is, they require that the upper bound in the margin assumption is tight. Under these assumptions they provide an upper bound on the Hamming excess risk and obtain a rate of convergence of order $\mathcal{O}((n/\log n)^{-1/2})$.

Constrained learning. Interestingly, all types of set-valued predictors (with all types of constraints) can be encompassed into the constrained estimation framework (Anbar, 1977; Brown and Low, 1996; Lepskii, 1990), 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 reads as a discrepancy measure, that is, the sum of error and size discrepancies. Let us point out that such risk can also be controlled in a minimax sense in the same way as the excess risk. This being said, we emphasize that our framework and the one of Sadinle, Lei, and Wasserman (2018) can be seen as an extension of the constrained estimation to the classification problems. In addition, a unified description of all constrained problems is studied in the PhD Thesis by Chzhen (2019, Section 1.1.3) where general optimal rules and excess risks are derived.

Semi-supervised learning. In a range of classical problems of statistics, the inference is solely governed by the behavior of the conditional distribution $\mathbb{P}_{Y|\boldsymbol{X}}$ (for instance regression or binary classification). Unlike those situations, the optimal set-valued classifier Γ_s^*, in this work, depends on the whole distribution $\mathbb{P}$. Precisely due to this reason, we allow to observe extra unlabeled data $\mathcal{D}_N^{\mathrm{U}}$ to better estimate the marginal distribution $\mathbb{P}_{\boldsymbol{X}}$ of $\boldsymbol{X}$. From the practical point of view, whenever unlabeled data are available, it is always reasonable to assume that $N \geq n$, since the labeling process is much more expensive than feature gathering process. Yet, we do not explicitly require this relation between N and n and we distinguish two global regimes: semi-supervised ($N > 0$) and supervised ($N = 0$).

Semi-supervised methods are studied in several papers (Bellec et al., 2018; Rigollet, 2007; Singh, Nowak, and Zhu, 2009; Vapnik, 1998) and references therein. Some contributions aim at improving a given supervised estimator with the help of unlabeled data and demonstrate this improvement empirically. In contrast, our work aims at understanding whether the semi-supervised methods should be considered superior to their supervised counterparts from minimax point of view. Our minimax analysis in Section 2.5 reveals that the semi-supervised approach might or might not outperform the supervised one

even in the context of the same problem. Similar conclusions were stated by Singh, Nowak, and Zhu (2009) in the context of learning under the cluster assumption (Rigollet, 2007).

Level sets estimation Minimax setup of the set-valued framework can also be related to the level set estimation problem (Hartigan, 1987; Polonik, 1995; Rigollet and Vert, 2009; Tsybakov, 1997). To draw this relation, we prove in Proposition 2.1, that under a mild assumption, for every $s \in (0, K)$, the set-valued classifier (s-Oracle) Γ_s^* is given by

$$\Gamma_s^*(\boldsymbol{x}) = \{k \in [K] \,:\, \mathbb{P}(Y = k \mid \boldsymbol{X} = \boldsymbol{x}) \geq \theta_s\} \quad ,$$

where θ_s is an *unknown* threshold, which depends on the distribution $\mathbb{P}$ of $(\boldsymbol{X}, Y)$. Similarly, the problem of level set estimation, focuses on the estimation of

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

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

2.7 Conclusion

In this chapter, we studied the set-valued classification problem with a controlled expected size. The theoretical analysis started with distribution-free and ERM-type results, then we introduced non-parametric assumptions and provided minimax analysis of the problem. We emphasized the role played by the unlabeled data – a key ingredient leading to fast rates of convergence. The use of unlabeled data for construction of estimators with desirable properties is a recurring theme of my research. In particular other statistical problems can benefit from similar analysis, which was partly described in the context of fairness in Chapter 1. We paid a particular attention to the continuity assumption on the CDF of posterior class probabilities and to the control of quantile-like quantities (*i.e.*, G^{-1}) using tools from empirical processes, rank statistics, and non-parametric theory.

Conclusion and Perspectives

The document summarizes part of the research I led to understand Lasso prediction, to build fair prediction, and I made a focus on the problem of set-valued classification. I chose to elaborate more on the set-valued classification since I introduced with C. Denis the setup of expected size constraint in the paper [MH-Journal9]. Later, E. Chzhen joined the project during his PhD and we developed in deep details the minimax analysis of the set-valued classifiers of controlled expected size in [MH-Preprint4]. Another reason why I chose this part of my work is that it can be linked to most of the other contributions: **i)** With C. Denis, I first started working on the binary classification with reject option [MH-Conf2]-[MH-Journal11] where semi-supervised approach is also considered. More recently, we extended this framework to the regression setting [MH-Conf6] with A. Zaoui, a PhD student I am co-advising with C. Denis; **ii)** Set-valued classification is related in some sense to the *fairness* problem since in both cases, we investigate minimization of risk under distribution dependent constraints as explained by Chzhen (2019). This led to three papers, one in fair classification [MH-Conf3] and the others two in regression [MH-Conf4]-[MH-Conf5]. However, these research areas remain quite far from each other and different questions arise in the different projects as well as they require different tools to tackle them. I present here some future research directions that I plan to investigate.

Fairness with sensitive feature out from prediction. In the papers [MH-Conf3]-[MH-Conf4]-[MH-Conf5], we developed fair predictors g that use the sensitive feature as input, that is $g : \mathbb{R}^d \times \mathcal{S} \to \mathcal{Y}$. However, despite the fact that the functions of the type $g : \mathbb{R}^d \times \mathcal{S} \to \mathcal{Y}$ are more accurate than the functions $g : \mathbb{R}^d \to \mathcal{Y}$ and can be made equally fair, the former might actually be illegal to use in certain domains. Hence, it is essential to be able to fairly predict without any information on the sensitive feature, that is, to develop a fair predictor $\bar{g} : \mathbb{R} \to \mathcal{Y}$ that only "sees" the input feature $\boldsymbol{X}$ and is unaware of the sensitive attribute. On the one hand, the methodology we developed in the papers [MH-Conf3]-[MH-Conf4] for classification and regression are robust enough to extend to this case even though not trivially. On the other hand, it seems completely unclear for the moment how the methodology we proposed in [MH-Conf5] can handle this situation. Already the main result of Theorem 1.3 which characterizes the fair opti-

mal prediction in terms of Wasserstein barycenters seems non-trivial. This extension is the core of a long project.

ε-fairness and Pareto frontier. The papers [MH-Conf3]-[MH-Conf4]-[MH-Conf5] focus on implementing a methods that obeys fairness in a strict sense, that is, $\mathcal{U}(g) = \sup_{t\in\mathbb{R}} \left|\mathbb{P}(g(X,S) \leq t \mid S = s) - \mathbb{P}(g(X,S) \leq t \mid S = s')\right| = 0$. Unless the regression function is already fair, imposing this constraint inevitably results in an accuracy deflation. Its is often preferred to ask for a less restrictive fairness condition: $\mathcal{U}(g) \leq \varepsilon$, which is referred to as ε-fairness constraint. It is likely that the methodology we built in our earlier papers exports well to this setting. As we start to deal with ε-fairness the question of the choice of the ε parameter becomes important. In particular, there is no physical meaning for ε so that its calibration is dictated by an application at hand. The study of the Pareto frontier related to the problem $\min\{\text{risk}(g) : \mathcal{U}(g) \leq \varepsilon\}$ is an interesting direction of future research to investigate a good choice of ε.

Other notions of fairness + other losses. In [MH-Conf4]-[MH-Conf5], we considered *Demographic Parity* as a measure of fairness for the problem of regression. It would be interesting to study extensions of our techniques to other notions of fairness such as *Equalized-Odds* and *Equality of Opportunity*. While in the context of classification, these notions are well defined, there is, to the best of my knowledge, no consensus in how to define them in the regression setting. The main technical difficulty with a possible extension of our machinery to such notions would come from the conditioning on Y. In the same vein, it would be interesting to observe how the objective function that we minimize under fairness affects the developed machinery. Indeed, the $\mathbb{L}_2$-risk was particularly convenient to translate our problem to a Wasserstein-2 barycenter problem. One question would be to understand whether $\mathbb{L}_q$-risk is to be linked to Wasserstein-q barycenters for all $q \in \mathbb{N}$ or if this result is linked only to the regression setting with $q = 2$.

Learning with one sample: Fairness + Set-valued classification. One of the main characteristics of the algorithms that we developed for the fairness and for the set-valued classification settings is their semi-supervised construction. Indeed, this allows us to use unlabeled data to make any preliminary estimator fair in one case, and to satisfy the size constraint in the other one. However, using two datasets or splitting one to estimate the conditional probabilities p_k and to satisfy the constraint seems to be a big limitation in some of the applications. This is mainly because real datasets often do not contain unlabeled samples. In that case, we need to split the labeled dataset in two and erase labels from one of the two, which intuitively seem to be wasteful. For this reason, one line of research is to develop a new algorithm for these two problems, and more generally for constraint learning problems of this type, with a single labeled dataset that is used

twice. The main challenge here is to bypass the independence assumption between the two datasets.

Minimax rate for ERM. In [MH-Journal9] we derived an upper bound on the excess risk and obtained a rate of convergence of order $(n/\log n)^{-\alpha/(\alpha+r)} + N^{-1/2}$, with r being a parameter that depends on the function ϕ and α being the margin parameter. This rate is slower than the rates obtained in [MH-Preprint4] (yet, they are not directly comparable), but also slower than those obtained in the standard classification framework. In particular, our bound provides no control on the risk when $\alpha = 0$ (no margin). A valuable project is to improve the rates of convergence for the ERM based set-valued classifier.

Dependency on K in set-valued classification and sparsity assumption. In [MH-Journal9]-[MH-Preprint4], the dependency in the number of labels K has not been considered. Yet, set-valued classification can face extreme classification scenarios. The prediction ability of set-valued classifiers should be investigated in a context where the number of labels is large, as well as the number of observations and features. By step, the first consideration should be to tackle the question of large number of labels. Techniques from high-dimensional statistic should be used, in particular, one could think of a new notion of sparsity assumption adapted to the set-valued classification. A possible direction is to introduce a margin assumption that involves a number $s^* \ll K$ which reflects, in the set-valued framework, the actual amount of relevant classes that should be included in an optimal set-valued classifier. In this case, one may expect only logarithmic deflation of the rate in terms of K, and maybe linear (or surperlinear) behavior in terms of the sparsity level s^* as it is typically the case in the high-dimensional statistics.

Neural Network under sparsity. One major difficulty when we deal with the statistical analysis of Neural Networks is the number of parameters that results from a large number of layers and/or nodes. In a recent contribution [MH-Preprint5] in collaboration with J. Lederer, I introduced a new notion of sparsity that acts at the level of layers referred to as *layer sparsity*. In particular we considered a new algorithm that can be seen as a ℓ_1-penalized least-squares method estimators where the penalty is well suited for exploiting the layer sparsity structure in the model. In future works I plan with J. Lederer to investigate statistical properties of this algorithm under the introduced assumption.

References

[MH-Journal1] C. Chesneau and M. Hebiri. Some theoretical results on the grouped variables Lasso. *Mathematical Methods of Statistics*, 17(4):317–326, 2008.

[MH-Journal2] M. Hebiri. Sparse Conformal Predictors. *Statistics and Computing*, 20(2):253–266, 2010.

[MH-Journal3] M. Hebiri and S. van de Geer. The Smooth-Lasso and other l1+l2-penalized methods. *Electron. J. Stat.*, 5:1184–1226, 2011.

[MH-Journal4] P. Alquier and M. Hebiri. Generalization of l1 constraints for high dimensional regression problems. *Statistics and Probability Letters*, 81:1760–1765, 2011.

[MH-Journal5] P. Alquier and M. Hebiri. Transductive versions of the LASSO and the Dantzig Selector. *Journal of Statistical Planning and Inference*, 142(9):2485–2500, 2012.

[MH-Journal6] M. Hebiri and J. Lederer. How Correlations Influence Lasso Prediction. *IEEE Trans. Inform. Theory*, 59(3):1846–1854, 2013.

[MH-Journal7] P. Alquier, C. Butucea, M. Hebiri, K. Meziani, and T. Morimae . Rank penalized estimation of a quantum system. *Physical Review A*, 88(3):032113, 2013.

[MH-Journal8] A. Dalalyan, M. Hebiri, and J. Lederer. On the Prediction Performance of the Lasso. *Bernoulli*, 23(1):552–581, 2017.

[MH-Journal9] C. Denis and M. Hebiri. Confidence sets with expected sizes for Multiclass Classification. *J. Mach. Learn. Res.*, 18(102):1–28, 2017.

[MH-Journal10] E. Chzhen, M. Hebiri, and J. Salmon. On Lasso refitting strategies. *Bernoulli*, 25(4A):3175–3200, 2019.

[MH-Journal11] C. Denis and M. Hebiri. Consistency of plug-in confidence sets for classification in semi-supervised learning. *J. Nonparametr. Stat.*, 32(1):42–72, 2020.

[MH-Conf1] A. Dalalyan, M. Hebiri, K. Meziani, and J. Salmon. Learning heteroscedastic models by convex programming under group sparsity. In *ICML*, pages 379–387, 2013.

[MH-Conf2] C. Denis and M. Hebiri. Confidence Sets for Classification. In *SLDS*: 301–312, 2015.

[MH-Conf3] E. Chzhen, C. Denis, M. Hebiri, L. Oneto, and M. Pontil. Leveraging labeled and unlabeled data for consistent fair binary classification. In *NeurIPS*, 2019.

[MH-Conf4] E. Chzhen, C. Denis, M. Hebiri, L. Oneto, and M. Pontil. Fair Regression via Plug-in Estimator and Recalibration With Statistical Guarantees. In *NeurIPS*, 2020.

[MH-Conf5] E. Chzhen, C. Denis, M. Hebiri, L. Oneto, and M. Pontil. Fair Regression with Wasserstein Barycenters. In *NeurIPS*, 2020.

[MH-Conf6] C. Denis, M. Hebiri, and A. Zaoui. Regression with reject option and application to kNN. In *NeurIPS*, 2020.

[MH-Preprint1] M. Hebiri. Regularization with the Smooth-Lasso procedure. Technical report, 2008.

[MH-Preprint2] M. Hebiri, J.-M. Loubes, and P. Rochet. Aggregation for Linear Inverse Problems. Technical report, 2014.

[MH-Preprint3] E. Chzhen, C. Denis, M. Hebiri, and J. Salmon. On the benefits of output sparsity for multi-label classification. Technical report, 2017.

[MH-Preprint4] E. Chzhen, C. Denis, M. Hebiri. Minimax semi-supervised confidence sets for multi-class classification. Accepted in Bernoulli, 2021.

[MH-Preprint5] M. Hebiri and J. Lederer. Layer Sparsity in Neural Networks. Submitted, 2020

[Ph.D. Thesis] M. Hebiri. *Quelques questions de sélection de variables autour de l'estimateur LASSO*. Ph.D. thesis, Université Paris Diderot, 2009.

External references

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., and Wallach, H. (2018). "A reductions approach to fair classification". *arXiv preprint arXiv:1803.02453* (p. 16).

Agarwal, A., Dudik, M., and Wu, Z. S. (2019). "Fair Regression: Quantitative Definitions and Reduction-Based Algorithms". *International Conference on Machine Learning* (p. 15).

Agueh, M. and Carlier, G. (2011). "Barycenters in the Wasserstein space". *SIAM Journal on Mathematical Analysis* 43.2, pp. 904–924 (p. 17).

Anbar, D (Jan. 1977). "A Modified Robbins-Monro Procedure Approximating the Zero of a Regression Function from Below". *Ann. Statist.* 5.1, pp. 229–234 (p. 47).

Audibert, J-Y. and Tsybakov, A. B. (2007). "Fast learning rates for plug-in classifiers". *Ann. Statist.* 35.2, pp. 608–633 (pp. 38–43, 45).

Bach, F., Jenatton, R., Mairal, J., and Obozinski, G. (2012). "Optimization with Sparsity-Inducing Penalties." *Foundations and Trends in Machine Learning* 4.1, pp. 1–106 (p. 9).

Barocas, S., Hardt, M., and Narayanan, A. (2019). *Fairness and Machine Learning*. http://www.fairmlbook.org. fairmlbook.org (p. 14).

Bartlett, P., Jordan, M., and McAuliffe, J. (2006). "Convexity, classification, and risk bounds". *J. Amer. Statist. Assoc.* 101.473, pp. 138–156 (pp. 32, 33, 35).

Bartlett, P. and Mendelson, S. (2006). "Empirical minimization". *Probab. Theory Related Fields* 135.3, pp. 311–334 (p. 35).

Bartlett, P. and Wegkamp, M. (2008). "Classification with a reject option using a hinge loss". *J. Mach. Learn. Res.* 9, pp. 1823–1840 (p. 46).

Bellec, P. C.., Dalalyan, A. S.., Grappin, E, and Paris, Q (2018). "On the prediction loss of the lasso in the partially labeled setting". *Electron. J. Statist.* 12.2, pp. 3443–3472 (p. 47).

Berk, R., Heidari, H., Jabbari, S., Joseph, M., Kearns, M., Morgenstern, J., Neel, S., and Roth, A. (2017). "A convex framework for fair regression". *Fairness, Accountability, and Transparency in Machine Learning* (p. 15).

Bickel, P., Ritov, Y., and Tsybakov, A. (2009). "Simultaneous analysis of lasso and Dantzig selector". *Ann. Statist.* 37.4, pp. 1705–1732 (pp. 10, 12).

Bobkov, S. and Ledoux, M. (2016). "One-dimensional empirical measures, order statistics and Kantorovich transport distances". *Memoirs of the American Mathematical Society* (pp. 21, 43, 45).

Brown, L. D. and Low, M. G. (1996). "A constrained risk inequality with applications to nonparametric functional estimation". *Ann. Statist.* 24.6, pp. 2524–2535 (p. 47).

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High Dimensional Data. Methods, Theory and Applications*. Springer (p. 10).

Bunea, F., Tsybakov, A., and Wegkamp, M. (2007a). "Aggregation for Gaussian regression". *Ann. Statist.* 35.4, pp. 1674–1697 (p. 9).
— (2007b). "Sparsity oracle inequalities for the Lasso". *Electron. J. Stat.* 1, 169–194 (electronic) (pp. 11, 12).
Cai, T., Wang, L., and Xu, G. (2010). "Shifting inequality and recovery of sparse signals". *IEEE Trans. Signal Process.* 58.3, part 1, pp. 1300–1308 (p. 12).
Calders, T., Karim, A., Kamiran, F., Ali, W., and Zhang, X. (2013). "Controlling attribute effect in linear regression". *IEEE International Conference on Data Mining* (p. 15).
Candès, E. and Plan, Y. (2009). "Near-ideal model selection by ℓ_1 minimization". *Ann. Statist.* 37.5A, pp. 2145–2177 (pp. 10, 12).
Candes, E. and Tao, T. (2007). "The Dantzig selector: statistical estimation when p is much larger than n". *Ann. Statist.* 35.6, pp. 2313–2351 (p. 12).
Chiappa, S., Jiang, R., Stepleton, T., Pacchiano, A., Jiang, H., and Aslanides, J. (2020). "A general approach to fairness with optimal transport". *AAAI* (p. 15).
Chow, C. (1957). "An optimum character recognition system using decision functions". *IRE Transactions on Electronic Computers* 4, pp. 247–254 (p. 46).
— (1970). "On optimum error and reject trade-off". *IEEE Trans. Inform. Theory* 16, pp. 41–46 (p. 46).
Chzhen, E. (Sept. 2019). "Plug-in methods in classification". Theses. Université Paris-Est. URL: https://tel.archives-ouvertes.fr/tel-02400552 (pp. 47, 49).
Dalalyan, A. and Tsybakov, A. (2007). "Aggregation by exponential weighting and sharp oracle inequalities". *Learning theory (COLT2007), Lecture Notes in Comput. Sci., Vol. 4539*, pp. 97–111 (pp. 9, 11).
— (2012a). "Mirror averaging with sparsity priors". *Bernoulli* 18.3, pp. 914–944 (pp. 9, 11).
— (2012b). "Sparse regression learning by aggregation and Langevin Monte-Carlo". *J. Comput. System Sci.* 78.5, pp. 1423–1443 (p. 11).
Del Barrio, E., Gordaliza, P., and Loubes, J.-M. (2020). "Review of Mathematical frameworks for Fairness in Machine Learning". *arXiv preprint arXiv:2005.13755* (p. 14).
Donini, M., Oneto, L., Ben-David, S., Shawe-Taylor, J. S., and Pontil, M. (2018). "Empirical risk minimization under fairness constraints". *Neural Information Processing Systems* (p. 15).
Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). "Least angle regression". *Ann. Statist.* 32.2. With discussion, and a rejoinder by the authors, pp. 407–499 (p. 9).
Fan, J. and Li, R. (2001). "Variable selection via nonconcave penalized likelihood and its oracle properties". *J. Amer. Statist. Assoc.* 96.456, pp. 1348–1360 (p. 9).
Fisher, R. (1936). "Design of experiments". *Br Med J* 1.3923, pp. 554–554 (p. 20).
Fitzsimons, Jack, Al Ali, AbdulRahman, Osborne, Michael, and Roberts, Stephen (2019). "A general framework for fair regression". *Entropy* 21.8, p. 741 (p. 15).
Freund, Y. and Schapire, R. E. (1997). "A decision-theoretic generalization of on-line learning and an application to boosting". *Journal of computer and system sciences* 55.1, pp. 119–139 (p. 32).

Friedman, J., Hastie, T. J., and Tibshirani, R. (2000). "Additive logistic regression: a statistical view of boosting". *Ann. Statist.* 28.2, pp. 337–407 (p. 32).
Gadat, S., Klein, T., and Marteau, C. (2016). "Classification in general finite dimensional spaces with the k-nearest neighbor rule". *Ann. Statist.* 44.3, pp. 982–1009 (p. 39).
Györfi, L., Kohler, M., Krzyżak, A., and Walk, H. (2002). *A distribution-free theory of nonparametric regression*. Springer Ser. Statist. New York: Springer-Verlag (p. 38).
Hardt, M., Price, E., and Srebro, N. (2016). "Equality of opportunity in supervised learning". *Neural Information Processing Systems* (p. 14).
Hartigan, J. A. (1987). "Estimation of a Convex Density Contour in Two Dimensions". *J. Amer. Statist. Assoc.* 82.397, pp. 267–270 (p. 48).
Herbei, R. and Wegkamp, M. (2006). "Classification with reject option". *Canad. J. Statist.* 34.4, pp. 709–721 (p. 46).
Hoeffding, W. (1952). "The large-sample power of tests based on permutations of observations". *The Annals of Mathematical Statistics*, pp. 169–192 (p. 20).
Jiang, R., Pacchiano, A., Stepleton, T., Jiang, H., and Chiappa, S. (2019). "Wasserstein fair classification". *arXiv preprint arXiv:1907.12059* (p. 15).
Juditsky, A. and Nemirovski, A. (2011). "Accuracy guarantees for ℓ_1-recovery". *IEEE Trans. Inform. Theory* 57.12, pp. 7818–7839 (p. 12).
Koltchinskii, V. (2011). *Oracle inequalities in empirical risk minimization and sparse recovery problems*. Vol. 2033. Lecture Notes in Mathematics. Heidelberg: Springer (p. 10).
Kpotufe, S. and Martinet, G. (2018). "Marginal Singularity, and the Benefits of Labels in Covariate-Shift". *Conference On Learning Theory*, pp. 1882–1886 (p. 39).
Lapin, M., Hein, M., and Schiele, B. (2015). "Top-k multiclass SVM". *Advances in Neural Information Processing Systems*, pp. 325–333 (p. 25).
Lei, J. (2014). "Classification with confidence". *Biometrika* 101.4, pp. 755–769 (pp. 42, 46).
Lei, J., Robins, J., and Wasserman, L. (2013). "Distribution-free prediction sets". *J. Amer. Statist. Assoc.* 108.501, pp. 278–287 (pp. 20, 30).
Lei, J. and Wasserman, L. (2014). "Distribution-free prediction bands for non-parametric regression". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76.1, pp. 71–96 (p. 20).
Lepskii, O (1990). "Asymptotic minimax estimation with prescribed properties". *Theory Probab. Appl.* 34.4, pp. 604–615 (p. 47).
Ma, C. and Robinson, J. (1998). "17 Approximations to distributions of sample quantiles". *Handbook of Statistics* 16, pp. 463–484 (p. 43).
Mammen, E. and Tsybakov, A. (1999). "Smooth discrimination analysis". *Ann. Statist.* 27.6, pp. 1808–1829 (pp. 35, 38).
Mammen, E. and van de Geer, S. (1997). "Locally adaptive regression splines". *Ann. Statist.* 25.1, pp. 387–413 (p. 12).
Massart, P. (1990). "The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality". *The Annals of Probability* 18.3, pp. 1269–1283 (p. 20).
Massart, P. and Nédélec, É (2006). "Risk bounds for statistical learning". *Ann. Statist.* 34.5, pp. 2326–2366 (p. 38).

Massias, M., Gramfort, A., and Salmon, J. (2018). "Celer: a Fast Solver for the Lasso with Dual Extrapolation". *ICML*. Vol. 80, pp. 3315–3324 (p. 9).
Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2019). "A survey on bias and fairness in machine learning". *arXiv preprint arXiv:1908.09635* (p. 14).
Oh, S. (2017). "Top-k hierarchical classification". *AAAI Conference on Artificial Intelligence* (p. 25).
Oneto, L. and Chiappa, S. (2020). "Fairness in Machine Learning". *Recent Trends in Learning From Data*. Springer, pp. 155–196 (p. 14).
Oneto, L., Donini, M., and Pontil, M. (2019). "General Fair Empirical Risk Minimization". *arXiv preprint arXiv:1901.10080* (p. 16).
Pérez-Suay, A., Laparra, V., Mateo-García, G., Muñoz-Marí, J., Gómez-Chova, L., and Camps-Valls, G. (2017). "Fair kernel learning". *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (p. 15).
Plečko, D. and Meinshausen, N. (2019). "Fair Data Adaptation with Quantile Preservation". *arXiv preprint arXiv:1911.06685* (p. 17).
Polonik, W. (1995). "Measuring Mass Concentrations and Estimating Density Contour Clusters-An Excess Mass Approach". *Ann. Statist.* 23.3, pp. 855–881 (pp. 35, 48).
Ramaswamy, H, Tewari, A, and Agarwal, S (2018). "Consistent algorithms for multiclass classification with an abstain option". *Electron. J. Stat.* 12.1, pp. 530–554 (p. 46).
Raskutti, G., Wainwright, M., and Yu, B. (2010). "Restricted eigenvalue properties for correlated Gaussian designs". *J. Mach. Learn. Res.* 11, pp. 2241–2259. ISSN: 1532-4435 (p. 12).
— (2011). "Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls". *IEEE Trans. Inform. Theory* 57.10, pp. 6976–6994 (p. 11).
Rigollet, P. (2007). "Generalization error bounds in semi-supervised classification under the cluster assumption". *J. Mach. Learn. Res.* 8.Jul, pp. 1369–1392 (pp. 47, 48).
Rigollet, P. and Tsybakov, A. (2011). "Exponential Screening and optimal rates of sparse estimation". *Ann. Statist.* 39.2, pp. 731–771 (pp. 9, 11).
— (2012). "Sparse estimation by exponential weighting". *Statist. Sci.* 27.4, pp. 558–575 (p. 9).
Rigollet, P. and Vert, R (Nov. 2009). "Optimal rates for plug-in estimators of density level sets". *Bernoulli* 4, pp. 1154–1178 (pp. 40, 41, 48).
Sadinle, M., Lei, J., and Wasserman, L. (2018). "Least ambiguous set-valued classifiers with bounded error levels". *J. Amer. Statist. Assoc.*, pp. 1–12 (pp. 25, 42, 46, 47).
Salmon, J. (2017). "Perspectives computationnelles et statistiques pour la régression en grande dimension". Habilitation à diriger des recherches (HDR). École Normale Supérieure Paris-Saclay (p. 9).
Santambrogio, F. (2015). *Optimal transport for applied mathematicians*. Springer (p. 17).
Singh, A., Nowak, R., and Zhu, J. (2009). "Unlabeled data: Now it helps, now it doesn't". *NIPS*, pp. 1513–1520 (pp. 47, 48).
Stone, C. (1977). "Consistent nonparametric regression". *Ann. Statist.*, pp. 595–620 (pp. 38, 42).

Stone, C. (1982). "Optimal global rates of convergence for nonparametric regression". *Ann. Statist.*, pp. 1040–1053 (p. 43).
Sun, T. and Zhang, C.-H. (2012). "Scaled sparse linear regression". *Biometrika* 99.4, pp. 879–898 (pp. 10–12).
Tibshirani, R. J. (2013). "The lasso problem and uniqueness". *Electron. J. Stat.* 7, pp. 1456–1490 (p. 9).
Tsybakov, A. (1986). "Robust reconstruction of functions by the local-approximation method". *Problemy Peredachi Informatsii* 22.2, pp. 69–84 (pp. 38, 42).
— (1997). "On nonparametric estimation of density level sets". *Ann. Statist.* 25.3, pp. 948–969 (p. 48).
— (2004). "Optimal aggregation of classifiers in statistical learning". *Ann. Statist.* 32.1, pp. 135–166 (p. 35).
— (2008). *Introduction to Nonparametric Estimation*. Springer Ser. Statist. Springer New York (pp. 39, 43).
Van de Geer, S. (2007). "The deterministic Lasso". *Proc. of Joint Statistical Meeting* (p. 12).
Van de Geer, S. and Bühlmann, P. (2009). "On the conditions used to prove oracle results for the Lasso". *Electron. J. Stat.* 3, pp. 1360–1392 (p. 12).
Van de Geer, S. and Lederer, J. (2013). "The Lasso, correlated design, and improved oracle inequalities". *From Probability to Statistics and Back: High-Dimensional Models and Processes–A Festschrift in Honor of Jon A. Wellner*. Institute of Mathematical Statistics, pp. 303–316 (pp. 10, 11).
Van der Vaart, A. (2000). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press (pp. 15, 20).
Vapnik, V. (1998). *Statistical learning theory*. Wiley (pp. 18, 32, 47).
Villani, C. (2003). *Topics in Optimal Transportation*. American Mathematical Society (p. 17).
Vovk, V. (2002a). "Asymptotic optimality of transductive confidence machine". *Algorithmic learning theory*. Vol. 2533. Lecture Notes in Comput. Sci. Berlin: Springer, pp. 336–350 (pp. 25, 30).
— (2002b). "On-line confidence machines are well-calibrated". *Proceedings of the Forty-Third Annual Symposium on Foundations of Computer Science*. Los Alamitos: CA. IEEE Computer Society, pp. 187–196 (p. 25).
Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic learning in a random world*. New York: Springer (pp. 20, 25, 29, 30, 46).
Wainwright, M. (2009). "Sharp Thresholds for High-Dimensional and Noisy Sparsity Recovery Using ℓ_1-Constrained Quadratic Programming (Lasso)". *IEEE Trans. Inf. Theory* 55.5, pp. 2183–2202 (p. 12).
Wegkamp, M. and Yuan, M. (2011). "Support vector machines with a reject option". *Bernoulli* 17.4, pp. 1368–1385 (pp. 33, 46).
Yang, Y. (1999). "Minimax nonparametric classification: Rates of convergence". *IEEE Transactions on Information Theory* 45.7, pp. 2271–2284 (p. 38).
Ye, F. and Zhang, C.-H. (2010). "Rate Minimaxity of the Lasso and Dantzig Selector for the ℓ_q loss in ℓ_r balls". *The Journal of Machine Learning Research* 11, pp. 3519–3540 (p. 12).

Yuan, M. and Wegkamp, M. (2010). "Classification methods with reject option based on convex risk minimization". *J. Mach. Learn. Res.* 11, pp. 111–130 (pp. 32, 33).
Zeni, G., Fontana, M., and Vantini, S. (2020). "Conformal Prediction: a Unified Review of Theory and New Challenges". *arXiv preprint arXiv:2005.07972* (p. 20).
Zhang, C.-H. (2010). "Nearly unbiased variable selection under minimax concave penalty". *Ann. Statist.* 38.2, pp. 894–942 (p. 9).
Zhang, T. (2004a). "Statistical Analysis of Some Multi-Category Large Margin Classification Methods". *J. Mach. Learn. Res.* 5, pp. 1225–1251 (p. 34).
— (2004b). "Statistical behavior and consistency of classification methods based on convex risk minimization". *Ann. Statist.* 32.1, pp. 56–85 (pp. 32, 33).
— (2009). "Some sharp performance bounds for least squares regression with L_1 regularization". *Ann. Statist.* 37.5A, pp. 2109–2144 (p. 12).